JMIR Medical Education

Impact Factor (2024): 3.2
Volume 11 (2025) ISSN 2369-3762 Editor-in-Chief: Blake J. Lesselroth, MD, MBI, FACP, FAMIA

Contents

Original Papers

Development and Validation of a Large Language Model–Based System for Medical History-Taking Training: Prospective Multicase Study on Evaluation Stability, Human-AI Consistency, and Transparency (e73419)
Yang Liu, Chujun Shi, Liping Wu, Xiule Lin, Xiaoqin Chen, Yiying Zhu, Haizhu Tan, Weishan Zhang. 16

Utility of Generative Artificial Intelligence for Japanese Medical Interview Training: Randomized Crossover Pilot Study (e77332)
Takanobu Hirosawa, Masashi Yokose, Tetsu Sakamoto, Yukinori Harada, Kazuki Tokumasu, Kazuya Mizuta, Taro Shimizu. 35

Leveraging Large Language Models for Simulated Psychotherapy Client Interactions: Development and Usability Study of Client101 (e68056)
Daniel Cabrera Lozoya, Mike Conway, Edoardo Sebastiano De Duro, Simon D'Alfonso. 46

Collaborative Development of Feedback Concept Maps for Virtual Patient–Based Clinical Reasoning Education: Mixed Methods Study (e57331)
Anja Mayer, Inga Hege, Andrzej Kononowicz, Anja Müller, Małgorzata Sudacka. 64

Exploring the Implementation of Multiple Telementoring ECHO Programs From an Institutional and Organizational Perspective: Qualitative Study (e75844)
M Pagé, Élise Develay, Annie Talbot, Rania Khemiri, Claire Wartelle-Bladou. 496

Investigating Learning Effects Through the Implementation of Teledermatology Consultations Among General Practitioners in Germany: Mixed Methods Process Evaluation (e65915)
Andreas Polanc, Inka Roesel, Elke Feil, Peter Martus, Stefanie Joos, Roland Koch. 509

Impact of Motivational Interviewing Education on General Practitioners' and Trainees' Learning and Diabetes Outcomes in Primary Care: Mixed Methods Study (e75916)
Isaraporn Thepwongsa, Pat Nonjui, Radhakrishnan Muthukumar, Poompong Sripa. 525

Using Web-Based Continuing Education to Improve New Diagnoses of Alzheimer Disease in Claims Data: Retrospective Case-Control Study (e72000)
Katie Lucero, Thomas Finnegan, Soo Borson. 545

Exploring Health Care Professionals' Perspectives on Education, Awareness, and Preferences for Digital Educational Resources to Support Transgender, Nonbinary, and Intersex Care: Interview Study (e67993)
Sravya Katta, Nadia Davoody. 553

A Web-Based Training Intervention for Primary Care Providers on Preparing Patients for Cancer Treatment Decisions and Conversations About Clinical Trials: Evaluation of a Pilot Study Using Mixed Methods and Follow-Up (e66892)
Naomi Parker, Margo Michaels, Carla Fisher, Alyssa Crowe, Elisa Weiss, Maria Sae-Hau, Jason Arnold, Andrea Cassells, Domenic Durante, Ji-Hyun Lee, Raymond Vega, Ana Natale-Pereira, Taylor Vasquez, Zhongyue Zhang, Carma Bylund. 576

Pharmacists' Attitudes, Perceptions, and Preferences Regarding Continuing Education: Cross-Sectional Study in Vietnam (e77013)
Trung Vo, Phuoc Le, Hien Tran, Hieu Nguyen, Thoai Nguyen, Trang Huynh, Bay Vo. 598

Evaluation of an Interdisciplinary Educational Program to Foster Learning Health Systems: Education Evaluation (e54152)
Sathana Dushyanthen, Nadia Zamri, Wendy Chapman, Daniel Capurro, Kayley Lyons. 615

Motivational Framing Strategies in Health Care Information Security Training: Randomized Controlled Trial (e73245)
Thomas Keller, Julia Warwas, Julia Klein, Richard Henkenjohann, Manuel Trenz, Simon Trang. 629

Using AI-Based Virtual Simulated Patients for Training in Psychopathological Interviewing: Cross-Sectional Observational Study (e78857)
Daniel García-Torres, César Fernández, José Mira, Alexandra Morales, María Vicente. 645

Assessing and Improving Study Skills Support in Medical Education Through a Student-Staff Partnership: Mixed Methods Approach (e65053)
Nicole Tay, Anaïs Deere, Dhivya Ilangovan, Carys Phillips, Emma Kelley. 662

Recruiting Medical, Dental, and Biomedical Students as First Responders in the Immediate Aftermath of the COVID-19 Pandemic: Prospective Follow-Up Study (e63018)
Nicolas Schnetzler, Victor Taramarcaz, Tara Herren, Eric Golay, Simon Regard, François Mach, Amanta Nasution, Robert Larribau, Melanie Suppan, Eduardo Schiffer, Laurent Suppan. 672

Comparison of Learning Outcomes Among Medical Students in Thailand to Determine the Right Time to Teach Forensic Medicine: Retrospective Study (e57634)
Ubon Chudoung, Wilaipon Saengon, Vichan Peonim, Wisarn Worasuwannarak. 685

Understanding Community Health Care Through Problem-Based Learning With Real-Patient Videos: Single-Arm Pre-Post Mixed Methods Study (e68743)
Kiyoshi Shikino, Kazuyo Yamauchi, Nobuyuki Araki, Ikuo Shimizu, Hajime Kasai, Tomoko Tsukamoto, Hiroshi Tajima, Yu Li, Misaki Onodera, Shoichi Ito. 692

Organizational Leaders' Views on Digital Health Competencies in Medical Education: Qualitative Semistructured Interview Study (e64768)
Humairah Zainal, Xin Xiao Hui, Julian Thumboo, Fong Kok Yong. 704

Educational Effectiveness of a 5-Country Virtual Exchange Program for Internationalization in Occupational Therapy Education: Mixed Methods Study (e77564)
Natsuka Suyama, Kaoru Inoue, Norikazu Kobayashi, Anuchart Kaunnil, Supatida Siangchin, Muhammad Sahid, Erayanti Saloko, Sk Moniruzzaman. 718

Exploration and Practice of the First Clinical Medical Postdoctoral Program in China: Retrospective, Nonrandomized, Controlled Study (e65622)
Lingda Zhang, Lianghong Sun, Honglei Li. 737

Comparison of Physician Assistant and Medical Students' Clinical Reasoning Processes Using an Online Patient Simulation Tool to Support Clinical Reasoning (eCREST): Mixed Methods Study (e68981)
Alistair Thorpe, Angelos Kassianos, Ruth Plackett, Vinodh Krishnamurthy, Maria Kambouri, Jessica Sheringham. 745

A Brief Web-Based Person-Centered Care Group Training Program for the Management of Generalized Anxiety Disorder: Feasibility Randomized Controlled Trial in Spain (e50060)
Vanesa Ramos-García, Amado Rivero-Santana, Wenceslao Peñate-Castro, Yolanda Álvarez-Pérez, Andrea Duarte-Díaz, Alezandra Torres-Castaño, María Trujillo-Martín, Ana González-González, Pedro Serrano-Aguilar, Lilisbeth Perestelo-Pérez. 757

Faculty Perceptions on the Roles of Mentoring, Advising, and Coaching in an Anesthesiology Residency Program: Mixed Methods Study (e60255)
Sydney Nykiel-Bailey, Kathryn Burrows, Bianca Szafarowicz, Rachel Moquin. 769

Performance of Plug-In Augmented ChatGPT and Its Ability to Quantify Uncertainty: Simulation Study on the German Medical Board Examination (e58375)
Julian Madrid, Philipp Diehl, Mischa Selig, Bernd Rolauffs, Felix Hans, Hans-Jörg Busch, Tobias Scheef, Leo Benning. 780

Visual Learning in Electrocardiography Training for Medical Residents: Comparative Intervention Study (e73328)
Heng-You Sung, Feng-Ching Liao, Shu-I Lin, Han-En Cheng, Chun-Wei Lee. 793

e-Learning in Phoniatrics and Speech-Language Pathology: Exploratory Analysis of Free Access Tools in Augmentative and Alternative Communication (e63392)
Jessica Büchs, Christiane Neuschaefer-Rube. 800

Evolution of Learning Styles in Surgery Comparing Residents and Teachers: Cross-Sectional Study (e64767)
Gabriela Gouvea Silva, Carlos da Silva Costa, Bruno Cardoso Gonçalves, Luiz Vianney Saldanha Cidrão Nunes, Emerson Roberto dos Santos, Natalia Almeida de Arnaldo Rodriguez Castro, Alba de Abreu Lima, Vânia Sabadoto Brienze, Antônio Oliani, Júlio André. 812

Exploring Connections Between Mental Health, Burnout, and Academic Factors Among Medical Students at an Iranian University: Cross-Sectional Questionnaire Study (e58008)
Elham Faghihzadeh, Ali Eghtesad, Muhammad Fawad, Xiaolin Xu. 820

Barriers to and Facilitators of Implementing Team-Based Extracorporeal Membrane Oxygenation Simulation Study: Exploratory Analysis (e57424)
Joan Brown, Sophia De-Oliveira, Christopher Mitchell, Rachel Cesar, Li Ding, Melissa Fix, Daniel Stemen, Krisda Yacharn, Se Wong, Anahat Dhillon. 831

Game-Based Assessment of Cognitive Abilities and Personality Characteristics for Surgical Resident Selection: A Preliminary Validation Study (e72264)
Noa Gazit, Gilad Ben-Gal, Ron Eliashar. 845

Trends in the Japanese National Medical Licensing Examination: Cross-Sectional Study (e78214)
Yuki Morimoto, Kiyoshi Shikino, Yukihiro Nomura, Shoichi Ito. 859

Evaluation of the Inverted Classroom Approach in a Case-Study Course on Antithrombotic Drug Use in a PharmD Curriculum: French Monocentric Randomized Study (e67419)
Georges Jourdi, Mayssa Selmi, Pascale Gaussem, Jennifer Truchot, Isabelle Margaill, Virginie Siguret. 878

Balancing Academics and Life: Qualitative Study of Health Professions Students' Perceptions of a Four-Day Academic Week in the United Arab Emirates (e67775)
Ashokan Arumugam, Jacqueline Dias, Sangeetha Narasimhan, Raneen Qadah, Reime Shalash, Taif Omran, Bashair Mussa, Basema Saddik, Nadia Al Mazrouei, Sivapriya Ramakrishnan. 891

Evaluating Tailored Learning Experiences in Emergency Residency Training Through a Comparative Analysis of Mobile-Based Programs Versus Paper- and Web-Based Approaches: Feasibility Cross-Sectional Questionnaire Study (e57216)
Hsin-Ling Chen, Chen-Wei Lee, Chia-Wen Chang, Yi-Ching Chiu, Tzu-Yao Hung. 904

Acceptance of AI-Powered Chatbots Among Physiotherapy Students: International Cross-Sectional Study (e76574)
Salwa El-Sobkey, Kerolous Kelini, Mahmoud ElKholy, Tayseer Abdeldayem, Mariam Abdallah, Dina Mohamed, Aya Fawzy, Yomna Ahmed, Ayman El Khatib, Hind Khalid, Balkhis Shaik, Ana Anjos, Mutasim Alharbi, Karim Fathy, Khaled Takey. 916

Paradox of AI in Higher Education: Qualitative Inquiry Into AI Dependency Among Educators in Palestine (e74947)
Anas Alhur, Zuheir Khlaif, Bilal Hamamra, Elham Hussein. 930

Preclinical Medical Students' Perspectives and Experiences With Structured Web-Based English for Medical Purposes Courses: Cross-Sectional Study (e65779)
Radhakrishnan Muthukumar, Isaraporn Thepwongsa, Poompong Sripa, Bangonsri Jindawong, Kamonwan Jenwitheesuk, Surapol Virasiri. . . 9 4 8

Health Workers' Perspectives on Mobile Health Care Learning Stickiness: Mixed Methods Study (e63827)
Sabila Nurwardani, Putu Handayani. 962

Training Gaps in Digital Skills for the Cancer Health Care Workforce Based on Insights From Clinical Professionals, Nonclinical Professionals, and Patients and Caregivers: Qualitative Study (e78490)
David Liñares, Theologia Tsitsi, Noemí López-Rey, Wilfredo Guanipa-Sierra, Susana Aldecoa-Landesa, Carme Carrión, Daniela Cabutto, Deborah Moreno-Alonso, Clara Madrid-Alejos, Andreas Charalambous, Ana Clavería. 983

Impact of Learner Autonomy on the Performance in Voluntary Online Cardiac Auscultation Courses: Prospective Self-Controlled Study (e78363)
Yudong Fang, Ligang Fang, Xue Lin. 999

Student Satisfaction in Social Media–Based Learning Environments: Development, Validation, and Psychometric Evaluation of the CuSAERS (Questionnaire of Satisfaction With Educational Activities Performed on Social Media) (e73805)
Roy La Touche, Álvaro Reina-Varona, Mónica Grande-Alonso, José León-Hernández, Joaquín Pardo-Montero, Néstor Requejo-Salinas, Raúl Ferrer-Peña, Alba Paris-Alemany. 1008

Factors Influencing Educators' Perspectives on Accepting Extended Reality in Health Care Education: Qualitative Study (e65042)
Zuheir Khlaif, Nisreen Salama, Bilal Hamamra, Allam Mousa. 1022

Assessing ChatGPT's Capability as a New Age Standardized Patient: Qualitative Study (e63353)
Joseph Cross, Tarron Kayalackakom, Raymond Robinson, Andrea Vaughans, Roopa Sebastian, Ricardo Hood, Courtney Lewis, Sumanth Devaraju, Prasanna Honnavar, Sheetal Naik, Jillwin Joseph, Nikhilesh Anand, Abdalla Mohammed, Asjah Johnson, Eliran Cohen, Teniola Adeniji, Aisling Nnenna Nnaji, Julia George. 1038

Virtual Reality Simulation for Undergraduate Nursing Students for Care of Patients With Infectious Diseases: Mixed Methods Study (e64780)
Wen Chang, Chun-Chih Lin, Julia Crilly, Hui-Ling Lee, Li-Chin Chen, Chin-Yen Han. 1051

Effectiveness of a 5G Local Area Network–Based Digital Microscopy Interactive System: Quasi-Experimental Design (e70256)
Jie Xu, Jihong Sha, Song Jia, Jiao Li, Lei Xu, Zhihua Shao. 1064

Pass/Fail Versus Tiered Grades and Academic Performance in Undergraduate Medical Education: Crossover Study (e74975)
Boris Modrau, Karina Kirk, Sinan Said, Carsten Bjarkam, Lone Sunde, Jacob Bodilsen, Jakob Dal, Jette Kristensen, Jeppe Emmersen, Mike Astorp, Stig Andersen. . . . 1075

Insights Into History and Trends of Teaching and Learning in Stomatology Education: Bibliometric Analysis (e66322)
Ziang Zou, Linna Guo. . . . 1084

Digital Literacy Training for Digitalization Officers ("Digi-Managers") in Outpatient Medical and Psychotherapeutic Care: Conceptualization and Longitudinal Evaluation of a Certificate Course (e70843)
Anne Mainz, Timo Neunaber, Paula D'Agnese, Alexander Eid, Tanja Galla, Christoph Ellers, Sven Meister. . . . 1099

Mapping the Evolution of China's Traditional Chinese Medicine Education Policies: Insights From a BERTopic-Based Descriptive Study (e72660)
Tao Yang, Fan Yang, Yong Li. . . . 1112

Mono-Professional Simulation-Based Obstetric Training in a Low-Resource Setting: Stepped-Wedge Cluster Randomized Trial (e54911)
Anne van Tetering, Ella de Vries, Peter Ntuyo, E van den Heuvel, Annemarie Fransen, M van der Hout-van der Jagt, Imelda Namagembe, Josaphat Byamugisha, S Oei. . . . 1129

Implementation Outcomes of Reusable Learning Objects in Health Care Education Across Three Malaysian Universities: Evaluation Using the RE-AIM Framework (e63882)
Hooi Lim, Chin Teo, Yew Lee, Ping Lee, Kuhan Krishnan, Zahiruddin Abu Hassan, Phelim Yong, Wei Yap, Renukha Sellappans, Enna Ayub, Nurhanim Hassan, Sazlina Shariff Ghazali, Nurul Nasharuddin, Puteri Jahn Kassim, Faridah Idris, Klas Karlgren, Natalia Stathakarou, Petter Mordt, Stathis Konstantinidis, Michael Taylor, Cherry Poussa, Heather Wharrad, Chirk Ng. . . . 1140

Resident Physician Recognition of Tachypnea in Clinical Simulation Videos in Japan: Cross-Sectional Study (e72640)
Kiyoshi Shikino, Yuji Nishizaki, Sho Fukui, Koshi Kataoka, Daiki Yokokawa, Taro Shimizu, Yu Yamamoto, Kazuya Nagasaki, Hiroyuki Kobayashi, Yasuharu Tokuda. . . . 1155

Awareness and Attitude Toward Artificial Intelligence Among Medical Students and Pathology Trainees: Survey Study (e62669)
Anwar Rjoop, Mohammad Al-Qudah, Raja Alkhasawneh, Nesreen Bataineh, Maram Abdaljaleel, Moayad Rjoub, Mustafa Alkhateeb, Mohammad Abdelraheem, Salem Al-Omari, Omar Bani-Mari, Anas Alkabalan, Saoud Altulaih, Iyad Rjoub, Rula Alshimi. . . . 1166

Reviewing Mobile Apps for Teaching Human Anatomy: Search and Quality Evaluation Study (e64550)
Guadalupe Rivera García, Miriam Cervantes López, Juan Ramírez Vázquez, Arturo Llanes Castillo, Jaime Cruz Casados. . . . 1174

Exploring Gender Perspectives in Medical Education: Latent Semantic Analysis of Israeli First-Year Medical Students' Reflections (e78371)
Rola Khamisy-Farah, Raymond Farah, Haneen Jabaly-Habib, Yara Nakhleh Francis, Nicola Bragazzi. . . . 1192

Enhancing Clinical Competencies Through Peer Role-Play in Oncology Graduate Students: Mixed Methods Study (e79771)
Yao Wang, Feixiang Wang, Gaojie Liu, Yuqing luo, Hongjun Ba, Jie Long. . . . 1210

Effectiveness of a Fully Online Scientific Research Works Peer Support Group Model for Research Capacity Building Through Conducting Systematic Reviews Among Health Care Professionals: Retrospective Cohort Studies (e78862)
Yuki Kataoka, Ryuhei So, Masahiro Banno, Yasushi Tsujimoto, SRWS-PSG Mentors. . . . 1218

Using Electronic Health Data to Deliver an Adaptive Online Learning Solution to Emergency Trainees: Mixed Methods Pilot Study (e65287)
Anna Janssen, Andrew Coggins, James Tadros, Deleana Quinn, Amith Shetty, Tim Shaw. . . . 1231

Engaging Undergraduate Medical Students With Introductory Research Training via an Educational Escape Room: Mixed Methods Evaluation (e71339)
Bastien Le Guellec, Victoria Gauthier, Rémi Lenain, Alexandra Nuytten, Luc Dauchet, Brigitte Bonneau, Erwin Gerard, Claire Castandet, Patrick Truffert, Marc Hazzan, Philippe Amouyel, Raphaël Bentegeac, Aghiles Hamroun. 1250

Gamified Learning in a Virtual World for Undergraduate Emergency Radiology Education: Quasi-Experimental Study (e68518)
Alba Pérez-Baena, Teodoro Rudolphi-Solero, Rocío Lorenzo-Álvarez, Miguel Ruiz-Gómez, Francisco Sendra-Portero. 1270

Integrated e-Learning for Shoulder Anatomy and Clinical Examination Skills in First-Year Medical Students: Randomized Controlled Trial (e62666)
Roland Koch, Lena Gassner, Navina Gerlach, Teresa Festl-Wietek, Bernhard Hirt, Stefanie Joos, Thomas Shiozawa. 1287

E-Learning for Pediatric Emergency Department Staff in Point-of-Care Electroencephalogram Interpretation: Prospective Cohort Study (e69395)
Leopold Simma, Maurice Schneeberger, Stefanie von Felten, Michelle Seiler, Georgia Ramantani, Bigna Bölsterli. 1305

Evaluation and Uptake of an Online ADHD Psychoeducation Training for Primary Care Health Care Professionals: Implementation Study (e59365)
Blandine French, Hannah Wright, David Daley, Elvira Perez Vallejos, Kapil Sayal, Charlotte Hall. 1317

Making Medical Education Courses Visible: Theory-Based Development of a National Database (e62838)
Andi Gashi, Monika Brodmann Maeder, Eva Hennel. 1326

Large Language Models in Biochemistry Education: Comparative Evaluation of Performance (e67244)
Olena Bolgova, Inna Shypilova, Volodymyr Mavrych. 1341

Medical Students' Acceptance of Tailored e–Mental Health Apps to Foster Their Mental Health: Cross-Sectional Study (e58183)
Catharina Grüneberg, Alexander Bäuerle, Sophia Karunakaran, Dogus Darici, Nora Dörrie, Martin Teufel, Sven Benson, Anita Robitzsch. . .
1 3 5 1

Integration of an Audiovisual Learning Resource in a Podiatric Medical Infectious Disease Course: Multiple Cohort Pilot Study (e55206)
Garrik Hoyt, Chandra Bakshi, Paramita Basu. 1362

Performance of ChatGPT-3.5 and ChatGPT-4 in the Taiwan National Pharmacist Licensing Examination: Comparative Evaluation Study (e56850)
Ying-Mei Wang, Hung-Wei Shen, Tzeng-Ji Chen, Shu-Chiung Chiang, Ting-Guan Lin. 1373

Guidelines for Patient-Centered Documentation in the Era of Open Notes: Qualitative Study (e59301)
Anita Vanka, Katherine Johnston, Tom Delbanco, Catherine DesRoches, Annalays Garcia, Liz Salmi, Charlotte Blease. 1383

Implementing the H&P 360 in Three Medical Institutions: Usability Study (e66221)
Rupinder Hayer, Joyce Tang, Julia Bisschops, Gregory Schneider, Kate Kirley, Tamkeen Khan, Erin Rieger, Eric Walford, Irsk Anderson, Valerie Press, Brent Williams. 1395

Comparing the Perceived Realism and Adequacy of Venipuncture Training on an in-House Developed 3D-Printed Arm With a Commercially Available Arm: Randomized, Single-Blind, Cross-Over Study (e71139)
Susan Brouwer de Koning, Amy Hofman, Sonja Gerber, Vera Lagerburg, Michelle van den Boorn. 1408

Resilience Training Web App for National Health Service Keyworkers: Pilot Usability Study (e51101)
Joanna Burrell, Felicity Baker, Matthew Bennion. 1419

Development of a Clinical Clerkship Mentor Using Generative AI and Evaluation of Its Effectiveness in a Medical Student Trial Compared to Student Mentors: 2-Part Comparative Study (e76702)
Hayato Ebihara, Hajime Kasai, Ikuo Shimizu, Kiyoshi Shikino, Hiroshi Tajima, Yasuhiko Kimura, Shoichi Ito. . . . 1426

Evaluating the Performance of DeepSeek-R1 and DeepSeek-V3 Versus OpenAI Models in the Chinese National Medical Licensing Examination: Cross-Sectional Comparative Study (e73469)
Weiping Wang, Yuchen Zhou, Jingxuan Fu, Ke Hu. . . . 1442

Evaluation of a Simulation Program for Providing Telenursing Training to Nursing Students: Cohort Study (e67804)
Ola Ali-Saleh, Layalleh Massalha, Ofra Halperin. . . . 1453

Global Trends in Cadaver Donation and Medical Education Research: Bibliometric Analysis Based on VOSviewer and CiteSpace (e71935)
Xianxian Zhou, Hua Xiong, Yi Wen, Fang Li, Dexi Hu. . . . 1463

Text Message (SMS) Microlearning for Tobacco Use Disorder: Pre-Post Pilot Study of Clinician Confidence (e73821)
Zehra Dhanani, Veena Dronamraju, Jamie Garfield. . . . 1486

Distance Learning During the COVID-19 Lockdown and Self-Assessed Competency Development Among Radiology Residents in China: Cross-Sectional Survey (e54228)
Peicheng Wang, Ziye Wu, Jingfeng Zhang, Yanrong He, Maoqing Jiang, Jianjun Zheng, Zhenchang Wang, Zhenghan Yang, Yanhua Chen, Jiming Zhu. . . . 1493

Comparison of an Emergency Medicine Asynchronous Learning Platform Usage Before and During the COVID-19 Pandemic: Retrospective Analysis Study (e58100)
Blake Briggs, Madhuri Mulekar, Hannah Morales, Iltifat Husain. . . . 1509

Innovative Mobile App (CPD By the Minute) for Continuing Professional Development in Medicine: Multimethods Study (e69443)
Peter Slinger, Maram Omar, Sarah Younus, Rebecca Charow, Michael Baxter, Craig Campbell, Meredith Giuliani, Jesse Goldmacher, Tharshini Jeyakumar, Inaara Karsan, Janet Papadakos, Tina Papadakos, Alexandra Rotstein, May-Sann Yee, Asad Siddiqui, Marcos Restrepo, Melody Zhang, David Wiljer. . . . 1518

Exploring Social Media Use Among Medical Students Applying for Residency Training: Cross-Sectional Survey Study (e59417)
Simi Jandu, Jennifer Carey. . . . 1533

Ethical Use of Social Media and Sharing of Patient Information by Medical Students at a University Hospital in Saudi Arabia: Cross-Sectional Survey (e57812)
Sara Farsi, Alaa Sabbahi, Deyala Sait, Raghad Kabli, Ghaliah Abduljabar. . . . 1543

Instagram as a Tool to Improve Human Histology Learning in Medical Education: Descriptive Study (e55861)
Alejandro Escamilla-Sanchez, Juan López-Villodres, Carmen Alba-Tercedor, María Ortega-Jiménez, Francisca Rius-Díaz, Raquel Sanchez-Varo, Diego Bermúdez. . . . 1553

Virtual Standardized Patients for Improving Clinical Thinking Ability Training in Residents: Randomized Controlled Trial (e73196)
Liyuan Xu, Qinrong Xu, Chunya Liu, Baozhen Chen, Chunxia Wang. . . . 1567

Exploring the Role of Immersive Virtual Reality Simulation in Health Professions Education: Thematic Analysis (e62803)
Jordan Talan, Molly Forster, Leian Joseph, Deepak Pradhan. . . . 1579

Knowledge Mapping and Global Trends in Simulation in Medical Education: Bibliometric and Visual Analysis (e71844)
Hongjun Ba, Lili Zhang, Xiufang He, Shujuan Li. . . . 1591

Global Disparities in Simulation-Based Learning Performance: Serial Cross-Sectional Mixed Methods Study (e52332)
Kashish Malhotra, Harshin Balakrishnan, Emily Warmington, Vina Soran, Francesca Crowe, Dengyi Zhou, SIMBA AND CoMICs Team, Punith Kempegowda. . . . 1601

Health Care Professionals' Knowledge, Attitude, Practice, and Infrastructure Accessibility for e-Learning in Ethiopia: Cross-Sectional Study (e65598)
Sophie Rossner, Muluken Gizaw, Sefonias Getachew, Eyerusalem Getachew, Alemnew Destaw, Sarah Negash, Lena Bauer, Eva Hermann, Abel Shita, Susanne Unverzagt, Pablo Santos, Eva Kantelhardt, Eric Kroeber. . . . 1612

Effectiveness of an Interactive Web-Based Clinical Practice Monitoring System on Enhancing Motivation in Clinical Learning Among Undergraduate Nursing Students: Longitudinal Quasi-Experimental Study in Tanzania (e45912)
Patricia Herman, Stephen M Kibusi, Walter C Millanzi. . . . 1627

Deconstructing Participant Behaviors in Virtual Reality Simulation: Ethnographic Analysis (e65886)
Daniel Loeb, Jamie Shoemaker, Kelly Ely, Matthew Zackoff. . . . 1647

A Virtual Simulator to Improve Weight-Related Communication Skills for Health Care Professionals: Mixed Methods Pre-Post Pilot Feasibility Study (e65949)
Fiona Quigley, Leona Ryan, Raymond Bond, Toni McAloon, Huiru Zheng, Anne Moorhead. . . . 1660

Media-Induced and Psychological Factors That Foster Empathy Through Virtual Reality in Nursing Education: 2×2 Between-Subjects Experimental Study (e59083)
Kuo-Ting Huang, Zexin Ma, Lan Yao. . . . 1672

Case-Based Virtual Reality Simulation for Severe Pelvic Trauma Clinical Skill Training in Medical Students: Design and Pilot Study (e59850)
Peng Teng, Youran Xu, Kaoliang Qian, Ming Lu, Jun Hu. . . . 1683

Extended Reality–Enhanced Mental Health Consultation Training: Quantitative Evaluation Study (e64619)
Katherine Hiley, Zanib Bi-Mohammad, Luke Taylor, Rebecca Burgess-Dawson, Dominic Patterson, Devon Puttick-Whiteman, Christopher Gay, Janette Hiscoe, Chris Munsch, Sally Richardson, Mark Knowles-Lee, Celia Beecham, Neil Ralph, Arunangsu Chatterjee, Ryan Mathew, Faisal Mushtaq. . . . 1700

Exploring the Impact of the COVID-19 Pandemic on Learning Experience, Mental Health, Adaptability, and Resilience Among Health Informatics Master's Students: Focus Group Study (e63708)
Nadia Davoody, Natalia Stathakarou, Cara Swain, Stefano Bonacina. . . . 1718

Feedback From Dental Students Using Two Alternate Coaching Methods: Qualitative Focus Group Study (e68309)
Lulwah Alreshaid, Rana Alkattan. . . . 1741

Alignment Between Classroom Education and Clinical Practice of Root Canal Treatment Among Dental Practitioners in China: Cross-Sectional Study (e65534)
XinYue Ma, JingShi Huang. . . . 1751

Enhancing Preclinical Training for Removable Partial Dentures Through Participatory 3D Simulation: Development and Usability Study (e71743)
Yikchi Siu, Hefei Bai, Jung-Min Yoon, Hongqiang Ye, Yunsong Liu, Yongsheng Zhou. . . . 1763

A Large-Scale Multispecialty Evaluation of Web-Based Simulation in Medical Microbiology Laboratory Education: Randomized Controlled Trial (e72495)
Lei Xu, Xichuan Deng, Tingting Chen, Nan Lu, Yuran Wang, Jia Liu, Yanan Guo, Zeng Tu, Yuxin Nie, Yeganeh Hosseini, Yonglin He. 1776

Open-Access Web-Based Gamification in Pharmacology Education for Medical Students: Quasi-Experimental Study (e73666)
Lujain Aloum, Halah Ibrahim, Senthil Rajasekaran, Eman Alefishat. 1787

Guidelines for Rapport-Building in Telehealth Videoconferencing: Interprofessional e-Delphi Study (e76260)
Paula Koppel, Jennie De Gagne, Michelle Webb, Denise Nepveux, Janelle Bludorn, Aviva Emmons, Paige Randall, Neil Prose. 1809

The Evolution of Medical Student Competencies and Attitudes in Digital Health Between 2016 and 2022: Comparative Cross-Sectional Study (e67423)
Paula Veikkolainen, Timo Tuovinen, Petri Kulmala, Erika Jarva, Jonna Juntunen, Anna-Maria Tuomikoski, Merja Männistö, Teemu Pihlajasalo, Jarmo Reponen. 1827

Refining Established Practices for Research Question Definition to Foster Interdisciplinary Research Skills in a Digital Age: Consensus Study With Nominal Group Technique (e56369)
Jana Sedlakova, Mina Staniki , Felix Gille, Jürgen Bernard, Andrea Horn, Markus Wolf, Christina Haag, Joel Floris, Gabriela Morgenshtern, Gerold Schneider, Aleksandra Zumbrunn Wojczy ska, Corine Mouton Dorey, Dominik Ettlin, Daniel Gero, Thomas Friemel, Ziyuan Lu, Kimon Papadopoulos, Sonja Schläpfer, Ning Wang, Viktor von Wyl. 1841

Global Health care Professionals' Perceptions of Large Language Model Use In Practice: Cross-Sectional Survey Study (e58801)
Ecem Ozkan, Aysun Tekin, Mahmut Ozkan, Daniel Cabrera, Alexander Niven, Yue Dong. 1856

Quantifying Emergency Medicine Residency Learning Curves Using Natural Language Processing: Retrospective Cohort Study (e82326)
Carl Preiksaitis, Joshua Hughes, Rana Kabeer, William Dixon, Christian Rose. 1869

AI-Generated "Slop" in Online Biomedical Science Educational Videos: Mixed Methods Study of Prevalence, Characteristics, and Hazards to Learners and Teachers (e80084)
Eric Jones, Jane Newman, Boyun Kim, Emily Fogle. 1882

Application of AI Communication Training Tools in Medical Undergraduate Education: Mixed Methods Feasibility Study Within a Primary Care Context (e70766)
Chris Jacobs, Hans Johnson, Nina Tan, Kirsty Brownlie, Richard Joiner, Trevor Thompson. 1901

Large Language Models for the National Radiological Technologist Licensure Examination in Japan: Cross-Sectional Comparative Benchmarking and Evaluation of Model-Generated Items Study (e81807)
Toshimune Ito, Toru Ishibashi, Tatsuya Hayashi, Shinya Kojima, Kazumi Sogabe. 1915

Novel Blended Learning on Artificial Intelligence for Medical Students: Qualitative Interview Study (e65220)
Zoe Oftring, Kim Deutsch, Daniel Tolks, Florian Jungmann, Sebastian Kuhn. 1926

Chatbots' Role in Generating Single Best Answer Questions for Undergraduate Medical Student Assessment: Comparative Analysis (e69521)
Enjy Abouzeid, Rita Wassef, Ayesha Jawwad, Patricia Harris. 1942

Role of Artificial Intelligence in Surgical Training by Assessing GPT-4 and GPT-4o on the Japan Surgical Board Examination With Text-Only and Image-Accompanied Questions: Performance Evaluation Study (e69313)
Hiroki Maruyama, Yoshitaka Toyama, Kentaro Takanami, Kei Takase, Takashi Kamei. 1953

Performance Evaluation of 18 Generative AI Models (ChatGPT, Gemini, Claude, and Perplexity) in 2024 Japanese Pharmacist Licensing Examination: Comparative Study (e76925)
Hiroyasu Sato, Katsuhiko Ogasawara, Hidehiko Sakurai. . . . 1963

Perception of Medical Undergraduates on Artificial Intelligence in Medical Education: Qualitative Exploration (e73798)
Thilanka Seneviratne, Kaumudee Kodikara, Isuru Abeykoon, Wathsala Palpola. . . . 1979

Impact of Prompt Engineering on the Performance of ChatGPT Variants Across Different Question Types in Medical Student Examinations: Cross-Sectional Study (e78320)
Ming-Yu Hsieh, Tzu-Ling Wang, Pen-Hua Su, Ming-Chih Chou. . . . 1990

ChatGPT in Medical Education: Bibliometric and Visual Analysis (e72356)
Yuning Zhang, Xiaolu Xie, Qi Xu. . . . 1999

AI's Accuracy in Extracting Learning Experiences From Clinical Practice Logs: Observational Study (e68697)
Takeshi Kondo, Hiroshi Nishigori. . . . 2027

ChatGPT's Performance on Portuguese Medical Examination Questions: Comparative Analysis of ChatGPT-3.5 Turbo and ChatGPT-4o Mini (e65108)
Filipe Prazeres. . . . 2041

Performance of ChatGPT-4 on Taiwanese Traditional Chinese Medicine Licensing Examinations: Cross-Sectional Study (e58897)
Liang-Wei Tseng, Yi-Chin Lu, Liang-Chi Tseng, Yu-Chun Chen, Hsing-Yu Chen. . . . 2049

Generative Artificial Intelligence in Medical Education—Policies and Training at US Osteopathic Medical Schools: Descriptive Cross-Sectional Survey (e58766)
Tsunagu Ichikawa, Elizabeth Olsen, Arathi Vinod, Noah Glenn, Karim Hanna, Gregg Lund, Stacey Pierce-Talsma. . . . 2066

Assessing Familiarity, Usage Patterns, and Attitudes of Medical Students Toward ChatGPT and Other Chat-Based AI Apps in Medical Education: Cross-Sectional Questionnaire Study (e63065)
Safia Elhassan, Muhammad Sajid, Amina Syed, Sidrah Fathima, Bushra Khan, Hala Tamim. . . . 2074

Performance Evaluation and Implications of Large Language Models in Radiology Board Exams: Prospective Comparative Analysis (e64284)
Boxiong Wei. . . . 2082

Factors Associated With the Accuracy of Large Language Models in Basic Medical Science Examinations: Cross-Sectional Study (e58898)
Naritsaret Kaewboonlert, Jiraphon Poontananggul, Natthipong Pongsuwan, Gun Bhakdisongkhram. . . . 2089

Enhancing Medical Student Engagement Through Cinematic Clinical Narratives: Multimodal Generative AI–Based Mixed Methods Study (e63865)
Tyler Bland. . . . 2099

Perceptions and Earliest Experiences of Medical Students and Faculty With ChatGPT in Medical Education: Qualitative Study (e63400)
Noura Abouammoh, Khalid Alhasan, Fadi Aljamaan, Rupesh Raina, Khalid Malki, Ibraheem Altamimi, Ruaim Muaygil, Hayfaa Wahabi, Amr Jamal, Ali Alhaboob, Rasha Assiri, Jaffar Al-Tawfiq, Ayman Al-Eyadhy, Mona Soliman, Mohamad-Hani Temsah. . . . 2111

Detecting Artificial Intelligence–Generated Versus Human-Written Medical Student Essays: Semirandomized Controlled Study (e62779)
Berin Doru, Christoph Maier, Johanna Busse, Thomas Lücke, Judith Schönhoff, Elena Enax- Krumova, Steffen Hessler, Maria Berger, Marianne Tokic. . . . 2123

AIFM-ed Curriculum Framework for Postgraduate Family Medicine Education on Artificial Intelligence: Mixed Methods Study (e66828)
Raymond Tolentino, Fanny Hersson-Edery, Mark Yaffe, Samira Abbasgholizadeh-Rahimi. . . . 2137

Automated Evaluation of Reflection and Feedback Quality in Workplace-Based Assessments by Using Natural Language Processing: Cross-Sectional Competency-Based Medical Education Study (e81718)
Jeng-Wen Chen, Hai-Lun Tu, Chun-Hsiang Chang, Wei-Chung Hsu, Pa-Chun Wang, Chun-Hou Liao, Mingchih Chen. . . . 2152

How AI Is Transforming Medical Education: Bibliometric Analysis (e75911)
Youyang Wang, Chuheng Chang, Wen Shi, Huiting Liu, Xiaoming Huang, Yang Jiao. . . . 2165

Evaluating ChatGPT-4o as an Educational Support Tool for the Emergency Management of Dental Trauma: Randomized Controlled Study Among Students (e80576)
Franziska Haupt, Tina Rödig, Paula Liersch. . . . 2180

Assessing Pharmacists' Use and Perception of AI Chatbots in Pharmacy Practice: Cross-Sectional Survey Study (e71767)
Anly Li, Amy Sheehan, Christopher Giuliano, Paul Dobry, Paul Walker, Jennifer Philips, Joseph Jordan. . . . 2192

Enhancing Large Language Models for Improved Accuracy and Safety in Medical Question Answering: Comparative Study (e70190)
Dingqiao Wang, Jinguo Ye, Jingni Li, Jiangbo Liang, Qikai Zhang, Qiuling Hu, Caineng Pan, Dongliang Wang, Zhong Liu, Wen Shi, Mengxiang Guo, Fei Li, Wei Du, Ying-Feng Zheng. . . . 2204

Large Language Model–Based Patient Simulation to Foster Communication Skills in Health Care Professionals: User-Centered Development and Usability Study (e81271)
Ahmed Elhilali, Andy Ngo, Daniel Reichenpfader, Kerstin Denecke. . . . 2218

Leveraging Datathons to Teach AI in Undergraduate Medical Education: Case Study (e63602)
Michael Yao, Lawrence Huang, Emily Leventhal, Clara Sun, Steve Stephen, Lathan Liou. . . . 2241

Effect of Immersive Virtual Reality Teamwork Training on Safety Behaviors During Surgical Cases: Nonrandomized Intervention Versus Controlled Pilot Study (e66186)
Lukasz Mazur, Logan Butler, Cody Mitchell, Shaian Lashani, Shawna Buchanan, Christi Fenison, Karthik Adapa, Xianming Tan, Selina An, Jin Ra. . . . 2254

Impact of Clinical Decision Support Systems on Medical Students' Case-Solving Performance: Comparison Study with a Focus Group (e55709)
Marco Montagna, Filippo Chiabrando, Rebecca De Lorenzo, Patrizia Rovere Querini, Medical Students. . . . 2275

Comparing the Effectiveness of Multimodal Learning Using Computer-Based and Immersive Virtual Reality Simulation–Based Interprofessional Education With Co-Debriefing, Medical Movies, and Massive Online Open Courses for Mitigating Stress and Long-Term Burnout in Medical Training: Quasi-Experimental Study (e70726)
Sirikanyawan Srikasem, Sunisa Seephom, Atthaphon Viriyopase, Phanupong Phutrakool, Sirhavich Khowinthaseth, Khuansiri Narajeenron, ER-VIPE Study Group. . . . 2285

Correlation Between Electroencephalogram Brain-to-Brain Synchronization and Team Strategies and Tools to Enhance Performance and Patient Safety Scores During Online Hexad Virtual Simulation-Based Interprofessional Education: Cross-Sectional Correlational Study (e69725)
Atthaphon Viriyopase, Khuansiri Narajeenron. . . . 2326

Reviews

Applications, Challenges, and Prospects of Generative Artificial Intelligence Empowering Medical Education: Scoping Review (e71125)

Yuhang Lin, Zhiheng Luo, Zicheng Ye, Nuoxi Zhong, Lijian Zhao, Long Zhang, Xiaolan Li, Zetao Chen, Yijia Chen. 78

How Learning Styles Characterize Medical Students, Surgical Residents, Medical Staff, and General Surgery Teachers While Learning Surgery: Scoping Review (e66766)

Gabriela Gouvea Silva, Marco Ribeiro Filho, Carlos da Silva Costa, Stela Pedroso Vilela Torres de Carvalho, Joao de Souza Menezes, Matheus Querino da Silva, William Donega Martinez, Bruno Cardoso Goncalves, Natália Almeida de Arnaldo Silva Rodriguez Castro, Luiz Vianney Cidrão Nunes, Emerson Santos, Helena Landim Gonçalves Cristóvão, Alexandre Lins Werneck, Alex Bertolazzo Quitério, Sonia Maciel Lopes, Denise Vaz-Oliani, Fernando Facio, Patrícia da Silva Fucuta, Alba de Abreu Lima, Vania Brienze, Heloisa Caldas, Julio Andre. 108

Technology Acceptance Model in Medical Education: Systematic Review (e67873)

Jason Lee, Jenelle Tan, Fernando Bello. 118

Virtual Simulation Tools for Communication Skills Training in Health Care Professionals: Literature Review (e63082)

Manuel Fernández-Alcántara, Silvia Escribano, Rocío Juliá-Sanchis, Ana Castillo-López, Antonio Pérez-Manzano, M Macur, Sedina Kalender-Smajlovi , Sofía García-Sanjuán, María Cabañero-Martínez. 141

Virtual Simulated Placements in Health Care Education: Scoping Review (e58794)

Juliana Samson, Marc Gilbey, Natasha Taylor, Rosie Kneafsey. 159

Multidisciplinary Oncology Education Among Postgraduate Trainees: Systematic Review (e63655)

Houman Tahmasebi, Gary Ko, Christine Lam, Idil Bilgen, Zachary Freeman, Rhea Varghese, Emma Reel, Marina Englesakis, Tulin Cil. 175

Bridging Gaps in Telemedicine Education in Romania to Support Future Health Care: Scoping Review (e66458)

Mircea Focsa, Virgil Rotaru, Octavian Andronic, Marius Marginean, Sorin Florescu. 192

AI in the Health Sector: Systematic Review of Key Skills for Future Health Professionals (e58161)

Javier Gazquez-Garcia, Carlos Sánchez-Bocanegra, Jose Sevillano. 204

Motivation Theories and Constructs in Experimental Studies of Online Instruction: Systematic Review and Directed Content Analysis (e64179)

Adam Gavarkovs, Erin Miller, Jaimie Coleman, Tharsiga Gunasegaran, Rashmi Kusurkar, Kulamakan Kulasegaram, Melanie Anderson, Ryan Brydges. 216

Online-Based and Technology-Assisted Psychiatric Education for Trainees: Scoping Review (e64773)

Mohd Mohd Kassim, Sidi Azli Shah, Jane Lim, Tuti Mohd Daud. 230

Gender Equality Training for Students in Higher Education: Scoping Review (e60061)

Claire Condron, Mide Power, Midhun Mathew, Siobhan Lucey, Patrick Henn, Tanya Dean, Michelle Kirrane Scott, Walter Eppich, Siobhan Lucey. 251

Applications of Artificial Intelligence in Psychiatry and Psychology Education: Scoping Review (e75238)

Julien Prégent, Van-Han-Alex Chung, Inès El Adib, Marie Désilets, Alexandre Hudon. 270

Effects of the Hidden Curriculum in Medical Education: Scoping Review (e68481)

Sebastian Parra Larrotta, Erwin Hernández Rincón, Daniela Niño Correa, Claudia Jaimes Peñuela, Alvaro Romero Tapia. 282

Evaluating the Potential and Accuracy of ChatGPT-3.5 and 4.0 in Medical Licensing and In-Training Examinations: Systematic Review and Meta-Analysis (e68070)

Anila Jaleel, Umair Aziz, Ghulam Farid, Muhammad Zahid Bashir, Tehmasp Mirza, Syed Khizar Abbas, Shiraz Aslam, Rana Sikander. 293

Barriers and Enablers to the Production of Open Access Medical Education Platforms: Scoping Review (e65306)
Ahmed Ahmed, Arushi Biswas, Nefti Bempong-Ahun, Ines Peri , Eric O'Flynn. 327

Viewpoints

What Are the Opportunities and Challenges of Using AI in Medical Education in Vietnam? (e77817)
Trung Nguyen, Thanh Nguyen, Duy Nguyen, Anh Vu, Khanh Dang, Nhu Le, Duy Ngo, Dang Nguyen, Van Hoang, Thanh Ngo. 344

The Need for Health Care Innovation Training in Medical Education (e79489)
Lily Zhu, Jeffrey Khong, Oren Wei, Katherine Chretien, Youseph Yazdi. 356

Leveraging Generative Artificial Intelligence to Improve Motivation and Retrieval in Higher Education Learners (e59210)
Noahlana Monzon, Franklin Hays. 363

Quo Vadis, AI-Empowered Doctor? (e70079)
Gary Takahashi, Laurentius von Liechti, Ebrahim Tarshizi. 374

Shaping the Future of Digital Health Education in Canada: Prioritizing Competencies for Health Care Professionals Using the Quintuple Aim (e75904)
Glynda Rees, Lorelli Nowell, Tracie Risling. 382

An Ecosystem Approach to Developing and Implementing a Cocreated Bachelor's Degree in Digital Health and Biomedical Innovation (e63903)
Patrícia Alves, Elisio Costa, Altamiro Costa-Pereira, Inês Falcão-Pires, João Fonseca, Adelino Leite-Moreira, Bernardo Sousa-Pinto, Nuno Vale. 397

Beyond Chatbots: Moving Toward Multistep Modular AI Agents in Medical Education (e76661)
Minyang Chow, Olivia Ng. 406

From Hype to Implementation: Embedding GPT-4o in Medical Education (e79309)
Sumaia Sabouni, Mohammad-Adel Moufti, Mohamed Taha. 413

Beyond Lectures: Reimagining Psychiatric Didactics for the Age of AI (e78110)
Laurent Elkrief, Alexandre Hudon, Giovanni Briganti, Paul Lespérance. 418

Digital Dentists: A Curriculum for the 21st Century (e54153)
Michelle Mun, Samantha Byrne, Louise Shaw, Kayley Lyons. 1734

Tutorials

Fostering Multidisciplinary Collaboration in Artificial Intelligence and Machine Learning Education: Tutorial Based on the AI-READI Bootcamp (e83154)
Taiki Nishihara, Fritz Kalaw, Adelle Engmann, Aya Motoyoshi, Paapa Mensah-Kane, Deepa Gupta, Victoria Patronilo, Linda Zangwill, Shahin Hallaj, Amirhossein Panahi, Garrison Cottrell, Bradley Voytek, Virginia de Sa, Sally Baxter. 424

Creation of the ECHO Idaho Podcast: Tutorial and Pilot Assessment (e55313)
Ryan Wiet, Madeline Casanova, Jonathan Moore, Sarah Deming, Russell Baker Jr. 436

Cardiac Implantable Electronic Device Educational Application for Cardiac Anesthesiology Trainees: Tutorial on App Development (e60087)
Ahmed Zaky, Aisha Waheed, Brittany Hatter, Srilakshmi Malempati, Sai Maremalla, Ragib Hasan, Yuliang Zheng, Scott Snyder. 444

Enhancing Access to Neuraxial Ultrasound Phantoms for Medical Education of Pediatric Anesthesia Trainees: Tutorial (e63682)
Leah Webb, Melissa Masaracchia, Kim Strupp. 459

Designing Personalized Multimodal Mnemonics With AI: A Medical Student's Implementation Tutorial (e67926)
Noor Elabd, Zafirah Rahman, Salma Abu Alinnin, Samiyah Jahan, Luciana Campos, Ovidiu Baltatu. 467

Faculty Retreats in Academic Medicine: Tutorial (e71622)
Rachel Skains, Julie Brown, Erin Shufflebarger, Justine McGiboney, Sherell Hicks, Laine McDonald, Katherine Griesmer, Christine Shaw, Emily Grass, Marie-Carmelle Elie, Lauren Walter. 475

Twelve Practical Tips for Integrating AI Into Medical Education: Tutorial to Support Educators Across Teaching, Research, Administration, and Ethical Domains (e81297)
Alireza Jalali, Kadidja Harbi Houssein, Salomon Fotsing. 490

Letters to the Editor

Authors' Reply: Enhancing AI-Driven Medical Translations: Considerations for Language Concordance (e71721)
Joyce Teng, Roberto Novoa, Maria Aleshin, Jenna Lester, Kira Seiger, Fiatsogbe Dzuali, Roxana Daneshjou. 1797

Author's Reply: Examining Multimodal AI Resources in Medical Education: The Role of Immersion, Motivation, and Fidelity in AI Narrative Learning (e72336)
Tyler Bland. 1799

Examining Multimodal AI Resources in Medical Education: The Role of Immersion, Motivation, and Fidelity in AI Narrative Learning (e72190)
Chris Jacobs. 1801

Enhancing AI-Driven Medical Translations: Considerations for Language Concordance (e70420)
Stephanie Quon, Sarah Zhou. 1803

Citation Accuracy Challenges Posed by Large Language Models (e72998)
Manlin Zhang, Tianyu Zhao. 1805

Authors' Reply: Citation Accuracy Challenges Posed by Large Language Models (e73698)
Mohamad-Hani Temsah, Ayman Al-Eyadhy, Amr Jamal, Khalid Alhasan, Khalid Malki. 1807

Corrigenda and Addenda

Correction: Comparing the Perceived Realism and Adequacy of Venipuncture Training on an in-House Developed 3D-Printed Arm With a Commercially Available Arm: Randomized, Single-Blind, Cross-Over Study (e89670)
Susan Brouwer de Koning, Amy Hofman, Sonja Gerber, Vera Lagerburg, Michelle van den Boorn. 2232

Editorial

Advantages of a Virtual Collaborative Research Dermatology Laboratory (e65697)
Natasha Barton, Kenny Ta, Angela Loczi-Storm, Cory Dunnick, Robert Dellavalle. 2234

Research Letters

Assessment of Large Language Model Performance on Medical School Essay-Style Concept Appraisal Questions: Exploratory Study (e72034)
Seysha Mehta, Eliot Haddad, Indira Burke, Alana Majors, Rie Maeda, Sean Burke, Abhishek Deshpande, Amy Nowacki, Christina Lindenmeyer, Neil Mehta. 2265

Perceptions and Intentions to Use Generative AI Among First-Year Medical Students in Japan: Cross-Sectional Survey Study (e77552)
Hiroshi Tajima, Hajime Kasai, Kiyoshi Shikino, Ikuo Shimizu, Shoichi Ito. 2269

Commentary

Transforming Medical Education to Make Patient Safety Part of the Genome of a Modern Health Care Worker (e68046)
Peter Lachman, John Fitzsimons. 2350

Development and Validation of a Large Language Model–Based System for Medical History-Taking Training: Prospective Multicase Study on Evaluation Stability, Human-AI Consistency, and Transparency

Yang Liu[1*], MM; Chujun Shi[1*], Prof Dr; Liping Wu[1], Prof Dr; Xiule Lin[1], MM; Xiaoqin Chen[1], MM; Yiying Zhu[1], MSc; Haizhu Tan[2*], Prof Dr, MD; Weishan Zhang[1*], MSc

[1]Medical Simulation Center, Shantou University Medical College, No. 22 Xinling Road, Shantou, China

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

[*]these authors contributed equally

Corresponding Author:
Weishan Zhang, MSc
Medical Simulation Center, Shantou University Medical College, No. 22 Xinling Road, Shantou, China

Abstract

Background: History-taking is crucial in medical training. However, current methods often lack consistent feedback and standardized evaluation and have limited access to standardized patient (SP) resources. Artificial intelligence (AI)–powered simulated patients offer a promising solution; however, challenges such as human-AI consistency, evaluation stability, and transparency remain underexplored in multicase clinical scenarios.

Objective: This study aimed to develop and validate the AI-Powered Medical History-Taking Training and Evaluation System (AMTES), based on DeepSeek-V2.5 (DeepSeek), to assess its stability, human-AI consistency, and transparency in clinical scenarios with varying symptoms and difficulty levels.

Methods: We developed AMTES, a system using multiple strategies to ensure dialog quality and automated assessment. A prospective study with 31 medical students evaluated AMTES's performance across 3 cases of varying complexity: a simple case (cough), a moderate case (frequent urination), and a complex case (abdominal pain). To validate our design, we conducted systematic baseline comparisons to measure the incremental improvements from each level of our design approach and tested the framework's generalizability by implementing it with an alternative large language model (LLM) Qwen-Max (Qwen AI; version 20250409), under a zero-modification condition.

Results: A total of 31 students practiced with our AMTES. During the training, students generated 8606 questions across 93 history-taking sessions. AMTES achieved high dialog accuracy: 98.6% (SD 1.5%) for cough, 99.0% (SD 1.1%) for frequent urination, and 97.9% (SD 2.2%) for abdominal pain, with contextual appropriateness exceeding 99%. The system's automated assessments demonstrated exceptional stability and high human-AI consistency, supported by transparent, evidence-based rationales. Specifically, the coefficients of variation (CV) were low across total scores (0.87% - 1.12%) and item-level scoring (0.55% - 0.73%). Total score consistency was robust, with the intraclass correlation coefficients (ICCs) exceeding 0.923 across all scenarios, showing strong agreement. The item-level consistency was remarkably high, consistently above 95%, even for complex cases like abdominal pain (95.75% consistency). In systematic baseline comparisons, the fully-processed system improved ICCs from 0.414/0.500 to 0.923/0.972 (moderate and complex cases), with all CVs ≤1.2% across the 3 cases. A zero-modification implementation of our evaluation framework with an alternative LLM (Qwen-Max) achieved near-identical performance, with the item-level consistency rates over 94.5% and ICCs exceeding 0.89. Overall, 87% of students found AMTES helpful, and 83% expressed a desire to use it again in the future.

Conclusions: Our data showed that AMTES demonstrates significant educational value through its LLM-based virtual SPs, which successfully provided authentic clinical dialogs with high response accuracy and delivered consistent, transparent educational feedback. Combined with strong user approval, these findings highlight AMTES's potential as a valuable, adaptable, and generalizable tool for medical history-taking training across various educational contexts.

(JMIR Med Educ 2025;11:e73419) doi:10.2196/73419

KEYWORDS

large language models; medical history-taking; structured evaluation; evaluation stability; human-AI consistency; evaluation transparency; virtual standardized patient; DeepSeek; Qwen; cross-model generalizability

Introduction

History-taking is fundamental to clinical practice and one of the clinicians' most frequently performed tasks [1]. Although technological advances in assessing patients have proliferated, history-taking remains the most crucial, cost-effective technique [2]. Therefore, enhanced training in medical history-taking is crucial for both improving disease diagnosis accuracy [3-5] and fostering the development of competent physicians [6].

Standardized patients (SPs) effectively teach and evaluate history-taking skills by ensuring structured learning experiences. The training process for SPs is rigorous, time-consuming, and resource-intensive [7,8]. Consequently, the availability of qualified SPs is limited [7]. During SP-based teaching and evaluation, subjective factors are an unavoidable influence [9]. Existing literature suggests SP feedback is highly variable in terms of its content and quality [10,11]. These factors pose a significant challenge to implementing effective one-on-one history-taking training.

The rapid evolution of artificial intelligence (AI) technology, especially with the emergence of large language models (LLMs), has demonstrated significant potential in medical education [12-16]. LLMs can act as virtual standardized patients (VSPs) [12,17-19] and create a human-like conversational experience [12,20]. In addition, the web-based system is accessible at any time, allows repeated practice, and significantly reduces teaching costs. However, feedback is the cornerstone of medical education and is crucial for the continuous learning of trainees [21]. Therefore, a system that only provides interactive practice without structured feedback may fail to meet the instructional needs.

Encouragingly, LLMs can also provide instant feedback. A recent single-case study has shown that an LLM-powered VSP can not only provide accurate interactions but also implement structured evaluations with high human-AI consistency for most assessment items, while identifying a subset of items where further alignment could significantly enhance performance [22]. The inherent scalability and personalization capabilities of LLMs may address the inefficiencies and inconsistencies associated with traditional feedback, holding the promise of democratizing high-quality learning experiences. These findings offer preliminary evidence for applying LLMs to evaluate medical history-taking training.

Despite these promising developments, implementing AI as VSPs presents several challenges and technical limitations, particularly in providing real-time educational feedback [23]. First, while some preliminary studies have explored LLMs in medical history-taking evaluation, there remains a gap in research investigating multiple cases of varying difficulty levels across different clinical scenarios. Specifically, given the diversity of disease types in clinical practice and the variations in content and evaluation standards for history-taking across different diseases, further exploration is required to assess their applicability and effectiveness in a broader range of clinical settings.

Second, LLMs may generate "hallucinations," producing information that appears reasonable but is incorrect [24], leading to different responses to the same query. Even the most advanced models may struggle to handle complex or highly specialized inputs, affecting the accuracy of their outputs. Furthermore, their decision-making process lacks transparency, resulting in users unable to understand the basis on which they draw conclusions, presenting a "black box" problem [25,26]. This black box characteristic raises doubts about their evaluation results. Research shows that the transparency and credibility of evaluation feedback significantly influence learner acceptance, particularly for AI-generated feedback, which requires clear evaluation criteria and frameworks [23].

Moreover, these risks highlight the importance of adding an extra layer of validation, especially for complex teaching tasks that directly affect patient diagnosis and treatment. It is evident that current LLMs cannot yet be solely relied upon in the fields of education and research. Therefore, to identify and prevent these safety risks, developing a systematic assessment system with broad adaptability is particularly important [27]. Consequently, when developing LLM-powered evaluation systems, special attention must be paid to making the evaluation process transparent and standardized. Finally, the stability of LLM-generated evaluations remains to be characterized, which is a critical factor for long-term application in teaching.

To address these gaps, we developed the AI-Powered Medical History-Taking Training and Evaluation System (AMTES). This study describes its development and rigorously evaluates its dialog quality, evaluation stability, human-AI consistency, and transparency across 3 clinical cases of varying complexity. We also validated our design through systematic baseline and cross-model comparisons.

Methods

System Design and Implementation of AMTES

Overview of AMTES

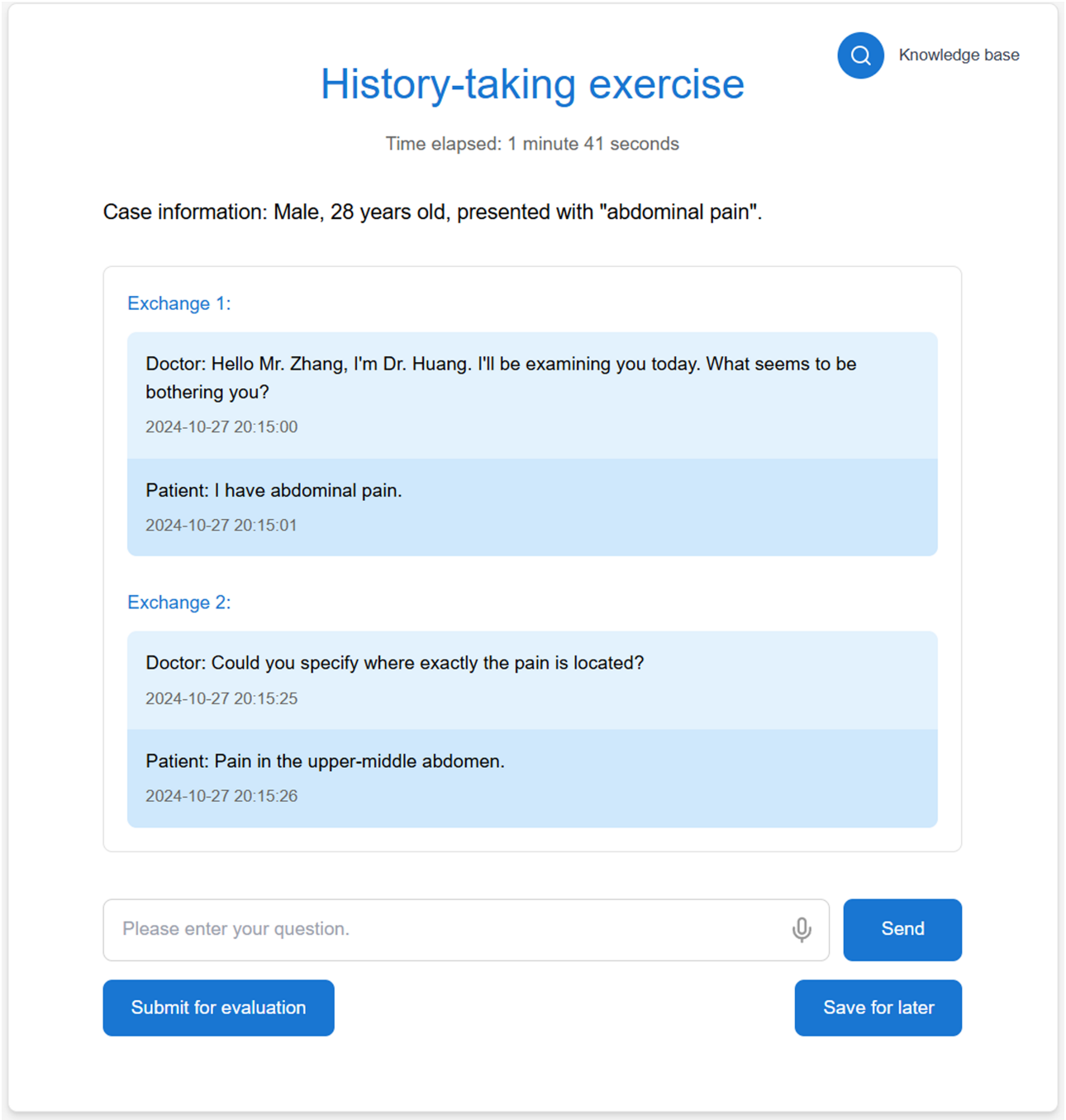
AMTES is a web-based system developed using the ASP.NET framework (Microsoft) with a Browser/Server architecture. The system seamlessly integrates the DeepSeek-V2.5 (DeepSeek) application programming interface (API), a sophisticated Chinese LLM. This model was selected after demonstrating superior performance for our specific medical dialog tasks in preliminary tests against other domestic models available during the study's implementation phase, thereby best fulfilling the project's technical and accessibility requirements. DeepSeek-V2.5 features an advanced architecture with 236 billion parameters and supports an extensive context window of up to 128,000 tokens, enabling robust natural language understanding and complex reasoning capabilities across extended conversations. The AMTES software has been granted

the Computer Software Copyright Registration Certificate by the National Copyright Administration of China (Registration No. 14651073).

System Modules

AMTES consists of two core modules: (1) the conversational Dialog Module (Figure 1), which enables multiround history-taking conversations with a VSP while recording all interactions for evaluation; (2) the Automated Evaluation Feedback Module, which analyzes dialog records using LLM technology to generate structured feedback. These 2 modules constitute the core of the system: simulation (conversational dialog) and assessment (automated evaluation).

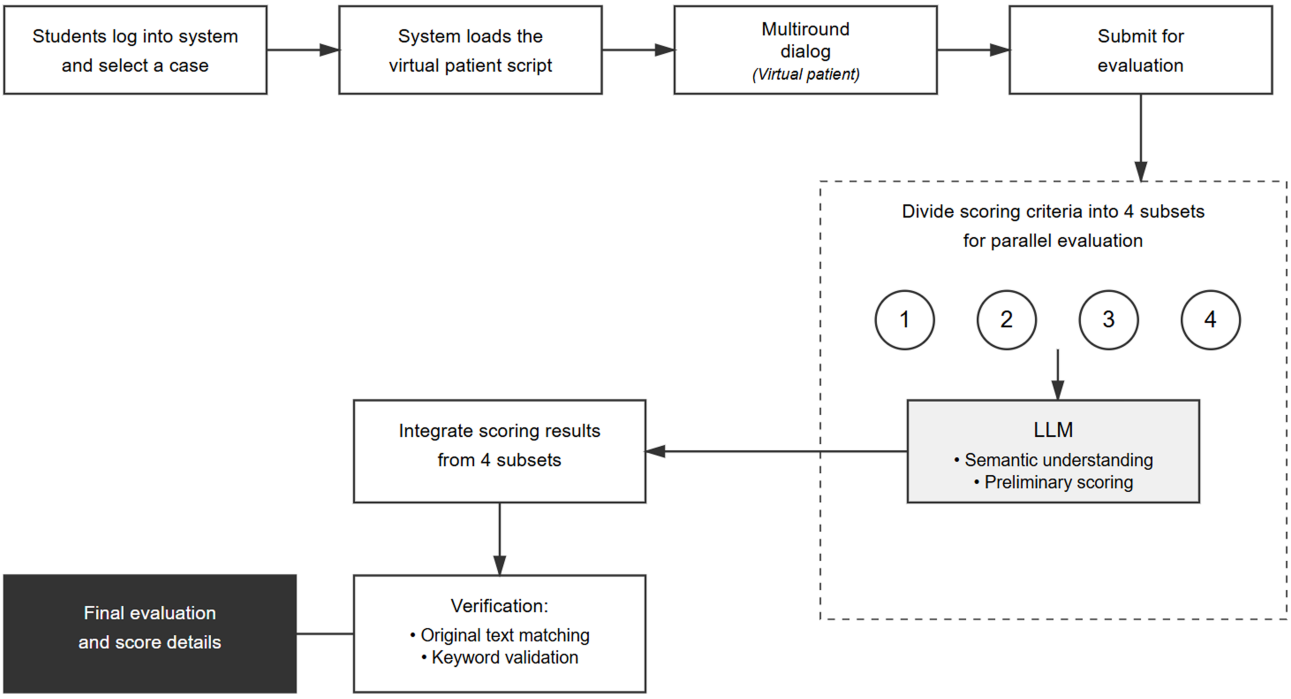
Figure 1. History-taking interface showing the chat environment where students interact with the virtual standardized patient.



System Workflow

The complete system workflow, from student login through final evaluation delivery, is illustrated in Figure 2. A granular, step-by-step description of each stage is provided in Multimedia Appendix 1.

Figure 2. System workflow diagram illustrating the complete medical history-taking training and evaluation process in the Artificial Intelligence–Powered Medical History-Taking Training and Evaluation System. LLM: large language model



System Design Strategies

This subsection elaborates on the underlying principles and innovations guiding AMTES's development.

Framework Objectives and Implementation Strategies

To ensure AMTES meets clinical education requirements, we implemented a comprehensive design framework with specific strategies aimed at achieving system reliability, human-AI consistency, and evaluation transparency (Table 1). System design framework and implementation strategies present this multifaceted framework in a structured format designed to clearly delineate each objective, the strategies used, and their specific implementation methods. These strategies were engineered specifically to address the key LLM challenges of transparency and hallucination identified in the introduction, ensuring the system's reliability and trustworthiness.

Table . System design framework and implementation strategies.

Core objective and strategy	Key methods	Specifications and examples
A. Reliability assurance (ensuring accurate and stable system outputs)		
Minimizing randomness	Lower the temperature parameter of the LLM's[a] API[b]	Dialog: set temperature as 0.05; assessment: set temperature as 0.0
Implementing multilevel verification[c]	1. Original-text matching: verifies whether the LLM-cited dialog exists verbatim in the original dialog 2. Keyword validation: confirms the presence of predefined mandatory keywords within the cited evidence	The specific strategy is in Multimedia Appendix 2
Parallelizing evaluations[d]	Split scoring items into 4 subsets for parallel LLM queries	Benefits: enables evidence-based scoring by circumventing token limits; reduces evaluation time through parallel processing
B. Human-AI Consistency (aligning AI evaluations with expert judgment)		
Decomposition for standardization	Complex scoring items decomposed into smaller, unambiguous sub-items to improve LLM execution accuracy	For example, original: "persistent moderate-to-severe burning abdominal pain"→ 3 sub-items: 1. Pain is burning in nature 2. Pain is moderate to severe 3. Pain is persistent
Disambiguation via guidelines	Detailed evaluation guidelines for contextual differentiation and terminology clarification	For example, context: "initial dull pain" versus "current burning pain"; terminology: "petechiae"=small red hemorrhagic spots
Few-shot examples	Uses in-prompt examples to demonstrate correct scoring	The complete prompt is available in Multimedia Appendix 3
C. Transparency enhancement (making evaluation processes traceable and verifiable)		
Evidence-based scoring[e]	Prompt engineering compels LLM to cite specific dialog text and explain the rationale for transparent decision-making	The complete prompt is available in Multimedia Appendix 3, "Final Output Format" section
Structured feedback	Organizes feedback hierarchically, from overall metrics down to item-level specifics	Hierarchy: overall performance → category analysis → item details → dialog links

[a]LLM: large language model

[b]API: application programming interface

[c]Multilevel verification mechanism (strategy A) represents a key innovation in handling large language model hallucinations through systematic validation protocols. This approach ensures reliability by implementing multiple checkpoints throughout the evaluation process (see Multimedia Appendix 2 for detailed implementation).

[d]Parallel evaluation architecture (strategy A) aims to improve computational efficiency while enabling comprehensive evidence-based scoring within token constraints.

[e]Evidence-based scoring framework ensures full traceability of artificial intelligence decision-making processes through mandatory citation requirements. Every evaluation decision must be supported by explicit dialog evidence.

Strategy Development and Refinement

The design strategies described above were not developed in isolation but emerged through an iterative development process. Over a 2-month period, a multidisciplinary team comprising 4 clinical instructors, 6 medical students, and 2 engineers extensively tested the system with 3 clinical cases. Through this collaborative process, the AMTES system itself, along with the dialog scripts, prompts, scoring standards, and evaluation guidelines, underwent continuous refinement based on practical insights and user feedback. This iterative approach resulted in a comprehensive set of evaluation rules and a well-structured bank of few-shot examples, establishing a solid foundation for the formal validation studies.

Construction of the Clinical Case Bank

Three representative cases were prepared by 5 senior clinical experts, aligned with the national medical licensing examination syllabus. They were chosen as the clinical conditions for the simulation due to their relevance to the material being taught at the time. This alignment ensured that the scenario was both clinically pertinent and integrated with the participants' ongoing coursework in basic sciences and clinical disciplines.

This bank includes 3 cases:

- Case 1: an 18-year-old male presenting with a chief complaint of recurrent cough. (This is a common respiratory system disease, characterized by a short course and typical symptoms. Difficulty: simple.)

- Case 2: a 27-year-old female presenting with a chief complaint of frequent urination. (This case pertains to a common urinary system disease, complicated by patient anxiety and a recurrent history from 6 months previous. Difficulty: moderate.)
- Case 3: a 28-year-old male presenting with a chief complaint of recurrent abdominal pain involving a gastrointestinal system disease. (This case involves a chronic digestive system disease, notable for its recurrent nature and the recent development of complications. Difficulty: complex.)

Each case includes detailed scripts covering the background introduction, patient profile, comprehensive medical history content, and proactively asked questions. The history content comprehensively covers the chief complaint, present illness, past history, personal history, marital history, reproductive history, family history, and, for female patients, menstrual history.

Scoring items were established and points allocated based on the clinical significance and teaching objectives of each medical history segment, with each scoring item clearly corresponding to a single evaluation criterion. The total score is 70 points, with the current medical history accounting for 45 - 50 points. Each scoring item is assigned different scores based on its diagnostic value, ranging from 2 points to 0.5 points. The number of scoring items for the 3 cases was 66 for the cough case, 59 for the frequent urination case, and 67 for the abdominal pain case.

Study Design and Validation Framework

Participants

Between September 2024 and November 2024, 31 third-year medical students (16 females, 15 males) from Shantou University Medical College undergoing diagnostic training were recruited. All participants had received theoretical training on medical history-taking. The inclusion criteria were possession of an electronic device and voluntary participation in the teaching trial. All participants provided informed consent, agreed to the use of their data for research purposes, and were provided with a login account with instructions on using AMTES. No participant was excluded from the final analysis.

Validation Approach

To rigorously evaluate AMTES, we conducted validation in 3 sequential phases, using 3 complementary strategies to comprehensively assess the system's performance, stability, and generalizability. Given the pedagogical imperative to protect students' learning experiences, all participants interacted with the fully processed system. Baseline and generalizability comparisons were conducted retrospectively using stored dialog records, ensuring educational quality while maintaining methodological rigor.

Implementation Phases

Phase 1 (Weeks 1 - 3)

Thrity-one students completed 3 history-taking sessions (cough → frequent urination → abdominal pain) and received immediate feedback from the fully-processed system. Posttraining questionnaires were administered to collect student feedback on system usability.

Phase 2

All 93 dialog records were collected, followed by 9 additional rounds of automated assessment to complete the 10 runs required for the comprehensive performance test. Subsequently, teachers not only manually scored the 93 dialog records based on the rubric but also evaluated the quality of the VSP's conversational dialog responses. The validation work was conducted by 2 senior teachers who were independent of the case development team in order to ensure the reliability of the evaluations.

Phase 3

We rescored the same 93 records to perform 2 main analyses. First, a Systematic Baseline Comparison Test was conducted across 3 versions of system: baseline, core-optimized, and fully processed systems. Second, a Cross-Model Generalizability Test was run using the Qwen-Max (version 20250409) model. Since our participants are native Chinese speakers, all interactions with AMTES were conducted in Chinese. The data and screenshots were then translated into English for presentation.

Validation Strategies Used

The following 3 strategies were used during the implementation phases described above.

Comprehensive Performance Test

The fully processed system was executed 10 times per student record to quantify evaluation stability, human-AI consistency, and transparency.

Systematic Baseline Comparison

To quantify the contribution of our distinct optimization layers, we conducted a baseline comparison. A 3-level approach was necessary because several of our core strategies are technically interdependent and could not be tested in isolation (eg, our evidence-based scoring strategy, which requires extensive dialog citation, is only feasible through parallel subevaluation to avoid exceeding the LLM's token limits; in turn, our final postverification stage actively uses this cited evidence to perform its validation checks). Therefore, we designed the study to compare 3 logically sequenced system versions on identical dialog data, representing distinct stages of optimization.

- Baseline system (minimal prompting): this represents the out-of-the-box performance of the LLM, using only a minimal prompt and sequential processing without any of our custom strategies.
- Core-optimized system (enhanced prompt and parallel processing): this version incorporates our full suite of optimization strategies that are applied during the LLM evaluation process. It includes our entire prompt architecture (eg, structured prompts, few-shot examples, evidence-based scoring; see Multimedia Appendix 3 for the complete integrated prompt) and the parallel subevaluation mechanism, which function as a synergistic whole. This stage is designed to isolate the impact of sophisticated

prompting and system architecture, but it explicitly excludes any postprocessing of the LLM's output.
- Fully processed system (with postprocessing verification): this represents the complete, fully processed AMTES system. It builds directly on core-optimized system by adding the final, critical layer of our multilevel verification mechanisms (ie, original-text matching and keyword validation) to the output generated in the previous stage. Comparing this to the previous system allows us to precisely measure the incremental gain achieved by our postprocessing verification strategies.

Cross-Model Generalizability Test

To assess the robustness and adaptability of our evaluation framework, we implemented the complete AMTES system using Qwen-Max, an alternative LLM, without modifying any prompts, evaluation strategies, or system parameters. This testing aimed to demonstrate that our design approach could generalize across different LLM platforms, which is crucial for educational institutions that may need to adapt to various AI technologies.

Outcome Metrics

To rigorously assess AMTES, we defined the following key outcome metrics and their measurement criteria.

Stability

This evaluates the consistency of AMTES evaluations across 10 repeated assessments. We calculated the coefficient of variation (CV) for total scores and the counts of scoring items where AMTES's evaluation matched human scoring (Human-AMTES Matched Item Counts). CV values ≤10% indicated minimal variation, 10%<CV≤20% indicated moderate variation, and CV >20% indicated significant variation.

Human-AI Consistency

This refers to the degree of agreement between evaluations by human experts and those generated by the AI system (AMTES) for the same student performance. We measured this consistency at 2 levels:

- Total score level: we assessed consistency by examining the intraclass correlation coefficient (ICC) and Pearson *r* between the overall scores assigned by human experts and the AI system, using human scoring as the benchmark. ICC values ≥0.75 were considered indicative of good reliability, and values ≥0.90 were regarded as highly consistent. For Pearson *r*, values of 0.50 - 0.70 were considered moderate, 0.70 - 0.90 indicated strong, and >0.90 indicated very strong.
- Item level: for human-AI consistency in item-level scoring, we quantified it using the following metrics.

$$\text{Mean Difference Items} = |\text{Human Scoring Items} - \text{AI Scoring Items}|$$

$$\text{Item-Level Consistency} = \text{Total Items} - \text{Mean Difference Items} \text{Total Items} \times 100\%$$

Transparency

Transparency was defined qualitatively per report: every scoring item had to (1) include a verbatim dialog citation with its rationale and (2) pass both stages of automated verification (original-text matching and keyword checks).

Statistical Analysis

The study data were analyzed using SPSS 24.0 (SPSS Inc). The Shapiro-Wilk normality test was conducted to examine whether the collected data followed a normal distribution. Values were presented as the mean (SD) for normally distributed data. In terms of dialog quality, the one-way ANOVA was performed to determine differences in the accuracy and appropriateness of AMTES responses across the 3 case scenarios. The CV was calculated to measure the system stability. The consistency at the total score level was assessed using the ICC and Pearson *r*. At the item level, the mean difference items and item-level consistency were calculated to evaluate the average discrepancy. Statistical significance for all tests was set at $P<.05$.

Ethical Considerations

This study received ethical approval from the Ethics Committee of Shantou University Medical College (approval number: SUMC-2024 - 079). All procedures were conducted in accordance with the principles of the Declaration of Helsinki and complied with relevant Chinese laws and institutional ethical standards. Prior to participation, all subjects provided written informed consent. Participants were fully informed of the study's purpose, procedures, and data handling methods and were explicitly told that they could withdraw from the study at any time without penalty. To ensure the privacy and confidentiality of participants, all data were anonymized. Personally identifiable information was removed from the research data, and all files were securely stored in password-protected documents accessible only to the research team. No compensation was provided to the participants for their involvement in this study.

Results

Participant Demographics

There were 16 (52%) females and 15 (48%) males in this study. All participants were enrolled in the same stage of their diagnostics curriculum and had the experience of learning theoretical knowledge but lacked practical experience in simulated patient history-taking. The study achieved a 100% completion rate, with no dropouts, and all participants were included in the final analysis.

Analysis of AMTES's Conversational Dialog Quality

A descriptive analysis was performed on all conversational dialog records between the 31 students and AMTES across 3 clinical case scenarios to assess the quality of AMTES's conversational dialog performance. The students completed a total of 93 history-taking sessions, generating 8606 questions (cough scenario: 2383; frequent urination: 2818; abdominal pain: 3405). The system's response rate was 100%.

The accuracy rates (respond in accordance with the case script) of AMTES's replies were as follows: 98.6% (SD 1.5%) for cough, 99.0% (SD 1.1%) for frequent urination, and 97.9% (SD 2.2%) for abdominal pain. The proportion of contextually appropriate responses from AMTES was consistently high: 99.74% (SD 0.67%) for cough, 99.09% (SD 1.14%) for frequent urination, and 99.36% (SD 1.03%) for abdominal pain.

Despite occasional participant errors, such as spelling mistakes, AMTES demonstrated the ability to accurately interpret and respond to the majority of these questions. Specifically, in the cough scenario, out of 28 erroneous questions, AMTES correctly interpreted and responded to 26; in frequent urination, out of 40 erroneous questions, 32 were correctly handled; and in abdominal pain, 12 out of 15 erroneous questions were correctly addressed by AMTES (Table 2).

Table . Quality analysis of Artificial Intelligence (AI)–Powered Medical History-Taking Training and Evaluation System conversational dialog.

Case	Cough (n=31)	Frequent urination (n=31)	Abdominal pain (n=31)	*P* value
Response accuracy (%)[a], mean (SD)	98.60 (1.5)	99 (1.1)	97.90 (2.2)	.04
Number of questions asked, mean (SD)	76.87 (21.22)	90.9 (21.56)	109.84 (28.83)	<.001
Information appropriateness rate (%)[b], mean (SD)	99.74 (0.67)	99.09 (1.14)	99.36 (1.03)	.03
Number of incorrect student questions, n	28	40	15	—[c]
Number of questions correctly understood and answered by AMTES[d, e], n	26	32	12	—

[a]Response accuracy (%) denotes the proportion of system responses evaluated as entirely correct relative to the total number of valid AI-generated answers in that scenario.

[b]Information appropriateness rate (%) refers to the percentage of system responses deemed relevant and contextually appropriate to the questions asked.

[c]Not available.

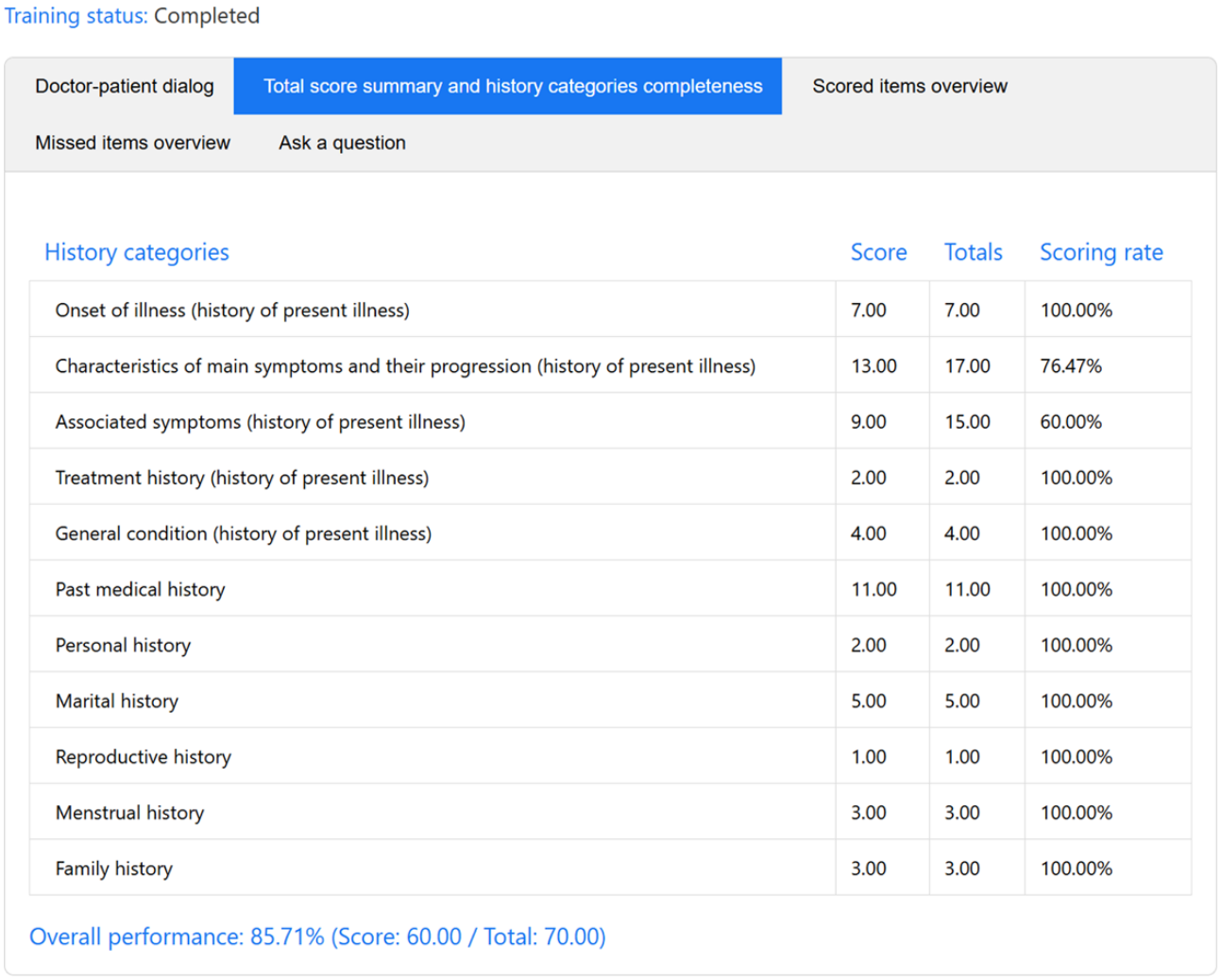
[d]AMTES: Artificial Intelligence–Powered Medical History-Taking Training and Evaluation System.

[e]The number of questions correctly understood and answered by AMTES shows how many of those erroneous questions were still accurately interpreted and properly answered by AMTES.

AMTES Provided Transparent and Structured Evaluation Reports

After automatically evaluating each student's history-taking session, AMTES generated comprehensive feedback reports that included the following components: doctor-patient dialog records, total score per attempt, completeness percentages for each history category, an overview of scored items, and an overview of missed items, thereby comprehensively presenting the evaluation feedback (Figure 3). The history categories included chief complaint, history of present illness, past medical history, personal history, menstrual history, reproductive history, family history, and other relevant sections.

Figure 3. Artificial Intelligence–Powered Medical History-Taking Training and Evaluation System comprehensive feedback report showing the structured evaluation components.



Importantly, across all 93 history-taking records, every feedback report met the predefined transparency criteria. The scoring items displayed specific scoring criteria with clear rationales, containing verbatim dialog citations along with explicit explanations of how the cited text satisfied the corresponding scoring standard (Figure 4). System logs confirmed that all displayed items passed both the original-text matching and keyword checks; items that lacked direct evidence or failed validation were automatically removed and were never shown to students.

Figure 4. Scoring items and rationale display showing how the system justifies scores.

Number	Scoring items	Category	Points earned	Scoring rationale
1	Chief complaint of abdominal pain	Onset of illness (history of present illness)	2	Doctor's question: "Hello Mr. Zhang, I'm Dr. Huang. I'll be examining you today. What seems to be bothering you?" Patient's response: "I have abdominal pain." Scoring rationale: Doctor explicitly inquired about the chief complaint, patient directly reported abdominal pain.
2	Precipitating factor: alcohol consumption at business dinner	Onset of illness (history of present illness)	2	Doctor's question: "Was there anything that might have triggered the abdominal pain a week ago? Patient's response: "I had been drinking heavily at a business dinner." Scoring rationale: Doctor explicitly asked about potential triggers, patient directly identified alcohol consumption.
3	Onset of present abdominal pain was one week ago	Characteristics of main symptoms and their progression (history of present illness)	1	Doctor's question: "When did the pain start?" Patient's answer: "It started a week ago." Scoring rationale: Doctor explicitly asked about onset timing, patient directly specified one week ago.
4	Character of pain described as burning	Characteristics of main symptoms and their progression (history of present illness)	2	Doctor's question: "Can you describe the nature of the pain? For example, is it a burning sensation or a sharp, cutting pain?" Patient's response: "It's a burning pain" Scoring rationale: Doctor explicitly asked about pain character, patient directly identified burning sensation.
5	Pain location in upper-middle abdomen	Characteristics of main symptoms and their progression (history of present illness)	1	Doctor's question: "Could you specify where exactly the pain is located?" Patient's answer: "Pain in the upper-middle abdomen." Patient's response: Doctor explicitly asked about pain location, patient directly specified upper-middle abdomen.
6	Pain intensity moderate to severe, rated 5-6/10	Characteristics of main symptoms and their progression (history of present illness)	2	Doctor's question: Doctor's question: "On a pain scale of 1 to 10, where 10 represents the most severe pain, how would you rate your pain?" Patient's response: "5 to 6" Scoring rationale: Doctor explicitly asked about pain intensity, patient directly rated it as 5-6/10.

Furthermore, AMTES implemented comprehensive logging of the evaluation process, capturing inputs, outputs, scoring results, and error data for each interaction. This logging facilitated subsequent verification, analysis, and system refinement. Therefore, the AMTES assessment process remained fully traceable and interpretable, providing learners with clear and reliable justifications for each awarded point.

AMTES Demonstrates High Stability and Repeatability in Evaluations

The stability and reliability of the system were confirmed by consistently low CVs at multiple levels of analysis. At the total score level, the average CVs were 0.87%, 1.12% and 1.07% for cough, frequent urination, and abdominal pain cases, respectively. At the item level, the CVs were exceptionally low, with averages of 0.55% (cough), 0.73% (frequent urination),

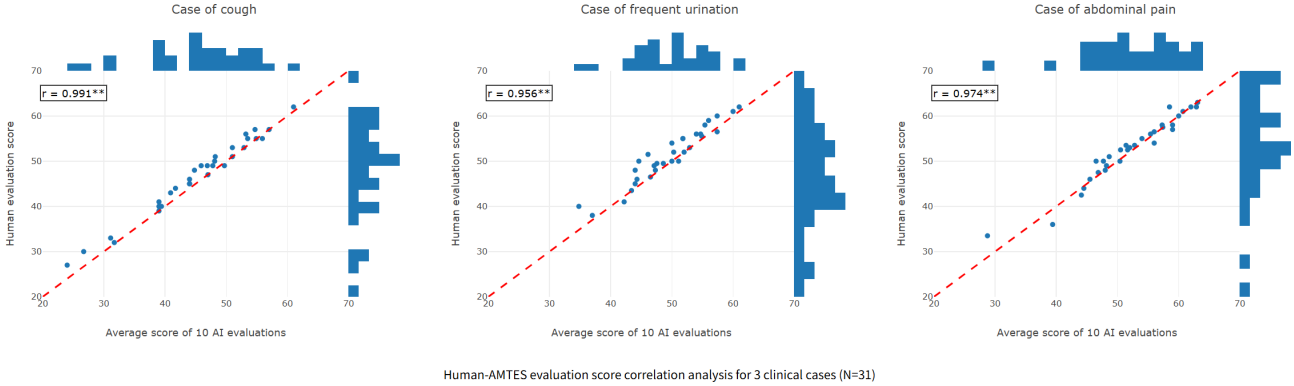
and 0.67% (abdominal pain) using human evaluations as the benchmark. At the specific history category level, the "chief complaint" category notably achieved a CV of 0 in both the cough and abdominal pain cases, indicating perfect consistency. Even the categories with the highest variability, such as "present history," maintained very low CVs (eg, 0.65% and 0.95%). These consistently low CV values across all levels of analysis robustly demonstrate that AMTES provides highly stable and reliable structured evaluations. All detailed CV data, including ranges and category-specific breakdowns, are presented in Multimedia Appendix 4.

Human-AI Consistency in Structured Evaluation

Human-AI Consistency at Total Score Level

Excellent consistency was observed between the total scores assigned by AMTES and human evaluators. The ICC exceeded 0.923 across all 3 clinical scenarios, indicating a high level of agreement between the AI and human experts. This strong positive relationship was further supported by high Pearson *r* (Figure 5). A detailed breakdown of the mean scores, SD, and specific ICC values with 95% CIs for each case is available in Multimedia Appendix 4.

Figure 5. Human-AMTES evaluation score correlation analysis showing strong positive correlations across all 3 cases. AI: artificial intelligence. AMTES: Artificial Intelligence–Powered Medical History-Taking Training and Evaluation System. ***P*<.01.



Human-AI Consistency at Item Level

The primary metric for evaluating human-AI consistency is the item-level consistency, which directly reflects the proportion of scoring items where AMTES and human evaluators agree. This metric provides a more accurate assessment of evaluation quality than total score comparisons, as it avoids the confounding effect where errors in opposite directions might cancel out. Across all 3 cases, not only was the mean difference items less than 3, but the item-level consistency was also remarkably high, exceeding 95% in all scenarios. Even in the most complex abdominal pain case, the system maintained an item-level consistency rate of 95.75%, demonstrating its robustness in nuanced evaluations (Table 3). Overall, AMTES demonstrated high consistency with human evaluations across multiple case scenarios.

Table . Discrepancy and consistency analysis of human-AMTES[a] matched item counts by case groups.

Case	Total items, n	Mean difference items, mean (SD)	Item-level consistency, %
Cough (n=31)	66	1.89 (1.49)	97.13
Frequent urination (n=31)	59	2.06 (1.36)	96.50
Abdominal pain (n=31)	67	2.85 (1.56)	95.75

[a]AMTES: Artificial Intelligence–Powered Medical History-Taking Training and Evaluation System.

Student Questionnaire

All 31 distributed questionnaires were returned (response rate: 100%). The results are displayed in Table 4. When considering whether the AMTES is helpful, a large proportion of students agreed (n=14, 45%) or strongly agreed (12, 39%). A significant portion (n=14, 45%) strongly agreed, and 11 (35%) agreed with the notion that the feedback and evaluation are very valuable. Furthermore, 11 (35%) students agreed and 15 (48%) strongly agreed that they would like to use the AMTES in the future. When asked whether they would like to recommend this AMTES to others, 11 (35%) agreed and 17 (55%) strongly agreed, while only 1 (3%) student disagreed and none strongly disagreed.

Table . Results of student questionnaire feedback.

Item	Students responding (n=31), n (%)				
	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
AMTES[a] is helpful as an additional tool	12 (39)	14 (45)	4 (13)	1 (3)	0 (0)
Feedback and evaluation are very valuable	14 (45)	11 (35)	5 (16)	1 (3)	0 (0)
Would like to use the AMTES in the future	15 (48)	11 (35)	4 (13)	1 (3)	0 (0)
Would like to recommend this AMTES to others	17 (55)	11 (35)	2 (6)	1 (3)	0 (0)

[a]AMTES: Artificial Intelligence–Powered Medical History-Taking Training and Evaluation System.

Baseline Comparison Analysis

The baseline comparison revealed substantial and consistent improvements across multiple dimensions, as detailed in Table 5.

Table . Baseline comparison of system performance across implementation levels and cross-model generalizability test results using Qwen-Max.

Metric and case	Baseline system	Core-optimized system	Fully-processed system	Relative change (%) baseline→fully-processed	LLM[a] backend Qwen-Max	Δ (Qwen - DeepSeek)
CV[b] (%), mean (range)						
Cough (n=31)	2.09 (0 - 8.66)	1.78 (0 - 5.71)	0.87 (0 - 2.78)	−58.3	1.71 (0 - 3.93)	+0.84
Frequent urination (n=31)	1.68 (0 - 8.22)	3.02 (0 - 14.99)	1.12 (0 - 7.64)	−33.3	2.90 (1.10 - 6.05)	+1.78
Abdominal pain (n=31)	0.57 (0 - 4.12)	0.98 (0 - 2.96)	1.07 (0 - 3.78)	+87.7	2.19 (0.63 - 4.91)	+1.12
Mean difference items, mean (SD)						
Cough (n=31)	4.83 (2.28)	4.60 (2.28)	1.89 (1.49)	−60.9	2.24 (1.14)	+0.35
Frequent urination (n=31)	7.92 (2.56)	4.83 (2.86)	2.06 (1.36)	−74.0	2.86 (1.59)	+0.80
Abdominal pain (n=31)	9.44 (3.86)	4.05 (2.17)	2.85 (1.56)	−69.8	3.71 (2.18)	+0.86
Item-level consistency, %						
Cough (n=31)	92.69	93.03	97.13	+4.8	96.60	−0.53
Frequent urination (n=31)	86.58	91.82	96.50	+11.5	95.14	−1.36
Abdominal pain (n=31)	85.92	93.96	95.75	+11.5	94.45	−1.30
ICC[c] (95% CI)						
Cough (n=31)	0.866 (0.785 - 0.942)	0.864 (0.747 - 0.931)	0.978 (0.955 - 0.989)	+12.9	0.970 (0.938 - 0.985)	-0.008
Frequent urination (n=31)	0.414 (0.023 - 0.702)	0.663 (0.445 - 0.815)	0.923 (0.849 - 0.962)	+122.9	0.893 (0.792 - 0.947)	+0.030
Abdominal pain (n=31)	0.500 (0.044 - 0.775)	0.897 (0.803 - 0.948)	0.972 (0.943 - 0.986)	+94.4	0.973 (0.945 - 0.987)	+0.001
Pearson *r*						
Cough (n=31)	0.953	0.944	0.991	+4.0	0.983	−0.008
Frequent urination (n=31)	0.768	0.785	0.956	+24.5	0.969	+0.013
Abdominal pain (n=31)	0.866	0.948	0.974	+12.5	0.973	−0.001

[a]LLM: large language model.

[b]CV: coefficients of variation.

[c]ICC: intraclass correlation coefficient.

Enhanced Evaluation Stability

For the cough case, CV saw a substantial 58.3% reduction, moving from a mean of 2.09% (baseline system) to 0.87% (fully processed system), indicating minimal variation. While frequent urination also showed a 33.3% reduction (1.68% to 1.12%), abdominal pain presented a unique trend, with CV increasing from 0.57% (baseline system) to 1.07% (fully processed system), suggesting increased variability in this most complex scenario despite overall alignment gains.

Significant Improvements in Human-AI Consistency

The optimization process yielded remarkable gains in human-AI alignment across multiple metrics. For detailed data, please refer to Table 5.

Item-Level Consistency

Our optimization efforts yielded significant gains in item-level human-AI consistency. Across all cases, the mean number of discrepant items saw reductions ranging from 60.9% to 74.0%. This was accompanied by a notable rise in the item-level consistency, which climbed from 85.92% - 92.69% (baseline) to a consistently high 95.75% - 97.13% (fully processed system). These improvements consistently showed incremental

gains from the baseline system to core-optimized system and then to the final fully processed system (Table 5).

Total Score-Level Consistency

Total score-level consistency demonstrated even more striking improvements, with both the ICC and Pearson *r* showing significant positive changes.

The ICC values experienced exceptional growth, particularly as case difficulty increased, showcasing the optimization's pronounced effect on alignment in more complex scenarios. For instance, the ICC for frequent urination surged by an impressive 122.9% (from 0.414 to 0.923), transforming from weak to highly consistent. Similarly, abdominal pain saw a 94.4% increase (from 0.500 to 0.972), also achieving high consistency. Even for the cough case, ICC improved by 12.9% (from 0.866 to 0.978), reaching near-perfect consistency. The progressive nature of these gains was clear across all versions of the systems (Table 5).

Correlation strength, as indicated by Pearson *r*, consistently improved across all cases. It moved from a strong correlation for cough (0.953) to nearly perfect (0.991), and from moderate and strong to very strong for frequent urination (0.768 to 0.956) and abdominal pain (0.866 to 0.974). This further confirms a much tighter alignment, with noticeable steps of improvement from each optimization layer (Table 5).

Cross-Model Generalizability Validation

Upon replacing DeepSeek-V2.5 in the AMTES system with Qwen-Max, the item-level consistency remained high (94.45% - 96.60%), confirming its excellent reliability. Notably, in the more challenging abdominal pain case, it demonstrated high human-AI consistency, with a Pearson *r* reaching 0.969. Despite a slight increase in the average number of differing items (18.5% - 38.8% increase, or 0.35 - 0.86 additional items) with Qwen-Max, the absolute difference remained very small, and the item-level consistency rate only saw a minor decrease of 0.53% - 1.36% (Table 5). Therefore, this cross-model validation provides strong evidence supporting the effectiveness of the AMTES system framework.

Discussion

Principal Findings and Methodological Innovations

In this study, we successfully developed the AMTES. Our findings demonstrate that through our comprehensive framework of integrated design strategies, AMTES effectively simulates patient interactions across 3 cases of varying difficulty, providing high-quality dialog and transparent, evidence-based feedback. Critically, its evaluations achieved exceptional stability (mean CV <1.20%) and high human-AI consistency (mean ICC >0.923). This remarkable stability and consistency demonstrate that AMTES holds significant potential as a history-taking training tool in medical education. The use of AMTES as a standardized patient offers a more accessible alternative to traditional human standardized patients, potentially enhancing access to medical training, especially in resource-limited settings. The positive student reception further supports its significant potential as an engaging history-taking training tool.

The key to achieving these robust results lies in our systematic approach, which extends beyond conventional prompt optimization to encompass a multistage strategic framework. Pre-evaluation, we implemented "Decomposition for Standardization" to break down complex tasks and "Disambiguation via Guidelines" to ensure input clarity. During the evaluation, we architected a "Parallelizing Evaluations" mechanism. This architecture segments the scoring task into multiple concurrent sub-queries, which not only circumvents token limit constraints in long-context scenarios but also significantly enhances processing throughput. Post-evaluation, a "Multi-level Verification" mechanism was deployed to cross-reference and validate the preliminary results, ensuring the accuracy and reliability of the final output. It is this organic integration of strategies across the entire workflow that provided the foundation for AMTES's superior performance.

Empirical Performance Results and Comparison With Previous Work

High-quality doctor-patient interaction is crucial in history-taking training. AMTES addresses the need for patient simulation through the integration of a LLM DeepSeek-V2.5. Through rule restrictions and multiple validations, AMTES mitigated the "hallucination" issue commonly associated with LLMs in complex dialogs, as well as the occurrence of unreliable answers stemming from their strong reasoning abilities, as noted in previous studies [12]. Our results show that response accuracy and information appropriateness are highest in the simplest cases among 3 different levels of difficulty, at 98.6% (SD 1.5%) and 99.74% (SD 0.67%). These response accuracy rates are on par with those of ChatGPT-powered systems [22,28]. In addition, our findings thus confirm that LLM-powered systems exhibit high accuracy and completeness, with accuracy slightly lower for higher difficulty compared to lower difficulty. This observation is consistent with findings from other studies [28].

Experimental data demonstrate that the fully processed system exhibits exceptional stability in repeated structured evaluations across 3 distinct cases. The CVs of total scores are low-cough: 0.87% (range 0% - 2.78%); frequent urination: 1.12% (range 0% - 7.64%); abdominal pain: 1.07% (range 0% - 3.78%). Moreover, the system shows low CVs in both item-level scoring and across history categories. Furthermore, AMTES demonstrates high consistency with human evaluations in both total score level and item-level assessments, thereby confirming the system's significant reliability and accuracy. This accuracy surpasses that reported for virtual patient systems in previous studies [29-31] and, in some aspects, exceeds the human-AI consistency of ChatGPT-4.0-driven systems [22].

Optimization Impact and Cross-Model Generalizability

Our baseline comparison analysis provides empirical evidence for the value of systematic optimization in LLM-powered educational tools. The progressive improvements from baseline system to fully processed system, particularly the dramatic enhancements in human-AI consistency for complex cases (ICC

improving from 0.500 to 0.972 for abdominal pain), demonstrate that sophisticated prompt engineering and verification mechanisms can transform unreliable LLM outputs into clinically acceptable evaluations. Most notably, the differential impact across case complexities, with the greatest improvements observed in the most challenging scenarios, suggests that our optimization strategies are particularly valuable for nuanced clinical assessments where raw LLMs struggle most. These findings offer practical guidance for institutions, rather than accepting out-of-the-box LLM performance, investing in comprehensive optimization can yield evaluation tools that approach human-level consistency.

Beyond demonstrating the importance of optimization, our framework also exhibits remarkable cross-model adaptability. Validation experiments with the new large-language model Qwen-Max demonstrated that, without any prompt modifications, our system could still provide transparent, stable, and highly accurate structured evaluations. The success of this experiment challenges the common assumption that LLM-powered systems require extensive customization for each model. This finding indicates that a well-designed evaluation framework can achieve a level of abstraction that transcends specific model architectures, and its cross-model generalizability suggests great potential for medical education applications. Educational institutions often face constraints in technology choices due to institutional policies, regional regulations, or resource limitations. Although our preliminary findings suggest that the evaluation framework may not be entirely dependent on a specific LLM, further validation across diverse platforms is needed to confirm this architectural flexibility. If fully realized, such adaptability could potentially facilitate broader adoption of AI-driven educational tools in varied educational settings. These findings suggest that a one-size-fits-all approach to implementing LLMs in education is suboptimal; instead, investing in a structured, multi-layered optimization and verification framework is critical to unlocking their full potential as reliable assessment tools.

Educational Value and Student Feedback

Evaluation and feedback are critical in clinical education [32], and particularly structured and procedural assessments, which positively impact teaching and student learning [33]. Therefore, through continuous prompt optimization, our fully processed system not only outputs structured scores but also provides detailed rationales for each item-level score by citing specific dialogs from the text, addressing the inherent opacity of scoring reasons in traditional SP programs and existing virtual patient systems. By providing clear evidence for each scoring decision, AMTES helps students understand not only what they missed but also why specific items are important for comprehensive history-taking. This transparency is crucial for building trust in AI-based educational tools and supporting effective learning.

Students who participated in the study provided positive feedback. Among them, 11 (35%) students agreed and 14 (45%) strongly agreed that the system's evaluation function is valuable. 11 (35%) students agreed and 15 (48%) strongly agreed to continue using the AMTES system for history-taking practice. Moreover, the majority of students were willing to recommend the system to their peers. The strong performance of AMTES across 3 cases of different difficulties and disease types, along with positive user feedback, highlights its potential adaptability to a broader range of clinical scenarios. This matches studies showing that AI in health care can help develop communication skills, critical thinking, and clinical reasoning abilities through good interactions and clear feedback [34]. In addition, studies confirm that practice in virtual environments helps improve skills and confidence in real-world clinical encounters [35,36], suggesting that tools like AMTES can optimize educational resources while maintaining educational quality.

AMTES Positioning and Application Prospects

Specifically, AMTES breaks through traditional training resource limitations and accessibility barriers by offering 24/7 learning support, personalized learning experiences, and tailored-structured feedback independent of standardized patient availability. This continuous accessibility and tailored response capability are key strengths of LLMs in medical education [37], potentially supporting student learning efficiency and skill acquisition. AMTES represents a significant advancement in history-taking education, but it is designed to complement rather than replace traditional standardized patient (SP) training. Given the focus of AMTES on evaluating the completeness of history-taking, it is evident that AMTES excels at providing continuous availability for early-stage skill development, allowing students to practice at their own pace and receive consistent, item-level feedback throughout their learning journey. However, the system currently lacks the ability to simulate nonverbal communication (such as facial expressions and body language) and cannot provide emotional understanding and ethical guidance in its feedback - elements which are unique strengths of human SP interactions. While LLMs demonstrate decision-making capabilities, their potential as a replacement for evidence-based professional teaching remains to be fully explored [37]. Therefore, the successful integration of LLMs in medical education feedback, as exemplified by AMTES, is unlikely to lead to the complete replacement of human educators; instead, it may facilitate a redistribution of human effort to areas where it's most impactful [38].

Limitations and Future Prospects

We acknowledge that this study has several limitations, which, in turn, provide clear directions for our future research.

First, from a methodological perspective, a key limitation is the retrospective nature of our baseline comparisons. This was a deliberate ethical choice, grounded in the pedagogical imperative to protect the student learning experience. Exposing learners to a potentially unoptimized baseline system risked undermining their motivation and trust, so we prioritized providing all participants with the most reliable and educationally beneficial version of the system. However, we acknowledge this precludes a direct, prospective comparison of learning outcomes between the different system versions. Future research could use a randomized controlled trial design to provide more definitive evidence on the educational impact of each optimization layer.

Second, the current system's inability to simulate or interpret nonverbal communication (eg, facial expressions and body

language) represents a significant shortcoming in achieving the full fidelity of patient-physician interactions. Maintaining a warm and friendly communication style and expressing empathy are particularly crucial for establishing effective doctor-patient relationships. Empathy, as a core competency in doctor-patient interactions [39], has proven highly effective in improving patient satisfaction, treatment outcomes [40], and generating positive health care results [41,42]. To address this gap, future work will focus on integrating cutting-edge multimodal technologies, such as virtual reality (VR), augmented reality (AR), and expressive speech synthesis, to create a more holistic and immersive simulation environment.

Finally, the scope of our validation needs to be broadened. The current evaluation was conducted not only with a limited sample size but was also confined to 2 prominent Chinese LLMs (DeepSeek and Qwen). To comprehensively establish the generalizability of our framework, a crucial future endeavor will be 2-fold: first, to expand our validation dataset with more diverse cases from multiple institutions; and second, to apply our framework to leading international models (eg, the GPT series [OpenAI], Claude [Anthropic]) and evaluate it in different linguistic contexts, such as English.

By systematically addressing these limitations, we are confident that the system has the potential to evolve into a more robust and comprehensive next-generation tool for medical education.

Conclusion

AMTES, built on a framework of transparent and verifiable evaluation, achieves high stability and human-AI consistency. To our knowledge, this is the first study to systematically evaluate an LLM-powered history-taking evaluation system across multiple disease scenarios while providing empirical validation of design strategies through baseline comparisons and demonstrating cross-model generalizability. By providing students with consistent, evidence-based feedback, AMTES is positioned as a valuable complementary tool in medical education, though further validation in diverse settings would strengthen these conclusions.

Acknowledgments

We acknowledge the support from the Teaching Reform Project in Guangdong Province (2024), the New Medical Education Reform Research Project for Undergraduate Universities in Guangdong Province (2023), and the Shantou University Medical College Teaching Reform and Research Project (2025).

Data Availability

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

Authors' Contributions

WZ was responsible for the conceptualization and project administration of the study, in addition to developing the software system and contributing to formal analysis. YL handled the development of the research methodology, conducted the investigation, performed data analysis, and prepared the original manuscript draft. CS contributed to the methodology, supervised the research team, managed the project, and secured funding. HT focused on developing the analytical methodology, performed formal statistical analysis, and was responsible for data visualization. LW contributed to the conceptualization and design of the study, also assisting with project management. XL and XC contributed to the conceptual framework and study design. YZ was responsible for data curation and contributed to the data analysis. YL and CS contributed equally to this work and are co-first authors; HT and WZ contributed equally to this work and are co-corresponding authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The Artificial Intelligence–Powered Medical History-Taking Training and Evaluation System workflow.
[PDF File, 73 KB - mededu_v11i1e73419_app1.pdf]

Multimedia Appendix 2
Multilevel verification.
[PDF File, 129 KB - mededu_v11i1e73419_app2.pdf]

Multimedia Appendix 3
Comprehensive prompt for medical history-taking scoring.
[PDF File, 77 KB - mededu_v11i1e73419_app3.pdf]

Multimedia Appendix 4

Detailed evaluation metrics.
[PDF File, 68 KB - mededu_v11i1e73419_app4.pdf]

References

1. Keifenheim KE, Teufel M, Ip J, et al. Teaching history taking to medical students: a systematic review. BMC Med Educ 2015 Sep 28;15:159. [doi: 10.1186/s12909-015-0443-x] [Medline: 26415941]
2. Palsson R, Kellett J, Lindgren S, et al. Core competencies of the European internist: A discussion paper. Eur J Intern Med 2007 Mar;18(2):104-108. [doi: 10.1016/j.ejim.2006.10.002] [Medline: 17338961]
3. Alharbi L, Almoallim H. History-taking skills in rheumatology. In: Almoallim H, Cheikh M, editors. Skills in Rheumatology: Springer; 2021:3-16.
4. Kantar A, Marchant JM, Song WJ, et al. History taking as a diagnostic tool in children with chronic cough. Front Pediatr 2022;10:850912. [doi: 10.3389/fped.2022.850912] [Medline: 35498777]
5. Steinkellner C, Schlömmer C, Dünser M. Medical history taking and clinical examination in emergency and intensive care medicine. Med Klin Intensivmed Notfmed 2020 Oct;115(7):530-538. [doi: 10.1007/s00063-020-00731-x] [Medline: 32885280]
6. Meng X, Zhang M, Ma W, Cheng X, Yang X. A clinical medicine level test at Jinan University School of Medicine reveals the importance of training medical students in clinical history-taking. PeerJ 2023;11:e15052. [doi: 10.7717/peerj.15052] [Medline: 37009162]
7. Aranda JH, Monks SM. Roles and responsibilities of the standardized patient director in medical simulation. In: StatPearls: StatPearls Publishing LLC; 2024.
8. Bauer D, Lahner FM, Huwendiek S, Schmitz FM, Guttormsen S. An overview of and approach to selecting appropriate patient representations in teaching and summative assessment in medical education. Swiss Med Wkly 2020 Nov 30;150:w20382. [doi: 10.4414/smw.2020.20382] [Medline: 33306811]
9. Kaplonyi J, Bowles KA, Nestel D, et al. Understanding the impact of simulated patients on health care learners' communication skills: a systematic review. Med Educ 2017 Dec;51(12):1209-1219. [doi: 10.1111/medu.13387] [Medline: 28833360]
10. Du J, Zhu X, Wang J, et al. History-taking level and its influencing factors among nursing undergraduates based on the virtual standardized patient testing results: Cross sectional study. Nurse Educ Today 2022 Apr;111:105312. [doi: 10.1016/j.nedt.2022.105312] [Medline: 35287063]
11. Bokken L, Linssen T, Scherpbier A, van der Vleuten C, Rethans JJ. Feedback by simulated patients in undergraduate medical education: a systematic review of the literature. Med Educ 2009 Mar;43(3):202-210. [doi: 10.1111/j.1365-2923.2008.03268.x] [Medline: 19250346]
12. Holderried F, Stegemann-Philipps C, Herschbach L, et al. A generative pretrained transformer (GPT)-powered chatbot as a simulated patient to practice history taking: prospective, mixed methods study. JMIR Med Educ 2024 Jan 16;10:e53961. [doi: 10.2196/53961] [Medline: 38227363]
13. Lee H. The rise of ChatGPT: Exploring its potential in medical education. Anat Sci Educ 2024;17(5):926-931. [doi: 10.1002/ase.2270] [Medline: 36916887]
14. Wong RSY, Ming LC, Raja Ali RA. The intersection of ChatGPT, clinical medicine, and medical education. JMIR Med Educ 2023 Nov 21;9:e47274. [doi: 10.2196/47274] [Medline: 37988149]
15. Russell RG, Lovett Novak L, Patel M, et al. Competencies for the use of artificial intelligence-based tools by health care professionals. Acad Med 2023 Mar 1;98(3):348-356. [doi: 10.1097/ACM.0000000000004963] [Medline: 36731054]
16. Wang S, Yang L, Li M, Zhang X, Tai X. Medical education and artificial intelligence: web of science-based bibliometric analysis (2013-2022). JMIR Med Educ 2024 Oct 10;10:e51411. [doi: 10.2196/51411] [Medline: 39388721]
17. Webb JJ. Proof of concept: using ChatGPT to teach emergency physicians how to break bad news. Cureus 2023 May;15(5):e38755. [doi: 10.7759/cureus.38755] [Medline: 37303324]
18. Cook DA, Overgaard J, Pankratz VS, Del Fiol G, Aakre CA. Virtual patients using large language models: scalable, contextualized simulation of clinician-patient dialogue with feedback. J Med Internet Res 2025 Apr 4;27:e68486. [doi: 10.2196/68486] [Medline: 39854611]
19. Or AJ, Sukumar S, Ritchie HE, Sarrafpour B. Using artificial intelligence chatbots to improve patient history taking in dental education (Pilot study). J Dent Educ 2024 Dec;88 Suppl 3(Suppl 3):1988-1990. [doi: 10.1002/jdd.13591] [Medline: 38783404]
20. Wang C, Li S, Lin N, et al. Application of large language models in medical training evaluation-using ChatGPT as a standardized patient: multimetric assessment. J Med Internet Res 2025 Jan 1;27:e59435. [doi: 10.2196/59435] [Medline: 39742453]
21. Bing-You R, Hayes V, Varaklis K, Trowbridge R, Kemp H, McKelvy D. Feedback for learners in medical education: what is known? A scoping review. Acad Med 2017 Sep;92(9):1346-1354. [doi: 10.1097/ACM.0000000000001578] [Medline: 28177958]

22. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, et al. A language model-powered simulated patient with automated feedback for history taking: prospective study. JMIR Med Educ 2024 Aug 16;10:e59213. [doi: 10.2196/59213] [Medline: 39150749]
23. Masters K, MacNeil H, Benjamin J, et al. Artificial intelligence in health professions education assessment: AMEE guide no. 178. Med Teach :1-15. [doi: 10.1080/0142159X.2024.2445037]
24. Choi HS, Song JY, Shin KH, Chang JH, Jang BS. Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer. Radiat Oncol J 2023 Sep;41(3):209-216. [doi: 10.3857/roj.2023.00633] [Medline: 37793630]
25. Kuzan BN, Meşe İ, Yaşar S, Kuzan TY. A retrospective evaluation of the potential of ChatGPT in the accurate diagnosis of acute stroke. Diagn Interv Radiol 2025 Apr 28;31(3):187-195. [doi: 10.4274/dir.2024.242892] [Medline: 39221691]
26. Nguyen T. ChatGPT in medical education: a precursor for automation bias? JMIR Med Educ 2024 Jan 17;10:e50174. [doi: 10.2196/50174] [Medline: 38231545]
27. Xu J, Lu L, Peng X, et al. Data set and benchmark (MedGPTEval) to evaluate responses from large language models in medicine: evaluation development and validation. JMIR Med Inform 2024 Jun 28;12:e57674. [doi: 10.2196/57674] [Medline: 38952020]
28. Goodman RS, Patrinely JR, Stone CA Jr, et al. Accuracy and reliability of chatbot responses to physician questions. JAMA Netw Open 2023 Oct 2;6(10):e2336483. [doi: 10.1001/jamanetworkopen.2023.36483] [Medline: 37782499]
29. Bond WF, Lynch TJ, Mischler MJ, et al. Virtual standardized patient simulation: case development and pilot application to high-value care. Simul Healthc 2019 Aug;14(4):241-250. [doi: 10.1097/SIH.0000000000000373] [Medline: 31116172]
30. Maicher K, Danforth D, Price A, et al. Developing a conversational virtual standardized patient to enable students to practice history-taking skills. Simul Healthc 2017 Apr;12(2):124-131. [doi: 10.1097/SIH.0000000000000195] [Medline: 28704290]
31. Zhang X, Zeng D, Wang X, et al. Analysis of virtual standardized patients for assessing clinical fundamental skills of medical students: a prospective study. BMC Med Educ 2024 Sep 10;24(1):981. [doi: 10.1186/s12909-024-05982-2] [Medline: 39256732]
32. van de Ridder JMM, Stokking KM, McGaghie WC, ten Cate OTJ. What is feedback in clinical education? Med Educ 2008 Feb;42(2):189-197. [doi: 10.1111/j.1365-2923.2007.02973.x] [Medline: 18230092]
33. Schut S, Heeneman S, Bierer B, Driessen E, van Tartwijk J, van der Vleuten C. Between trust and control: Teachers' assessment conceptualisations within programmatic assessment. Med Educ 2020 Jun;54(6):528-537. [doi: 10.1111/medu.14075] [Medline: 31998987]
34. Younis HA, Eisa TAE, Nasser M, et al. A systematic review and meta-analysis of artificial intelligence tools in medicine and healthcare: applications, considerations, limitations, motivation and challenges. Diagnostics (Basel) 2024 Jan 4;14(1):109. [doi: 10.3390/diagnostics14010109] [Medline: 38201418]
35. Borja-Hart NL, Spivey CA, George CM. Use of virtual patient software to assess student confidence and ability in communication skills and virtual patient impression: A mixed-methods approach. Curr Pharm Teach Learn 2019 Jul;11(7):710-718. [doi: 10.1016/j.cptl.2019.03.009] [Medline: 31227094]
36. Fidler BD. Use of a virtual patient simulation program to enhance the physical assessment and medical history taking skills of doctor of pharmacy students. Curr Pharm Teach Learn 2020 Jul;12(7):810-816. [doi: 10.1016/j.cptl.2020.02.008] [Medline: 32540042]
37. Li L, Li P, Wang K, Zhang L, Ji H, Zhao H. Benchmarking state-of-the-art large language models for migraine patient education: performance comparison of responses to common queries. J Med Internet Res 2024 Jul 23;26:e55927. [doi: 10.2196/55927] [Medline: 38828692]
38. Pérez-Esteve C, Guilabert M, Matarredona V, et al. AI in home care-evaluation of large language models for future training of informal caregivers: observational comparative case study. J Med Internet Res 2025 Apr 28;27:e70703. [doi: 10.2196/70703] [Medline: 40294407]
39. Ennab F. Teaching clinical empathy skills in medical education: Can ChatGPT assist the educator? Med Teach 2023 Dec;45(12):1440-1441. [doi: 10.1080/0142159X.2023.2247144] [Medline: 37591768]
40. Di Blasi Z, Harkness E, Ernst E, Georgiou A, Kleijnen J. Influence of context effects on health outcomes: a systematic review. Lancet 2001 Mar 10;357(9258):757-762. [doi: 10.1016/s0140-6736(00)04169-6] [Medline: 11253970]
41. MacPherson H, Mercer SW, Scullion T, Thomas KJ. Empathy, enablement, and outcome: an exploratory study on acupuncture patients' perceptions. J Altern Complement Med 2003 Dec;9(6):869-876. [doi: 10.1089/107555303771952226] [Medline: 14736359]
42. Bikker AP, Mercer SW, Reilly D. A pilot prospective study on the consultation and relational empathy, patient enablement, and health changes over 12 months in patients going to the Glasgow Homoeopathic Hospital. J Altern Complement Med 2005 Aug;11(4):591-600. [doi: 10.1089/acm.2005.11.591] [Medline: 16131282]

Abbreviations

AI: artificial intelligence
AMTES: Artificial Intelligence–Powered Medical History-Taking Training and Evaluation System

CV: coefficient of variation
ICC: intraclass correlation coefficient
LLM: large language model
SP: standardized patient
VSP: virtual standardized patient

Edited by J Eriksen; submitted 04.03.25; peer-reviewed by C Ma, HB Burke; revised version received 10.07.25; accepted 28.07.25; published 29.08.25.

Please cite as:
Liu Y, Shi C, Wu L, Lin X, Chen X, Zhu Y, Tan H, Zhang W
Development and Validation of a Large Language Model–Based System for Medical History-Taking Training: Prospective Multicase Study on Evaluation Stability, Human-AI Consistency, and Transparency
JMIR Med Educ 2025;11:e73419
URL: https://mededu.jmir.org/2025/1/e73419
doi:10.2196/73419

Utility of Generative Artificial Intelligence for Japanese Medical Interview Training: Randomized Crossover Pilot Study

Takanobu Hirosawa[1], MD, PhD; Masashi Yokose[1], MD, PhD; Tetsu Sakamoto[1], MD; Yukinori Harada[1], MD, PhD; Kazuki Tokumasu[2], MD, PhD; Kazuya Mizuta[3], MD; Taro Shimizu[1], MD, MSc, MPH, MBA, PhD

[1]Department of Diagnostic and Generalist Medicine, Dokkyo Medical University, 880 Kitakobayashi, Mibu-cho, Shimotsuga, Japan

[2]Department of General Medicine, Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama University, Okayama, Japan

[3]Department of Intensive Care Medicine, Kameda Medical Center, Chiba, Japan

Corresponding Author:
Takanobu Hirosawa, MD, PhD
Department of Diagnostic and Generalist Medicine, Dokkyo Medical University, 880 Kitakobayashi, Mibu-cho, Shimotsuga, Japan

Abstract

Background: The medical interview remains a cornerstone of clinical training. There is growing interest in applying generative artificial intelligence (AI) in medical education, including medical interview training. However, its utility in culturally and linguistically specific contexts, including Japanese, remains underexplored. This study investigated the utility of generative AI for Japanese medical interview training.

Objective: This pilot study aimed to evaluate the utility of generative AI as a tool for medical interview training by comparing its performance with that of traditional face-to-face training methods using a simulated patient.

Methods: We conducted a randomized crossover pilot study involving 20 postgraduate year 1‐2 physicians from a university hospital. Participants were randomly allocated into 2 groups. Group A began with an AI-based station on a case involving abdominal pain, followed by a traditional station with a standardized patient presenting chest pain. Group B followed the reverse order, starting with the traditional station for abdominal pain and subsequently within the AI-based station for the chest pain scenario. In the AI-based stations, participants interacted with a GPT-configured platform that simulated patient behaviors. GPTs are customizable versions of ChatGPT adapted for specific purposes. The traditional stations involved face-to-face interviews with a simulated patient. Both groups used identical, standardized case scenarios to ensure uniformity. Two independent evaluators, blinded to the study conditions, assessed participants' performances using 6 defined metrics: patient care and communication, history taking, physical examination, accuracy and clarity of transcription, clinical reasoning, and patient management. A 6-point Likert scale was used for scoring. The discrepancy between the evaluators was resolved through discussion. To ensure cultural and linguistic authenticity, all interviews and evaluations were conducted in Japanese.

Results: AI-based stations scored lower across most categories, particularly in patient care and communication, than traditional stations (4.48 vs 4.95; *P*=.009). However, AI-based stations demonstrated comparable performance in clinical reasoning, with a nonsignificant difference (4.43 vs 4.85; *P*=.10).

Conclusions: The comparable performance of generative AI in clinical reasoning highlights its potential as a complementary tool in medical interview training. One of its main advantages lies in enabling self-learning, allowing trainees to independently practice interviews without the need for simulated patients. Nonetheless, the lower scores in patient care and communication underline the importance of maintaining traditional methods that capture the nuances of human interaction. These findings support the adoption of hybrid training models that combine generative AI with conventional approaches to enhance the overall effectiveness of medical interview training in Japan.

Trial Registration: UMIN-CTR UMIN000053747; https://center6.umin.ac.jp/cgi-open-bin/ctr_e/ctr_view.cgi?recptno=R000061336

(JMIR Med Educ 2025;11:e77332) doi:10.2196/77332

KEYWORDS

artificial intelligence; generative artificial intelligence; medical interview training; mock patient; simulation education

Introduction

Medical Interview Training

Medical interview training is an essential part of medical education, significantly influencing clinical competence, patient satisfaction, and treatment outcomes [1-5]. Effective medical interviewing skills are crucial not only for accurate diagnosis but also for establishing trust and rapport among health care professionals, patients, and their families [6-11]. For example, several studies revealed that proper diagnoses can often be made based mainly on an effective medical interview rather than investigations [12,13]. These findings highlighted the pivotal role of communication skills in clinical practice.

Barriers to Medical Interview Training

Despite its importance, medical interview training often faces several barriers [14]. For instance, traditional training methods typically involve simulated patient interactions, which are resource-intensive, requiring substantial time commitments from both medical trainees and educators [15]. While simulation training can provide valuable experiential learning [16-18], its scalability is often limited by resource and financial constraints [19-22]. Consequently, medical students and junior physicians may not receive sufficient opportunities for comprehensive and repeated practice, limiting their development of essential communication and clinical reasoning skills [23,24].

Potential of Artificial Intelligence for Medical Interview Training

In response to these challenges, artificial intelligence (AI) has emerged as a promising tool in medical education [25-28]. Until recent breakthroughs, AI performance remained inadequate due to technical limitations [29]. However, the current development of suitable technologies, including Compute Unified Device Architecture and advanced graphics processing units, has remarkably enhanced AI capabilities [30-33]. AI-driven platforms offer scalable, consistent, and flexible training experiences that allow trainees to practice extensively [34]. These tools have the potential to bridge gaps in access to traditional training by enabling frequent, independent practice [35,36].

Potential of Generative AI for Medical Interview Training

Generative AI, a subset of AI that generates human-like responses and interactions [37,38], presents exciting potential for medical interview training [39,40]. It often incorporates natural language processing and large language models, which enable it to generate and respond to human dialogue in contextually appropriate ways [41,42]. Unlike traditional training methods, generative AI can simulate diverse and complex patient scenarios, providing interactive, responsive, and personalized feedback [43]. This capability not only enhances clinical reasoning but also facilitates self-learning, allowing students to practice repeatedly at their convenience [44-46].

Prior Work

Recent studies have explored the application of generative AI in medical interview training, particularly in the context of Objective Structured Clinical Examinations (OSCEs). For example, research in Japan reported that GPT-4 (legacy) based stations outperformed traditional stations for medical students [47]. However, direct comparison with previous work is limited by differences in AI versions, participant populations, clinical cases, and study designs. Further, earlier studies found that previous versions of GPT occasionally generated implausible responses [48,49]. Additionally, the comparative performance between ChatGPT-4 (legacy) and human physicians in conducting medical interviews revealed comparable aggregate scores across 5 components on the 5-Likert scale (15/25 vs 15/25; $P<.28$) [50].

Research Gap and Aim of the Study

Despite these advances, there is still a lack of research evaluating the utility of generative AI tools in Japanese clinical contexts. Cultural and linguistic nuances, including Japanese, play a significant role in effective communication [51-53]. However, there is a lack of enough research evaluating the effectiveness and adaptability of generative AI tools within the Japanese clinical context. To the best of our knowledge, there is limited research regarding the effectiveness and applicability of generative AI-driven training tools for Japanese medical trainees [47]. Therefore, this study aimed to evaluate the utility and limitations of generative AI by comparing AI-driven medical interview scenarios with traditional mock patient interactions among postgraduate physicians in Japan.

Methods

Setting

This pilot study was conducted in the Department of Diagnostic and Generalist Medicine (general internal medicine [GIM]) at Dokkyo Medical University, Tochigi, Japan.

To minimize variability in participants' medical interview skills, a randomized crossover design was used [54]. All interviews and evaluations were conducted in Japanese to preserve cultural and linguistic integrity. The study consisted of 3 main components: participant recruitment, medical interview implementation, and interview evaluation. This study adhered to the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) guidelines (the CONSORT-EHEALTH checklist is provided in Checklist 1).

Ethical Considerations

Ethics approval was obtained from the Institutional Review Board at Dokkyo Medical University Hospital (number R-79 - 14J). The research adhered strictly to the Helsinki Declaration guidelines to ensure ethical conduct in human participant research.

Participant Inclusion

Participants included postgraduate year 1 - 2 physicians rotating through the GIM department at Dokkyo Medical University

Hospital between April 2024 and January 2025. All eligible physicians during this period were invited to participate. Exclusion criteria included hearing loss or unwillingness to attend the research. Before enrollment, all participants received detailed explanations regarding the study's objectives, procedures, and confidentiality protocols from researchers. Written informed consent was obtained from each participant.

Medical Interview

Overview

Participants were randomly allocated into 2 groups through block randomization to ensure an equal group size [55]. The random allocation sequence was generated by an independent researcher (KM) using Microsoft Excel. This ensured balanced distribution and minimized potential confounding from individual differences.

Each participant completed 2 types of medical interview stations—an AI-based station using the GPTs platform and a traditional station with face-to-face interviews with a trained actor simulating the patient (simulated patient). The 2 stations covered separate clinical cases: abdominal pain and chest pain. In the AI-based stations, participants typed their questions and responses into a laptop computer to interact with the GPTs platform. In the traditional stations, participants engaged in spoken conversation with a simulated patient to conduct the medical interview.

Participants in Group A started with the AI-based interview on abdominal pain, followed by the traditional interview on chest pain. Group B began with the traditional interview on abdominal pain and proceeded to the AI-based interview on chest pain.

Station Structure

Both the AI-based and traditional stations followed an identical structure based on The OSCE [56]. Initially, participants reviewed the simulated patient's basic information for 1 minute. The medical interview, including questions relevant to physical examination, was conducted over 15 minutes. Physical examinations were not actually performed in either station due to maintaining consistency with the text-based interaction in the AI-based station. Following the medical interview, participants had 6 minutes to formulate an assessment and plan. Brief feedback and learning points were then provided for several minutes, after which the participants moved to the next station.

GPTs Setting

GPTs are custom versions of ChatGPT that we can adjust for a specific purpose without programming [57]. In this study, the systems were configured to simulate a patient based on detailed case information provided in Japanese. Importantly, the GPTs were not trained or fine-tuned in the Japanese medical language. The systems did not provide a final diagnosis, even if participants asked. Furthermore, if a participant inputted medical jargon [58], GPTs responded with queries such as "What is XXX?" to simulate realistic patient confusion. Additional configuration with translation in English details is provided in Multimedia Appendix 1.

Simulated Patient

The traditional simulated patient interviews were conducted by researcher TH, who was trained to ensure consistency in responses and demeanor. This approach was chosen because the researcher serves not only as a trained actor simulating symptoms but also as an educator providing brief feedback to the participants at the end of each session. Identical clinical scenarios were used across both groups, based on a widely used and standardized textbook for medical interview training [59].

Evaluation for Medical Interview

Traditional stations were video-recorded and transcribed. AI-based stations used the saved text logs. For consistency in evaluation, the transcriptions were refined to match the same structures between stations. For example, headers labeled as "GPTs" in the AI-based stations were changed to "Patient." Self-introduction parts were removed. The corresponding text files were also anonymized. Sample transcript with translation in English is available in Multimedia Appendix 2.

Two experienced physicians, MK and TSa, independently evaluated the transcripts. The evaluators did not take part in the previous participant recruitment and medical interview implementation. Evaluators used a structured scoring system using a 6-point Likert scale, where 1 is inferior and 6 is excellent. Assessments were based on six key domains: (1) patient care and communication skills, (2) thoroughness of history-taking, (3) physical examination proficiency, (4) accuracy and clarity of transcription, (5) clinical reasoning capability, and (6) overall patient management strategies. The discrepancy was resolved through discussion. Evaluators were blinded to interview methods and participant identity. They assessed transcripts in random order. The scoring system is also based on The OSCE [56,60].

Statistical Analysis

Outcome

The primary outcomes were the comparison of mean scores between AI-based and traditional stations for the whole and each assessment component. The secondary outcome measures involved comparisons within each clinical case, abdominal pain, and chest pain, by interview style.

Data Collection

Baseline characteristics data were collected, including years since obtaining a degree in medicine and sex. All medical interviews were also recorded to ensure accurate transcription: traditional stations were video-recorded, and AI-based stations preserved the conversation logs as text.

Analysis

For both primary and secondary outcomes, scores on the 6-point Likert scale were presented as mean with 95% CIs. To assess the appropriateness of statistical tests, the normality of the paired score differences between AI-based and traditional stations was checked using the Shapiro-Wilk test [61]. As the score differences were not normally distributed, the Mann-Whitney *U* test was used as the primary method for comparing paired outcomes between the 2 stations. A *P* value <.05 was considered

statistically significant. For reference, the 95% CIs are provided to supplement the *P* values (Multimedia Appendix 3 contains detailed normality test results and detailed mean difference).

Continuous variable related to participant characteristics is presented as medians and IQRs and compared using the Mann-Whitney *U* test. The categorical variable was compared using the Fisher exact test. All statistical analyses were conducted using R (version 4.2.2; The R Foundation for Statistical Computing) for MacOS X.

Results

Participants Characteristics

A total of 20 postgraduate physicians were enrolled (Figure 1). Among them, 11 (56%) physicians were first year after graduation, while 9 (45%) physicians were in their second year. Two (10%) female participants were included. There were no statistical differences in participant characteristics between group A and group B, as shown in Table 1.

Figure 1. The flow chart includes participants and allocating the groups.

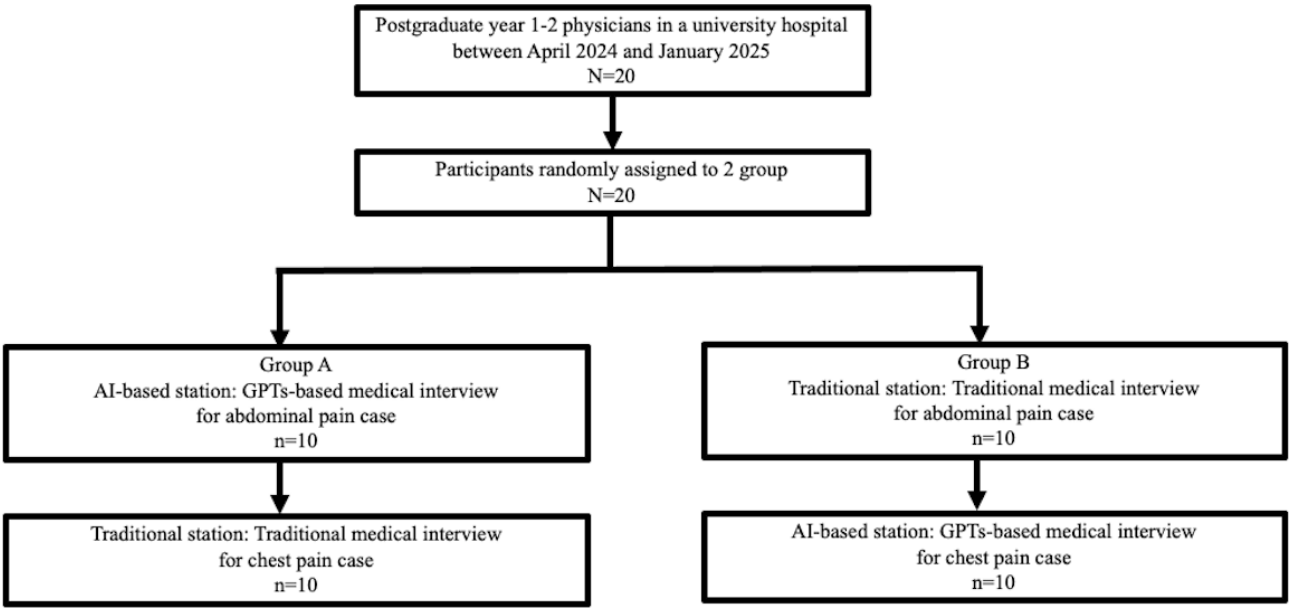


Table . Participants' characteristics.

Variable	Group A (N=10)	Group B (N=10)	*P* value
Female, n (%)	0 (0)	2 (20)	.47[a]
Years after graduation (years), median (IQR)	1.5 (1.0)	1.0 (1.0)	.69[b]

[a]Fisher exact test.

[b]Mann-Whitney *U* test.

Evaluation Outcomes

Performance scores were compared between the AI-based and traditional stations across overall and 6 assessment domains, as shown in Table 2. Overall, the total score was 4.89 in the AI-based stations compared with 5.47 in the traditional stations (*P*<.001).

Table . Performance scores were compared between the artificial intelligence–based and traditional stations across overall and 6 assessment domains.

Scoring system with a 6-point Likert scale	Artificial intelligence–based (GPTs) stations (N=20[a]), 95% CI	Traditional stations (N=20[a]), 95% CI	*P* value[b]
Overall	4.89 (4.74 - 5.04)	5.47 (5.35 - 5.58)	<.001
Patient care and communication	5.05 (4.73 - 5.37)	5.45 (5.06 - 5.84)	.04
History taking	4.90 (4.69 - 5.11)	5.30 (4.96 - 5.65)	.04
Physical examination	5.10 (4.73 - 5.47)	5.80 (5.61 - 5.99)	.001
Accuracy and clarity of transcription	4.70 (4.36 - 5.05)	5.40 (5.16 - 5.64)	.002
Clinical reasoning	4.75 (4.23 - 5.27)	5.30 (4.96 - 5.64)	.13
Management	4.85 (4.34 - 5.36)	5.55 (5.31 - 5.79)	.02

[a]Crossover participants with 10 chest paincasese and 10 abdominal paincasese.

[b]Mann-Whitney *U* test.

AI-based stations yielded slightly lower scores in patient care and communication (mean score: 5.05 vs 5.45; *P*=.04). Scores in other domains such as history taking (4.90 vs 5.30; *P*=.04), physical examination (5.10 vs 5.80*; P*=.001), accuracy and clarity of transcription (4.70 vs 5.40; *P*=.002), and management (4.85 vs 5.55; *P*=.02) also trended lower for the AI-based stations. In contrast, the domain of clinical reasoning showed no significant difference between AI-based and traditional stations (4.75 vs 5.30; *P*=.13).

Subgroup Analysis

Overview

Subgroup analyses were performed to compare the AI-based and traditional stations for each clinical case individually. The initial case presented to participants was abdominal pain, followed sequentially by a chest pain case.

Abdominal Pain Cases

For the abdominal pain case, as shown in Table 3, the overall score was significantly lower in the AI-based stations compared with the traditional stations (4.70 vs 5.48; *P*<.001). Notably, scores for clinical reasoning (4.30 vs 5.50; *P*=.01) and accuracy and clarity of the transcript (4.40 vs 5.40; *P*=.009) were significantly lower in the AI-based stations. While other domains such as patient care and communication (5.00 vs 5.60; *P*=.06), physical examination (5.20 vs 5.80; *P*=.06), and management (4.60 vs 5.50; *P*=.07) were lower in the AI-based stations than the traditional stations, these did not reach statistical significance.

Table . Subgroup analysis for abdominal pain cases compared the artificial intelligence-based and traditional stations across overall and 6 assessment domains.

Scoring system with a 6-point Likert scale	Artificial intelligence-based (GPTs) stations (N=10), 95% CI	Traditional stations (N=10), 95% CI	*P* value[a]
Overall	4.70 (4.47 - 4.93)	5.48 (5.31 - 5.66)	<.001
Patient care and communication	5.00 (4.52 - 5.48)	5.50 (4.80 - 6.20)	.06
History taking	4.70 (4.35 - 5.05)	5.20 (4.54 - 5.86)	.17
Physical examination	5.20 (4.64 - 5.76)	5.80 (5.50 - 6.10)	.06
Accuracy and clarity of transcription	4.40 (3.78 - 5.00)	5.40 (5.03 - 5.77)	.009
Clinical reasoning	4.30 (3.54 - 5.06)	5.50 (5.12 - 5.88)	.01
Management	4.60 (3.70 - 5.50)	5.50 (5.12 - 5.88)	.07

[a]Mann-Whitney *U* test.

Chest Pain Cases

In the case of chest pain, as shown in Table 4, the AI-based stations scored slightly lower in overall scores compared with those in the traditional stations (5.08 vs 5.45; *P*=.004). Physical examination skills were also significantly lower in the AI-based stations (5.00 vs 5.80; *P*=.009). Other domains, including patient care and communication (5.10 vs 5.40; *P*=.37), history taking (5.10 vs 5.40; *P*=.14), and transcription clarity (5.00 vs 5.40; *P*=.09), demonstrated trends in favor of the traditional stations but did not reach significance. Clinical reasoning scores were comparable between the 2 stations (5.10 vs 5.20; *P*=.72), indicating consistent reasoning performance regardless of the interview modality.

Table . Subgroup analysis for chest pain cases compared the artificial intelligence–based and traditional stations across overall and 6 assessment domains.

Scoring system with a 6-point Likert scale	Artificial intelligence-based (GPTs) stations (N=10), 95% CI	Traditional stations (N=10), 95% CI	*P* value[a]
Overall	5.08 (4.90 - 5.27)	5.45 (5.29 - 5.61)	.004
Patient care and communication	5.10 (4.57 - 5.63)	5.40 (4.90 - 5.90)	.37
History taking	5.10 (4.87 - 5.33)	5.40 (5.03 - 5.77)	.14
Physical examination	5.00 (4.42 - 5.58)	5.80 (5.50 - 6.10)	.009
Accuracy and clarity of transcription	5.00 (4.66 - 5.34)	5.40 (5.03 - 5.77)	.09
Clinical reasoning	5.20 (4.46 - 5.94)	5.10 (4.47 - 5.73)	.72
Management	5.10 (4.47 - 5.73)	5.60 (5.23 - 5.97)	.20

[a]Mann-Whitney *U* test.

Discussion

Principal Findings

This study evaluated the utility of generative AI in medical interview training compared with traditional simulated patient interactions among postgraduate physicians in Japan. The principal findings indicate that while AI-based stations provide alternative training methods, they generally yield lower performance scores across several critical domains, including patient care and communication, thoroughness of history-taking, physical examination proficiency, accuracy and clarity of transcription, and management. Participants may have found it difficult to express empathy or engage in natural conversation through typed exchanges [62], limiting the development of interpersonal skills in the GPT stations. While generative AI demonstrates the potential for medical interview training, our findings suggest that it is best suited as a supplementary tool rather than a replacement for traditional simulated patient interactions. The lower performance observed in domains dependent on human interaction—such as communication and patient care—highlights current limitations in AI's ability to simulate empathy and nonverbal cues. Traditional stations, facilitated by trained actors or simulated patients, remain essential for developing advanced interpersonal and communication skills.

A key methodological aspect of this study was configuring the GPT instance to realistically simulate Japanese patient interactions. The GPTs were set up to operate entirely in Japanese, with patient cases, and presented in culturally appropriate language. To enhance authenticity, the system was instructed to respond using typical expressions. Furthermore, the GPTs were directed to avoid using medical terminology.

Despite the limitations in interpersonal skill development, domains such as clinical reasoning remained comparable between GPTs and traditional stations. This finding reinforces the potential of AI-based stations in supporting cognitive aspects of clinical assessment. This result highlights the enduring value of traditional stations, where human dynamics and emotional responsiveness can be authentically practiced and assessed.

Subgroup analyses further demonstrated these differences across specific clinical scenarios. In the abdominal pain case, AI-based stations scored significantly lower in overall performance, clinical reasoning, and transcription clarity. Although other domains like patient care and physical examination were also lower, they did not reach statistical significance. For the chest pain case, while the overall scores were also lower in the GPT stations, the difference was narrower, with physical examination skills showing the most significant disparity. Interestingly, a sub-analysis of abdominal pain cases revealed a significantly lower clinical reasoning score in the AI-based station. This disparity may be attributed to differences in case complexity or the broader differential diagnoses associated with abdominal presentations. In particular, abdominal pain may demand a nuanced interpretation of information [63], suggesting that the limited interactivity of the AI-based format may have constrained diagnostic reasoning. This finding, which was not apparent in the overall analysis, provides an important supplementary insight. It highlights the need to account for case-specific characteristics when selecting cases or designing AI-driven educational tools [64].

Limitations

Several limitations must be acknowledged. First, this study was designed as a feasibility and exploratory trial and was not fully powered or intended for formal hypothesis testing. The small sample size (n=20) and limited number of stations constrain the generalizability of the findings. The primary goal was to assess the feasibility and gather preliminary data to inform future larger-scale studies. Second, the study only included postgraduate physicians from a single institution, potentially restricting the diversity and representativeness of the findings. Results may not be directly applicable to undergraduate medical students, other health care professionals, or participants from different institutions or backgrounds. Third, the mode of interaction differed between AI, typed input, and traditional stations, spoken conversation, which may have inherently biased communication-related scores. Furthermore, physical examinations were not really performed in either station to unify the format for the text-based interaction in the AI-based station, which could have influenced how this domain was assessed. Fourth, the blinded evaluators may have been able to discern the interview modality indirectly, potentially introducing bias. Fifth, it should also be noted that there was some difference in difficulty between the abdominal pain and chest pain cases.

This discrepancy arose because it is inherently challenging to create cases of identical complexity based on different primary concerns. Such differences in case difficulty may have influenced performance results and should be considered when interpreting subgroup analyses. Finally, the study was conducted in a single language using only one generative AI platform, GPTs, limiting its applicability to other languages, cultural contexts, and AI technologies.

Comparison With Prior Work

The current findings expand upon the existing literature. Previous research on OSCEs in Japan found that GPT-4 (legacy) based stations outperformed traditional stations of medical students, with significantly higher total scores across 5 components of a 6-point Likert scale (28.1/31, vs 27.1/31; *P*=.01) [47]. Several differences between the previous study and the current findings limit direct comparison. These include variations in the AI versions used (GPT-4 legacy vs GPTs), participant demographics (medical students vs physicians), cases, and study design (nonrandomized vs randomized crossover).

In relation to the quality of simulated patient responses, previous research on GPT-3.5 and GPT-4 (legacy) indicated implausible response rates of 2% (14/842) and 0.7% (13/1894), respectively [48,49]. In this study using the latest GPTs, responses were almost entirely plausible, with only one instance where GPTs prematurely revealed full physical exam results. This highlights rare but relevant issues in prompt sensitivity.

These findings are particularly promising for resource-limited settings or educational scenarios where access to trained professionals for mock interviews is constrained [65]. However, caution remains warranted in extrapolating these outcomes to real-world clinical environments.

Future Direction

To expand the utility of generative AI in medical interview training, future research should aim for broader validation across diverse educational settings, languages, and digital technology platforms. Improvements in multimodal AI and the integration of voice-based interactions may enhance the realism and interpersonal aspects of AI simulations [66]. Multimodal AI processes and understands information from different types of data, including text, images, audio, video, and sometimes even sensor data [67]. Future investigations should also explore the longitudinal impacts of repeated practice with AI-driven tools to better evaluate the long-term benefits [68]. Additionally, studies comparing hybrid models—such as AI-assisted interviews followed by human debriefing—may offer insights into how best to combine the strengths of both methods [69,70].

Conclusions

This study provides important proof-of-concept evidence for the use of generative AI, specifically GPTs, as a tool in medical interview training among postgraduate physicians. While the AI-based (GPT) station underperformed compared with traditional stations across several domains, including patient care and communication, the performance in clinical reasoning was comparable. These results suggested that generative AI could serve as a supplemental tool for medical education in cognitive components of clinical assessment.

The practical implications for medical education are important. Generative AI can enable self-directed, scalable, and accessible medical interview practice. However, the current findings also reinforce the value of human interaction in developing nuanced communication and empathy. Therefore, the adoption of hybrid educational models may be particularly effective. This approach is the unique strength of combining AI and human educators in simulation-based learning environments.

Nevertheless, these conclusions are preliminary. The small sample size, single-institution setting, and limited number of clinical cases restrict the generalizability of our findings. The crossover design, differences in case complexity, modality of interaction (typed vs spoken), and the use of a single AI language model and language all further limit broad application. These feasibility findings warrant cautious interpretation and highlight the need for larger, multicenter, and longitudinal studies to establish comparative effectiveness and assess the long-term educational impact of AI-assisted training.

Future research should explore the integration of multimodal AI systems to enhance the realism and authenticity of patient simulations. Additionally, multiple institutional collaborations, broader participant demographics, and studies in other languages and contexts are needed to determine the true potential and limitations of AI in medical education.

Acknowledgments

This study was made possible using the resources from the Department of Diagnostic and Generalist Medicine, Dokkyo Medical University.

Authors' Contributions

TH, MY, TSa, YH, KT, KM, and TSh contributed to the study's conceptualization and design. TH served as a simulated patient, and MY was responsible for participant allocation using block randomization. MK and TSa independently evaluated the interview transcripts. TH conducted the statistical analyses and drafted the manuscript. YH, KT, and TSh provided critical revisions to the manuscript for intellectual content. All authors reviewed and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Details of GPTs setting for artificial intelligence (AI)–based medical interview training.
[DOCX File, 26 KB - mededu_v11i1e77332_app1.docx]

Multimedia Appendix 2
An example of transcription.
[DOCX File, 25 KB - mededu_v11i1e77332_app2.docx]

Multimedia Appendix 3
Supplementary statistical analysis.
[DOCX File, 27 KB - mededu_v11i1e77332_app3.docx]

Checklist 1
CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist.
[PDF File, 824 KB - mededu_v11i1e77332_app4.pdf]

References

1. Lipkin M Jr, Quill TE, Napodano RJ. The medical interview: a core curriculum for residencies in internal medicine. Ann Intern Med 1984 Feb;100(2):277-284. [doi: 10.7326/0003-4819-100-2-277] [Medline: 6362513]
2. Stoeckle JD, Billings JA. A history of history-taking: the medical interview. J Gen Intern Med 1987;2(2):119-127. [doi: 10.1007/BF02596310] [Medline: 3550009]
3. Seitz T, Raschauer B, Längle AS, Löffler-Stastka H. Competency in medical history taking-the training physicians' view. Wien Klin Wochenschr 2019 Jan;131(1-2):17-22. [doi: 10.1007/s00508-018-1431-z] [Medline: 30569233]
4. Keifenheim KE, Teufel M, Ip J, et al. Teaching history taking to medical students: a systematic review. BMC Med Educ 2015 Sep 28;15:159. [doi: 10.1186/s12909-015-0443-x] [Medline: 26415941]
5. Lichstein PR. The medical interview. In: Walker HK, Hall WD, Hurst JW, editors. Clinical Methods: The History, Physical, and Laboratory Examinations, 3rd edition: Butterworths; 1990.
6. Novack DH, Dubé C, Goldstein MG. Teaching medical interviewing. a basic course on interviewing and the physician-patient relationship. Arch Intern Med 1992 Sep;152(9):1814-1820. [doi: 10.1001/archinte.152.9.1814] [Medline: 1520048]
7. Eggly S. Physician-patient co-construction of illness narratives in the medical interview. Health Commun 2002;14(3):339-360. [doi: 10.1207/S15327027HC1403_3] [Medline: 12186492]
8. Derksen F, Bensing J, Lagro-Janssen A. Effectiveness of empathy in general practice: a systematic review. Br J Gen Pract 2013 Jan;63(606):e76-e84. [doi: 10.3399/bjgp13X660814] [Medline: 23336477]
9. Hatem DS, Barrett SV, Hewson M, Steele D, Purwono U, Smith R. Teaching the medical interview: methods and key learning issues in a faculty development course. J Gen Intern Med 2007 Dec;22(12):1718-1724. [doi: 10.1007/s11606-007-0408-9] [Medline: 17952511]
10. Foronda C, MacWilliams B, McArthur E. Interprofessional communication in healthcare: an integrative review. Nurse Educ Pract 2016 Jul;19:36-40. [doi: 10.1016/j.nepr.2016.04.005] [Medline: 27428690]
11. Dang BN, Westbrook RA, Njue SM, Giordano TP. Building trust and rapport early in the new doctor-patient relationship: a longitudinal qualitative study. BMC Med Educ 2017 Feb 2;17(1):32. [doi: 10.1186/s12909-017-0868-5] [Medline: 28148254]
12. Hampton JR, Harrison MJ, Mitchell JR, Prichard JS, Seymour C. Relative contributions of history-taking, physical examination, and laboratory investigation to diagnosis and management of medical outpatients. BMJ 1975 May 31;2(5969):486-489. [doi: 10.1136/bmj.2.5969.486]
13. Peterson MC, Holbrook JH, Von Hales D, Smith NL, Staker LV. Contributions of the history, physical examination, and laboratory investigation in making medical diagnoses. West J Med 1992 Feb;156(2):163-165. [Medline: 1536065]
14. Oliveira Franco RL, Martins Machado JL, Satovschi Grinbaum R, Martiniano Porfírio GJ. Barriers to outpatient education for medical students: a narrative review. Int J Med Educ 2019 Sep 27;10:180-190. [doi: 10.5116/ijme.5d76.32c5] [Medline: 31562805]
15. Purva M, Baxendale B, Scales E, Anderson A, Nicklin J, Howes S, et al. Simulation-based education in healthcare standards framwork and guidance. Association for Simulated Practice in Healthcare.: NHS Health Education England URL: https://aspih.org.uk/wp-content/uploads/2017/07/standards-framework.pdf [accessed 2023-04-20]
16. Higham H. Simulation past, present and future-a decade of progress in simulation-based education in the UK. BMJ Simul Technol Enhanc Learn 2021;7(5):404-409. [doi: 10.1136/bmjstel-2020-000601] [Medline: 35515719]
17. Beal MD, Kinnear J, Anderson CR, Martin TD, Wamboldt R, Hooper L. The effectiveness of medical simulation in teaching medical students critical care medicine: a systematic review and meta-analysis. Simul Healthc 2017 Apr;12(2):104-116. [doi: 10.1097/SIH.0000000000000189] [Medline: 28704288]

18. Kononowicz AA, Woodham LA, Edelbring S, et al. Virtual patient simulations in health professions education: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019 Jul 2;21(7):e14676. [doi: 10.2196/14676] [Medline: 31267981]
19. Cleland JA, Abe K, Rethans JJ. The use of simulated patients in medical education: AMEE Guide No 42. Med Teach 2009 Jun;31(6):477-486. [doi: 10.1080/01421590903002821] [Medline: 19811162]
20. Bosse HM, Nickel M, Huwendiek S, Schultz JH, Nikendei C. Cost-effectiveness of peer role play and standardized patients in undergraduate communication training. BMC Med Educ 2015 Oct 24;15:183. [doi: 10.1186/s12909-015-0468-1] [Medline: 26498479]
21. Al Odhayani A, Ratnapalan S. Teaching communication skills. Can Fam Physician 2011 Oct;57(10):1216-1218. [Medline: 21998240]
22. Maloney S, Haines T. Issues of cost-benefit and cost-effectiveness for simulation in health professions education. Adv Simul (Lond) 2016;1:13. [doi: 10.1186/s41077-016-0020-3] [Medline: 29449982]
23. Elendu C, Amaechi DC, Okatta AU, et al. The impact of simulation-based training in medical education: a review. Medicine (Baltimore) 2024 Jul 5;103(27):e38813. [doi: 10.1097/MD.0000000000038813] [Medline: 38968472]
24. Abe K, Suzuki T, Fujisaki K, Ban N. Demographic characteristics of standardized patients (SPs) and their satisfaction and burdensome in Japan: the first report of a nationwide survey. Igaku Kyoiku 2007;38(5):301-307.
25. Chen L, Chen P, Lin Z. Artificial intelligence in education: a review. IEEE Access 2020;8:75264-75278. [doi: 10.1109/ACCESS.2020.2988510]
26. Li R, Wu T. Evolution of artificial intelligence in medical education from 2000 to 2024: bibliometric analysis. Interact J Med Res 2025 Jan 30;14:e63775. [doi: 10.2196/63775] [Medline: 39883926]
27. Lee J, Wu AS, Li D, Kulasegaram KM. Artificial intelligence in undergraduate medical education: a scoping review. Acad Med 2021 Nov 1;96(11S):S62-S70. [doi: 10.1097/ACM.0000000000004291] [Medline: 34348374]
28. Adamopoulou E, Moussiades L. Chatbots: history, technology, and applications. Mach Learn Appl 2020 Dec;2:100006. [doi: 10.1016/j.mlwa.2020.100006]
29. Delipetrev B, Tsinaraki C, Kostic U. Historical Evolution of Artificial Intelligence: Publications Office of the European Union; 2020.
30. Jeon W, Ko G, Lee J, Lee H, Ha D, Ro WW. Deep learning with gpus. In: Advances in Computers: Elsevier; 2021, Vol. 122:167-215. [doi: 10.1016/bs.adcom.2020.11.003]
31. Pandey M, Fernandez M, Gentile F, et al. The transformational role of GPU computing and deep learning in drug discovery. Nat Mach Intell 2022;4(3):211-221. [doi: 10.1038/s42256-022-00463-x]
32. Akkisetty PK. An overview of AI platforms, frameworks, libraries, and processors. In: AMR PRC, Colby R, Nagasubramanian G, Ranganath S, editors. Model Optimization Methods for Efficient and Edge AI: Federated Learning Architectures, Frameworks and Applications: Wiley; 2024:43-55. [doi: 10.1002/9781394219230]
33. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Commun ACM 2017;60(6):84-90. [doi: 10.1145/3065386]
34. Tu T, Schaekermann M, Palepu A, et al. Towards conversational diagnostic artificial intelligence. Nature New Biol 2025 Jun;642(8067):442-450. [doi: 10.1038/s41586-025-08866-7] [Medline: 40205050]
35. Stamer T, Steinhäuser J, Flägel K. Artificial intelligence supporting the training of communication skills in the education of health care professions: scoping review. J Med Internet Res 2023 Jun 19;25:e43311. [doi: 10.2196/43311] [Medline: 37335593]
36. Okonkwo CW, Ade-Ibijola A. Chatbots applications in education: a systematic review. Comput Educ Artif Intell 2021;2:100033. [doi: 10.1016/j.caeai.2021.100033]
37. Sai S, Gaur A, Sai R, Chamola V, Guizani M, Rodrigues JJPC. Generative AI for transformative healthcare: a comprehensive study of emerging models, applications, case studies, and limitations. IEEE Access 2024;12:31078-31106. [doi: 10.1109/ACCESS.2024.3367715]
38. de Vere Hunt IJ, Jin KX, Linos E. A framework for considering the use of generative AI for health. NPJ Digit Med 2025 May 21;8(1):297. [doi: 10.1038/s41746-025-01695-y] [Medline: 40399429]
39. Sardesai N, Russo P, Martin J, Sardesai A. Utilizing generative conversational artificial intelligence to create simulated patient encounters: a pilot study for anaesthesia training. Postgrad Med J 2024 Mar 18;100(1182):237-241. [doi: 10.1093/postmj/qgad137] [Medline: 38240054]
40. Abdelnabi AAB, Soykan B, Bhatti D, Rabadi G. Usefulness of large language models (LLMs) for student feedback on H&P during clerkship: artificial intelligence for personalized learning. ACM Trans Comput Healthcare 2025. [doi: 10.1145/371229]
41. Otter DW, Medina JR, Kalita JK. A survey of the usages of deep learning for natural language processing. IEEE Trans Neural Netw Learning Syst 2020;32(2):604-624. [doi: 10.1109/TNNLS.2020.2979670]
42. Chang Y, Wang X, Wang J, et al. A survey on evaluation of large language models. ACM Trans Intell Syst Technol 2024 Jun 30;15(3):1-45. [doi: 10.1145/3641289]
43. White CB, Wendling A, Lampotang S, Lizdas D, Cordar A, Lok B. The role for virtual patients in the future of medical education. Acad Med 2017 Jan;92(1):9-10. [doi: 10.1097/ACM.0000000000001487] [Medline: 28027092]

44. Parente DJ. Generative artificial intelligence and large language models in primary care medical education. Fam Med 2024 Oct;56(9):534-540. [doi: 10.22454/FamMed.2024.775525] [Medline: 39207784]
45. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ 2023 Mar 6;9:e46885. [doi: 10.2196/46885] [Medline: 36863937]
46. Mohammad B, Supti T, Alzubaidi M, et al. The pros and cons of using ChatGPT in medical education: a scoping review. Stud Health Technol Inform 2023 Jun 29;305:644-647. [doi: 10.3233/SHTI230580] [Medline: 37387114]
47. Yamamoto A, Koda M, Ogawa H, et al. Enhancing medical interview skills through AI-simulated patient interactions: nonrandomized controlled trial. JMIR Med Educ 2024 Sep 23;10:e58753. [doi: 10.2196/58753] [Medline: 39312284]
48. Holderried F, Stegemann-Philipps C, Herschbach L, et al. A generative pretrained transformer (GPT)-powered chatbot as a simulated patient to practice history taking: prospective, mixed methods study. JMIR Med Educ 2024 Jan 16;10:e53961. [doi: 10.2196/53961] [Medline: 38227363]
49. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, et al. A language model-powered simulated patient with automated feedback for history taking: prospective study. JMIR Med Educ 2024 Aug 16;10:e59213. [doi: 10.2196/59213] [Medline: 39150749]
50. Huang TY, Hsieh PH, Chang YC. Performance comparison of junior residents and ChatGPT in the objective structured clinical examination (OSCE) for medical history taking and documentation of medical records: development and usability study. JMIR Med Educ 2024 Nov 21;10:e59902. [doi: 10.2196/59902] [Medline: 39622713]
51. Schouten BC, Meeuwesen L. Cultural differences in medical communication: a review of the literature. Patient Educ Couns 2006 Dec;64(1-3):21-34. [doi: 10.1016/j.pec.2005.11.014] [Medline: 16427760]
52. Hydén LC, Mishler EG. Language and medicine. Ann Rev Appl Linguist 1999 Jan;19:174-192. [doi: 10.1017/S0267190599190093]
53. Meyer A, Riese J, Streichert T. Comparison of the performance of GPT-3.5 and GPT-4 with that of medical students on the written German medical licensing examination: observational study. JMIR Med Educ 2024 Feb 8;10:e50965. [doi: 10.2196/50965] [Medline: 38329802]
54. Dale MacLaine T, Lowe N, Dale J. The use of simulation in medical student education on the topic of breaking bad news: a systematic review. Patient Educ Couns 2021 Nov;104(11):2670-2681. [doi: 10.1016/j.pec.2021.04.004]
55. Broglio K. Randomization in clinical trials: permuted blocks and stratification. JAMA 2018 Jun 5;319(21):2223-2224. [doi: 10.1001/jama.2018.6360] [Medline: 29872845]
56. Madrazo L, Lee CB, McConnell M, Khamisa K. Self-assessment differences between genders in a low-stakes objective structured clinical examination (OSCE). BMC Res Notes 2018 Jun 15;11(1):393. [doi: 10.1186/s13104-018-3494-3] [Medline: 29903050]
57. Introducing gpts 2023. OpenAI. URL: https://openai.com/index/introducing-gpts [accessed 2025-05-12]
58. Hersh L, Salzman B, Snyderman D. Health literacy in primary care practice. Am Fam Physician 2015 Jul 15;92(2):118-124. [Medline: 26176370]
59. Le T, Bhushan V, Sheikh-Ali M, Shahin FA. First Aid for the USMLE Step 2 CS, 4th edition: McGraw-Hill Medical; 2012.
60. Organization CAT. Earning and assessment items related to the skills and attitudes required of students participating in clinical participatory clinical practice (version 42). CATO. 2022. URL: https://www.cato.or.jp/pdf/osce_42.pdf [accessed 2025-05-12]
61. Mardia KV. 9 tests of unvariate and multivariate normality. In: Handbook of Statistics: Elsevier; 1980, Vol. 1:279-320. [doi: 10.1016/S0169-7161(80)01011-5]
62. Limpanopparat S, Gibson E, Harris DA. User engagement, attitudes, and the effectiveness of chatbots as a mental health intervention: a systematic review. Comput Hum Behav Artifi Hum 2024 Aug;2(2):100081. [doi: 10.1016/j.chbah.2024.100081]
63. Cartwright SL, Knudson MP. Evaluation of acute abdominal pain in adults. Am Fam Physician 2008 Apr 1;77(7):971-978. [Medline: 18441863]
64. Lafleur A, Côté L, Leppink J. Influences of OSCE design on students' diagnostic reasoning. Med Educ 2015 Feb;49(2):203-214. [doi: 10.1111/medu.12635] [Medline: 25626751]
65. Dangi RR, Sharma A, Vageriya V. Transforming healthcare in low-resource settings with artificial intelligence: recent developments and outcomes. Public Health Nurs 2025;42(2):1017-1030. [doi: 10.1111/phn.13500] [Medline: 39629887]
66. Kalyan KS, Sangeetha S. SECNLP: a survey of embeddings in clinical natural language processing. J Biomed Inform 2020 Jan;101:103323. [doi: 10.1016/j.jbi.2019.103323] [Medline: 31711972]
67. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nat Med 2022 Sep;28(9):1773-1784. [doi: 10.1038/s41591-022-01981-2] [Medline: 36109635]
68. Feigerlova E, Hani H, Hothersall-Davies E. A systematic review of the impact of artificial intelligence on educational outcomes in health professions education. BMC Med Educ 2025 Jan 27;25(1):129. [doi: 10.1186/s12909-025-06719-5] [Medline: 39871336]
69. Duan W, Zhou S, Scalia MJ, et al. Understanding the evolvement of trust over time within human-AI teams. Proc ACM Hum-Comput Interact 2024 Nov 7;8(CSCW2):1-31. [doi: 10.1145/3687060]

70. Raisch S, Fomina K. Combining human and artificial intelligence: hybrid problem-solving in organizations. AMR 2025 Apr;50(2):441-464. [doi: 10.5465/amr.2021.0421]

Abbreviations

AI: artificial intelligence
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
GIM: general internal medicine
OSCE: Objective Structured Clinical Examination

Edited by J Gentges; submitted 12.05.25; peer-reviewed by RT Potla, V Izquierdo-Alvarez; revised version received 08.06.25; accepted 12.06.25; published 01.08.25.

Please cite as:
Hirosawa T, Yokose M, Sakamoto T, Harada Y, Tokumasu K, Mizuta K, Shimizu T
Utility of Generative Artificial Intelligence for Japanese Medical Interview Training: Randomized Crossover Pilot Study
JMIR Med Educ 2025;11:e77332
URL: https://mededu.jmir.org/2025/1/e77332
doi:10.2196/77332

Leveraging Large Language Models for Simulated Psychotherapy Client Interactions: Development and Usability Study of Client101

Daniel Cabrera Lozoya[1], MSc; Mike Conway[1], PhD; Edoardo Sebastiano De Duro[2], MSc; Simon D'Alfonso[1], PhD

[1]School of Computing and Information Systems, The University of Melbourne, Gratham St, Parkville VIC, Melbourne, Australia

[2]Department of Psychology and Cognitive Science, University of Trento, Trento, Italy

Corresponding Author:
Simon D'Alfonso, PhD
School of Computing and Information Systems, The University of Melbourne, Gratham St, Parkville VIC, Melbourne, Australia

Abstract

Background: In recent years, large language models (LLMs) have shown a remarkable ability to generate human-like text. One potential application of this capability is using LLMs to simulate clients in a mental health context. This research presents the development and evaluation of Client101, a web conversational platform featuring LLM-driven chatbots designed to simulate mental health clients.

Objective: We aim to develop and test a web-based conversational psychotherapy training tool designed to closely resemble clients with mental health issues.

Methods: We used GPT-4 and prompt engineering techniques to develop chatbots that simulate realistic client conversations. Two chatbots were created based on clinical vignette cases: one representing a person with depression and the other, a person with generalized anxiety disorder. A total of 16 mental health professionals were instructed to conduct single sessions with the chatbots using a cognitive behavioral therapy framework; a total of 15 sessions with the anxiety chatbot and 14 with the depression chatbot were completed. After each session, participants completed a 19-question survey assessing the chatbot's ability to simulate the mental health condition and its potential as a training tool. Additionally, we used the LIWC (Linguistic Inquiry and Word Count) tool to analyze the psycholinguistic features of the chatbot conversations related to anxiety and depression. These features were compared to those in a set of webchat psychotherapy sessions with human clients—42 sessions related to anxiety and 47 related to depression—using an independent samples *t* test.

Results: Participants' survey responses were predominantly positive regarding the chatbots' realism and portrayal of mental health conditions. For instance, 93% (14/15) considered that the chatbot provided a coherent and convincing narrative typical of someone with an anxiety condition. The statistical analysis of LIWC psycholinguistic features revealed significant differences between chatbot and human therapy transcripts for 3 of 8 anxiety-related features: negations (t_{56}=4.03, P=.001), family (t_{56}=–8.62, P=.001), and negative emotions (t_{56}=–3.91, P=.002). The remaining 5 features—sadness, personal pronouns, present focus, social, and anger—did not show significant differences. For depression-related features, 4 of 9 showed significant differences: negative emotions (t_{60}=–3.84, P=.003), feeling (t_{60}=–6.40, P<.001), health (t_{60}=–4.13, P=.001), and illness (t_{60}=–5.52, P<.001). The other 5 features—sadness, anxiety, mental, first-person pronouns, and discrepancy—did not show statistically significant differences.

Conclusions: This research underscores both the strengths and limitations of using GPT-4-powered chatbots as tools for psychotherapy training. Participant feedback suggests that the chatbots effectively portray mental health conditions and are generally perceived as valuable training aids. However, differences in specific psycholinguistic features suggest targeted areas for enhancement, helping refine Client101's effectiveness as a tool for training mental health professionals.

(JMIR Med Educ 2025;11:e68056) doi:10.2196/68056

KEYWORDS

medical education; mental health; chatbots; psychotherapy training; virtual client

Introduction

Background

Psychotherapy training requires a comprehensive approach, encompassing practical skill development through supervised sessions with clients or peer role-playing, along with the analysis and discussion of therapy sessions from experienced psychologists [1]. Optimal psychotherapy training ideally needs abundant practice opportunities coupled with immediate performance-based feedback [2]. However, both training methods, working with clients or using role-play, present significant challenges. The use of clients raises ethical concerns regarding patient welfare, particularly when inexperienced psychotherapists provide treatment, which can result in risks to

vulnerable individuals receiving suboptimal care. However, role-play scenarios, while more controlled, are resource-intensive. They face challenges such as finding suitable peers for role-playing and ensuring a consistent learning experience due to varying skill levels among participants.

The integration of natural language processing (NLP) in the field of mental health training offers potential solutions to challenges in client availability. Chatbots leveraging large language models (LLMs) can simulate human dialogue and offer a structured framework for task-oriented interactions [3]. Their evolving conversational abilities allow them to actively engage in dialogues, making them a promising educational tool in fields such as health care and medical education [4]. The use of chatbots as virtual patients in mental health training holds significant potential. Through prompt engineering techniques, chatbots can be configured to simulate diverse mental health conditions and behaviors. This capability makes them valuable assets for psychotherapy training in a controlled environment, thereby mitigating the risks associated with using real clients for training purposes. Additionally, chatbots offer the advantage of being readily accessible platforms for simulated therapeutic interactions at any time.

In this study, we present Client101, a web-based conversational platform that uses LLMs as chatbots to simulate the behavior of mental health clients. We used 2 distinct prompt configurations to generate 2 types of virtual clients: one simulating the experience of depression and the other representing individuals coping with generalized anxiety disorder (GAD). A total of 16 individuals with a background in psychotherapy (ie, clinical psychologists and qualified counselors) used the platform to evaluate it. The participant therapists were tasked with conducting single sessions with each of the 2 chatbots before completing a questionnaire. While no specific instructions were given, it was suggested to the participants that they use something like a single-session integrated cognitive behavioral therapy (SSI-CBT) approach [5].

Furthermore, we used the LIWC-22 (Linguistic Inquiry Word Count) software to measure psycholinguistic features of the sessions. This enabled us to examine whether the chatbots used linguistic indicators commonly associated with depression and GAD. Additionally, we conducted a comparative analysis between the psycholinguistic features of the sessions conducted with the virtual clients and those from single webchat therapy sessions typical of an online Australian mental health support service.

Study Aim

This study pursued 2 primary objectives. First, it aimed to assess psychotherapists' perceptions of the chatbot's ability to simulate client characteristics during sessions, using the questionnaire completed by the therapists. Second, it sought to assess the degree of divergence between synthetic and organic psychotherapy transcripts by identifying any statistically significant differences in specific psycholinguistic indicators. These 2 aims guided our investigation of the following research questions: (1) How well does Client101 simulate the language of psychotherapy clients? (2) How effective is Client101 as an educational tool for training therapists? and (3) What are Client101's limitations, and how can they be addressed?

Thus, our main contributions in this paper are as follows: (1) we built Client101, a web-based conversational platform that uses LLMs as chatbots to simulate the behavior of mental health clients; (2) we present a prompt engineering methodology to generate and evaluate counseling transcripts for simulated psychotherapy client interactions; (3) we performed a psycholinguistic analysis comparing depression and anxiety dimensions between therapy transcripts obtained from an online counseling service and therapy sessions using Client101; and (4) we present results from a preliminary questionnaire that gathered the perceptions and feedback of participant therapists via a set of Likert and open-ended items.

Related Work

Conversational Agents for Mental Health

The connection between chatbots and psychology can be traced to the creation of ELIZA, developed by computer scientist Weizenbaum [6] in the mid-1960s. The ELIZA system used reassembly and decomposition rules to act as a Rogerian psychotherapist (ie, based on the approach of humanistic psychology pioneer Carl Rogers). However, ELIZA was not intended to be a therapy chatbot. Rather, Weizenbaum developed ELIZA to explore interactions between humans and chatbots and to ultimately demonstrate what he saw as the superficiality of such interactions. It is in recent years, partly due to advances in NLP and partly due to the rise of digital mental health, that we have seen the emergence of conversational agents as mental health interventions [3].

Contrary to the prevalence of chatbots that serve as virtual therapists, there is scant lineage of chatbots that instead simulate an individual with mental health issues. In 1972, psychiatrist Colby et al [7] developed PARRY, a chatbot that simulated a person with paranoid schizophrenia. However, PARRY was not intended to serve as a therapy training tool. Rather, its purpose was to model and understand the thought processes and verbal expressions of paranoia, aiming to aid in psychiatric research and study. Since PARRY, little work has been done on the idea of developing chatbots to simulate individuals with mental health issues, particularly for training purposes.

ELIZA, PARRY, and many of the currently available chatbots that rely on predefined rules undermine their potential for therapeutic services by operating through simplistic pattern-matching mechanisms [8]. These systems generate responses by rigidly mapping user inputs to predetermined templates, which result in repetitive, noncontextual interactions that fail to capture the nuanced, dynamic nature of human communication [9]. Unlike humans who can intuitively interpret emotional subtexts, recognize implicit meanings, and respond with genuine empathy, these chatbots provide surface-level, algorithmic outputs that lack the depth, adaptability, and emotional intelligence critical for meaningful dialogue [9]. More recent NLP models, such as recurrent neural networks with long short-term memory (LSTM), generate text by learning from a large corpus of examples rather than relying on hand-crafted rules [10]. Tanana et al [2] developed ClientBot, a web-based

system that uses machine-based feedback for training counseling skills. The underlying NLP component of ClientBot was an LSTM recurrent neural network model trained on 2 distinct datasets. The first dataset consisted of a vast collection of movie and TV show subtitles, encompassing 1689 bitexts and totaling 2.6 billion sentences across 60 languages [11]. The second dataset included psychotherapy transcripts published by Alexander Street Press [12]. Since ClientBot was trained on a corpus of movie transcripts, its responses were sometimes contextually incoherent and often lacked the depth and length typical of a client in a psychotherapy session. As observed by Zhang et al [13], the lack of coherency and consistency in conversational agents results in an unsatisfying overall experience for human users.

Conversational agents struggle to keep an engaging conversation due to a lack of consistent personality [11] and the absence of an explicit long-term memory [10]. To address this issue, Zhang et al [13] developed the dataset "persona-chat dataset," a collection of 164,356 written utterances between crowd workers who were asked to communicate with each other while playing the part of a specific persona. While training LSTM models with this dataset enhances engagement during conversations between humans and chatbots, it was not designed to address the challenge of simulating a realistic mental health client.

Transformer-based generative models leverage self-attention mechanisms to capture long-range dependencies and contextual relationships, allowing them to produce coherent responses [14]. As a result, these models have been used as chatbots for educational applications, including medical training via simulated patient interactions [15]. Efficient prompt engineering, the process of iterating a generative artificial intelligence (AI) prompt to improve its effectiveness, plays a crucial role in creating a conversational agent that accurately mimics a patient. In the study by Stapleton et al [16], prompts were designed for GPT-3.5 Turbo to simulate a patient experiencing suicidal ideation. To achieve realistic conversations, their prompts built a persona detailing the patient's age, past experiences, and intrusive thoughts. When designing the prompts, the authors did not rely on licensed psychologists or psychiatrists, but instead they used the lived experiences of suicidal people as a reference. However, Demasi et al [17] stress the importance of engaging specialized users, such as mental health professionals, in system development and evaluation to achieve successful results. Hence, to achieve engaging and realistic conversations that accurately mimic mental health patients, we collaborated with mental health professionals in designing the prompts and incorporated a memory system into the chatbot to ensure cohesive and consistent interactions over long conversations.

Psycholinguistic Features

A crucial limitation of previous conversational agents for mental health is the lack of validation and grounding in psychological theory for the outputs they generate. Although NLP models have demonstrated remarkable performance in text generation, their effectiveness in mental health applications remains uncertain due to insufficient grounding in psychological theory [18]. To address this gap, our research involved evaluating the quality of the model-generated responses with the assistance of licensed psychologists. Additionally, we used the LIWC-22 [19], a tool designed to assess various psychosocial constructs within text, to evaluate if psycholinguistic features associated with depression and anxiety were present in the sessions generated by the chatbots. A statistical analysis was conducted to observe if there is a statistically significant difference between the psycholinguistic features of the Client101 simulated sessions and single webchat therapy sessions.

LIWC text analysis software has been used to identify linguistic markers of depression and anxiety. Eichstaedt et al [20] used language from Facebook posts of consenting individuals to predict depression recorded in electronic medical records, specifically major depression (*ICD* [*International Classification of Diseases*] codes 296.2) and depressive disorder (*ICD* codes 311). Using different linguistic markers, they could identify depressed patients with fair accuracy: area under the curve=0.69, approximately matching the accuracy of screening surveys benchmarked against medical records. The LIWC negative emotions, feel, sadness, anxiety, health, illness, mental, first-person singular, and discrepancy dictionaries were significantly associated with future depression status. LIWC psycholinguistic features have also shown promising results for remote screening of GAD. Rook et al [21] used linguistic features to predict clinically validated behavioral measures for GAD and self-reported sensitivity in behavioral avoidance or inhibition and behavioral approach. Significant positive correlations were found between scores from the GAD-7 measure and the following LIWC categories: negations, sadness, personal pronouns, present focus, social processes, family, negative emotions, and anger.

Methods

Client101 Platform

The Client101 platform was built using the Django framework and is hosted on an Amazon Web Services EC2 T2 Medium instance (2 vCPUs, 4 GiB RAM) running Ubuntu, with 32 GB of storage. The platform features 2 chatbots that users can engage with to simulate psychotherapy sessions. Each chatbot is tailored to mimic a client with a distinct mental health issue: one represents a person experiencing depression, while the other portrays someone struggling with anxiety. Once a chatbot is selected, psychologists can commence the therapy session using a standard chat interface, akin to those used for online messaging platforms. Alongside the chat interface, there is a text field available to enter any session notes, which can be used for reference and feedback. Additionally, there is a numerical field provided to rate the chatbot's session quality. The session transcript, along with a summary of the session, is stored in a database. If the therapist wishes to pause and then resume the session, the platform can retrieve the session and restart it from where the therapist left off. All session records are stored in an SQLite relational database on the Amazon Web Services instance. Refer to Figure 1 for an example of the user interface. Figure 2 presents excerpts from simulated therapy sessions between a user and chatbots programmed to simulate patients with anxiety (Alice) and depression (Luke).

Figure 1. Client101 user interface.

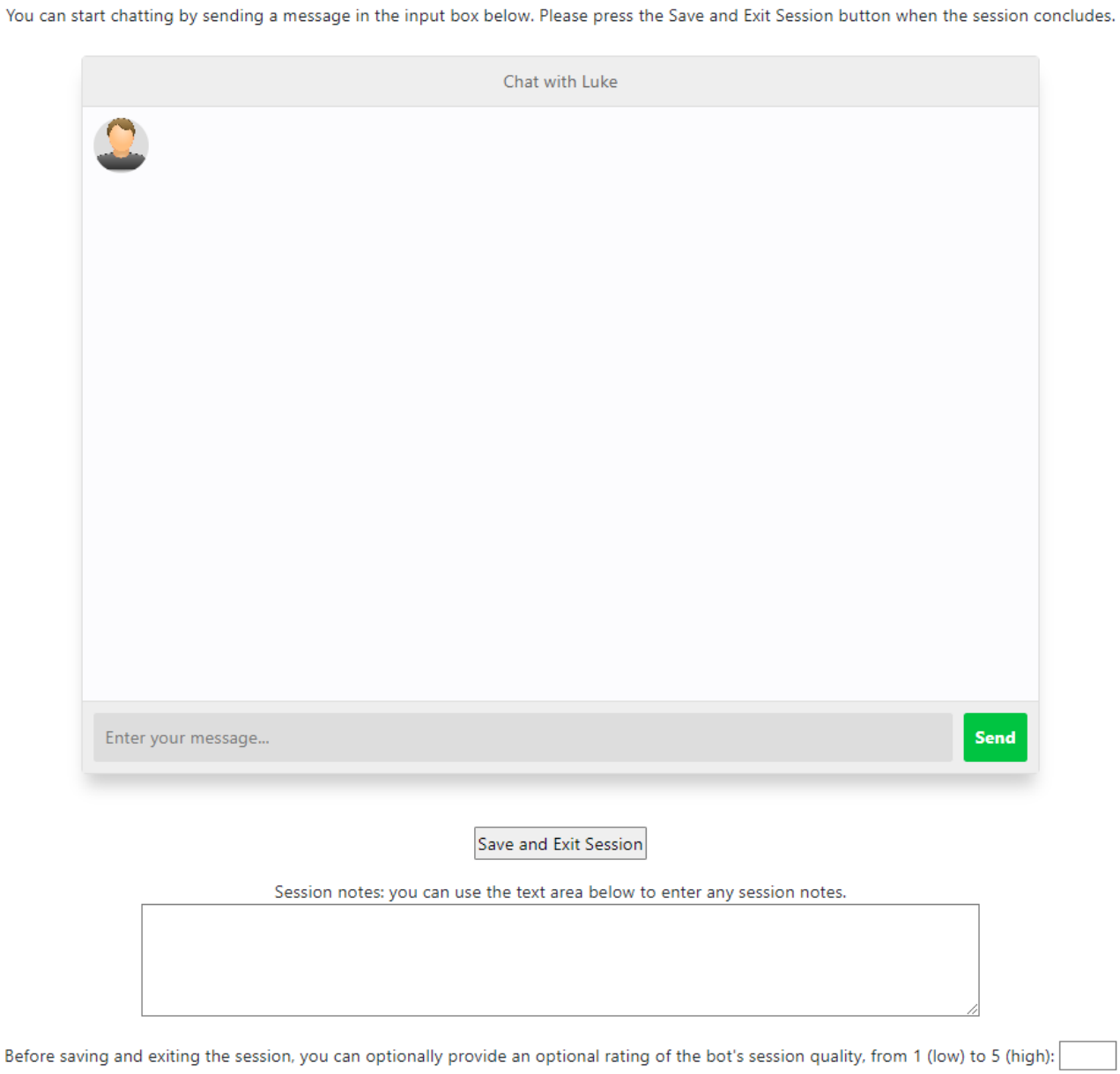
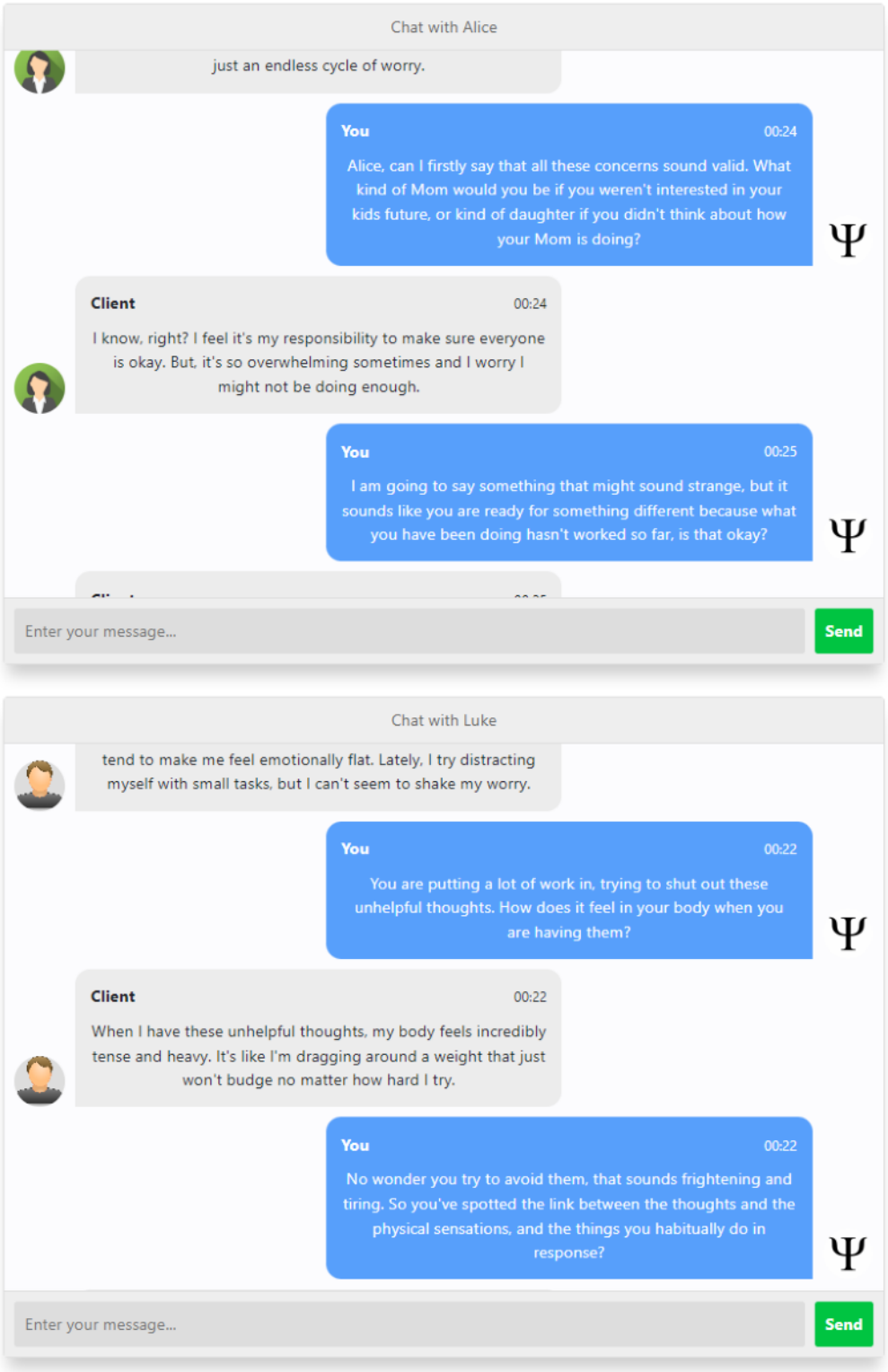


Figure 2. Simulated therapy sessions featuring the top image, of a therapist interacting with Alice, a chatbot portraying a patient with anxiety, and the bottom image, of a therapist engaging with Luke, a chatbot representing a patient with depression.



LLM-Driven Chatbots

Overview

The chatbots available on the Client101 platform used GPT-4 [22] as their LLM, accessed via OpenAI's paid application programming interface. The hyperparameters set for these models include a temperature of 1.0 and a presence penalty of 0. To aid the chatbots in simulating authentic patient interactions, we crafted prompts derived from psychotherapy case illustrations. To enhance the performance of our LLMs, we used the following prompt engineering techniques:

Memory-Assisted Prompt Editing

Inspired by Madaan et al [23], we paired our generative model with an external memory of previous messages. Previous messages were stored in the SQLite database and interfaced

with the LLM via a buffer. Such a memory allowed the LLMs to create better responses by referencing information from previous dialogues. The memory system included a summarizer module that condensed the content of prior messages, ensuring that the length of previous dialogues fit within the context window of the LLM.

Prompt Curation

The prompt design followed the role-play prompting technique [24], where the assigned role or persona provides context for the LLM's identity and background. The chatbots' personas were developed based on clinical vignettes from a psychology textbook [25], guidelines from the National Institute for Health and Care Excellence [26], and feedback from 2 colleagues with experience in psychotherapy. In collaboration with the psychologist, we conducted iterative prompt curation to develop high-quality prompts that effectively simulate mental health clients. A total of 5 sessions with the therapist were held to ensure optimal prompt quality and relevance. The finalized prompts for each chatbot were then used to configure the message-roles parameter in the chat completions application programming interface, with the role type set to developer. The prompts for each chatbot are available in Multimedia Appendix 1.

Human Evaluation

Overview

To assess the mental health professionals' perception of the chatbot's ability to simulate mental health client characteristics, we used a survey designed per the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) [27]:

Design

This study targeted professionals in psychotherapy, specifically clinical psychologists, psychiatrists, and counselors. The survey was a postsession evaluation form completed by participants after interacting with the chatbots. Participants completed a 19-question survey evaluating the general quality of the chatbot, its ability to simulate the mental health condition in question (anxiety or depression), its characteristics in terms of acting as a client, and its suitability for use as a potential training tool.

Development and Testing

Questions 1 to 4 were adapted from the chatbot usability questionnaire [28], with modifications to specify that the chatbots represented psychotherapy clients. Questions 5 to 8 were custom-designed for this survey: questions 5 and 6 assessed whether therapists correctly identified the chatbot's simulated mental health condition, while questions 7 and 8 evaluated the chatbot's usability as an educational tool for trainee psychologists. Guided by the SSI-CBT framework [5], we formulated questions 9 to 18 to assess client characteristics typically observed in SSI-CBT sessions. These questions were reviewed by a mental health professional to ensure their relevance and clarity. Finally, question 19 was an open-ended prompt, allowing mental health professionals to provide additional insights, concerns, or suggestions not covered in the structured questions. Eight of these questions used a Likert scale with response options including "strongly agree," "agree," "neutral," "disagree," "strongly disagree," and "not applicable." Psychologists also had the option to provide additional comments regarding their responses to each question. All the questions are enumerated in Multimedia Appendix 2.

Recruitment Process and Sample Access

The survey was a closed survey, distributed exclusively via email to selected participants. Invitations were sent through direct email contact, with no open online advertisement or public recruitment. To ensure consistency in chatbot interactions, participants were provided with session procedure guidelines (Multimedia Appendix 3).

Survey Administration

The survey was administered electronically after each chatbot session. The questionnaire consisted of 19 questions, including 8 Likert scale-based questions and open-ended response options.

Response Rates

A total of 16 individuals participated, with 12 being clinical psychologists, 2 psychiatrists, and 2 counselors. Further, 13 participants completed sessions with both chatbots, while 3 engaged with only 1 chatbot. The total number of synthetic therapy transcripts generated was 29.

Preventing Multiple Entries

As participants were invited individually, 2 authors of this paper were able to track and verify that no participant submitted multiple entries.

Analysis

All completed surveys were analyzed, including those from participants who interacted with only 1 chatbot.

Therapy Transcript Dataset

To compare the Client101 sessions against therapy sessions involving human clients and therapists, we obtained psychotherapy transcripts from a 2020 study in which an on-demand online mental health chat service was embedded into a mental health web platform [29].

A total of 200 therapy transcripts were collected from this study. From this dataset, we selected sessions if they met one of the following conditions: (1) clients self-identified as experiencing symptoms of anxiety, depression, or both during the therapy session; (2) clients are participating or encouraged to participate in the anxiety or the depression pathways programs, each designed to offer comprehensive support and resources for managing their respective conditions; and (3) clients indicated in the pretherapy session questionnaire that their visit was related to anxiety or depression.

From this dataset, 42 sessions were related to anxiety and 47 sessions were related to depression.

LIWC Analysis

We used the LIWC software to extract psycholinguistic features from the synthetic and organic therapy sessions. An independent samples *t* test was conducted to compare the psycholinguistic features of the sessions with Client101 to those of the single webchat therapy sessions. The selected LIWC features were

based on studies that have identified them as markers of depression and anxiety [20,21]. These *t* tests aim to determine how realistic the Client101 content is in terms of psycholinguistic feature prevalence.

The psycholinguistic features corresponding to each mental health condition were obtained from the following LIWC-22 dictionaries: (1) depression: negative emotions, feel, sadness, anxiety, health, illness, mental, first-person singular, and discrepancy; (2) anxiety: negations, sadness, personal pronouns, present focus, social, family, negative emotions, and anger.

A power analysis was conducted to calculate the sufficient number of samples for a *t* test. With a statistical power of 0.8, a large effect size of 0.8, a type I error of 0.05, and an allocation ration of 2.4 between the organic therapy transcripts and the synthetic therapy transcripts, G*Power (version 3.1.9.6; Heinrich-Heine-Universität Düsseldorf) [30,31] determined that a *t* test would require 34 organic therapy transcripts and 14 synthetic therapy transcripts. We performed 1 *t* test for each psycholinguistic feature associated with each mental health condition, resulting in 9 *t* tests for the depression condition and 8 for the anxiety condition. To account for multiple comparisons, we applied the Bonferroni correction to adjust the *P* values.

Ethical Considerations

Ethics approval was obtained from the low and negligible risk 3B committee at the University of Melbourne Office of Research Ethics and Integrity (2024-27815-51746-4). Upon being provided with a plain language statement outlining the study and signing the consent form, participants were sent an email with guidelines on how to test and evaluate the platform. Session lengths were left to the discretion of the participants, resulting in an average session duration of 43 minutes. Participants were made aware that after the session, they needed to fill out a 19-question survey. To prevent multiple entries, the questionnaire responses contained participant names. This data was handled by investigator SD who had already contacted participants and was responsible for recruiting them. Questionnaire responses were subsequently anonymised for further usage by replacing individual names with identifiers of the form Px.

Results

Overview

In this section, we present our statistical findings derived from comparing the synthetic therapy transcripts from Client101 with the organic therapy transcripts that we obtained. We also report the survey responses from participants who used Client101. The subsequent section is dedicated to a comprehensive discussion and analysis of the implications arising from these outcomes.

Figure 3 compares the LIWC feature distributions related to anxiety in organic and synthetic therapy transcripts. Among the 8 anxiety-related features, 3 of them, negations (t_{56}=4.03, *P*=.001), family (t_{56}=–8.62, *P*=.001), and negative emotions (t_{56}=–3.91, *P*=.002), showed statistically significant differences between the organic and synthetic therapy transcripts. The remaining 5 features, including sadness, personal pronouns, present focus, social, and anger, did not show statistically significant differences. Figure 4 illustrates the survey responses from mental health professionals for the Alice chatbot.

Figure 3. Comparison of LIWC feature distributions associated with anxiety: Organic versus synthetic therapy transcripts. LIWC: Linguistic Inquiry and Word Count.

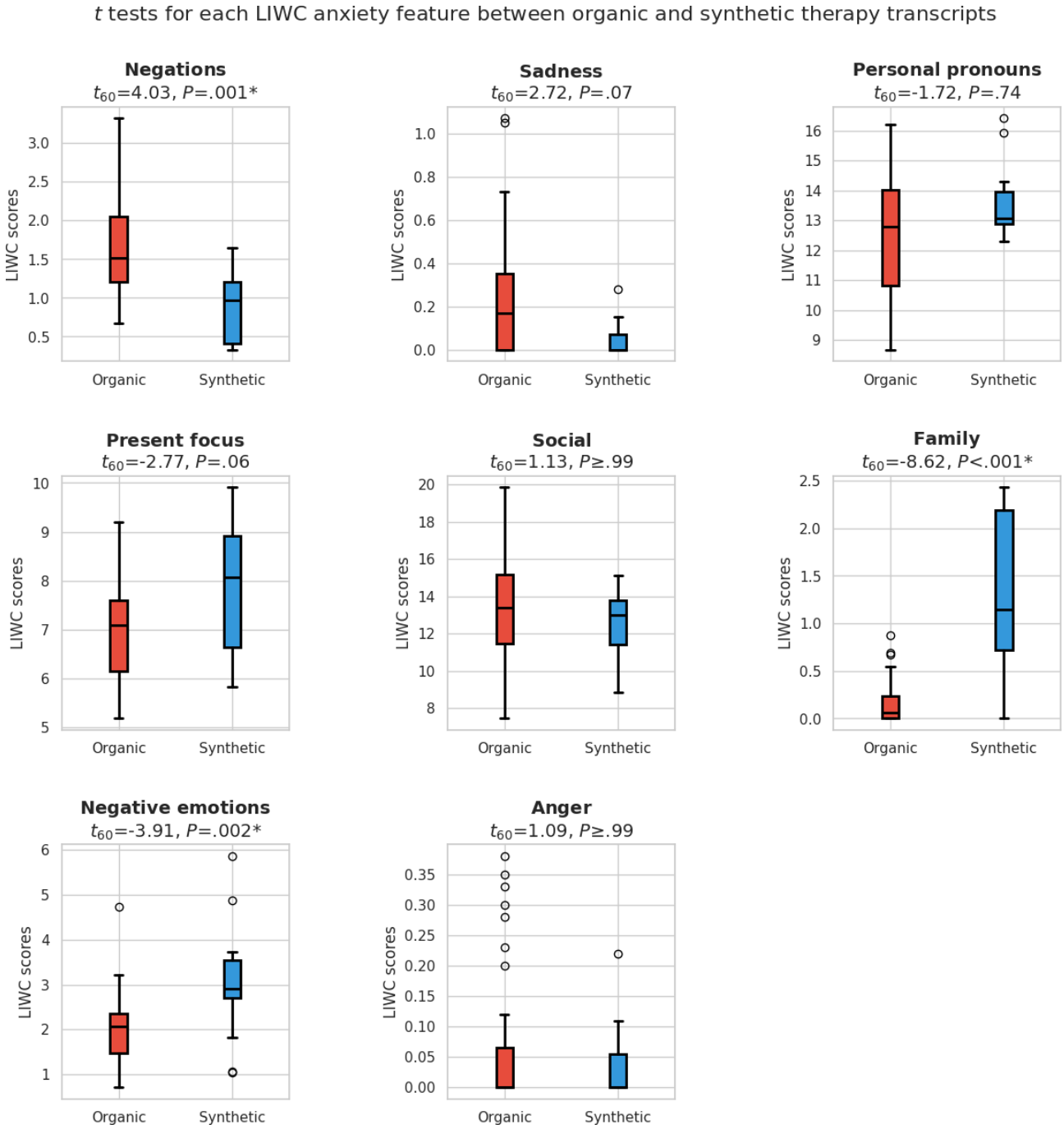


Figure 4. Likert scale responses for the Alice chatbot.

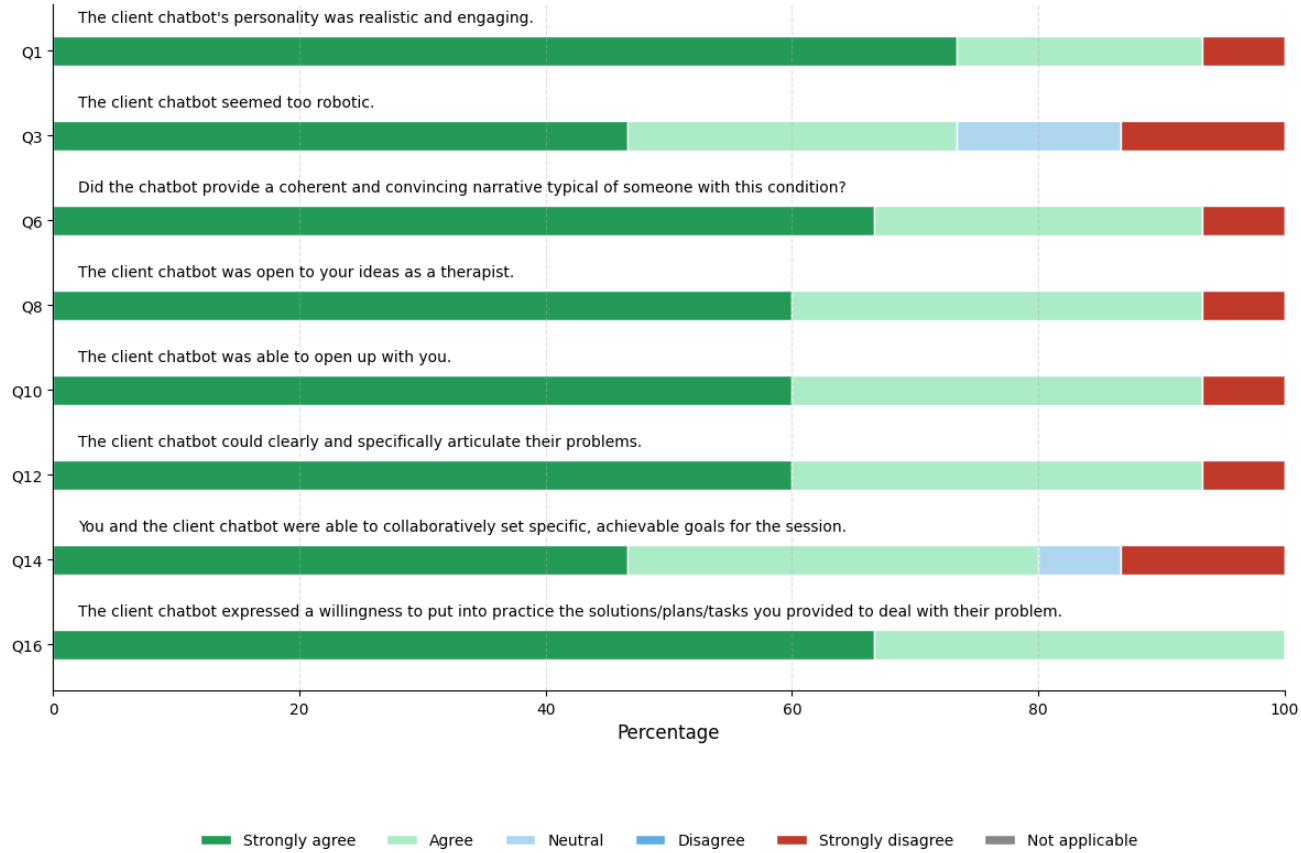


Figure 5 compares the LIWC feature distributions related to depression in organic and synthetic therapy transcripts. Among the 9 depression-related features, 4 of them, negative emotions (t_{60}=–3.84, P=.003), feeling (t_{60}=–6.40, P<.001), health (t_{60}=–4.13, P=.001), and illness (t_{60}=–5.52, P<.001), showed statistically significant differences between the organic and synthetic therapy transcripts. The remaining 5 features, including sadness, anxiety, mental, first personal pronouns, and discrepancy, did not show statistically significant differences. Figure 6 illustrates the survey responses from mental health professionals for the Luke chatbot.

Figure 5. Comparison of LIWC feature distributions associated with depression: Organic versus synthetic therapy transcripts. LIWC: Linguistic Inquiry and Word Count.

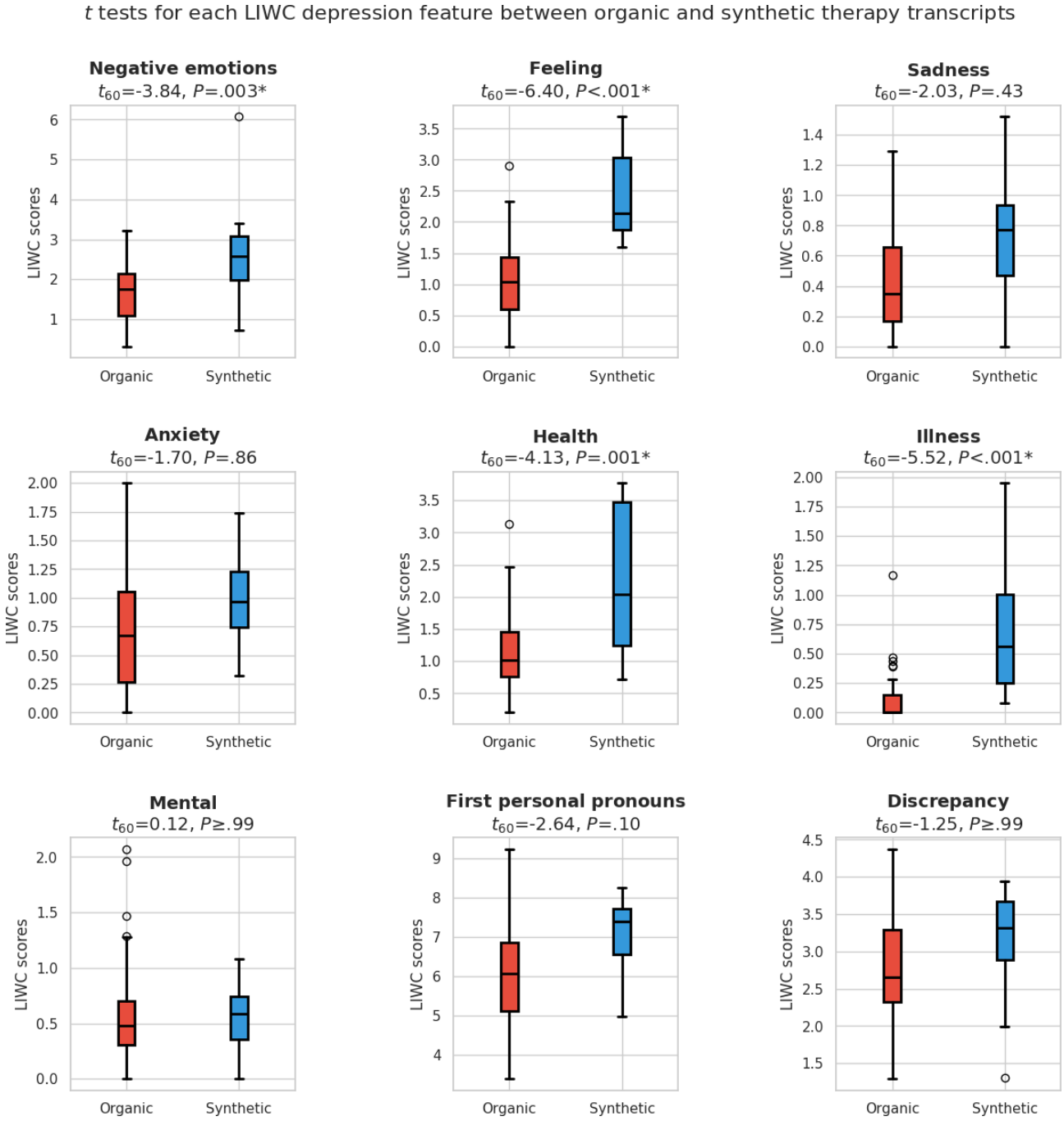
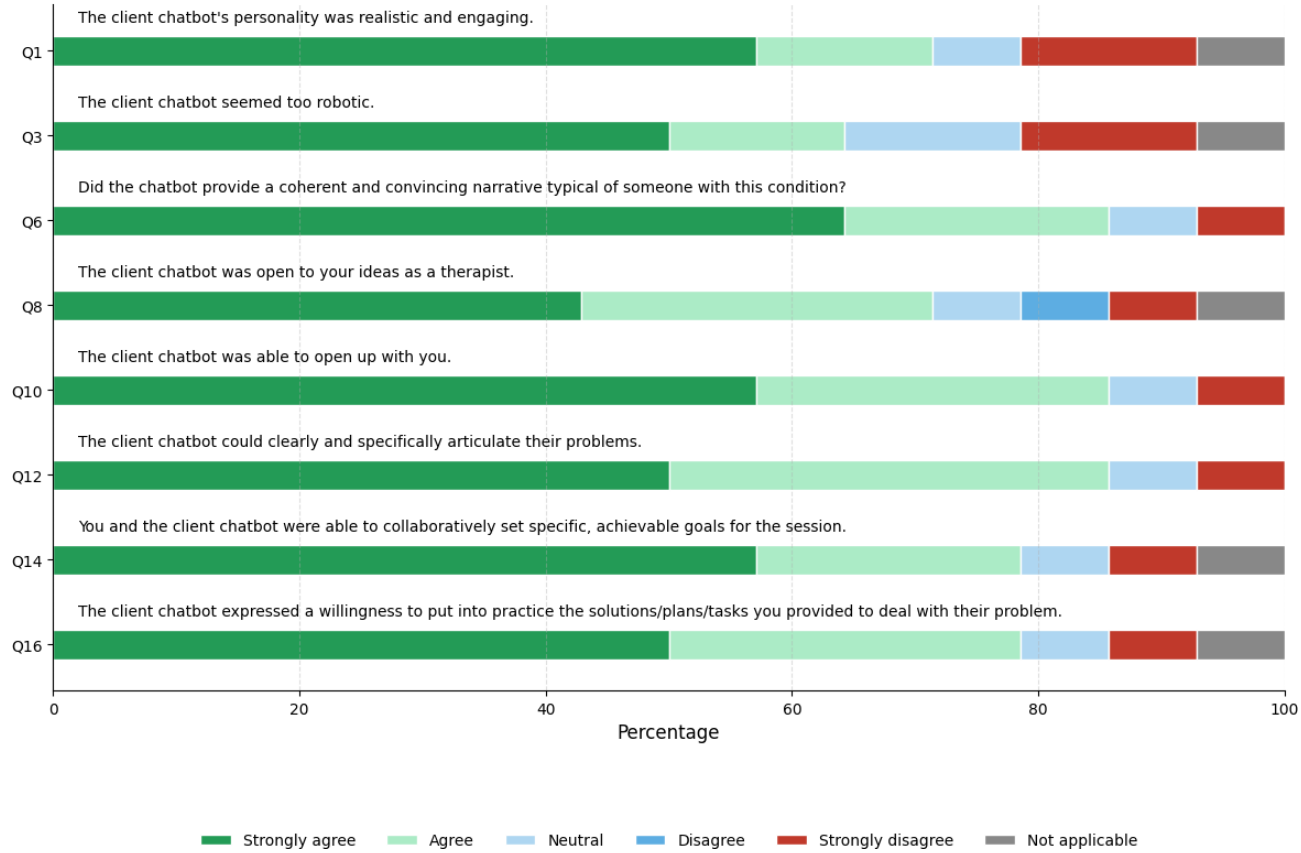


Figure 6. Likert scale responses for the Luke chatbot.



Alice Chatbot

When examining the anxiety-related LIWC features that showed statistically significant differences between organic and synthetic transcripts, the synthetic data exhibited higher values for all features except for negations. Chat-based language models are designed to be helpful and tend to comply with safe queries [32]. In contrast, patients sometimes refuse to engage in certain activities and may say no to therapists' requests or suggestions. This cooperative behavior in chatbots is further supported by survey results for questions 8, 10, 14, and 16, where 93% (14/15) of therapists agreed or strongly agreed for questions 8 and 10, 80% (12/15) for question 14, and 100% (15/15) for question 16 that the Alice chatbot is receptive and willing to collaborate. This is further corroborated by open-ended survey responses:

> *The client chatbot seemed too ready to engage, overly compliant*
>
> *The responses of the bot were appropriate in response to questions, however it was hard to tell whether the bot had learned to be very compliant and easy with the therapist, or if it was the personality of Alice.*

The *t* statistic with the largest magnitude of difference was observed for the family feature, which we attribute to the prompting components within the clinical vignette used to construct Alice's persona. This vignette includes references to concerns about a mother and her children. A common practice in single-session psychotherapy sessions is setting an agenda. This involves the therapist asking the patient what topics they would like to discuss during the session. When responding to such agenda-setting prompts, the Alice chatbot frequently mentions family members, as they are part of its prompt. This frequent reference to family contributes to a higher family score. This openness and specificity in discussing family issues is further supported by survey results for questions 10 and 12, where 93% (14/15) of psychologists agree or strongly agree that Alice can clearly articulate family issues. The following open-ended responses support these claims:

> *I found Alice's language sufficiently realistic for training purposes. Where Alice lacked realism was that she guided me back to the main issue more than a real client might if I took them off track.*
>
> *The bot was pretty good (always felt responsive, didn't get tripped up by spelling errors, or when I hit enter too soon) but felt a bit too clear, compliant, as well as a bit too repetitive. It represents a pretty nice client who can appropriately summarise their thoughts and experiences in an extremely clear way and then is eager to try whatever you give them.*

Despite the 3 anxiety features that showed a statistically significant difference between synthetic and organic transcripts, participants still found that the chatbots provided a coherent and convincing narrative typical of someone with anxiety. This is evident from survey results for questions 1 and 6, where approximately three-fourths of therapists agreed or strongly agreed that the chatbot conducted a realistic therapy session. The following answers support the realism demonstrated by the Alice chatbot:

> *The answers it gave were very human, including paraphrases and reflections. I found it hard to distinguish its responses from human.*
>
> *The bot reminded me of some challenging clients I have worked with who engage in rigid thinking and are still developing insight.*

However, there is still room for improvement, as indicated by the results for question 3, where 73% (11/15) of therapists felt the chatbot seemed too robotic, as illustrated by the following feedback:

> *Some of the responses were robotic. For example, repeating key words too often, which, as mentioned at 4, brought me (the clinician) back to the focal topic in a way that a real client would not.*

Luke Chatbot

Analysis of depression-related LIWC features revealed statistically significant differences in health and illness features between synthetic and organic transcripts. This discrepancy can be attributed to specific elements within the clinical vignette used to construct Luke's persona. The vignette's inclusion of an ankle injury reference influenced the chatbot's responses, particularly when addressing questions about preferred discussion topics for the therapy session. As a result, the Luke chatbot consistently and repeatedly mentioned the injury when responding to inquiries about setting the session agenda, leading to elevated scores in both health and illness categories. This openness and specificity in discussing health issues is further supported by survey results for questions 10 and 12, where 86% (12/14) of psychologists agree or strongly agree that Luke can clearly articulate issues. The following answers from participants support this point:

> *It felt quite repetitive. This isn't necessarily atypical of someone with depression/anxiety though so may not be an issue.*
>
> *Luke tends to be repetitive with the reason for consultation but doesn't elaborate beyond that. Typically, people aren't as repetitive with their symptoms; humans describe their symptoms in various ways, drawing on their biographical memory and recounting specific situations. Luke doesn't do this.*

Despite the 4 depression features that showed statistically significant differences between synthetic and organic transcripts, mental health professionals still found that the chatbots provided a coherent and convincing narrative typical of someone enduring depression. This is evident from survey results for questions 1 and 6, where over two-thirds of therapists agree or strongly agree that the chatbot conducted a realistic therapy session. The following answers support the realism demonstrated by the Luke chatbot:

> *Reminded me of significantly depressed clients that I've previously worked with.*
>
> *The client was realistic in that he was resistant to intervention in a similar way to a real depressed client. It was hard work.*

However, there is still room for improvement, as indicated by the results of question 3, where 64% (9/14) of therapists felt the chatbot appeared too robotic. This sentiment is reflected in the following feedback:

> *The bot struggled to understand some of my questions/responses in ways that very much seemed robotic. One example was when trying to provide validation (ie, not asking a question) it responded with "I don't understand your question.*"

Client 101

Overview

The clinicians' feedback on the Client101 platform reveals a promising tool for training psychologists, while also highlighting areas for enhancement to increase its efficacy and realism. The platform's strengths and potential improvements can be summarized as follows:

Strengths and Training Potential

Of skill development, multiple participants emphasized the platform's value in developing crucial skills. One participant noted:

> *I see it as a valuable instrument capable of honing skills in posing precise questions and adhering to protocols during psychological consultations.*

Of diagnostic practice, participants highlighted its utility in differential diagnosis and criteria assessment. As 1 participant stated:

> *[it serves as a] useful tool for conducting differential diagnoses and developing proficiency in assessing diagnostic criteria, akin to using a checklist.*

Of protocol adherence, the ability to practice following established protocols was highlighted, with 1 participant commenting:

> *I think it's a valuable tool that potentially enables the development of skills for asking the right questions or learning to follow protocols in a psychological consultation.*

Areas for Improvement

Despite its strengths, the participants' feedback indicates significant room for improvement in enhancing the chatbot's human-like qualities. Two key areas for refinement emerged from the responses: response timing and proactive engagement.

Of response timing, therapists noted that the chatbot's near-instantaneous response generation created a perceptible sense of artificiality. As 1 participant observed:

> *Responses were very quick! Could some delay be added to feel more "human"?*

Of proactive engagement, therapists also emphasized the need for the chatbot to initiate conversations or offer insights without requiring user prompts. As 1 clinician observed:

> *Not having an in-person session always makes it less human-like. Also, having to first ask something before*

getting a response makes it feel more like a chatbot than an actual real person.

Discussion

Principal Findings

Our key findings for each research question are:

1. When analyzing the statistical differences in LIWC features between organic and synthetic transcripts, we found that more features associated with thematic content exhibited statistical significance differences than features related to writing style.
2. Mental health professionals highlighted the platform training value, particularly for developing critical clinical skills in formulating precise questions, practicing differential diagnosis, assessing diagnostic criteria, and following established therapeutic protocols during psychological consultation.
3. Mental health professionals highlighted areas for improvement in the chatbot's human-like qualities, specifically in response timing, proactive engagement, and the chatbot's responses, which were perceived as robotic.

Implications of LIWC Findings

Interestingly, LIWC features associated with writing style, such as the use of personal pronouns and present-focused language, did not show a statistically significant difference between the organic and synthetic data. Additionally, psychological process features, including mental and social dimensions, also did not reveal significant differences. These results provide a foundation for future work that explores an equivalence or similarity test for these features, as the lack of significant differences may suggest that GPT-4 powered chatbots not only captured surface-level linguistic patterns but also effectively modeled deeper psychological constructs typically associated with depressive and anxious states.

However, when inspecting the 2 writing style features that did show significant differences, negation and negative emotions, the findings illustrate important implications for using Client101 as a training tool. The greater use of negation in the organic data may indicate a limitation in the types of clients the chatbots could model. For instance, in the organic transcripts, patients sometimes refused to follow advice or suggestions from the therapist, whereas the chatbot consistently complied with the suggestions. This difference suggests that while chatbots can simulate certain aspects of human communication, they may not fully replicate the more complex, resistance-based interactions seen with real clients, thus impacting their effectiveness as a training tool in some therapeutic scenarios.

Whereas the greater use of negative emotions in the synthetic data may indicate a limitation in how well the chatbots can represent the full emotional range experienced by clients. This could suggest that the chatbot models may lean toward portraying more extreme emotional states more constantly, potentially limiting their ability to simulate the nuanced emotional dynamics that occur in real-life therapeutic interactions.

In general, all the LIWC features tend to have higher values for the synthetic data. This finding implies that GPT-4 may be amplifying or intensifying certain linguistic characteristics associated with depression and anxiety. This amplification could be attributed to a pattern of over-generalization by the model, in which it has identified key linguistic patterns associated with depression and anxiety and applied them more consistently or intensely than typically observed in organic data, where individual variation is more prominent. Yet, this variation is not achieved due to the prompt influence, which constantly instructs the model to follow the persona of a depressed or anxious patient.

In the case of the LIWC features associated with thematic content that presented a significant difference, such as illness and family, we attribute these differences to the specific prompts provided to the LLMs. These prompts inherently contained elements associated with family dynamics and health concerns, thus influencing the thematic content of the generated text. While the prompts used to instruct the LLMs may reflect the diversity of topics encountered in 1 real therapeutic session, this could become a limitation if the platform were used to simulate multiple sessions with the chatbot portraying a patient. Over time, the topics could become stale and repetitive, as the chatbot would continually focus on the same themes, whereas real clients typically bring up a broader and more varied range of issues based on their individual experiences. This potential lack of diversity in simulated sessions could limit the platform's effectiveness in training mental health professionals to handle the evolving nature of a client's concerns and the wide array of therapeutic situations encountered in real-world practice.

Implications of Survey Findings

Therapist feedback revealed a consensus that both chatbots exhibited robotic interactions. When chatbots fail to replicate authentic human communication, they limit trainees' opportunities to practice essential therapeutic competencies, such as building rapport, responding to emotional cues, or managing the unpredictability of real client interactions. This rigidity risks creating a distorted representation of therapeutic dialogue, potentially conditioning trainees to approach client interactions with a structured, algorithmic mindset rather than the fluid, empathetic engagement required in mental health practice.

Responses to the open-ended survey questions highlighted key factors contributing to this robotic perception, including response timing, proactive engagement, and repetitiveness. These insights provide valuable guidance for improving the platform's human-like qualities. For instance, implementing a variable delay mechanism, where response times fluctuate within a natural human range, could significantly enhance the authenticity of interactions by eliminating the unnatural immediacy of automated replies. Similarly, programming the chatbot to occasionally initiate meaningful follow-up questions or offer unsolicited insights could create a more dynamic and human-like interaction.

The repetitiveness partly arises from the use of static prompts to configure message-role parameters, which rigidly instruct the LLM in the same manner for every interaction, constraining

natural variability in responses. To address this while preserving the chatbot's core therapeutic persona, a dynamic prompting approach could be implemented. This method would retain fixed persona elements that uphold the chatbot's essential identity while integrating session-specific details into the prompt.

Another option to reduce the robotic perception of the chatbot responses is exploring different prompt engineering techniques that aim to enhance the emotional generation aspect of LLMs. One such prompt engineering technique is Emotional Chain-of-Thought [33], which aligns LLMs with human emotional intelligence guidelines to enhance emotional generation capabilities. Emotional Chain-of-Thought has been shown to improve the quality of emotionally charged responses, making interactions with AI more relatable and human-like.

Strengths, Limitations, and Future Work

To strengthen the claims and results obtained in our analysis, this research could benefit from a larger sample size of synthetic therapy sessions. Due to practicalities, we limited the number of participants to 16, with the aim that each conduct 2 virtual sessions, 1 with each of the 2 chatbots. Future studies could increase the number of mental health professionals involved or the number of sessions each psychologist performs, thereby enhancing the robustness and generalizability of the findings.

The next step for Client101 is to evaluate its effectiveness as an educational tool for training and assessing psychologists. We are currently running a study in which a sample of psychology students is testing out Client101. Via their questionnaire responses and thematic analysis of their follow-up interviews, we are gathering preliminary feedback and perspectives from such student end users.

Client101 will subsequently be, as a trial, embedded into the curricula of 2 psychology subjects at the University of Melbourne in 2025. In these subjects, students are required to report on the delivery of psychological intervention to a client. However, this assignment poses challenges related to the equivalency of client presentations, the severity of client issues, and the number of client sessions each student can conduct. The incorporation of Client101 aims to address these issues by providing a standardized and equitable platform for students to demonstrate their intervention skills. By using Client101, students will engage in a prescribed number of sessions with the chatbot, ensuring a controlled and comparable environment for skill demonstration. This approach will enhance the ability to assess student competency more effectively. Students will be required to download the transcripts of their sessions with Client101 and critically appraise and evaluate their performance. This process will facilitate a more consistent and objective assessment of their intervention skills, ultimately improving the training and evaluation of future psychologists.

Future research could explore adapting the Client101 platform to other languages and using the LIWC-22 dictionaries in those languages to evaluate the psycholinguistic features of the generated transcripts. If therapy transcript databases in these languages are available, it would be valuable to measure the similarities between synthetic and organic therapy transcripts across different languages. Furthermore, comparing and contrasting the psycholinguistic features of therapy transcripts in various languages could provide valuable insights into how LLMs capture and reflect language-specific nuances, cultural contexts, and psychological dynamics. Beyond measuring LIWC-22 features, it is crucial to involve mental health professionals who are native speakers of the corresponding languages to evaluate the quality of the sessions. Assessing the quality of non-English synthetic text is particularly important, as multiple studies have shown that the performance of LLMs tends to be worse in non-English tasks [34].

In this research, we used clinical vignettes to construct prompts for simulating patients with depression or anxiety. Future research could develop new clinical vignettes that represent a broader range of backgrounds and demographics than those used in our study. Additionally, future studies could explore the simulation of other types of mental health conditions, such as obsessive-compulsive disorder, schizophrenia, or comorbid conditions such as anxiety and depression. Other types of therapies, such as dialectical behavior therapy and motivational interviewing, could also be tested to evaluate how well an LLM can simulate a patient in those therapeutic contexts. Group therapies could also be explored, with one or multiple LLMs simulating different patients. This approach could provide valuable training environments by reducing the challenges and costs associated with hiring actors to portray mental health patients.

Further development of the Client101 platform remains an open task. The minor improvement of adding organic response delays, as mentioned earlier, is 1 simple example. More generally, there are opportunities to incorporate more advanced mechanisms into the chatbot architecture. For example, using expanded prompt constructions or more advanced LLM-related information storage techniques, such as retrieval-augmented generation, could enable the construction of a more detailed set of client details and the accumulation of facts revealed during chat sessions. Ultimately, the goal is to have chatbots that can sustain a succession of therapy sessions and demonstrate a therapeutic evolution as their sessions progress.

Finally, exploring the use of open LLMs could offer significant benefits for the Client 101 platform, particularly in overcoming the limitations of third-party services such as OpenAI's models. One key drawback of proprietary models is the presence of built-in guardrails that may restrict certain responses, which poses a challenge when simulating mental health patients. For example, outputs related to sensitive topics such as suicide or self-harm may be blocked, limiting the realism of the simulation. Additionally, proprietary LLMs require paid access, which could constrain the number of simulated sessions due to cost limitations. In contrast, open-source models can be deployed without these restrictions, provided users have the necessary computational resources to run them locally. Another significant advantage of open-source LLMs is enhanced data security, as no information would be transmitted to third parties, further strengthening the platform's privacy and safety measures.

Other Ethical Considerations

While using conversational agents powered by LLMs offers significant potential as a tool for training psychologists, they also raise notable ethical considerations.

In our experimental approach, Client101 was designed to closely mimic patients with mental health problems. However, further research is needed to assess potential biases that the model could be exhibiting when simulating mental health patients. LLMs can manifest various types of biases in their outputs, including gender, racial, and religious biases [35-37]. Studies have shown that these stereotypes and biases exert adverse effects on mental health treatment outcomes [38,39]. Therefore, it is essential to examine these models to prevent the inadvertent propagation of such biases.

Additionally, a primary concern in using Client101 is the potential for boundary transgressions, traditionally observed in human therapist-client interactions. In conventional therapy, therapists may inadvertently cross boundaries through behaviors such as excessive self-disclosure, forming dual relationships, or displaying undue affection toward clients [40]. The introduction of chatbot clients raises questions about whether trainee clinicians might exhibit similar behaviors when interacting with AI entities. For instance, the "ELIZA effect"—the tendency to unconsciously attribute human-like qualities to computer programs—could lead trainees to engage in inappropriate self-disclosure or develop misplaced trust in the chatbot's responses [41].

Furthermore, the use of AI-driven chatbots in psychotherapy training necessitates a thorough understanding of their limitations and ethical implications. Counselors must receive adequate training to comprehend these limitations and ensure that AI serves as a supportive tool rather than a replacement for human judgment. This approach aligns with ethical guidelines emphasizing the importance of professional competence and the responsible integration of technology in clinical settings [42].

Moreover, while chatbots can simulate certain aspects of client interactions, they lack the genuine emotional depth and autonomy of human clients. Relying heavily on chatbot interactions during training may impede the development of essential relational skills, such as empathy and the ability to navigate complex human emotions, which are crucial for establishing effective therapeutic relationships. Therefore, it is imperative to assess whether trainees can effectively use chatbots as an initial training tool without experiencing adverse effects on their ability to engage authentically with human clients in real-world scenarios.

Conclusions

We developed Client101, a web conversational platform featuring LLM-driven chatbots designed to simulate mental health clients. To evaluate the chatbots' performance, participants engaged in single sessions with the chatbots and provided questionnaire responses. Comparative analysis was undertaken to examine the psycholinguistic features of the sessions conducted with Client101 and single webchat therapy sessions obtained from an actual online counseling service. This comparison aimed to evaluate the resemblance between the AI-generated and human-generated transcripts, focusing on linguistic and psychosocial indicators associated with depression and GAD.

Notably, LIWC features related to thematic content showed a significant difference between the organic and the synthetic data, likely due to the explicit inclusion of these themes in the prompts. Interestingly, no significant differences were detected in the means of psychosocial process features, such as mental and social dimensions, within the limits of our sample size and data variability. The absence of significant differences suggests potential for future work that explores equivalence or similarity testing for these features, offering further insights into the capacity of GPT-4, and LLMs more generally, to model deeper psychological constructs.

These findings, along with feedback from therapists, have important implications for the utility of Client101 as a simulation tool, its potential as a training resource for mental health professionals, and its applications in psycholinguistic and computational psychiatry research. The platform's value is enhanced by its availability and accessibility compared to established training methods, such as using actors to simulate mental health patients for training psychologists. Moreover, Client101 offers flexibility in creating a wide variety of mental health personas, capable of simulating different types of mental health disorders. This versatility, combined with its ability to resemble actual therapy sessions, makes Client101 a promising tool for mental health education, training, and research.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Virtual patient's prompt.
[DOCX File, 16 KB - mededu_v11i1e68056_app1.docx]

Multimedia Appendix 2
Questionnaire.
[DOCX File, 16 KB - mededu_v11i1e68056_app2.docx]

Multimedia Appendix 3
Single-session integrated CBT. CBT: cognitive-behavioral therapy.
[DOCX File, 17 KB - mededu_v11i1e68056_app3.docx]

References

1. Rønnestad MH, Ladany N. The impact of psychotherapy training: introduction to the special section. Psychother Res 2006 May;16(3):261-267. [doi: 10.1080/10503300600612241]
2. Tanana MJ, Soma CS, Srikumar V, Atkins DC, Imel ZE. Development and evaluation of clientbot: patient-like conversational agent to train basic counseling skills. J Med Internet Res 2019 Jul 15;21(7):e12529. [doi: 10.2196/12529] [Medline: 31309929]
3. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. Can J Psychiatry 2019 Jul;64(7):456-464. [doi: 10.1177/0706743719828977] [Medline: 30897957]
4. Frangoudes F, Hadjiaros M, Schiza EC, Matsangidou M, Tsivitanidou O, Neokleous K. An overview of the use of chatbots in medical and healthcare education. In: Zaphiris P, Ioannou A, editors. Learning and Collaboration Technologies: Games and Virtual Environments for Learning: Springer; 2021:170-184. [doi: 10.1007/978-3-030-77943-6_11]
5. Dryden W. Single-Session Integrated CBT: Routledge; 2021. [doi: 10.4324/9781003214557]
6. Weizenbaum J. ELIZA—a computer program for the study of natural language communication between man and machine. Commun ACM 1966 Jan;9(1):36-45. [doi: 10.1145/365153.365168]
7. Colby KM, Weber S, Hilf FD. Artificial paranoia. Artif Intell 1971;2(1):1-25. [doi: 10.1016/0004-3702(71)90002-6]
8. Abd-Alrazaq AA, Alajlani M, Ali N, Denecke K, Bewick BM, Househ M. Perceptions and opinions of patients about mental health chatbots: scoping review. J Med Internet Res 2021 Jan 13;23(1):e17828. [doi: 10.2196/17828] [Medline: 33439133]
9. Adamopoulou E, Moussiades L. An overview of chatbot technology. In: Maglogiannis I, Iliadis L, Pimenidis E, editors. Artificial Intelligence Applications and Innovations: Springer; 2020:373-383. [doi: 10.1007/978-3-030-49186-4_31]
10. Vinyals O, Le Q. A neural conversational model. arXiv. Preprint posted online on Jun 19, 2015. [doi: 10.48550/arXiv.1506.05869]
11. Lison P, Tiedemann J. OpenSubtitles2016: extracting large parallel corpora from movie and TV subtitles. In: Calzolari N, Choukri K, Declerck T, Goggi S, Grobelnik M, Maegaard B, et al, editors. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) Portorož: European Language Resources Association (ELRA); 2016:923-929.
12. Alexander Street Press. Counseling and psychotherapy transcripts: volume I. : Redivis; 2023. [doi: 10.57761/C9DA-ZQ22]
13. Zhang S, Dinan E, Urbanek J, Szlam A, Kiela D, Weston J. Personalizing dialogue agents: i have a dog, do you have pets too? Presented at: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Jul 15-20, 2018; Melbourne, Australia. [doi: 10.18653/v1/P18-1205]
14. Zhang Y, Sun S, Galley M, et al. DIALOGPT: large-scale generative pre-training for conversational response generation. Presented at: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; Jul 6-8, 2020 p. 270-278 URL: https://www.aclweb.org/anthology/2020.acl-demos [accessed 2025-06-04] [doi: 10.18653/v1/2020.acl-demos.30]
15. Li Y, Zeng C, Zhong J, Zhang R, Zhang M, Zou L. Leveraging large language model as simulated patients for clinical education. arXiv. Preprint posted online on Apr 13, 2024. [doi: 10.48550/arXiv.2404.13066]
16. Stapleton L, Taylor J, Fox S, Wu T, Zhu H. Seeing seeds beyond weeds: green teaming generative AI for beneficial uses. arXiv. Preprint posted online on May 30, 2023. [doi: 10.48550/arXiv.2306.03097]
17. Demasi O, Hearst MA, Recht B. Towards augmenting crisis counselor training by improving message retrieval. Presented at: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology; Jun 6-8, 2019; Minneapolis, Minnesota. [doi: 10.18653/v1/W19-3001]
18. Chancellor S, De Choudhury M. Methods in predictive techniques for mental health status on social media: a critical review. NPJ Digit Med 2020;3(1):43. [doi: 10.1038/s41746-020-0233-7] [Medline: 32219184]
19. Boyd RL, Ashokkumar A, Seraj S, Pennebaker JW. The development and psychometric properties of LIWC-22. : The University of Texas at Austin; 2022. [doi: 10.13140/RG.2.2.23890.43205]
20. Eichstaedt JC, Smith RJ, Merchant RM, et al. Facebook language predicts depression in medical records. Proc Natl Acad Sci U S A 2018 Oct 30;115(44):11203-11208. [doi: 10.1073/pnas.1802331115] [Medline: 30322910]
21. Rook L, Mazza MC, Lefter I, Brazier F. Toward linguistic recognition of generalized anxiety disorder. Front Digit Health 2022;4:779039. [doi: 10.3389/fdgth.2022.779039] [Medline: 35493530]
22. OpenAI, Achiam J, Adler S. GPT-4 technical report. arXiv. Preprint posted online on Mar 15, 2023. [doi: 10.48550/arXiv.2303.08774]
23. Madaan A, Tandon N, Clark P, Yang Y. Memory-assisted prompt editing to improve GPT-3 after deployment. Presented at: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing; Dec 7-11, 2022; Abu Dhabi, United Arab Emirates. [doi: 10.18653/v1/2022.emnlp-main.183]

24. Kong A, Zhao S, Chen H, et al. Better zero-shot reasoning with role-play prompting. Presented at: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics; Jun 16-21, 2024; Mexico City, Mexico. [doi: 10.18653/v1/2024.naacl-long.228]
25. Sipe WEB, Eisendrath SJ. Chapter 3 - mindfulness-based cognitive therapy for treatment-resistant depression. In: Baer RA, editor. Mindfulness-Based Treatment Approaches, 2nd edition: Academic Press; 2014:61-76. [doi: 10.1016/B978-0-12-416031-6.00003-7]
26. NICE. Generalised Anxiety Disorder and Panic Disorder in Adults: Management: London: National Institute for Health and Care Excellence (NICE); 2019.
27. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
28. Holmes S, Moorhead A, Bond R, Zheng H, Coates V, Mctear M. Usability testing of a healthcare chatbot: can we use conventional methods to assess conversational user interfaces? Presented at: ECCE '19: Proceedings of the 31st European Conference on Cognitive Ergonomics; Sep 10, 2019; Belfast, United Kingdom. [doi: 10.1145/3335082.3335094]
29. Alvarez-Jimenez M, Rice S, D'Alfonso S, et al. A novel multimodal digital service (moderated online social therapy+) for help-seeking young people experiencing mental ill-health: pilot evaluation within a national youth e-mental health service. J Med Internet Res 2020 Aug 13;22(8):e17155. [doi: 10.2196/17155] [Medline: 32788151]
30. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007 May;39(2):175-191. [doi: 10.3758/BF03193146]
31. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 2009 Nov;41(4):1149-1160. [doi: 10.3758/BRM.41.4.1149]
32. Brahman F, Kumar S, Balachandran V, et al. The art of saying no: contextual noncompliance in language models. Presented at: 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks; Dec 10-15, 2024; Vancouver, Canada.
33. Li Z, Chen G, Shao R, Xie Y, Jiang D, Nie L. Enhancing emotional generation capability of large language models via emotional chain-of-thought. arXiv. Preprint posted online on Jan 12, 2024. [doi: 10.48550/arXiv.2401.06836]
34. Lozoya D, Berazaluce A, Perches J, Lúa E, Conway M, D'Alfonso S. Generating mental health transcripts with SAPE (spanish adaptive prompt engineering). In: Duh K, Gomez H, Bethard S, editors. Presented at: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics; Jun 16-21, 2024; Mexico City, Mexico. [doi: 10.18653/v1/2024.naacl-long.285]
35. Lozoya DC, D'Alfonso S, Conway M. Identifying gender bias in generative models for mental health synthetic data. Presented at: 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI); Jun 26-29, 2023; Houston, TX. [doi: 10.1109/ICHI57859.2023.00109]
36. Abid A, Farooqi M, Zou J. Persistent anti-muslim bias in large language models. Presented at: AIES '21; May 19-21, 2021 URL: https://dl.acm.org/doi/proceedings/10.1145/3461702 [accessed 2025-06-04] [doi: 10.1145/3461702.3462624]
37. Shihadeh J, Ackerman M, Troske A, Lawson N, Gonzalez E. Brilliance bias in GPT-3. Presented at: 2022 IEEE Global Humanitarian Technology Conference (GHTC); Sep 8-11, 2022; Santa Clara, CA. [doi: 10.1109/GHTC55712.2022.9910995]
38. Wirth JH, Bodenhausen GV. The role of gender in mental-illness stigma. Psychol Sci 2009 Feb;20(2):169-173. [doi: 10.1111/j.1467-9280.2009.02282.x]
39. Chatmon BN. Males and mental health stigma. Am J Mens Health 2020;14(4):1557988320949322. [doi: 10.1177/1557988320949322] [Medline: 32812501]
40. Aravind VK, Krishnaram VD, Thasneem Z. Boundary crossings and violations in clinical settings. Indian J Psychol Med 2012 Jan;34(1):21-24. [doi: 10.4103/0253-7176.96151] [Medline: 22661802]
41. Natale S. The ELIZA effect: Joseph Weizenbaum and the emergence of chatbots. In: Deceitful Media: Oxford Academic; 2021:50-67. [doi: 10.1093/oso/9780190080365.003.0004]
42. Meadi MR, Sillekens T, Metselaar S, van Balkom A, Bernstein J, Batelaan N. Exploring the ethical challenges of conversational ai in mental health care: scoping review. JMIR Ment Health 2025 Feb 21;12:e60432. [doi: 10.2196/60432] [Medline: 39983102]

Abbreviations

AI: artificial intelligence
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
GAD: generalized anxiety disorder
***ICD*:** *International Classification of Diseases*
LIWC: Linguistic Inquiry Word Count
LLM: large language model
LSTM: long short-term memory
NLP: natural language processing
SSI-CBT: single-session integrated cognitive behavioral therapy

Edited by B Lesselroth; submitted 27.10.24; peer-reviewed by A Shirazi, F Shojaei, M Beg; revised version received 29.03.25; accepted 06.05.25; published 31.07.25.

Please cite as:
Cabrera Lozoya D, Conway M, Sebastiano De Duro E, D'Alfonso S
Leveraging Large Language Models for Simulated Psychotherapy Client Interactions: Development and Usability Study of Client101
JMIR Med Educ 2025;11:e68056
URL: https://mededu.jmir.org/2025/1/e68056
doi:10.2196/68056

Collaborative Development of Feedback Concept Maps for Virtual Patient–Based Clinical Reasoning Education: Mixed Methods Study

Anja Mayer[1], MPH; Inga Hege[2], MSc, MD; Andrzej A Kononowicz[3], PhD; Anja Müller[1], MD; Małgorzata Sudacka[4], MD

[1]Medical Education Sciences, University of Augsburg, Augsburg, Germany

[2]Institute for Research in Health Science Education, Brandenburg Medical School Theodor Fontane, Neuruppin, Germany

[3]Department of Bioinformatics and Telemedicine, Jagiellonian University Medical College, Kraków, Poland

[4]Department of Medical Education, Center for Innovative Medical Education, Jagiellonian University Medical College, Kraków, Poland

Corresponding Author:
Anja Mayer, MPH
Medical Education Sciences, University of Augsburg, Augsburg, Germany

Abstract

Background: Concept maps are a suitable method for teaching clinical reasoning (CR). For example, in a concept map, findings, tests, differential diagnoses, and treatment options can be documented and connected to each other. When combined with virtual patients, automated feedback can be provided to the students' concept maps. However, as CR is a nonlinear process, feedback concept maps that are created together by several individuals might address this issue and cover perspectives from different health professionals.

Objective: In this study, we aimed to develop a collaborative process for creating feedback concept maps in virtual patient–based CR education.

Methods: Health professionals of different specialties, nationalities, and levels of experience in education individually created concept maps and afterward reached a consensus on them in structured workshops. Then, medical students discussed the health professionals' concept maps in focus groups. We performed a qualitative content analysis of the transcribed audio records and field notes and a descriptive comparison of the produced concept maps.

Results: A total of 14 health professionals participated in 4 workshops, each with 3 - 4 participants. In each workshop, they reached a consensus on 1 concept map, after discussing content and presentation, as well as rationales, and next steps. Overall, the structure of the workshops was well-received. The comparison of the produced concept maps showed that they varied widely in their scope and content. Consensus concept maps tended to contain more nodes and connections than individual ones. A total of 9 medical students participated in 2 focus groups of 4 and 5 participants. Their opinions on the concept maps' features varied widely, balancing between the wish for an in-depth explanation and the flexibility of CR.

Conclusions: Although the number of participating health professionals and students was relatively low, we were able to show that consensus workshops are a constructive method to create feedback concept maps that include different perspectives of health professionals with content that is useful to and accepted by students. Further research is needed to determine which features of feedback concept maps are most likely to improve learner outcomes and how to facilitate their construction in collaborative consensus workshops.

(JMIR Med Educ 2025;11:e57331) doi:10.2196/57331

KEYWORDS

clinical reasoning; consensus building process; concept map; consensus map; virtual patient; international collaboration; health professionals' education; undergraduate; collaborative; development; feedback; content analysis; health professional; medical student; mixed method; Europe; questionnaire; descriptive analysis

Introduction

Background

"Clinical reasoning encompasses health professionals thinking and acting in assessment, diagnostic, and management processes in clinical situations taking into account the patient's specific circumstances and preferences" [1]. It is evident that health professionals in different disciplines (eg, physicians and nurses) differ in their reasoning approaches [2], and there are differences between novices and experts [3]. Even experienced health professionals of the same discipline do not follow the same

diagnostic process, even when they are confronted with the same medical case, and ultimately arrive at the same diagnosis [4,5]. A study by Charlin et al [6] showed that experts' case solutions also varied depending on the situation, for example, whether they were asked to give answers as an examinee or as a panel member.

The variety and nonlinearity of possible clinical reasoning (CR) approaches make CR training and assessment a highly complex matter [4,7]. Therefore, concept maps have been suggested as a useful method for training the CR skills of medical students [4,8], especially in terms of problem representation [9].

Concept mapping is a method used to represent concepts and their relationships in a visual diagram, using explanatory terms to relate concepts to each other [10]. A typical use in health education is to present students with a case scenario and have them create a concept map to represent their thought process as the case unfolds [9,11]. They can record relevant findings, tests, differential diagnoses, and treatment options and connect concepts to each other to visualize their CR process [8]. Torre et al [12] show that concept maps promote the connection between theory and practice and facilitate knowledge integration and critical thinking.

Teachers can ask students to create concept maps in different forms, depending on the purpose, such as freely from scratch or in a preconstructed form [10,11]. Because creating a comprehensive and accurate concept map is time-consuming and students need some time to learn how to do it, Daley and Torre [8] suggest the use of semistructured concept maps.

Concept maps have also been found to be suitable for measuring learning outcomes [13], and various ways of assessing and scoring concept maps, both qualitatively and quantitatively, have been described in the literature [14-17]. A study by Morse and Jutras [18] showed that working with concept maps had an effect on the students' problem-solving performance only when feedback was provided. However, in order to provide students with feedback on their concept map, some form of "expert concept map" is needed to compare students' results with [19], which can then be provided in real time in digital environments. Such "expert concept maps" can be created by a single teacher or by a panel of professionals or experts [19-21]. In their systematic review of different methods for assessing CR skills, Daniel et al [9] concluded that "using written cases, expert consensus is the most prevalent method" used to create concept maps as feedback for students. However, little is known about the process and challenges involved when health professionals are asked to reach a consensus on a concept map for teaching CR.

Recent studies suggest that virtual patients (VPs) are an appropriate method for training CR [22-24], especially for some components of this process, such as collecting data, generating differential diagnoses, or developing a treatment plan [25,26]. VPs are computer-based patient case scenarios that students can interact with [27]. Often, such scenarios are designed so that the cases gradually lead the student to the final diagnosis by providing more and more information over time [28,29]. VPs provide a safe environment, in which mistakes can be made without harming real patients [30]. It has been suggested that combining concept map activities with VPs can reinforce the educational effect of VPs in CR outcomes [31]. The importance of VPs has increased over the years [32], especially since the beginning of the COVID-19 pandemic, when direct patient contact and opportunities for CR training were limited [33].

Objectives

In this study, we aimed to develop a collaborative process for creating feedback concept maps in VP-based CR education. From this, we derive the following research questions: (1) What are the similarities and differences of concept maps for teaching CR that have been created by individual health professionals and groups? (2) What themes emerge when health professionals are asked to jointly create a concept map in a consensus workshop? (3) What are the challenges and benefits of such consensus workshops? (4) What aspects of the consensus concept maps do medical students find helpful in learning CR?

Methods

Study Design

This study followed a convergent mixed methods approach. First, we asked health professionals from different disciplines to individually create concept maps for 2 VPs that would serve as feedback for medical students. We then conducted structured digital workshops for those health professionals in which they reached a consensus on the concept maps. After the workshops were finished, we conducted focus groups with medical students to discuss which aspects of the professionals' concept maps they found helpful for learning CR.

Ethical Considerations

The study was approved by the institutional review board of the Ludwig-Maximilians-University, Munich, Germany (21 - 0941), and adhered to ethical guidelines. Informed consent was obtained from all participants prior to their participation in the study, with assurances of anonymity and confidentiality. Participants were informed of the objectives of the study and how the data collected would be used. In addition, strict measures were taken to protect the privacy and confidentiality of the study data. Students who participated in the focus groups received a US $16 voucher as compensation for their time.

Data Collection

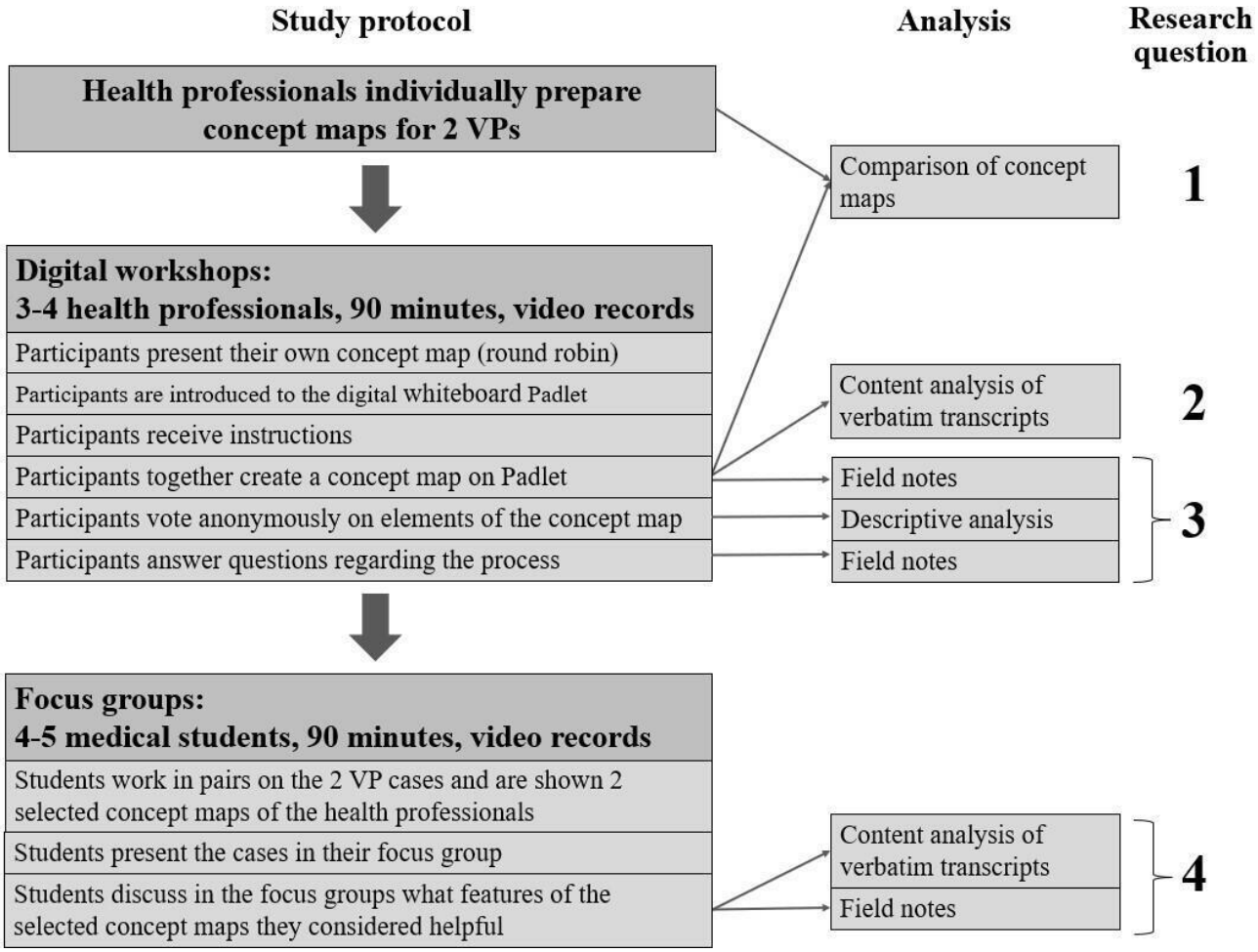
Between November 2021 and January 2022, we sent out emails to invite health professionals of different specialties, nationalities, and levels of experience in medical education to participate in our study. The email included study information and a written informed consent form. After returning the signed written consent by email, participants were asked to create concept maps for 2 VP cases. We used the software platform CASUS (Instruct gGmbH), which is a VP player and authoring environment with integrated concept map functionality [31]. The study participants were told that these concept maps would serve as feedback for medical students. We carefully chose the VPs with regard to their sociodemographic features, key symptoms, and difficulty levels. They were a 19-year-old female student with mononucleosis and a 58-year-old male nurse with hepatitis E. We chose them for providing patients of different

sex, age, and profession. Both were heterosexual and of Caucasian origin. We deliberately chose these VPs because they provided different levels of difficulty for the students but had key symptoms that are common in daily practice and easily recognizable by health professionals of different specialties. The 2 VPs can be found on the CASUS platform and are part of a collection of over 200 freely available VPs in 6 languages [34]. These VPs include a semistructured concept map that students fill out while solving the case, covering 4 categories: findings, differential diagnoses, tests or examinations, and treatments [31].

Participants were also asked to complete a web-based questionnaire that included personal data and level of experience in concept mapping and teaching CR. We used a convenience sampling strategy, inviting partners from 2 recent Erasmus+ projects, iCoViP (International Collection of Virtual Patients) and DID-ACT (Developing, Implementing, and Disseminating an Adaptive Clinical Reasoning Curriculum for Healthcare Students and Educators) [35,36], who were interested in teaching CR with concept maps. As the VPs are available in multiple languages, we valued the international composition of the study group that would reduce local bias in clinical practice. Participants were given 10 days to create the individual concept maps and were reminded of the task 3 days before the workshop. After that, we held structured digital workshops of 90 minutes, where they met in groups of 3 - 4 to reach a consensus on a common concept map. The workshops took place on the Zoom platform (Zoom Video Communications) and were video recorded. A Mayer and MS facilitated the workshops, following a predefined structure according to the nominal group technique [37,38] (Figure 1): first, all participants explained their own concept maps to the others in a round robin and described their reasoning. Then, A Mayer and MS introduced them to the digital whiteboard Padlet (Wallwisher Inc), and the participants had the opportunity to try it out. Once they felt comfortable with the tool, A Mayer and MS gave them instructions on how to create a concept map together. They then created a new concept map on Padlet based on their individual concept maps. A Mayer and MS answered participants' questions, kept track of the timeline, and reminded participants of the original assignment if they strayed from the topic. When the concept map was complete, A Mayer and MS provided the opportunity to anonymously rate the concepts and connections with a thumbs up or down mechanism on Padlet. Afterward, they asked the participants about their experience of creating the concept map together. IH and AAK attended the workshops as neutral observers and, together with A Mayer and MS, took field notes, which they all discussed immediately after the workshop. The study was piloted as a face-to-face workshop in October 2021. Afterward, we decided that web-based meetings would be equally feasible and made minor changes to the study protocol, such as adding an anonymous voting round.

After all workshops were completed, IH and MS selected 4 individual and 4 consensus concept maps to be discussed by medical students in focus groups. For this purpose, 9 international medical students were recruited to participate in 90-minute focus groups during a transnational meeting of the iCoViP project. Written informed consent was obtained prior to participation. In the beginning, the students were asked to work in small teams (2 - 3 students) and solve 1 of the 2 VP cases together. Afterward, the teams were shown 2 of the selected concept maps from the workshops to compare and decide which one they would prefer to have as feedback for their case and why. Then, 2 teams of students who had worked on different cases were brought together as a focus group. They presented their cases to the others and then started a group discussion, facilitated by A Mayer and MS, about which of the presented concept maps they found most helpful and how different features of the concept maps could improve their CR process.

Figure 1. Flowchart of study protocol and analysis plan with regard to the research questions. VP: virtual patient.



Data Analysis

This study used a convergent mixed methods design. In the quantitative part, a descriptive statistical analysis of the questionnaire, concept maps, and votes was performed using Microsoft Excel. The individual and consensus concept maps were analyzed for scope (number of nodes and connections), agreement (number of "likes" of nodes or connections in the consensus phases of the group concept map authors), and content (number of times a particular concept, eg, "fever," appeared in the individual and consensus concept maps). We extracted information from the concept maps and compared them separately for each of the 2 cases.

The qualitative part of the study involved the thematic analysis of the transcripts and field notes from the workshops and focus groups. It was conducted in several steps. The recordings of the workshops were transcribed verbatim and anonymized. Two authors (A Mayer and A Müller) performed a thematic analysis of the transcripts, following the 6 steps for qualitative content analysis proposed by Kuckartz [39]. Using an inductive approach, they independently created codes for the first 2 workshops and reached a consensus on an initial coding framework. They then coded 1 workshop at a time, applying and refining the coding framework in an iterative process. They used MAXQDA software (version Analytics Pro 2022; VERBI GmbH) for coding and discussed discrepancies until a consensus was reached. A Mayer, A Müller, and IH then grouped similar codes into themes. Throughout the process, AAK and MS reviewed the coding framework and emerging themes and provided feedback; discrepancies were discussed until a consensus was reached.

We analyzed field notes taken during the workshops and participants' responses during the round of questions for challenges and benefits. Student focus group recordings were transcribed verbatim and anonymized. Two authors (A Mayer and A Müller) independently extracted statements from the transcripts about what students found helpful in the selected concept maps, grouped them into themes, and discussed discrepancies until a consensus was reached. Finally, we looked for confirmation or discrepancies of the results obtained from the mixed methods.

Results

Participants

A total of 14 health professionals from 6 European countries participated in our study, of whom 9 were female and 5 were male. On average, participants were 37 (SD 10) years of age and had 10 (SD 9) years of professional experience. Participants worked in different disciplines (Table 1) and had an average of 6 (SD 5) years of experience in health education. Participants differed only slightly in their teaching experience with concept maps or CR.

Table . Characteristics of participating health professionals (N=14).

Characteristics	Values
Age (years), mean (SD)	37 (10)
Sex, n (%)	
Female	9 (64)
Male	5 (36)
Country (place of work), n (%)	
France	1 (7)
Germany	3 (21)
Poland	2 (14)
Portugal	1 (7)
Spain	4 (29)
Sweden	3 (21)
Specialty, n (%)	
Internal medicine	4 (29)
Nursing	2 (14)
Biochemistry	2 (14)
Rheumatology	2 (14)
Family medicine	1 (7)
Neurology	1 (7)
Paramedic	1 (7)
Occupational medicine	1 (7)
Professional experience of physicians (n=9), n (%)	
Resident	6 (67)
Consultant	3 (33)
Working experience (years), mean (SD)	10 (9)
Experience in health teaching (years), mean (SD)	6 (5)
Experience in teaching with concept maps, n (%)	
None	9 (64)
Some	5 (36)
Much	0 (0)
Experience in teaching clinical reasoning, n (%)	
None	5 (36)
Some	9 (64)
Much	0 (0)

Participants created 13 individual concept maps prior to the workshops. We held 4 digital workshops with 3‐4 participants each, resulting in 4 consensus concept maps (2 hepatitis E and 2 mononucleosis). We also conducted 2 focus groups with 4 and 5 medical students, respectively. The students were in their final year of study (sixth year), with an average age of 24 (SD 0.5) years. In total, 8 students were female, and 1 was male. We chose students from Portugal (n=5) and Poland (n=4) because these countries represent educational systems from different parts of Europe.

Research Question 1: Comparison of Individual and Consensus Concept Maps

The individual concept maps varied widely from each other regarding scope and content. We found most similarities in the final diagnoses and treatment options and only a few similarities regarding findings, differential diagnoses, and tests. The same was true when comparing the consensus concept maps.

When we compared the consensus concept maps to the individual versions, we found that they all had a bigger scope than the individual concept maps, as can be seen in Table 2 and

in the examples given in Figure 2 (original images are provided in Multimedia Appendix 1). We also found that most of the nodes from the individual concept maps were present in the consensus versions, and only in a few cases were nodes left out or new nodes added during the workshops. Altogether, the consensus versions showed higher similarities to the underlying individual versions than to each other. All consensus concept maps included connections, while these were missing in 5 of the individual versions.

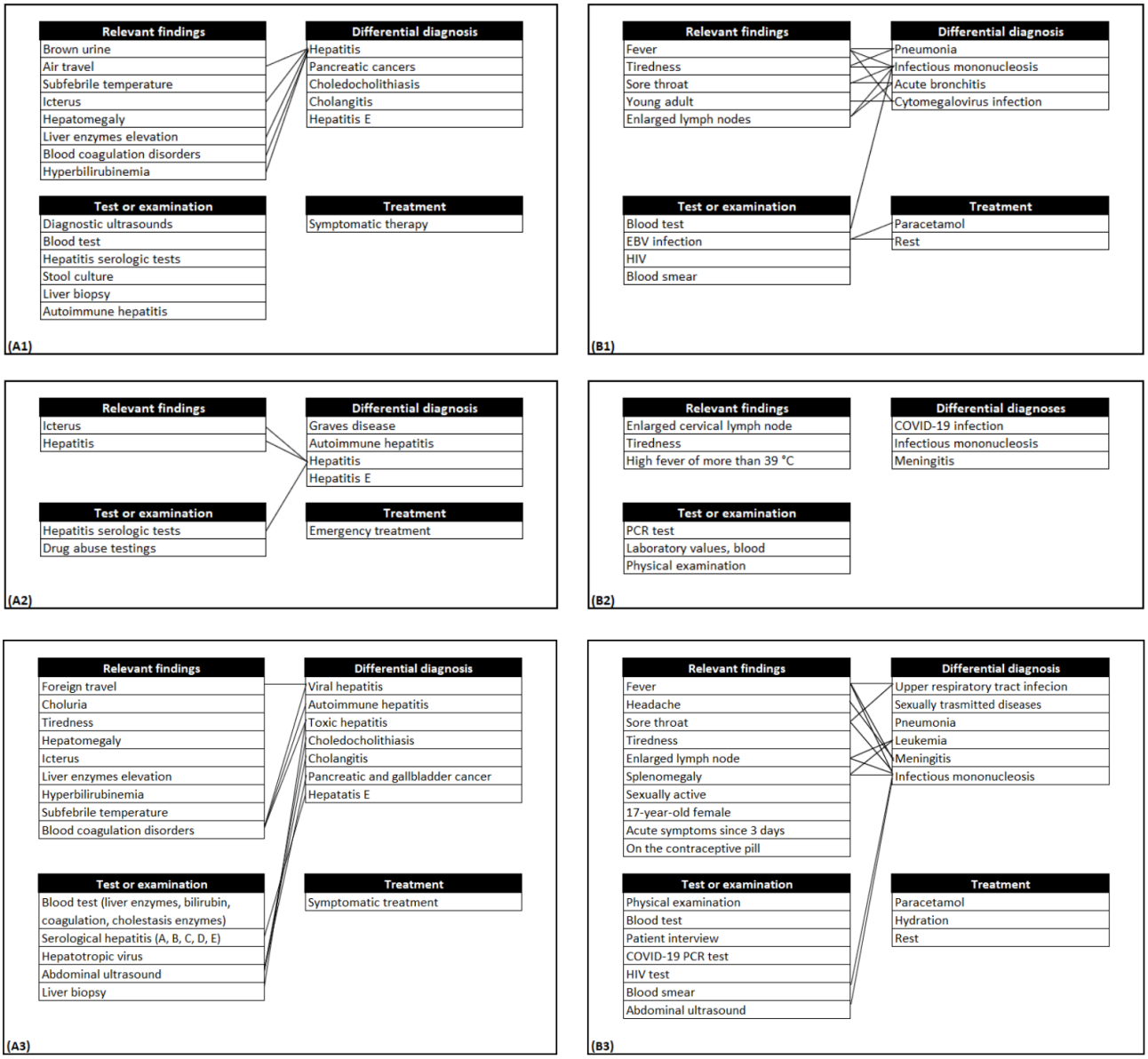
Table . Comparison of number of elements in consensus and individual concept maps.

	Hepatitis E				Mononucleosis			
	Workshop 1		Workshop 2		Workshop 3		Workshop 4	
	n (GRP)[a] (%)	Rn (IND)[b]	n (GRP) (%)	Rn (IND)	n (GRP) (%)	Rn (IND)	n (GRP) (%)	Rn (IND)
Total	56 (100)	10-29	33 (100)	11-25	39 (100)	13-22	43 (100)	9-11
Nodes								
Findings	9 (16)	6-8	9 (27)	1-8	10 (26)	3-6	12 (28)	3-4
Examinations or tests	11 (20)	2-8	5 (15)	2-5	7 (18)	2-4	7 (16)	0-3
Differential diagnoses	11 (20)	1-8	7 (21)	3-6	7 (18)	4-7	6 (14)	3-7
Treatments	1 (2)	1-1	1 (3)	1-1	3 (8)	0-2	2 (5)	0-0
Connections	24 (43)	0-13	11 (33)	3-5	12 (31)	0-13	16 (37)	0-0

[a]n (GRP): number of elements in the group consensus concept map.

[b]Rn (IND): range of element number in the individual concept maps.

Figure 2. Examples of individual concept maps and consensus versions. (A1 and A2) Individual concept map (hepatitis E), (A3) consensus concept map (hepatitis E), (B1 and B2) individual concept map (mononucleosis), and (B3) consensus concept map (mononucleosis). EBV: Epstein-Barr virus; PCR: polymerase chain reaction



Research Question 2: Themes Emerging During the Consensus Process

From the qualitative content analysis of the creation phase, we identified 4 themes: the first theme covered the content of the consensus concept maps, that is, participants discussed which findings, examinations or tests, differential diagnoses, treatments, and connections should be included. Related to this, the second theme was the rationales they gave during their discussion, that is, why they thought something should (not) be part of the consensus concept map. The third theme covered the presentation of the consensus concept map, that is, participants discussed how to present the content. In the fourth theme, participants discussed their next steps, that is, how to approach the creation of the consensus concept map. Table 3 shows the 4 themes and associated subthemes, including sample quotes, and their frequency in the verbatim transcripts of the workshops.

Table . Themes and associated subthemes derived from the qualitative content analysis.

Themes and subthemes	Description	Sample quotes	Values, n (%)
Content of the consensus concept maps (n=677)			
Relevant findings	Which findings (not) to include in the concept map	“Sore throat” [Workshop 4]	227 (41)
Differential diagnoses	Which differentials (not) to include in the concept map	“Should we add EBV and CMV as a differential?” [Workshop 1]	200 (30)
Tests or examinations	Which tests or examinations (not) to include in the concept map	“Liver biopsy could be relevant” [Workshop 2]	170 (25)
Treatment	Which treatments (not) to include in the concept map	“Paracetamol I think is perfect in this case” [Workshop 3]	25 (4)
Connections	Which connections (not) to include in the concept map	“Maybe this splenomegaly should be also connected?“ [Workshop 3]	55 (8)
Presentation of the content (n=89)			
Layout	Visual aspects of the concept map or possible actions such as highlighting, crossing out, rearranging, merging, splitting, enlarging, or reducing nodes	“I’m just putting this in a nice order” [Workshop 3], “Can I change the color of the connections?” [Workshop 1]	40 (45)
Categorization	Which heading a node should be assigned to	“Could we add other headings, for example ‘recommendations’?” [Workshop 4]	6 (7)
Phrasing	Use of synonyms or abbreviations	“Is writing ‘STDs’ appropriate or should I use ‘sexually transmitted disease’?” [Workshop 3]	32 (36)
Level of granularity	Detail level of the concepts	“Viral hepatitis, autoimmune hepatitis or [just] hepatitis?” [Workshop 2]	11 (12)
Rationales (n=318)			
Medical relations	Medical relations between the concepts, including the probability of differential diagnoses	“I think about it because it’s a young female on the pill” [Workshop 3], “We think about it because it’s quite common” [Workshop 1]	96 (30)
Relevance	Highlighting the medical urgency or indicating that most participants are of the same opinion	“It’s a potentially dangerous situation for our patient” [Workshop 3], “And also COVID-19, I think we all agree” [Workshop 4]	42 (13)
Individual concept maps	Referring to individual concept maps or clinical reasoning process when creating them	“In the individual mapping, we have PCR-test, someone wrote that” [Workshop 4], “Was this something you came up with now during this process or [when creating your concept map]?” [Workshop 1]	62 (19)
Referring to the case	Referring to the case by quoting or repeating facts	“He’s not saying that he takes any drugs” [Workshop 2], “The case has provided us a biopsy” [Workshop 1]	65 (20)
Professional experience	What participants have experienced in daily practice or what they are accustomed to doing	“This is usually the first serology I order” [Workshop 1]	19 (6)
Common knowledge	General phenomena in society or “universal truths”	“People lie – he might be an alcoholic” [Workshop 2] “One would expect that this nurse is already immunized” [Workshop 1]	5 (2)

Themes and subthemes	Description	Sample quotes	Values, n (%)
Encounter setting	Regional standards or differences between facilities (hospital, general practice, etc)	"How [do] you have it in Spain or Germany?" [Workshop 4], "I was wondering whether I would have done urine analysis in the [general] practice" [Workshop 1]	9 (3)
Hindsight	Assumption that the consensus process might be unconsciously guided by already knowing the final diagnosis	"[We think so because] we already know that it's mononucleosis" [Workshop 3]	6 (2)
Didactical aspects	What could be helpful for the students or is the content appropriate for their level of knowledge, etc	"It could be a good training for students, to think what can cause hepatitis" [Workshop 1], "I think it's too specialistic" [Workshop 3]	8 (3)
Functionality of the VP[a] platform	Features, navigation, or structure of the CASUS platform	"I don't know if this is possible on CASUS" [Workshop 2]	6 (2)
Next steps (n=107)			
Developing a strategy	How to approach the creation of the concept map	"[Let's] do differentials first before adding anything to tests" [Workshop 3]	78 (73)
Referring to facilitators	Referring to instructions given by facilitators or directly asking them for advice	"[It depends on] what is wanted or what is expected" [Workshop 1]	29 (27)

[a]VP: virtual patient.

Research Question 3: Challenges and Benefits of the Workshops

The results presented here are a summary of the field notes from the creation phase and the final round of questions, expanded by a descriptive analysis of the voting round. From a technical point of view, there were some problems due to the digital format, for example, weak network signal, low audio quality, or some participants feeling uncomfortable using Padlet for the first time. Since none of the participants were native English speakers, some struggled to find the right terms or misunderstood what others were saying due to a lack of vocabulary or the speaker's accent.

Regarding the different disciplines, it seemed that participants who had worked in their specialty for many years were somewhat biased by their daily experiences and had difficulty seeing the cases from a student's point of view. Some of the participants who were not physicians by training struggled to find the right diagnosis and expressed their uncertainty about certain medical terms or conditions. It was noticeable that topics such as didactic purpose, uncertainty (probability of differential diagnoses), or logical arrangement of nodes were hardly discussed.

All participants were cooperative and reached a consensus on the concept maps in an amicable manner. For about 10% (n=10) of the nodes, half (or more) of the participants abstained from voting or gave a thumbs down. We compared these nodes with the verbatim transcripts and found that for 6 nodes, there was no evidence in the discussion that any of the participants disagreed.

In the final round of questions, participants reported that creating a concept map with others was a complex task. On the other hand, participants found the group work helpful in stimulating their reflection and that it was constructive to create concept maps that included perspectives of different health professionals. In general, the structure of the workshops and the given timeline for the different parts were well-received. The round robin was seen as a useful introduction that helped them to understand the reasoning of other participants. Some participants mentioned that it was difficult to create a concept map for a case that they had not developed themselves or that they struggled with the fact that the case evolved over time, which made it more difficult to agree on a final version. Participants had mixed feelings regarding the usefulness of the consensus concept maps. While some were satisfied with the final concept maps and expected them to be helpful for students, others found the concept maps too messy or crowded in the end.

Research Question 4: What Students Considered as Helpful

When the medical students were asked whether they preferred the individual or consensus concept maps, there was a slight tendency toward the consensus versions as they contained more findings, which the students found helpful for their own CR process.

Regarding the content and scope of the concept maps, there was agreement that there should not be too many connections between nodes, as this was seen as more confusing than helpful. However, the students expressed contradictory opinions regarding the nodes. While some preferred the concept maps with only the most relevant nodes, others preferred those with a wider scope, as these would contain "the most details that we also agreed on while we were solving the case."

The same was true for the presentation of the content. Some students suggested having more layout features, such as "some type of colors" or dropdown functions, while others preferred a clear design and simple structure. Regarding the granularity of the nodes, some suggested that "the feedback map should be more [general.] To give us freedom" and should use broad terms such as "blood test." Others said that in their medical school, they "can't just say 'do blood test,' [but] must be very specific"; therefore, more specific terms would be helpful in the concept maps.

Discussion

Main Findings

In this study, we described the process of collaborative authoring of concept maps to serve as feedback in CR education using VPs. The participants regarded the collective process stimulating for reflection and helpful to understand the perspectives of the other health professional groups. We were able to find confirmation for this qualitative finding quantitatively by showing that the consensus concept maps contained more nodes and connections than the individual ones. This can also have negative aspects, as in the consensus workshops, participants tended to collect all nodes from the individual concept maps into the consensus version instead of selecting only the most relevant ones, paying little attention to didactic aspects. The structure of the workshops was well-received, participants appreciated working in interprofessional groups and easily reached a consensus, supporting their additions to the concept maps by high scoring of the concept map elements. However, there were some challenges, such as technical problems or participants being biased by their daily practice as specialists.

The final-year medical students in our focus groups preferred a variety of features of the concept maps, most of which were contradictory. As a result, it remains unclear which features can improve learners' outcomes and whether consensus concept maps are more suitable for teaching CR than individual ones.

Implications of the Findings

Our research suggests that there are a few approaches to help health professionals reach a consensus on a concept map. The procedure we used for the workshops served its purpose and was well-received by the participants. Thus, the results of this study can be seen as an important step toward establishing a sound consensus concept map protocol, informing about the benefits and challenges, and leading to the following recommendations for improving the process in the future:

1. Regarding the technical aspects of the workshops, we recommend that participants be given access to the digital whiteboard prior to the workshop so that those who wish to can familiarize themselves with the tool in advance.
2. Since didactic aspects played a minor role in the creation of the consensus concept maps, we recommend that an independent person with experience in didactics and concept mapping participates at the workshop. An alternative would be to prepare a pedagogical guide or checklist to be considered when developing concept maps for teaching CR. If such an opportunity arises, addressing the pedagogical aspects of concept map development would be a helpful element of faculty development courses on VP authoring.
3. When considering concept maps for VPs that address general CR skills in medicine, such as the one in the iCoViP project repository, workshops should preferably involve only internal medicine or family medicine physicians to avoid specialty bias. This would be different if the goal of the VPs was to achieve learning objectives for specialty or interprofessional education from the outset.

In terms of real-world implementation, we consider this study an important step in providing more diverse feedback to students working on CR concept maps in the context of VPs. This study contributed to this by showing that the concept maps created by consensus groups were more elaborate, both in terms of representing many viewpoints and in terms of the number of concepts and connections. However, this study also showed that the consensus groups should be more effectively encouraged to discuss the pedagogical aspects of the concept maps, such as how to adjust the complexity to the level of knowledge or cognitive load of the students.

Limitations

Our study has several limitations. First, the number of concept maps underlying the quantitative analysis was limited, so that the corresponding results should be interpreted with caution.

Second, it is possible that the results of the workshop are not applicable to "real-world" situations, in which colleagues work together on a concept map without being observed. The participants in our workshops were very polite to each other and tended to avoid disagreements, probably because most of them did not know each other. On the other hand, we were able to include the perspectives of professionals from different disciplines.

Third, we had a limited number of students in the focus groups. Our data suggest that the effect of different features of concept maps on individual learning and preferences may vary considerably from student to student. Future research is needed to explore this in more depth.

Fourth, the sample size of VPs and workshop participants was limited, which might make our findings less generalizable. However, we did not see any new themes emerging in the subsequent workshops and focus groups, suggesting that the qualitative analysis had reached its saturation point.

Comparison With Prior Work

There is a large body of literature on the so-called "group concept mapping" [14,40], including approaches to optimizing group compositions [41] or to identifying different cognitive styles [42]. However, to the best of our knowledge, most of these studies only include undergraduate students. Therefore, our study can be considered unique in proposing a novel approach to consensus concept mapping for health professionals.

The structure of the workshops, derived from the nominal group technique and adjusted to the needs of digital education, can be seen as a major strength. First, the round robin allows participants to gain insight into one another's CR approach.

Second, participants found the consensus creation of the concept maps useful and inspiring. Third, anonymous voting at the end facilitates the interpretation of the final concept map, as it gives the participants' view on each individual element without the need to openly criticize someone else's ideas.

When we compared the individual concept maps, we found a common tendency but also a great deal of variation, with most having only the final diagnosis and treatment in common. This is supported by the work of McGaghie et al [43,44], which shows the wide variety of approaches to a concept map of pulmonary physiology, even among experts in the same field. Therefore, the consensus process in our approach increased the universality of the feedback concept maps. Another positive aspect is that the process contributed to a rational increase in the number of connections in the concept maps. As previous research has shown, well-chosen connections are an important element of this form of knowledge representation, which is helpful in CR education [45].

We did not exclude from the study professionals without teaching experience, as we did not see clear evidence that this might be a limiting factor in creating meaningful concept maps. This is supported by a study by Charlin et al [46], who found that teaching and nonteaching physicians were similarly well suited to be part of the reference panel for concordance tests used to assess complexity and ambiguity in CR.

While some authors suggest the use of concept maps for CR assessment [47], most researchers in the field are ambivalent on this issue [8,9]. This is consistent with our findings, which suggest that the complexity and variety of the CR process make it very challenging to generate "expert concept maps" that can be used as a gold standard against which student versions can be compared.

Participants reported that they found the consensus workshops useful for reflecting on their individual concept maps and CR approach. Therefore, such workshops could also be a suitable tool for improving the concept map development of the individual participants. Further research is needed to determine the impact of the workshops on participants' ability to develop concept maps.

Our study focused on the creation of concept maps for medical students. Future research should investigate how this can be applied to other professions, as a recent meta-analysis suggests that concept maps are also an appropriate method for improving critical thinking skills in nursing students [48].

Conclusions

By providing feedback concept maps that illustrate the complexity and diversity of the CR process, we aim to support students in reflecting on their own thinking. The collaborative creation of concept maps for teaching CR is an opportunity to integrate different perspectives of health professionals and to account for individual differences in the reasoning process. In our study, we described a process for developing such collaborative concept maps and identified themes that emerged in workshops using this process. The resulting consensus concept maps tended to contain more nodes and connections than those created by individual health professionals and were well-received by students. We consider this study an important step in establishing a robust method for collaboratively creating effective concept maps in CR education. Future studies will focus on streamlining the process and identifying the most effective pedagogical features of feedback concept maps.

Acknowledgments

The creation of these resources has been partially funded by the Erasmus+ grant program of the European Union (grant 2020-1-DE01-KA226-005754). Neither the European Commission nor the project's national funding agency DAAD (German Academic Exchange Service) are responsible for the content or liable for any losses or damage resulting from the use of these resources. The authors would like to thank all the participants who took part in the workshops and student focus groups. In addition, the authors would like to thank Harriet Bergman for the language and grammar review.

Authors' Contributions

A Mayer, AAK, MS, and IH conceptualized the study. A Mayer and MS facilitated the workshops, which AAK and IH attended as observers, taking field notes. A Mayer and A Müller conducted the qualitative content analysis of the workshop transcripts, developing the coding framework and deriving themes from the data. AAK, MS, and IH reviewed the coding framework and the themes. MS and A Mayer facilitated the student focus groups. A Mayer and A Müller extracted themes from the focus group transcripts. A Mayer and AAK conducted the descriptive analysis of the questionnaire and the concept maps. A Mayer prepared all figures and the first draft of the paper. All authors reviewed and edited the draft and agreed on the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Original images of individual concept maps and consensus versions.
[PNG File, 450 KB - mededu_v11i1e57331_app1.png]

References

1. Huesmann L, Sudacka M, Durning SJ, et al. Clinical reasoning: what do nurses, physicians, and students reason about. J Interprof Care 2023 Nov 2;37(6):990-998. [doi: 10.1080/13561820.2023.2208605] [Medline: 37190790]
2. Vreugdenhil J, Somra S, Ket H, et al. Reasoning like a doctor or like a nurse? A systematic integrative review. Front Med (Lausanne) 2023;10:1017783. [doi: 10.3389/fmed.2023.1017783] [Medline: 36936242]
3. Cuthbert L, duBoulay B, Teather D, Teather B, Sharples M, duBoulay G. Expert/novice differences in diagnostic medical cognition - a review of the literature. : University of Sussex; 1999 URL: https://users.sussex.ac.uk/~bend/papers/csrp508.pdf [accessed 2025-01-09]
4. Durning SJ, Lubarsky S, Torre D, Dory V, Holmboe E. Considering "nonlinearity" across the continuum in medical education assessment: supporting theory, practice, and future research directions. J Contin Educ Health Prof 2015;35(3):232-243. [doi: 10.1002/chp.21298] [Medline: 26378429]
5. Grant J, Marsden P. Primary knowledge, medical education and consultant expertise. Med Educ 1988 May;22(3):173-179. [doi: 10.1111/j.1365-2923.1988.tb00002.x] [Medline: 3405111]
6. Charlin B, Gagnon R, Pelletier J, et al. Assessment of clinical reasoning in the context of uncertainty: the effect of variability within the reference panel. Med Educ 2006 Sep;40(9):848-854. [doi: 10.1111/j.1365-2929.2006.02541.x] [Medline: 16925634]
7. Charlin B, Lubarsky S, Millette B, et al. Clinical reasoning processes: unravelling complexity through graphical representation. Med Educ 2012 May;46(5):454-463. [doi: 10.1111/j.1365-2923.2012.04242.x] [Medline: 22515753]
8. Daley BJ, Torre DM. Concept maps in medical education: an analytical literature review. Med Educ 2010 May;44(5):440-448. [doi: 10.1111/j.1365-2923.2010.03628.x] [Medline: 20374475]
9. Daniel M, Rencic J, Durning SJ, et al. Clinical reasoning assessment methods: a scoping review and practical guidance. Acad Med 2019;94(6):902-912. [doi: 10.1097/ACM.0000000000002618] [Medline: 30720527]
10. Novak J, Cañas A. Theoretical origins of concept maps, how to construct them, and uses in education. Reflect Educ 2007;3 [FREE Full text]
11. Vink S, van Tartwijk J, Verloop N, Gosselink M, Driessen E, Bolk J. The articulation of integration of clinical and basic sciences in concept maps: differences between experienced and resident groups. Adv Health Sci Educ Theory Pract 2016 Aug;21(3):643-657. [doi: 10.1007/s10459-015-9657-2] [Medline: 26692262]
12. Torre DM, Daley B, Stark-Schweitzer T, Siddartha S, Petkova J, Ziebert M. A qualitative evaluation of medical student learning with concept maps. Med Teach 2007 Nov;29(9):949-955. [doi: 10.1080/01421590701689506] [Medline: 18158670]
13. Hay DB. Using concept maps to measure deep, surface and non‐learning outcomes. Stud Higher Educ 2007 Feb;32(1):39-57. [doi: 10.1080/03075070601099432]
14. Torre D, German D, Daley B, Taylor D. Concept mapping: an aid to teaching and learning: AMEE Guide No. 157. Med Teach 2023 May;45(5):455-463. [doi: 10.1080/0142159X.2023.2182176] [Medline: 36862077]
15. Anohina A, Grundspenkis J. Scoring concept maps. Presented at: Proceedings of the International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing; Jun 18-19, 2009; Ruse, Bulgaria.
16. Ruiz-Primo MA, Schultz SE, Li M, Shavelson RJ. Comparison of the reliability and validity of scores from two concept-mapping techniques. J Res Sci Teach 2001 Feb;38(2):260-278. [doi: 10.1002/1098-2736(200102)38:2<260::AID-TEA1005>3.0.CO;2-F]
17. Kinchin IM, Hay DB, Adams A. How a qualitative approach to concept map analysis can be used to aid learning by illustrating patterns of conceptual development. Educ Res 2000 Jan;42(1):43-57. [doi: 10.1080/001318800363908]
18. Morse D, Jutras F. Implementing concept-based learning in a large undergraduate classroom. CBE Life Sci Educ 2008;7(2):243-253. [doi: 10.1187/cbe.07-09-0071] [Medline: 18519616]
19. Torre DM, Durning SJ, Daley BJ. Twelve tips for teaching with concept maps in medical education. Med Teach 2013;35(3):201-208. [doi: 10.3109/0142159X.2013.759644] [Medline: 23464896]
20. Cutrer WB, Castro D, Roy KM, Turner TL. Use of an expert concept map as an advance organizer to improve understanding of respiratory failure. Med Teach 2011;33(12):1018-1026. [doi: 10.3109/0142159X.2010.531159] [Medline: 22225439]
21. McGaghie WC, McCrimmon DR, Mitchell G, Thompson JA, Ravitch MM. Quantitative concept mapping in pulmonary physiology: comparison of student and faculty knowledge structures. Adv Physiol Educ 2000 Jun;23(1):72-81. [doi: 10.1152/advances.2000.23.1.S72] [Medline: 10902530]
22. Plackett R, Kassianos AP, Kambouri M, et al. Online patient simulation training to improve clinical reasoning: a feasibility randomised controlled trial. BMC Med Educ 2020 Jul 31;20(1):245. [doi: 10.1186/s12909-020-02168-4] [Medline: 32736583]
23. Dekhtyar M, Park YS, Kalinyak J, et al. Use of a structured approach and virtual simulation practice to improve diagnostic reasoning. Diagnosis (Berl) 2021 Jul 12;9(1):69-76. [doi: 10.1515/dx-2020-0160] [Medline: 34246202]
24. Watari T, Tokuda Y, Owada M, Onigata K. The utility of virtual patient simulations for clinical reasoning education. Int J Environ Res Public Health 2020 Jul 24;17(15):5325. [doi: 10.3390/ijerph17155325] [Medline: 32722097]
25. Plackett R, Kassianos AP, Timmis J, Sheringham J, Schartau P, Kambouri M. Using virtual patients to explore the clinical reasoning skills of medical students: mixed methods study. J Med Internet Res 2021 Jun 4;23(6):e24723. [doi: 10.2196/24723] [Medline: 34085940]

26. Botezatu M, Hult H, Fors UG. Virtual patient simulation: what do students make of it? A focus group study. BMC Med Educ 2010 Dec 4;10(1):91. [doi: 10.1186/1472-6920-10-91] [Medline: 21129220]
27. Kononowicz AA, Woodham L, Georg C, Edelbring S. Virtual patient simulations for health professional education. Cochrane Database Syst Rev 2018;6:CD012194. [doi: 10.1002/14651858.CD012194.pub2]
28. Hege I, Dietl A, Kiesewetter J, Schelling J, Kiesewetter I. How to tell a patient's story? Influence of the case narrative design on the clinical reasoning process in virtual patients. Med Teach 2018 Jul;40(7):736-742. [doi: 10.1080/0142159X.2018.1441985] [Medline: 29490538]
29. Kononowicz AA, Narracott AJ, Manini S, et al. A framework for different levels of integration of computational models into web-based virtual patients. J Med Internet Res 2014 Jan 23;16(1):e23. [doi: 10.2196/jmir.2593] [Medline: 24463466]
30. Edelbring S, Dastmalchi M, Hult H, Lundberg IE, Dahlgren LO. Experiencing virtual patients in clinical learning: a phenomenological study. Adv Health Sci Educ 2011 Aug;16(3):331-345. [doi: 10.1007/s10459-010-9265-0]
31. Hege I, Kononowicz AA, Adler M. A clinical reasoning tool for virtual patients: design-based research study. JMIR Med Educ 2017 Nov 2;3(2):e21. [doi: 10.2196/mededu.8100] [Medline: 29097355]
32. Lang VJ, Kogan J, Berman N, Torre D. The evolving role of online virtual patients in internal medicine clerkship education nationally. Acad Med 2013 Nov;88(11):1713-1718. [doi: 10.1097/ACM.0b013e3182a7f28f] [Medline: 24072116]
33. Hege I, Sudacka M, Kononowicz AA, et al. Adaptation of an international virtual patient collection to the COVID-19 pandemic. GMS J Med Educ 2020;37(7):Doc92. [doi: 10.3205/zma001385] [Medline: 33364371]
34. Mayer A, Da Silva Domingues V, Hege I, et al. Planning a collection of virtual patients to train clinical reasoning: a blueprint representative of the European population. Int J Environ Res Public Health 2022 May 19;19(10):6175. [doi: 10.3390/ijerph19106175] [Medline: 35627711]
35. International Collection of Virtual Patients. URL: http://icovip.eu [accessed 2025-01-09]
36. Project results. DID-ACT. 2022. URL: https://did-act.eu/results [accessed 2025-01-09]
37. Humphrey-Murto S, Varpio L, Gonsalves C, Wood TJ. Using consensus group methods such as Delphi and Nominal Group in medical education research. Med Teach 2017 Jan;39(1):14-19. [doi: 10.1080/0142159X.2017.1245856] [Medline: 27841062]
38. McMillan SS, King M, Tully MP. How to use the Nominal Group and Delphi techniques. Int J Clin Pharm 2016 Jun;38(3):655-662. [doi: 10.1007/s11096-016-0257-x] [Medline: 26846316]
39. Kuckartz U. Qualitative Text Analysis: A Guide to Methods, Practice & Using Software: SAGE; 2002.
40. Rosas SR. Group concept mapping methodology: toward an epistemology of group conceptualization, complexity, and emergence. Qual Quant 2017 May;51(3):1403-1416. [doi: 10.1007/s11135-016-0340-3]
41. Kinchin I, Hay D. Using concept maps to optimize the composition of collaborative student groups: a pilot study. J Adv Nurs 2005 Jul;51(2):182-187. [doi: 10.1111/j.1365-2648.2005.03478.x] [Medline: 15963190]
42. Stoyanov S, Jablokow K, Rosas SR, Wopereis IGJH, Kirschner PA. Concept mapping—an effective method for identifying diversity and congruity in cognitive style. Eval Program Plann 2017 Feb;60:238-244. [doi: 10.1016/j.evalprogplan.2016.08.015]
43. McGaghie WC, Boerger RL, McCrimmon DR, Ravitch MM. Agreement among medical experts about the structure of concepts in pulmonary physiology. Acad Med 1994 Oct;69(10 Suppl):S78-S80. [doi: 10.1097/00001888-199410000-00049] [Medline: 7916837]
44. McGaghie WC, McCrimmon DR, Mitchell G, Thompson JA. Concept mapping in pulmonary physiology using pathfinder scaling. Adv Health Sci Educ Theory Pract 2004;9(3):225-240. [doi: 10.1023/B:AHSE.0000038299.79574.e8]
45. Kononowicz AA, Torre D, Górski S, Nowakowski M, Hege I. The association between quality of connections and diagnostic accuracy in student-generated concept maps for clinical reasoning education with virtual patients. GMS J Med Educ 2023;40(5):Doc61. [doi: 10.3205/zma001643] [Medline: 37881522]
46. Charlin B, Gagnon R, Sauvé E, Coletti M. Composition of the panel of reference for concordance tests: do teaching functions have an impact on examinees' ranks and absolute scores? Med Teach 2007 Feb;29(1):49-53. [doi: 10.1080/01421590601032427] [Medline: 17538834]
47. Radwan A, Abdelnasser A, Elaraby S, Talaat W. Correlation between concept maps and clinical reasoning for final year medical students at the faculty of medicine—Suez Canal University. QJM 2018 Dec 1;111(Suppl 1). [doi: 10.1093/qjmed/hcy200.108]
48. Yue M, Zhang M, Zhang C, Jin C. The effectiveness of concept mapping on development of critical thinking in nursing education: a systematic review and meta-analysis. Nurse Educ Today 2017 May;52:87-94. [doi: 10.1016/j.nedt.2017.02.018] [Medline: 28273528]

Abbreviations

CR: clinical reasoning
DID-ACT: Developing, Implementing, and Disseminating an Adaptive Clinical Reasoning Curriculum for Healthcare Students and Educators
iCoViP: International Collection of Virtual Patients

VP: virtual patient

Edited by B Lesselroth; submitted 23.02.24; peer-reviewed by N Rohani, S Jung; revised version received 18.06.24; accepted 23.11.24; published 30.01.25.

Please cite as:
Mayer A, Hege I, Kononowicz AA, Müller A, Sudacka M
Collaborative Development of Feedback Concept Maps for Virtual Patient–Based Clinical Reasoning Education: Mixed Methods Study
JMIR Med Educ 2025;11:e57331
URL: https://mededu.jmir.org/2025/1/e57331
doi:10.2196/57331

Applications, Challenges, and Prospects of Generative Artificial Intelligence Empowering Medical Education: Scoping Review

Yuhang Lin[1*]; Zhiheng Luo[2*]; Zicheng Ye[1]; Nuoxi Zhong[2]; Lijian Zhao[1]; Long Zhang[3]; Xiaolan Li[1], PhD; Zetao Chen[1], PhD; Yijia Chen[1], PhD

[1]Guangdong Provincial Key Laboratory of Stomatology, Hospital of Stomatology, Guanghua School of Stomatology, Sun Yat-sen University, No. 56, Lingyuan Road West, Guangzhou, China

[2]Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, China

[3]School of Government, Sun Yat-sen University, Guangzhou, China

[*]these authors contributed equally

Corresponding Author:
Yijia Chen, PhD
Guangdong Provincial Key Laboratory of Stomatology, Hospital of Stomatology, Guanghua School of Stomatology, Sun Yat-sen University, No. 56, Lingyuan Road West, Guangzhou, China

Abstract

Background: Nowadays, generative artificial intelligence (GAI) drives medical education toward enhanced intelligence, personalization, and interactivity. With its vast generative abilities and diverse applications, GAI redefines how educational resources are accessed, teaching methods are implemented, and assessments are conducted.

Objective: This study aimed to review the current applications of GAI in medical education; analyze its opportunities and challenges; identify its strengths and potential issues in educational methods, assessments, and resources; and capture GAI's rapid evolution and multidimensional applications in medical education, thereby providing a theoretical foundation for future practice.

Methods: This scoping review used PubMed, Web of Science, and Scopus to analyze literature from January 2023 to October 2024, focusing on GAI applications in medical education. Following PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines, 5991 articles were retrieved, with 1304 duplicates removed. The 2-stage screening (title or abstract and full-text review) excluded 4564 articles and a supplementary search included 8 articles, yielding 131 studies for final synthesis. We included (1) studies addressing GAI's applications, challenges, or future directions in medical education, (2) empirical research, systematic reviews, and meta-analyses, and (3) English-language articles. We excluded commentaries, editorials, viewpoints, perspectives, short reports, or communications with low levels of evidence, non-GAI technologies, and studies centered on other fields of medical education (eg, nursing). We integrated quantitative analysis of publication trends and Human Development Index (HDI) with thematic analysis of applications, technical limitations, and ethical implications.

Results: Analysis of 131 articles revealed that 74.0% (n=97) originated from countries or regions with very high HDI, with the United States contributing the most (n=33); 14.5% (n=19) were from high HDI countries, 5.3% (n=7) from medium HDI countries, and 2.2% (n=3) from low HDI countries, with 3.8% (n=5) involving cross-HDI collaborations. ChatGPT was the most studied GAI model (n=119), followed by Gemini (n=22), Copilot (n=11), Claude (n=6), and LLaMA (n=4). Thematic analysis indicated that GAI applications in medical education mainly embody the diversification of educational methods, scientific evaluation of educational assessments, and dynamic optimization of educational resources. However, it also highlighted current limitations and potential future challenges, including insufficient scene adaptability, data quality and information bias, overreliance, and ethical controversies.

Conclusion: GAI application in medical education exhibits significant regional disparities in development, and model research statistics reflect researchers' certain usage preferences. GAI holds potential for empowering medical education, but widespread adoption requires overcoming complex technical and ethical challenges. Grounded in symbiotic agency theory, we advocate establishing the resource-method-assessment tripartite model, developing specialized models and constructing an integrated system of general large language models incorporating specialized ones, promoting resource sharing, refining ethical governance, and building an educational ecosystem fostering human-machine symbiosis, enabling deep tech-humanism integration and advancing medical education toward greater efficiency and human-centeredness.

(JMIR Med Educ 2025;11:e71125) doi:10.2196/71125

KEYWORDS

generative artificial intelligence; GAI; large language model; ChatGPT; medical education; human-machine symbiosis

Introduction

Background

The 21st century has seen accelerated advancement in information technology and artificial intelligence (AI), significantly altering lifestyles and work paradigms. With progress in deep learning and large-scale data processing, generative artificial intelligence (GAI) has emerged as an influential innovation. GAI rapidly expands into diverse applications, enabling content generation across text, images, and audio through the analysis of extensive datasets [1]. Its market demonstrates notable growth, with a 2024 global valuation of ~US $16.8 billion and a projected 37.6% compound annual growth rate (CAGR) from 2025 to 2030 [2], reflecting its significance in commercial and academic domains.

GAI's development is driven by advances in natural language processing (NLP), particularly the Transformer architecture, which enables the generation of complex content. Large language models (LLMs) serve as core technical implementations of GAI. Models like GPT-3, GPT-4, Copilot, and LLaMA 3 have expanded GAI applications from basic automation to sophisticated tasks including content creation, data analysis, and intelligent question-answering systems [3]. These transformer-based LLMs exemplify how conceptual GAI frameworks are operationalized via model architectures and engineering practices.

With technological advancements, GAI has gradually infiltrated more specialized fields, with medical education a prime example. This domain faces challenges due to its knowledge-intensive and highly practical characteristics: traditional teaching methods struggle to replicate clinical scenarios efficiently, and increasingly scarce clinical teaching specimens and patient resources limit the clinical practice training of medical students, all of which are not conducive to the cultivation of medical talents with both clinical thinking and practical ability [4]. In this content, GAI may empower medical education through its enhancement effects on 3 core educational elements: improving resource generation efficiency, optimizing the interactivity of pedagogical approaches, and enhancing the automation level of assessment processes [5,6]. Nevertheless, the accompanying integration risks include potential biases and inaccuracies in generated content [7] and possible inhibition of critical thinking through over-reliance [8]. Thus, optimal implementation strategies warrant further investigation.

Current GAI integration in medical education involves rapid technological iteration and shifting research paradigms [1,9,10]. Prior reviews exhibit three limitations: (L1) Overreliance on single-model analyses (predominantly ChatGPT) [9,10], (L2) insufficient examination of geographical disparities in adoption patterns, and (L3) fragmented assessment of GAI's impact across 3 core dimensions of medical education. These dimensions include resources (teaching support materials like GAI-generated clinical cases and pathological images), methods (instructional strategies like adaptive learning pathways and simulated decision-making), and assessment (automated evaluation of learner performance, such as automated short-answer scoring). Crucially, studies before 2023 were constrained by the technology's maturity, missing the recent shift from theoretical exploration to operational implementation [1]. Therefore, a new round of scoping review is urgently needed to focus on the critical evolution period between January 2023 and October, 2024 (before the completion of this scoping review), construct a multidimensional analytical framework (encompassing resources, methods, and assessment), and clarify the complex picture of the deep interaction between GAI and medical education. To guide this investigation, this study discusses the multifaceted landscape of GAI adoption in medical education through 3 interconnected lines of inquiry. First, it aims to examine whether regional disparities exist in GAI implementation and how researchers exhibit preferences for specific LLMs (eg, ChatGPT). We posit that adoption patterns will demonstrate significant stratification aligned with national development levels and reflect preferential usage of widely accessible general-purpose models. Second, it seeks to map the current state of GAI applications across educational resources, methods, and assessment dimensions. We hypothesize that effectiveness will vary substantially across these domains due to differences in technical implementation requirements and inherent task complexities. Third, it intends to identify current limitations and future challenges, positing that technical deficiencies, including ethical risks such as compromised academic integrity and data hallucinations, will constitute the most significant barriers to sustainable integration.

Theoretical Framework

Theoretical Model: The Theory of Symbiotic Agency

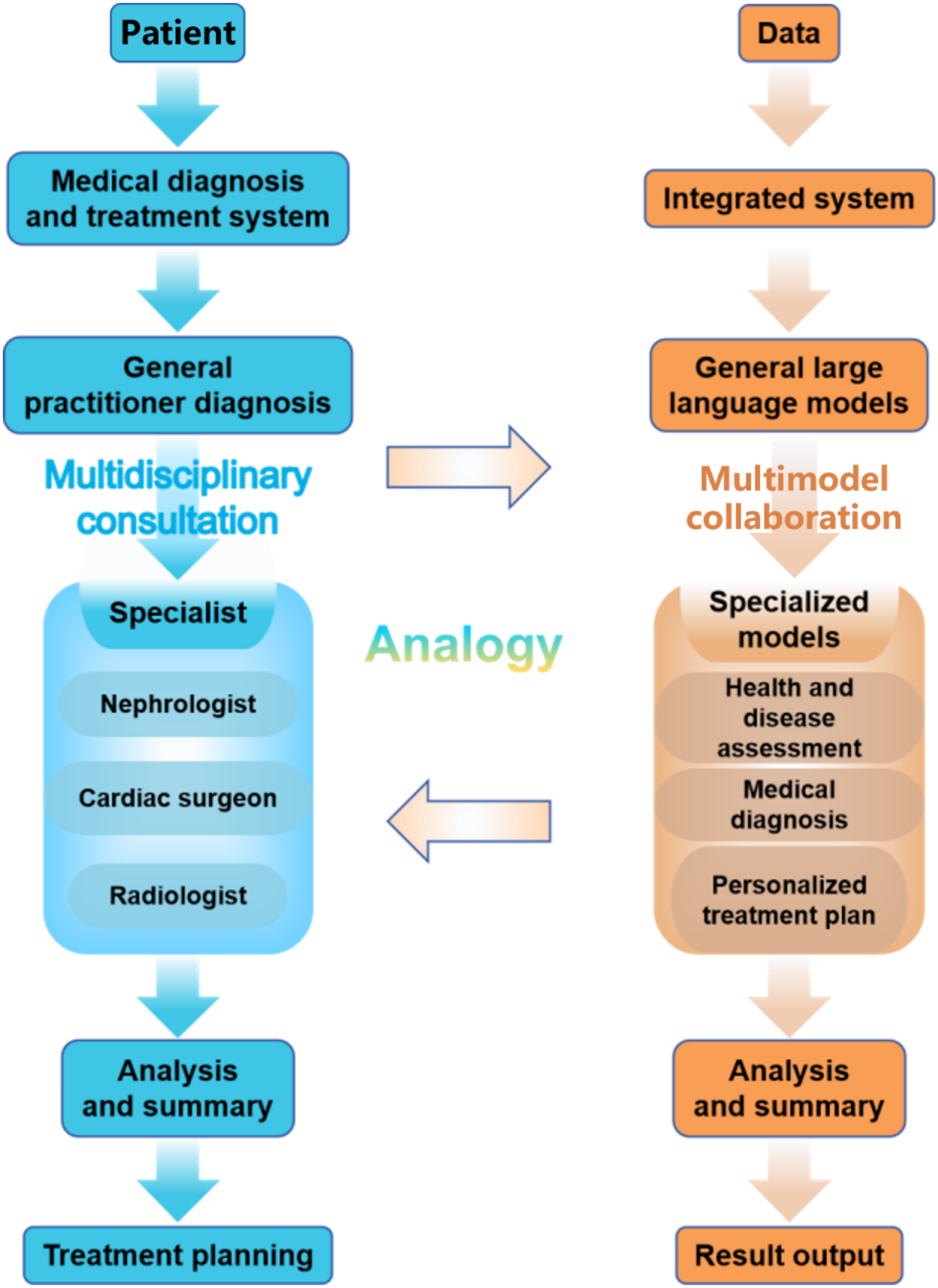
Based on the theoretical framework of symbiotic agency [11], a theory emphasizing interdependent and collaborative relationships between humans and technology, this study conceptualizes human-technology relations as a process of mutual constitution. Technology functions neither as a passive instrument dominated by humans nor as an autonomous replacement for human agency. Instead, it develops in tandem with humans through interdependent interactions: technology enhances human efficacy by expanding cognitive boundaries and enabling novel multimodal interactions, while humans legitimize technological practice by embedding ethical norms and conducting context-specific interpretations such as weighting clinical decisions. This symbiosis transcends traditional master-servant dichotomies by establishing a responsibility-sharing network. Within this network, technology acts as a co-agent in human activity systems, collectively enhancing capabilities rather than substituting human roles. This perspective provides the foundational understanding needed to maintain a dynamic balance in human-technology interdependence within medical education, forming the basis of our conceptual model.

Conceptual Model 1: Specialized Models Integrated System Based on General Large Language Models

Building upon the analytical framework established in Table S1 in Multimedia Appendix 1, which systematically compares general-purpose and domain-specialized GAI models across 3 critical dimensions (knowledge representation fidelity, task compatibility, and ethical constraint mechanisms), this study deconstructs technological heterogeneity to avoid conflating "GAI" as a homogeneous entity. The models in Multimedia Appendix 1 (see Table S1) were selected via multisource evidence synthesis, including peer-reviewed studies, industry reports (eg, Global Large Language Model (LLM) Market Research Report 2024) [12-20], and empirical validation in educational contexts, based on four criteria: (1) technological representativeness of core advancements (multimodality, reasoning, and domain adaptation); (2) broad academic and practical relevance in medical education; (3) functional diversity covering text, image, video, and domain-specific tasks; and (4) market prevalence, wide recognition, technical maturity, and development by prominent AI companies. Notably, models like Perplexity, DeepSeek, Notebook LM, and Midjourney, though used by clinicians and students in specific scenarios, were not included due to limited evaluative data and insufficient supporting information in the referenced reports.

Within this ecosystem, general LLMs serve as multitasking hubs, leveraging cross-domain adaptability and natural language interaction, while specialized models achieve context-specific efficacy through the embedding of deep medical knowledge. To resolve their complementary yet fragmented coexistence, we propose a specialized model integration system anchored to general LLMs, inspired by symbiotic agency theory and hospital diagnostic workflows [21] (see Figure 1). This architecture establishes a 3-tiered clinical analog: general LLMs serve as primary coordinators, managing task orchestration; specialized models act as domain experts, executing depth-specific processing; and protocol-based collaboration enables online consultation through knowledge distillation and output validation. This hierarchical integration embodies symbiotic agency principles: general models extend the applicability of specialized techniques by transcending domain boundaries, while specialized models enhance system depth by reinforcing medical logical rigor. Through functional complementarity and role differentiation, they form a synergistic symbiont exceeding individual capability limits, establishing an intelligent foundation for medical education characterized by adaptability, expertise, and reliability.

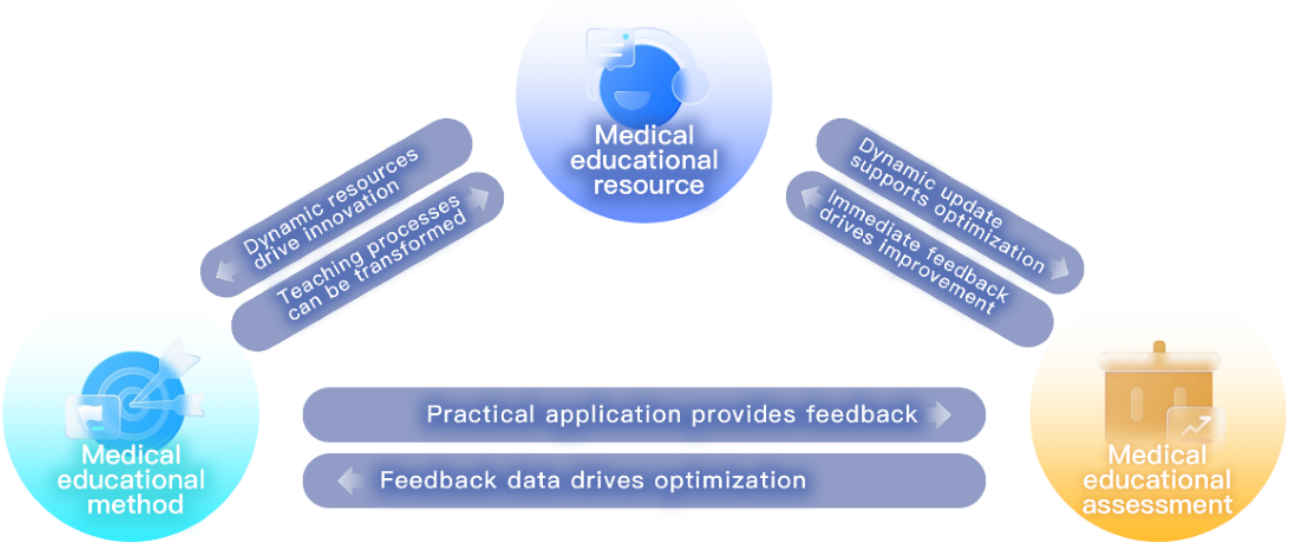
Figure 1. Specialized model integration system based on general large language models.



Conceptual Model 2: Tripartite Synergistic Integration Model for Medical Education Resources, Methods, and Assessment

The tripartite synergy paradigm, rooted in complex systems management theory and evidenced across domains from political governance to integrated health care systems (eg, the mission alignment model by Peek et al [22]), establishes our resource-method-assessment (RMA) framework (see Figure 2) as the core analytical structure [22,23]. This framework defines three interdependent dimensions: (1) resources encompassing dynamic content provisioning mechanisms, (2) methods designing knowledge-to-practice training pathways, and (3) assessment managing outcome monitoring and feedback generation. Their cyclical optimization forms an integrated whole, as resource renewal enables pedagogical innovation, method implementation yields evaluative data, and assessment outputs drive resource refinement and method calibration. Within this architecture, GAI operates as a collaborative instrument executing content generation, interaction support, and data analysis under educator-directed goal design, ethical governance, and critical intervention. The established framework provides essential categorization criteria for subsequent empirical analysis: it defines 3 dimensions—resources, methods, and assessment—directly corresponding to 3 primary research domains in GAI applications for medical education. By consolidating fragmented literature within a unified analytical structure, this framework systematically addresses cognitive limitations arising from isolated examinations of technological functions, thereby elucidating the intrinsic operational logic of technology-enabled educational transformation.

Figure 2. The model of integrated and collaborative development of medical education methods, resources, and assessment.



Methods

Review

With the rapid development of GAI, its applications in medical education have garnered considerable attention and have become a significant research focus. We conducted a preliminary search using the keyword combination of “generative artificial intelligence” and “medical education” across PubMed, Web of Science, and Scopus. Our goal was to analyze the publication trend regarding the applications and challenges of GAI in medical education over the past 5 years (from January 2020 to October 2024). The literature search was limited to sources published between January 2023 and October 2024 for the following reasons: (1) Technological progression: The 2023‐2024 period coincides with a shift from theoretical proposals (pre-2023) to empirical studies on GAI implementation in medical education. (2) Scope alignment: The review prioritizes analysis of current applications, identified limitations (eg, output inaccuracies and integrity concerns), and near-future developments rather than historical trends. (3) Avoiding redundancy: Pre-2023 literature is excluded to prevent overlap with existing syntheses and focus on emergent applications (eg, automated assessment and adaptive resources) evidenced in the sampled literature (n=131). (4) Practical relevance: This timeframe reflects consolidated evidence on operational challenges and benefits relevant to contemporary pedagogical decision-making.

Search Strategy

We used Boolean operators to combine GAI and medical education keywords, creating the final search strategy (see Multimedia Appendix 2). A thorough search was conducted across 3 major databases: PubMed, Web of Science, and Scopus, focusing only on English-language articles published from January 2023 to October 2024.

Inclusion and Exclusion Criteria

This study included research articles focusing on the applications, challenges, and future development of GAI in medical education applications. Articles were excluded if they were commentaries, editorials, viewpoint, perspective, and short reports or communications with low level of evidence or did not discuss GAI within medical education. Studies focusing on non-GAI forms such as predictive analytics and natural language processing or those centered on other fields of medical education (eg, nursing) were also excluded. We excluded nursing based on fundamental educational differences. Clinical and dental

education follow structured undergraduate curricula focused on acute care, diagnostics, and technical skills within hospital settings. Nursing emphasizes community practice, longitudinal relationships, and chronic disease management [24]. Including nursing would introduce significant heterogeneity in learning outcomes, GAI applications, and educational contexts. This methodological exclusion preserves thematic coherence and internal validity for analyzing GAI's role in comparable, technology-driven medical education environments.

Initially, YL and ZL conducted a preliminary screening of titles and abstracts from 3 databases. With the help of Zotero 7.0.13 (64-bit), a document management software (it is a project of Digital Scholar and developed by a global community), ZL detected duplicates of the initially screened articles according to title, author, abstract, and other information and removed duplicates. Following this initial phase, YL and ZL independently reviewed the full texts for a second round of evaluation. In cases of disagreement, ZY and NZ were consulted to mediate and make the final determination regarding inclusion.

Data Extraction Protocol

To ensure the systematicity, transparency, and reproducibility of this scoping review, a detailed data extraction protocol was developed and rigorously followed.

Data Point Definition and Protocol Development

Before comprehensive data extraction, a structured data extraction form was collaboratively developed by all authors. This iterative process was guided by our research questions and the predefined thematic framework outlined in Table 1, which focused on the applications, challenges, and prospects of GAI in medical education applications. The form was designed to systematically capture key information from each included article, encompassing: bibliographic details (eg, authors, publication year, journal, and country or region), study characteristics (eg, research design, objectives, and population), specific GAI models used (eg, ChatGPT [OpenAI] and Gemini [Google]), application scope (single-model vs multimodel), analysis type (performance comparison across models or examination of synergistic enhancement through model integration), detailed descriptions of identified applications, challenges, and future directions of GAI application in medical education categorized exclusively through our tripartite Trinity Framework and quantitative performance metrics (reported accuracy rates, percentages, mean scores, standard deviations, and *P* values related to GAI model performance in various tasks). This granular definition of data points ensured that all relevant information pertinent to our broad research inquiry was systematically collected.

Table . A systematic thematic analysis of applications and challenges of generative artificial intelligence (GAI) in medical education.

Category and theme		Subtheme
Medical educational assessment		
	Scoring short answers automatically.	—[a]
	Evaluating articles.	—
Medical educational resources		
	Providing standard answers.	• The performance of different question types. • The performance of different difficulty questions. • The performance of questions at different cognitive levels.
	Generating diverse clinical cases.	—
	Digital interaction and communication training.	—
	Sharing educational resources.	—
	Generating clinical images.	—
Medical educational methods		
	Curriculum design.	—
	Generating customized teaching aids.	—
	Generating explanations for MCQ[b].	—
	Personalized learning support.	—
	Medical decision aid.	—
	Multidisciplinary knowledge acquisition.	—
	Academic writing optimization.	—
Existing defects at this stage		
	Insufficient scene adaptability.	• Poor ability to handle complex clinical scenarios. • Lack of local background in specific regions. • Language adaptability issues. • Lack of nontextual information analysis skills.
	Data quality and information bias	• Hallucination phenomena. • Lack of details on output content. • Lack of personalization. • Dataset dependency.
Potential issues in the future		
	Overreliance	• Impaired critical thinking. • Decreased creativity. • Decreased teamwork ability. • Decreased practical problem-solving ability.
	Ethical controversy	• Authenticity of the test results. • Academic misconduct. • Lack of clinical interaction and emotional resonance. • Resource inequality. • Ownership of intellectual property rights. • “Black box” problem and attribution of responsibility.

[a]Not available.

[b]MCQ: multiple choice question.

To better understand the global research landscape in this field, we analyzed the countries or regions of origin for the 131 selected articles. For those without a precise location, we assigned them according to the country or region of the corresponding author's institution. To analyze the distribution of research based on the countries or region's development level, we used the Human Development Index (HDI) classification. The latest HDI data categorizes countries or regions into 4 tiers: very high, high, medium, and low human development with higher HDI scores correlating with greater national development. We also investigated cross-level HDI collaborations, which refer to partnerships between countries from different HDI categories [25].

Protocol Testing and Quality Control

To validate the comprehensiveness and clarity of the data extraction form, a pilot test was independently conducted by 2 reviewers, YL and ZL, on a randomly selected subset of 10 included articles. During this pilot phase, any discrepancies in data extraction or ambiguities within the form were identified and discussed. Based on these discussions, the data extraction form underwent iterative revisions to refine categories, clarify definitions, and ensure consistent interpretation of data points among reviewers. Following this refinement, YL and ZL independently extracted data from all 131 included articles. In cases of disagreement between the 2 independent extractors, consensus was initially sought through discussion. If a consensus could not be reached, a third and fourth reviewer, ZY and NZ, were consulted to mediate and make final determinations regarding the applicability and extraction of the data.

Synthesis of Results

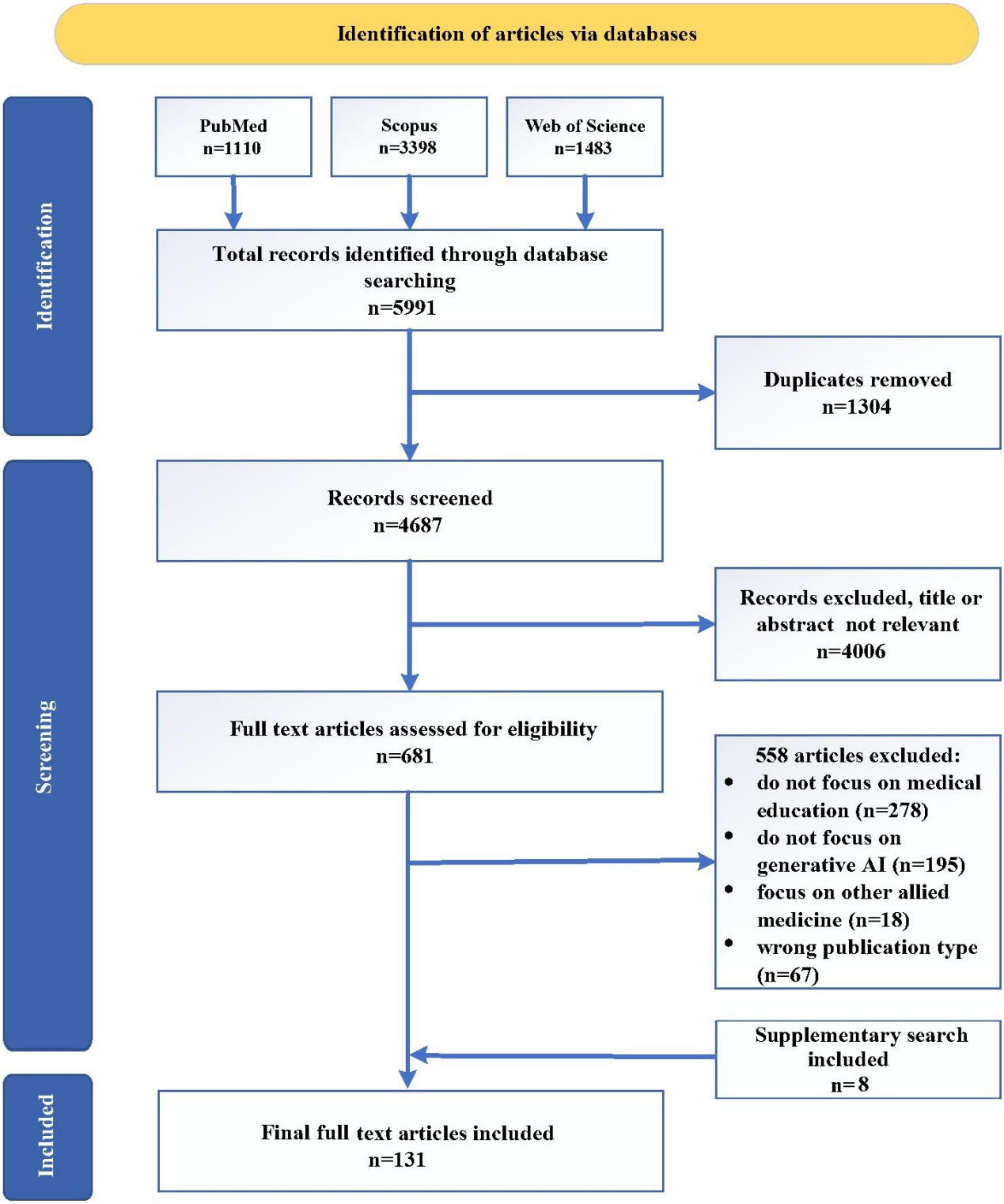
ZY subsequently compiled and reorganized the extracted data, assigning new identifiers for easier reference. This organized dataset was then categorized according to the predefined themes and subthemes (see Table 1), forming the basis for the subsequent descriptive summary and analysis. Our analysis employed a theory-driven, top-down approach anchored in a tripartite conceptual model of medical education: resource generation, method innovation, and assessment upgrade. The following sections will present a descriptive summary of the extracted data.

Results

Overview

Following our search strategy, we retrieved 5991 articles, of which 1304 were duplicates, leaving 4687 articles. In the first round of screening, 4006 irrelevant articles were excluded based on titles and abstracts, leaving 681 articles. In the second round, we excluded 558 articles after full-text review, including 278 nonmedical education articles, 195 non-GAI articles, 18 focused on other medical fields (eg, nursing), and 67 of different types (eg, commentaries). During the paper preparation, we conducted a supplementary search for 8 systematic reviews and meta-analyses. Ultimately, 131 articles were included in the final review (see Figure 3). Among the 131 included studies, the distribution of research designs was as follows: 83 cross-sectional studies, 5 randomized controlled trials (RCTs), 2 quasi-experimental studies, 1 cohort study, 1 quasi-randomized controlled trial, 8 systematic reviews and meta-analyses, 5 mixed-methods studies, and 1 case study. The remaining 25 studies were categorized as "other" with nonstandardized research designs, which were not fitting typical epidemiological or evidence-based medicine classifications. Collectively, cross-sectional studies (descriptive research designs) constituted the majority (n=83), reflecting the emerging state of GAI in medical education, where most research focuses on initial application explorations, feasibility assessments, and user experience descriptions rather than hypothesis-driven experimental designs. Other study types, including RCTs, cohort studies, and systematic reviews, provided supplementary evidence on intervention effects, longitudinal trends, and synthesized findings, respectively.

Figure 3. Article screening flow chart. AI: artificial intelligence.



Analysis of Literature Source and Human Development Levels

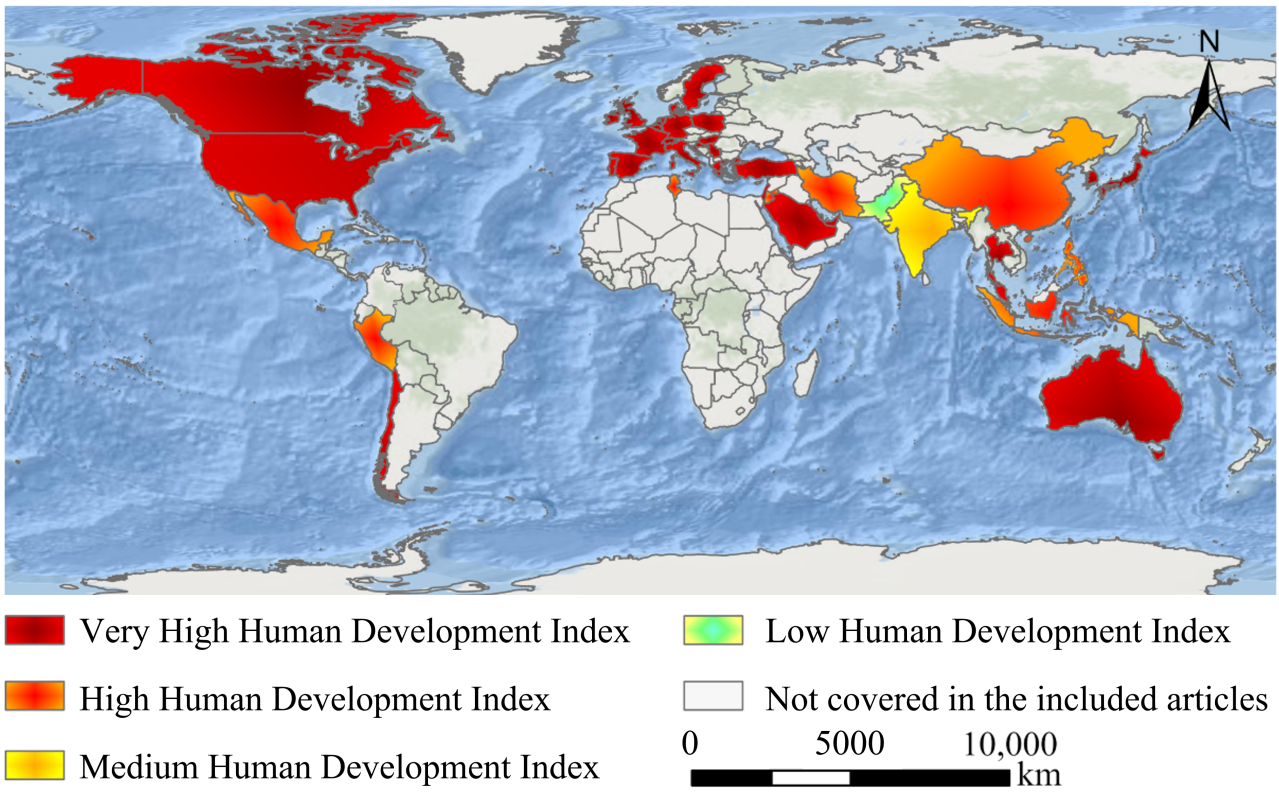
Based on the countries or regions of origin for the included articles and HDI classification, we analyzed the distribution of related studies. The results are illustrated in Table 2 and Figure 4. A significant portion (74%, n=97 articles) of the research came from countries or regions with very high human development, with the United States contributing 33 studies. High human development countries or regions produced 15% (n=19 articles), with China contributing 13 studies. Medium human development countries or regions contributed 5% (n=7 articles), mainly from India, while low human development countries or regions accounted for only 2% (n=3 articles). Furthermore, 4% of the studies (n=5 articles) involved cross-level collaborations, primarily between very high and medium or low HDI countries or regions.

Table . Distribution of countries or regions of origin for generative artificial intelligence (GAI) research in medical education (categorized by the HDI[a]).

HDI classification	Portion, n (%)
Very high human development	97 (74.0)
High human development	19 (14.5)
Medium human development	7 (5.3)
Low human development	3 (2.2)
Cross-level HDI collaboration	5 (3.8)

[a]HDI: Human Development Index.

Figure 4. Geographical distribution of countries or regions of origin for generative artificial intelligence (GAI) research in medical education.



Applications of GAI in Medical Education

Medical Educational Assessment

Scoring Short Answers Automatically

A recent study examined GPT-4 (OpenAI) and Gemini 1.0 Pro in automated short answer grading using 2288 student responses from 12 undergraduate medical courses across 3 languages, with instructor-provided rubrics or sample solutions as reference standards. GPT-4 showed high precision (0.91) in identifying fully correct answers, though its scores were significantly lower than human graders, while Gemini 1.0 Pro had no significant difference from human evaluations, with a mean normalized score of 0.68 (SD 0.32) and median of 0.75, similar to humans. Both models demonstrated high consistency across repeated evaluations, especially with high-quality standard responses, and these findings are specific to undergraduate medical education contexts [26].

Evaluating Articles

Liu et al [27] reported that in their study of 50 rehabilitation-related original articles, 50 sections (introductions, discussions, and conclusions) were generated by ChatGPT-3.5 and 50 were corresponding AI-rephrased versions using Wordtune Originality.ai, achieved 100% accuracy in detecting both AI-generated and AI-rephrased content. ZeroGPT correctly identified 96% of AI-generated texts and 88% of rephrased ones. The study focused specifically on rehabilitation medicine with analyzed content limited to partial article sections rather than full texts. It is notable that such high detection rates have not been widely observed across other disciplines or with newer large language model versions. The specialized nature of medical writing, including technical terminology use, may also influence these outcomes in ways not seen in broader academic contexts, which should be considered when evaluating the generalizability of these findings. Another study comparing automatic scoring systems (ChatGPT-3.5 and ChatGPT-4) with manual scoring for article quality assessment found no

significant difference between GPT-4-based scoring and human grading. This demonstrates the considerable potential of GAI to enhance the quality evaluation of articles [28].

Medical Educational Resources

Providing Standard Answers

The performance of different question types: The studies encompassed a range of question types, including multiple-choice questions (MCQs), single-choice questions, short-answer questions (SAQs), true or false questions, open-ended short-answer questions (SOAQs), long-answer questions, clinical case analysis questions (CAQs), and image-text integrated questions [29-37]. An exploratory study conducted by a research team from Qatar University evaluated ChatGPT's performance across various assessment formats relevant to undergraduate dental education. The study included 50 assessment items covering 50 different learning outcomes, with 10 items for each of the 5 formats: MCQs, SAQs, short essay questions (SEQs), single true or false questions, and fill-in-the-blank items. These items were based on core clinical topics in dental education, such as restorative dentistry, periodontics, endodontics, and oral surgery, aligned with the learning outcomes expected of undergraduate dental students. In this study, ChatGPT demonstrated 90% accuracy for SAQs, SEQs, and fill-in-the-blank items and notably achieved 100% accuracy in single true or false questions [31]. However, other studies have revealed a significant decline in accuracy for CAQs, as low as 17%, which require strong logical reasoning and lack predefined options [34].

Regarding MCQs, a study reported that GPT-4 and Microsoft Bing achieved top scores (76%) on the University of Antwerp medical licensing MCQ exam, outperforming medical students. However, ChatGPT's accuracy fell considerably when tackling Chinese-language medical MCQs with an accuracy of 37%. In addition, another study reported that in the Chinese Master's Degree Entrance Examination, ChatGPT's accuracy for single-choice questions (A1 type) was 56%, whereas for MCQs, it dropped to 33% [15]. These findings suggest that GAI's performance is not uniformly robust across all MCQ types and is influenced by factors such as question structure, subject domain, difficulty, language, and the presence of clinical vignettes or images.

The performance of different difficulty questions: In terms of difficulty, questions were generally classified as "easy," "medium," or "difficult." For example, the difficulty levels are defined based on the performance indicators of the historical question bank: "Difficult" (P<.30; less than 30% of the students answered correctly), "Medium" (P=.30 to .80), and "Easy" (P>.80) [35]. ChatGPT-4 demonstrated strong performance on easy questions, with accuracy rates reaching 97.4%. Yet, even in this category, ChatGPT-4's performance lagged behind that of residents [35,38-42]. In contrast, ChatGPT-4 excelled on medium and difficult questions, outperforming residents by 25.4 and 24.4 percentage points, respectively [38]. Across all models, performance tended to decline with increased difficulty, especially for higher-level questions that required multistep reasoning, where accuracy dropped markedly [39-45].

The performance of questions at different cognitive levels: 2 studies investigated the performance of ChatGPT-4 on questions categorized by Bloom's taxonomy, which includes 6 cognitive levels: remembering, understanding, applying, analyzing, evaluating, and creating [46]. These studies found that ChatGPT-4 consistently performed well across all cognitive levels, with an average correct answer rate of 71.96% for each cognitive level [47,48].

Generating Diverse Clinical Cases

By collaborating with instructors, GAI can quickly generate comprehensive clinical cases, including patient history, physical examination results, lab data, and differential diagnoses tailored to predefined learning objectives (eg, chest pain and joint pain). This reduces the time instructors spend developing such cases [49,50]. Furthermore, GAI-generated cases can integrate various contextual factors such as race, occupation, and lifestyle, significantly enriching the diversity of teaching materials [51]. For example, when creating a case based on a disease profile specific to a region, the ethnicity of the generated patient can be adjusted accordingly. In the context of type 2 diabetes, modifications can be made to the age range and weight distribution. In addition, randomized prompts for urine analysis may be included in urinary tract infection cases. Both patient presentations and examination findings can be randomized, and symptom expression can be customized to meet specific learning needs [51].

In Smith and colleagues' [52] study, GAI was assigned the task of creating a case of an immigrant with mental health concerns, as this group may require specialized social psychiatry interventions. The results indicated that GAI was able to produce a case that met fundamental educational objectives. However, it included several signs of emotional disorders, highlighting a need for further refinement.

Digital Interaction and Communication Training

Studies have shown that GAI is effective in promoting interactive learning and providing practice in communication skills. GAI-powered simulation tools simulate changes in clinical conditions in scenarios such as advanced cardiac life support (ACLS) and intensive care unit (ICU) sepsis, prompting students to critically analyze whether their decisions are correct [53]. In addition, conversational GAI-created digital patients provide anesthetists with valuable training for patient interactions, reducing reliance on human actors while enhancing the flexibility and consistency of the training process [54]. These digital interactions create a safe space for repeated practice, providing dynamic learning experiences that traditional textbooks cannot match [52,55]. Furthermore, conversational GAI models, such as chatbots, can simulate the role of a professor, offering critical evaluations of literature and distilling complex research into easily understandable key findings, thus fostering simulated discussions between students and experts in the field [56]. However, besides experiences and qualitative observations, formal evaluations of the reliability and validity of such GAI-generated information in a professor-like capacity are still needed.

Sharing Educational Resources

By generating accessible public health information, GAI enhances the public's understanding of essential health issues, such as infectious disease prevention and vaccination, ultimately leading to improved health literacy [30]. Furthermore, GAI-generated clinical cases can be disseminated as Open Educational Resources (OERs), providing medical educators with globally adaptable teaching materials that are customized to local contexts [51].

Generating Clinical Images

GAI tools such as Adobe Firefly, DALL·E 2 (OpenAI), Bing Image Creator, and generative adversarial networks (GANs) can create clinical images displaying various pathological features based on textual descriptions, potentially addressing the shortage of authentic pathological images in traditional medical education due to medical confidentiality and patient privacy restrictions [51,57-59]. For instance, images of retinal disease generated by the stable diffusion model enhance students' learning opportunities in ophthalmic pathology, greatly enhancing the availability of visual teaching resources [57]. However, their reliability and accuracy vary significantly across models and tasks. For example, DALL·E 2 demonstrated an overall clinical accuracy rate of 22.2% in aligning generated images with textual prompts across 15 semantic relations (eg, spatial and action-based relationships), with only 3 relationships (touching, helping, and kicking) achieving moderate consistency above 25%. In a medical education context, DALL·E 2 achieved 78% accuracy for soft-tissue tumor images but produced inconsistent results for wound images, with 65% of generated wound images containing anatomical inaccuracies or irrelevant elements [58]. A comparative study of DALL·E 2, Midjourney, and Blue Willow for generating skin ulcer images showed DALL·E 2 performed best with an average score of 3.2/5 (scale 1 - 5) but still produced irrelevant content (eg, X-rays instead of pressure ulcers) in 20% of cases. Midjourney generated stylized, exaggerated features in 40% of images, while Blue Willow produced images with little relevance to prompts in 70% of attempts [59].

Medical Educational Methods

Curriculum Design

GAI shows potential in the early stages of curriculum development, aiding in quickly creating course objectives, learning strategies, and frameworks. For instance, in a study on integrated pharmacotherapy of infectious disease education modules, ChatGPT helped design curriculum goals (eg, "describe mechanisms of antibiotic resistance") with an average expert rating of 92% for appropriateness and accuracy, supporting educators—especially in designing foundational courses [60].

Generating Customized Teaching Aids

Researchers have developed derivative applications based on classical models, such as Glass AI (a powerful AI-driven knowledge management system developed by Glass Health, focusing on organizing and retrieving health-related information efficiently). It integrates GPT-4 with evidence-based, peer-reviewed clinical guidelines to generate differential diagnoses and clinical plans based on textual input of clinical cases, enabling students to interact with it and experience the GAI-driven diagnostic process for cases [61]. Similarly, an MCQ generator based on ChatGPT-generated cases offers a dynamic platform for personalized learning assessments [62].

Generating Explanations for Multiple-Choice Questions

Research shows that when GAI is used to answer MCQs, the explanations generated by GAI can better convey key knowledge points and achieve good accuracy and degree of matching with teachers' explanations. Of the 81 questions explained by the teacher and correctly answered by ChatGPT, 92.6% of the explanations were accurate and included at least part of the teacher's explanation. However, the research also highlights that if an initial response is incorrect, the likelihood of subsequent errors increases significantly (P<.001), indicating that an early mistake may lead to systematic inaccuracies in later explanations [63]. Complementing this, in a systematic review, the broader literature reviewed showed that the majority of studies (5/8, 62.5%) indicate the effectiveness of AI in generating valid MCQs, with a preference for the latest GPT-4 models (6/8, 75%) [64].

Personalized Learning Support

Studies demonstrate that GAI boosts students' learning efficiency across multiple stages by offering personalized feedback and customized content. This includes support for exam preparation [55,65-68], optimizing learning paths and review strategies [52,69-71], clarifying medical concepts [68,72-76], and assisting in the development of tailored career plans [77]. For instance, in physiological case analysis, GAI offers precise responses and contextually relevant feedback. A cross-sectional study tested 77 physiology case vignettes (covering diverse physiological and pathophysiological scenarios, designed for undergraduates) on ChatGPT 3.5, Google Bard, and Microsoft Bing. Rated by two physiologists on a 0 - 4 scale, ChatGPT scored highest at 3.19 (SD 0.3), outperforming Bard (2.91, SD 0.5) and Bing (2.15, SD 0.6) with P<.001. ChatGPT's precision accelerates task completion, helping students grasp medical knowledge in practical scenarios more effectively [78]. Furthermore, a study found that in cases of initial incorrect responses, GPT-4 was able to self-correct and provide accurate answers after simple follow-up questions or hints, mimicking pedagogical interactions observed in residency programs. This dynamic learning approach, coupled with rapid information processing, positions GPT-4 as an important asset for personalized learning [79].

Medical Decision Aid

GAI uses its ability to analyze complex, domain-specific knowledge to support the diagnosis of rare and intricate diseases. In addition to diagnosis, it can generate differential diagnoses tailored to the unique characteristics of each disease, providing health care professionals with precise decision-making support [80-83]. For common pathological issues and basic data analysis, GAI tools are efficient and accurate, helping pathologists organize their thought processes and expedite the initial diagnostic phases [84]. The impact of domain-specific training is profound. For instance, refined datasets in the surgical and

anesthesiology fields enhance GAI's clinical decision-making capabilities. In scenarios such as a "30-year-old pregnant woman requiring an emergency appendectomy," GAI suggests not only tailored surgical strategies but also factors in critical anesthesia protocols [85]. Furthermore, in the field of traditional Chinese medicine, when combined with such tools, GAI can effectively create knowledge maps that organize entities, attributes, and their relationships to traditional Chinese medicine through graphical structures. GAI provides unique support for teaching traditional Chinese medicine and disease diagnosis and treatment decisions [86].

Multidisciplinary Knowledge Acquisition

GAI demonstrates potential in multidisciplinary knowledge acquisition within medical education by providing high-quality knowledge across various medical subfields [87-94]. GAI demonstrates adaptability across disciplines, including shoulder and elbow surgery, sports medicine, and oncology [91]. Research further indicates that GAI models such as ChatGPT-4 excel in internal medicine, pediatrics, obstetrics and gynecology, surgery, emergency care, and public health [88-90,92-94]. Notably, a study assessing ChatGPT-4's performance in the American Board of Family Medicine (ABFM) certification examination demonstrated its significant proficiency, with both the custom robot version (embedded in a specialized subenvironment designed to mimic examination conditions and given extensive preparation resources) and the regular version (standard ChatGPT-4) achieving high correct response rates of 88.67% and 87.33% respectively, well above the passing threshold. This further highlights GAI's value in enhancing medical education within a multidisciplinary framework, making it a powerful learning support tool across a wide range of fields, including family medicine [95]. A meta-analysis of ChatGPT-3.5/4 across medical, pharmacy, dentistry, and nursing licensing exams revealed an overall accuracy of 70.1% (95% CI 65% - 74.8%; $P<.001$). Performance varied significantly by field (Q=15.334; $P=.002$), with pharmacy having the highest rate (71.5%, 95% CI 66.3% - 76.2%) and nursing having the lowest rate (61.8%, 95% CI 58.7% - 64.9%). These results demonstrate GAI's potential to provide multidisciplinary learning support in health professions [96]. It is crucial to note that the evidence presented in this section highlights the individual learner's ability to access and comprehend information across disciplines. This review's existing evidence has not yet extensively covered GAI's direct support for complex interdisciplinary teamwork, closed-loop communication, or the cultivation of specific professional behaviors within collaborative learning environments.

Academic Writing Optimization

A study shows that GAI excels in creating article outlines and editing formatting, which alleviates common writing challenges related to poor organization and grammatical mistakes [28]. In addition, GAI can significantly improve the quality and standardization of academic writing, allowing medical educators and students to express their ideas more accurately and clearly [28,55,97]. Furthermore, GAI assists students in organizing and generating literature content while writing their thesis [98]. The content produced by GAI maintains consistency in language and includes appropriate academic terminology and logical structure, helping students present themselves more professionally in their academic writing [55]. Furthermore, GAI supports many non-native English speakers in overcoming language barriers during the academic writing process, which enables them to engage more confidently in academic communication [71].

Statistical Analysis of the Application of GAI Models

The models discussed in 131 articles include ChatGPT, Gemini (formerly known as Bard), Copilot (formerly known as Bing), Claude, and LLaMA, as well as other types of models such as StyleGAN2-ADA, Stable Diffusion, and customized chatbots.

Among the various models studied, ChatGPT stands out due to its advanced natural language processing capabilities. Of the 131 articles, 119 (89.5%) focused on ChatGPT, which was applied in diverse educational contexts, including simulating doctor-patient conversations, generating exam questions, and providing personalized learning support. These applications highlight their flexibility and adaptability in medical education. Notably, research had shown that as versions have iterated, ChatGPT-4 has significantly improved in both performance and scope compared to ChatGPT-3.5 [26,94,99-101].

Gemini was mentioned in 22 articles, accounting for 16.5% of the total. Copilot was mentioned in 11 articles, primarily due to its integration with the Microsoft ecosystem, making it ideal for educational management and resource development. Claude was cited in 6 articles. LLaMA, referenced in 4 articles, stands out for its ability to run locally, making it suitable for environments with limited resources. In addition, StyleGAN2-ADA, Stable Diffusion, and Convai were discussed in individual studies, mainly for their use in image generation and visualizing doctor-patient interactions.

The performance assessment of two or more models was compared in 26 articles. In comparative studies within the articles, numerous models have undergone head-to-head research, including ChatGPT-4 with Gemini 1.0 Pro [26], ChatGPT-4 with ChatGPT-3.5 [96], ChatGPT 3.5 with Google Bard and Microsoft Bing [78], DALL-E 2 with Midjourney and Blue Willow [59], and Originality.ai with ZeroGPT [27]. Based on these head-to-head investigations, different models demonstrate proficiency in specific tasks: ChatGPT-4 performs better in handling complex tasks, providing accurate medical knowledge, generating exam questions, and offering personalized learning support, especially in English-language medical licensing examinations; Gemini 1.0 Pro is noted for its strong contextual understanding and multimodal capabilities; ChatGPT-3.5 excels in simulating doctor-patient conversations, generating exam questions, and providing personalized learning support; Microsoft Bing achieved top scores alongside GPT-4 in medical licensing MCQ exams; DALL-E 2 shows potential in creating clinical images with specific pathological features from textual descriptions; and Originality.ai achieves high accuracy in detecting both AI-generated and AI-rephrased medical writing.

Challenges of GAI in Medical Education

Existing Defects at This Stage

Insufficient Scene Adaptability

Insufficient scene adaptability is due to the following factors.

First is the poor ability to handle complex clinical scenarios. GAI faces substantial limitations when handling complex clinical scenarios, particularly in cases requiring multistep reasoning, intricate calculations, and recognition of atypical clinical symptoms [45,87,102,103]. For instance, studies have shown that GAI struggles with MCQs, X-type problems, and tasks demanding deep reasoning. This underscores its limited ability to perform the nuanced decision-making required in medical judgments [29,30,39,41,44,47,48,66,67,84,89-92,104-107]. Furthermore, GAI-generated clinical scenarios often lack flexibility and fail to replicate the diversity and complexity of real-life clinical environments, thereby limiting learners' exposure to the spectrum of challenging cases [108,109]. GAI also faces technical limitations in generating simulated images for complex diseases, resulting in images that fail to depict atypical manifestations accurately [57]. Furthermore, GAI models demonstrate uneven knowledge depth, exemplified by an ophthalmology meta-analysis: accuracy was 78% in "Pathology" but significantly lower in foundational or clinical areas, such as "Ophthalmology fundamentals" (52%), "Clinical ophthalmology" (57%), and "Refractive surgery" (59%) [110].

Second is the lack of local background in specific regions. Numerous studies have shown that GAI often struggles to adapt effectively to a specific region's unique background and needs when dealing with medical content related to that region, thereby undermining its universal applicability in multicultural settings [33,38,102,111]. For example, ChatGPT often responds to public health issues in India with a Western-centric perspective, overlooking local situations and cultural differences [33]. Similarly, ChatGPT struggles to accurately comprehend and adapt to the local regulatory environment when addressing medical policies specific to China, largely due to the limited representation of Chinese data in its training set [102,112].

Third is language adaptability issues. Currently, GAI exhibits significant limitations in processing languages, particularly in non-English medical education environments. The accuracy of GAI models like ChatGPT often varies greatly when handling languages such as Chinese, Korean, and Polish, resulting in incorrect outcomes in these contexts [29,30,34,38,106,113-115]. A meta-analysis quantified disparity: GPT-3.5 achieved 57% accuracy (95% CI 52% - 62%; P<.01) in English-speaking countries and 58% (95% CI 52% - 64%; P<.01) in non-English-speaking countries (P=.72). GPT-4 scored 86% (95% CI 82% - 89%; P<0.01) in English-speaking countries versus 80% (95% CI 76% - 83%; P<.01) in non-English-speaking countries (P=.02), demonstrating the adaptability issues of GAI models across different linguistic and regional contexts [116].

Fourth is a lack of nontextual information analysis skills. Current GAI tools like ChatGPT and Bard struggle to handle image-based queries, limiting their application in fields such as dentistry, neurosurgery, and nuclear medicine, where visual analysis of images and tissue samples is crucial for clinical decision-making [31,36,42,67,73-75,117].

Data Quality and Information Bias

Data quality issues and information bias occur due to the following factors.

First is the hallucination phenomenon. In GAI applications, hallucinations occur when the content generated by GAI diverges from factual accuracy or contradicts itself, remaining a prevalent issue. In total, 3 primary types of hallucinations have been identified: input-conflicting hallucination, context-conflicting hallucination, and fact-conflicting hallucination. Input-conflicting hallucination occurs when the GAI-generated content contradicts the initial information provided by the user. This can mislead learners and hinder their understanding of specific concepts [51,65,118]. Context-conflicting hallucination arises when the GAI offers contradictory responses to the same or similar questions. This inconsistency is particularly evident in complex case analyses [71,90,119,120]. Fact-conflicting hallucination occurs when the GAI reports facts that contradict established information, often with a high confidence level, which can easily mislead learners [54,121-137].

Second is the lack of details on output content. Numerous studies have highlighted that GAI often generates overly simplified or vague responses, lacking essential details and knowledge necessary for a comprehensive understanding [31,53,56,60,63,73,114,118,120,135,136,138,139]. For instance, evaluations of GAI in cardiology have revealed that it fails to specify the types of heart murmurs associated with valve diseases. In addition, GAI-generated descriptions of pathophysiology and epidemiology tend to be overly general, often including vague statements such as "certain age groups are at higher risk" without specifying the specific conditions. Furthermore, GAI often produces incomplete or inaccurate information when generating case study materials, which can lead to misleading students. For example, GAI-generated learning materials on melanoma have been known to omit crucial tumor markers like S-100 or the latest treatment for BRAF (B-Raf proto-oncogene, serine, or threonine kinase) mutations [63,138]. The same problems are evident in academic writing assistance, where GAI may create basic article structures but often lacks the depth, detail, and critical citations found in human-generated content [120].

Third is the lack of personalization. The content generated by GAI lacks personalization tailored to individual needs. This limitation mainly manifests in the generated text, which often adopts similar writing patterns and standardized language, struggling to incorporate personalized perspectives or creative expressions [28]. In a medical environment, GAI-generated treatment plans, although generally reasonable, often fail to consider individual patient characteristics, such as the severity of the disease, lifestyle, and personal preferences [105].

Fourth is dataset dependency. The performance of GAI is significantly influenced by the quality and diversity of its training data. If the data is insufficient or skewed, it may lead

to potential biases and limitations in practical applications, causing underperformance in less-represented areas [33,59,73,82,85,86,111,117,122,126,140-143]. In addition, the cutoff date for the training data means that GAI may lack knowledge of the latest research, leading to outdated or inaccurate recommendations [26,32,41,66,67,74,80,89,92,94,109,129,136,141,144]. For example, when advising on treatment for bipolar disorder in pregnant women, ChatGPT-4 failed to incorporate the latest studies and instead suggested outdated methods [89]. Furthermore, the data bias present in GAI during the training process cannot be overlooked [51-53,58,61,71,72,78,107,108,132,139,145]. Such biases often arise from the intrinsic imbalances within the dataset, which subsequently permeate the generated content. These biases manifest as stereotypes, mainly depicting certain professions or physical attributes. For instance, some occupations may be associated with higher BMIs, while the French ethnicity is stereotypically linked to the profession of "wine connoisseur" [51].

Potential Issues in the Future

Overreliance

Overreliance can be caused due to the following factors.

First is impaired critical thinking. The rapid feedback provided by GAI may reduce students' time for deep thinking, weakening their ability to analyze problems and independently engage in critical learning. This phenomenon is particularly evident in medical education, where students often rely on the answers provided by GAI when solving complex problems rather than relying on their logical reasoning and knowledge accumulation for analysis and resolution [35,39,40,50,55,69,70,72,74,75,77,98,99,124,136,146-152].

Second is decreased creativity. When students use GAI tools like ChatGPT, they often receive writing suggestions that lack the creativity and depth of human-generated content. Thus, prolonged reliance on such tools may weaken their independent writing skills and hinder their ability to engage with complex topics that require critical thinking and practical expertise [28]. Similarly, educators who overly depend on GAI for content creation may stifle their curricular innovation, limit diversity and depth in teaching materials, and ultimately diminish the overall quality of education [60].

Third is decreased teamwork ability. Overreliance on GAI tools such as ChatGPT can weaken students' communication skills and ability to engage actively in collaborative teamwork [72,152]. Furthermore, the frequent use of these tools limits opportunities for meaningful interpersonal interaction with peers and mentors, hindering the development of essential teamwork and communication skills [152].

Fourth is decreased practical problem-solving ability. Practical problem-solving is essential for clinical decision-making and patient management. However, the convenience of GAI tools may lead students to rely on preexisting solutions, neglecting the deeper analysis and logical reasoning necessary to develop personalized answers [52,55,74,75,77,87,151-153]. Furthermore, using these tools may reduce interaction with mentors and peers, limiting students' opportunities to gain diverse perspectives through collaborative discussions and approach problems from multiple angles [147].

Ethical Controversy

Ethical controversies can occur due to the following factors.

First is the authenticity of the test results, as the integration of GAI in testing and assessment may compromise the accuracy and effectiveness of traditional methods used to evaluate students' actual capabilities. GAI-generated responses or GAI-assisted evaluations risk reflecting the performance of GAI itself rather than students' authentic abilities. This issue is evident in various exams, such as medical licensing and specialty exams, presenting new ethical challenges in medical education [27,50,78,99,144,146,154].

Second is academic misconduct, since GAI-generated content often evades traditional plagiarism detection tools, making it easier for students to exploit GAI tools to complete assignments or write papers without being detected, thus jeopardizing academic integrity [31,154]. In addition, the ease of using GAI to generate answers cultivates students' mindset of overreliance on such tools for academic tasks, which may increase their future likelihood of academic misconduct [70-72,124,155]. This issue extends beyond individual students and poses a broader threat to academic ethics, as GAI-generated content can be misinterpreted as original work, distorting academic evaluations [125,150].

Third is a lack of clinical interaction and emotional resonance. When addressing complex ethical or emotional medical issues, GAI lacks the empathy and emotional responsiveness inherent in human physicians, potentially undermining trust in the doctor-patient relationship [98,131]. This limitation is supported by a General Medicine In-Training Examination (GM-ITE) study comparing GPT-4 and Japanese residents. In the GM-ITE, "medical interview and professionalism" category assesses patient communication, ethics, and professionalism. It uses scenario-based questions (eg, addressing a terminally ill patient's anxiety or resolving treatment ethics). Responses are scored 0‐10 based on communication appropriateness, empathy depth, and ethical application, with top marks for nuanced, human-centric judgment. Notably, GPT-4 scored 8.6 points lower here than residents [38]. Furthermore, because GAI tools do not provide an authentic, interactive experience or situational awareness, they may struggle to simulate the behavior and reactions of real patients accurately. This limitation makes it challenging for students to fully appreciate the importance of empathy and its application in doctor-patient interactions, which affects their development of communication and empathy skills development [38,54,76,77,101,107,147,152].

Fourth is resource inequality, which is most evident in the unequal access to technology and data. Datasets used for training GAI often exhibit biases, particularly involving data from different racial or socioeconomic backgrounds. This can worsen existing health care disparities. Furthermore, developing high-quality LLMs requires substantial computational resources, creating significant access barriers, especially for educational institutions or students with limited financial means. Hence,

subscription fees and hardware limitations restrict their access to these GAI tools [67,74,85,134].

Fifth is the ownership of intellectual property rights. The widespread use of GAI in medical education raises numerous intellectual property concerns, particularly regarding copyright disputes related to the medical data used during AI training [113]. In addition, the legal status of GAI-generated content remains unclear, as current copyright laws do not adequately address the ownership of GAI-generated images and texts. This leaves the ownership of such content unclear, complicating the determination of whether the rights belong to the user, the developer, or other stakeholders [27,50,58,59,124].

Sixth is the "black box" problem and the attribution of responsibility. The application of GAI in medical education faces a significant challenge known as the "black box" problem. This issue arises from the lack of transparency and interpretability of GAI models, which directly affects the safety and reliability of these applications in medical settings. This lack of transparency makes it hard to understand how GAI reaches specific conclusions, especially when results are erroneous or biased, complicating efforts to trace and correct mistakes [86,88,148]. Furthermore, when GAI is used for diagnostic or clinical decision support, any errors or biases in its generated results can make it difficult to establish accountability. Trust in the doctor-patient relationship is built on clear responsibility. However, the lack of transparency in GAI models undermines this trust, leaving patients and physicians uncertain about the safety and reliability of GAI-driven decisions [36,114].

Discussion

Principal Findings

This scoping review systematically identifies 3 core characteristics of GAI in medical education through an analysis of 131 included studies: pronounced regional disparities, empowerment potential via RMA synergy, and unresolved technical and ethical challenges. These findings must be contextualized within the field's evolving landscape: Our initial screening retrieved 5991 articles, a striking number reflecting both the opportunities and challenges of this emerging domain. This vast volume can be attributed to GAI's rapid evolution as a nascent technology, where relevant concepts remain loosely defined and inconsistent. Consequently, keyword usage lacks standardization, often resulting in the inclusion of tangentially related cross-disciplinary studies. Furthermore, GAI's inherently interdisciplinary nature broadens the scope of relevant literature. While this abundance highlights widespread interest and diverse applications, it also emphasizes the lack of conceptual clarity and consistency in frameworks. Therefore, although research is progressing, the field remains in a transitional stage, moving from "conceptual standardization" to "unified frameworks." To propel the field forward, the academic community needs to reach a consensus on GAI-related definitions and application structures. Achieving this standardization will enable better tracking of emerging trends and facilitate the effective use of new insights.

Against this backdrop, regional distribution analysis reveals marked concentration of GAI research in very high HDI regions (74%), with minimal contributions from low-HDI regions (2%) and scarce cross-regional collaborations (4%), highlighting structural inequities in global technology diffusion. Model use patterns further demonstrate ChatGPT's dominant adoption (89.5%), driven by its superior performance in multifaceted educational tasks: (1) iterative version advancements (eg, GPT-4's significant improvements in reasoning accuracy and error reduction over GPT-3.5); (2) proven efficacy across diverse applications including clinical simulation, exam question generation, and personalized tutoring; and (3) robust multilingual support despite variability in non-English contexts. This technical versatility explains its preferential adoption by researchers. The disproportionately high usage rate of general LLMs over specialized models, coupled with a predominant focus on cross-model comparisons rather than synergistic integration, reflects insufficient exploration of technical adaptability and system interoperability within current research.

Within the RMA tripartite framework established in this study, GAI reshapes medical education through coordinated optimization across 3 dimensions. In resource provisioning, it effectively mitigates traditional constraints of specimen scarcity and privacy limitations through the efficient generation of diverse clinical cases and pathological images. Methodologically, it facilitates the transition from standardized instruction to personalized education through interdisciplinary knowledge integration and targeted learning support. For assessment, high concordance in automated scoring and academic integrity monitoring provides scalable solutions for educational quality assurance. This closed-loop optimization mechanism, which encompasses resource allocation, pedagogical implementation, and evaluative feedback, validates the framework's explanatory power for technology-enabled educational transformation.

Nevertheless, profound barriers impede deeper GAI integration. Current technical deficiencies manifest as: inadequate contextual adaptation (eg, limitations in complex clinical reasoning and MCQ processing), data quality flaws (including hallucinatory outputs and deficient nontextual information analysis), and linguistic or regional biases (particularly performance degradation in non-English contexts). Long-term risks include erosion of critical thinking and creativity due to overreliance, alongside ethical governance dilemmas that encompass ambiguous accountability, inequitable resource distribution, and deficient clinician-patient emotional engagement. These dual challenges constitute fundamental barriers to implementing human-AI collaboration paradigms.

Comparison With Existing Literature

This scoping review specifically focuses on the period between January 2023 and October 2024, a critical transitional phase where GAI in medical education shifted from theoretical exploration to practical implementation. By capturing this transformative era, it addresses the gap in previous reviews [1,9] that lacked coverage of the latest advancements. While building on the foundational insights of earlier studies, this review extends their scope by identifying emerging trends and practical

applications that have emerged with GAI's maturation in educational contexts.

Our observation of pronounced regional disparities starkly aligns with and quantifies the well-documented "digital divide" prevalent in global health technology diffusion [156]. However, this study provides concrete, GAI-specific evidence within medical education, highlighting the extreme concentration and the critical scarcity of cross-tier collaboration, thereby reinforcing concerns about equity in accessing transformative educational technologies and potentially exacerbating global health workforce inequities.

Regarding model use, the overwhelming dominance of ChatGPT mirrors its widespread popularity in GAI application studies [157]. Yet, our analysis delves deeper than mere prevalence reports or bibliometric study [158-160], specifically attributing this dominance to its rapid iteration (eg, GPT-4's improvements), proven versatility across key educational tasks (clinical sim, QG, and tutoring), and relatively robust (though imperfect) multilingual support, which are crucial for adoption in the diverse contexts of medical education research.

Our development of the RMA tripartite framework represents a key theoretical departure. While existing research acknowledges GAI's impact on discrete educational facets (resource provision, teaching methodologies, and evaluative processes), a unifying framework that binds these elements into a synergistic, closed-loop optimization mechanism is conspicuously absent from the current discourse [1,9,10]. Such a framework uniquely conceptualizes these three dimensions as an interdependent, dynamic closed-loop system essential for understanding GAI's holistic transformative potential. Crucially, the empirical identification of significant RMA imbalance (robust exploration of educational methods and resources vs sparse focus on learner assessment) does not imply that assessment is under-prioritized in education broadly, but rather reflects a current skew in GAI-medical education integration—with research disproportionately focusing on resource enrichment and methodological optimization, while lagging in the development of learner assessment applications [161]. This imbalance, viewed through our novel integrative lens, offers a structured diagnostic for the systemic gap in aligning GAI capabilities with the specific needs of learner assessment within medical education.

The unresolved technical-ethical challenges documented (eg, contextual limitations, hallucinations, biases, erosion of critical thinking, and concerns about empathy) resonate strongly with growing critiques of LLMs in healthcare [162,163]. Our review explicitly maps these well-recognized limitations onto the sensitive context of medical education, highlighting their manifestation and potential impact in shaping future clinicians. This reinforces concerns raised elsewhere but grounds them firmly in the educational domain.

Another distinctive contribution of this review lies in revealing a critical technological imbalance: the overwhelming focus on general-purpose LLMs like ChatGPT contrasts sharply with the lack of systematic development of specialized medical models and the near absence of research on multimodel collaborative mechanisms within medical education [10,164]. This finding highlights a gap in the current technological approach, which hinders depth and clinical authenticity. While previous studies used available tools, our synthesis highlights this specific limitation as a barrier to deeper integration.

Implications of the Findings

Implications for Educational Practice

This study makes a key contribution to pedagogical practice by establishing the RMA tripartite framework and revealing its developmental imbalances, thereby providing a practical paradigm for the integration of GAI into medical education. The core value of this framework lies in elucidating the dynamic closed-loop nature of technology-enhanced education, wherein resource provision establishes the pedagogical foundation, methodological innovation activates knowledge transformation, and assessment feedback drives systemic evolution; these 3 components constitute an interlocking educational mechanism [165].

As evidenced in the results section, the current imbalance, characterized by rich exploration in GAI-supported educational resources and teaching methods yet relatively limited progress in GAI-driven automated evaluation of learner performance, stems from an overemphasis on short-term efficiency in early technology adoption. This has led to systemic neglect of assessment's role as an optimization tool. For example, GAI is widely used to generate diverse clinical cases and pathological images to enrich educational resources and design adaptive learning pathways to innovate teaching methods. However, in learner assessment, most GAI tools still rely on simple automated scoring of knowledge-based quizzes, with few leveraging GAI to evaluate higher-order competencies such as clinical reasoning or diagnostic accuracy [26]. Another instance is that many researchers use GAI to create interactive simulation scenarios as a methodological advancement but fail to integrate automated assessment features that track learners' decision-making processes in these scenarios [53]. This misses opportunities to use assessment data to refine the scenarios themselves. Overreliance on GAI for resource and method innovation without matching progress in automated learner assessment risks disconnecting what is taught or provided from what learners need to master, ultimately limiting GAI's ability to drive meaningful change in medical education.

Achieving optimal integration requires establishing a bidirectional enhancement cycle centered on assessment. Automated assessment data capturing learning bottlenecks should guide the real-time expansion of clinical case libraries' pathological spectra and difficulty calibration [166], shifting resource provision from one-size-fits-all to demand responsiveness. Simultaneously, the focus on core competencies (such as clinical reasoning and problem-solving) emphasized in teaching methods must be integrated into new assessment dimensions [167], driving teaching methods to evolve from mere knowledge transmission to competency development. Within this cycle, assessment functions not merely as a quality monitoring tool, but as the central nexus for the co-evolution of resources and methods.

Realizing this vision necessitates educators reconceptualizing operational logic [165]. This involves using assessment data to inform the development of educational resources, specifically leveraging insights into learners' knowledge gaps and skill deficiencies to dynamically adjust the complexity of clinical cases [168], embedding real-time, practical, and contextual feedback mechanisms within high-order teaching activities like simulated diagnostics to optimize pedagogical strategies [169], and establishing adaptive rules enabling cross-dimensional interaction to facilitate systemic iteration [170,171]. Collectively, this structural transformation elevates the tripartite framework into an organic educational operating system.

However, technological integration inherently presents dual challenges, highlighting the importance of upholding core principles of human-AI collaboration. Generating educational resources without clinical context review risks reinforcing data biases [172]; methodological innovation overly reliant on algorithmic decisions may erode critical thinking [9]; and automated assessment replacing human judgment may overlook students' psychological needs, reducing course engagement and well-being scores [173]. These manifestations of technological alienation arise from the partial ceding of human agency. Resolution lies in upholding a human-AI symbiotic vision: recognizing GAI as a collaborator, not a replacement, in educational evolution. Specifically at the resource layer, clinicians and educators must oversee the development of educational resources (eg, clinical cases) to balance efficiency, ethics, and clinical authenticity [174,175]. At the method layer, educators should direct learning path design to integrate technological augmentation with pedagogical wisdom [176]. At the assessment layer, institutions should implement verification systems that combine human evaluation with machine automation, ensuring assessments balance efficiency with humanistic dimensions [173,177]. This reconfiguration of responsibilities positions technology as a tool and reaffirms human stewardship of education.

Implications for Technological Development

This study identifies a technological imbalance in the application of GAI within medical education. This imbalance is characterized by the dominance of large general language models, while the development of specialized models for specific medical disciplines has lacked systematic progress. This limitation restricts the depth of technology-enabled education and indicates a neglect of multimodel collaborative mechanisms within current research paradigms.

The study proposes an integrated system using general LLMs alongside specialized medical models, employing a hierarchical collaborative architecture to reshape the technological ecosystem of medical education. The core operational logic establishes a 3-tiered functional division: general models act as the central hub for teaching interactions, handling basic task parsing and process orchestration; medical specialized models, drawing on vertical domain knowledge bases, execute high-complexity core teaching tasks such as clinical reasoning and medical image generation; and a cross-model validation mechanism forms a closed-loop quality control system. This architecture adapts the hospital's multidisciplinary team approach to AI in education, aligning technological capabilities with the requirements of medical education for expertise, reliability, and contextual authenticity.

Within medical education, this integrated system can facilitate 3 key changes. First, it addresses limitations in specialized knowledge depth inherent in traditional general models, improving training efficacy for advanced clinical reasoning. Second, it leverages GAI's multimodal capabilities, which integrate text and image data, to address key issues in medical imaging education including shortages of teaching resources like rare pathological images and the limits of static materials in showing dynamic anatomical relationships. This support helps evolve pathology visualization from static atlases to interactive 3D simulations, letting students explore spatial structures and pathological changes more intuitively [178]. Third, it establishes a cross-model knowledge validation chain to automatically identify and correct typical logical inconsistencies and factual errors in general models, ensuring the academic rigor of teaching content. These changes collectively represent a paradigm shift from tool-assisted learning to intelligent teaching partnership systems [179].

Supporting the effective operation of this system requires targeted solutions to key technical challenges. The primary task involves developing specialized models with medical context adaptive capabilities, specifically enhancing their semantic parsing of unstructured clinical texts to address performance variability in complex case analysis [180,181]. Concurrently, it is necessary to construct dynamically evolving medical education datasets that incorporate cross-regional case spectra and multilingual clinical literature to systematically mitigate cultural biases and time-lag effects in training data [182]. Integrating privacy-preserving computation techniques like federated learning can enable secure data collaboration among institutions, continuously optimizing model localization and adaptation while safeguarding patient information security [183,184].

Implications for Policies and Governance

This study reveals a pronounced regional disparity in the application of GAI within the field of medical education. Specifically, regions with a very high HDI dominate research output in this domain, while contributions from the low-HDI areas account for only 2%. The scarcity of cross-tier collaboration between very high- and low-HDI areas further exacerbates this structural inequity in resource distribution. This imbalance epitomizes systemic inequalities within global knowledge production systems, rooted in 3 compounding barriers: inadequate computational infrastructure in resource-constrained settings impedes technological localization, proprietary restrictions on core models under patent regimes limit feasible technology transfer, and excessive reliance on clinical data from high-income countries compromises model adaptability to regional health care priorities. Without deliberate intervention, this self-reinforcing Matthew Effect cycle risks intensifying the global fragmentation of medical educational resources [185].

Addressing this complex challenge necessitates a multitiered governance framework. At the international level, binding

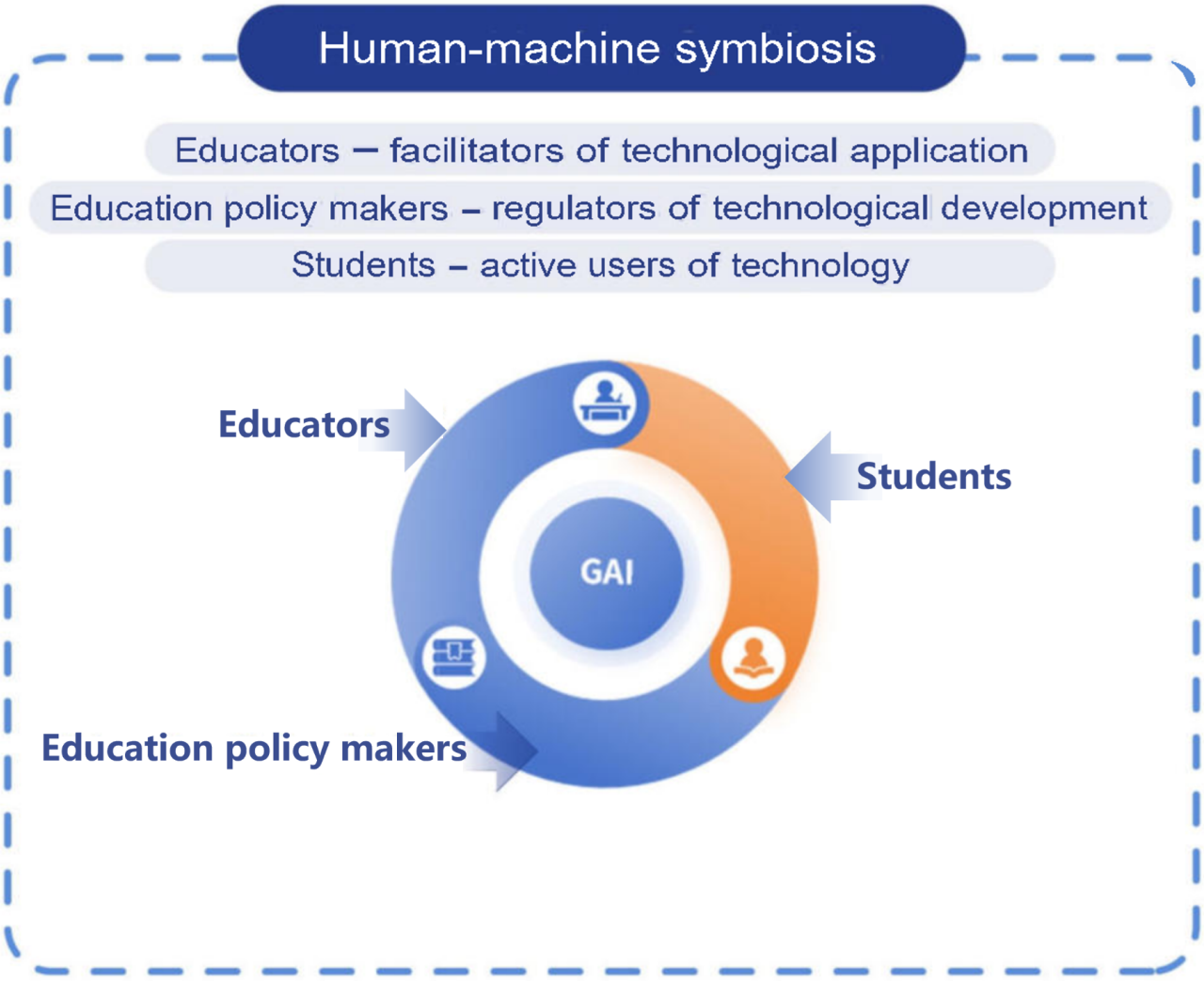
technology-sharing agreements should request that holders of advanced models provide architectural access under fair-use principles, emphasizing the need to balance innovation with equitable access, while emulating open-source paradigms as a reference model [186]. Concurrently, the World Health Organization could coordinate multinational efforts to develop nonprofit medical corpora incorporating disease spectra prevalent in low-HDI regions, such as tropical and endemic diseases [187]. Nationally, ministries of education should integrate computational infrastructure into public medical education budgets [188] and similar to the Medical Education Partnership Initiative (MEPI) [189], establish dedicated funds for cross-border institutional partnerships to co-develop localized pedagogical tools that address specific regional educational needs. Institutionally, medical schools should adopt algorithmic transparency protocols, requiring deployed GAI tools to provide auditable model documentation that details the demographics and geographical coverage of training data. Fairness assessments of these tools should be carried out by multidisciplinary committees, which include clinicians, ethicists, and community representatives [190].

Simultaneously, institutional responses must address secondary risks through integrated technical, educational, and regulatory safeguards. To counter academic misconduct, educational institutions should implement dual-track verification systems that require GAI-assisted submissions to be accompanied by generation logs and validated through detection tools [191]. Academic journals must establish clear authorship standards declaring proportional human-GAI contributions [192]. Mitigating critical thinking erosion requires curriculum committees to incorporate GAI-free clinical reasoning assessments, such as on-site case analyses evaluating independent diagnostic and management planning capabilities as prerequisites for professional certification [193].

Technical deficiencies demand targeted interventions. Reducing model hallucinations requires dynamic fact-checking systems linking GAI outputs to authoritative medical knowledge bases, with confidence levels displayed during teaching platform usage [194]. To address the opacity of algorithms, where the process by which GAI models derive conclusions remains unclear, it is necessary to document the diagnostic reasoning processes of these models. Such documentation allows instructors to review the reasoning, helps determine accountability when inconsistencies occur, and can be integrated into resident training evaluations to strengthen oversight of GAI-assisted decision-making [195].

Fundamentally, governance paradigms must transition from a technocentric approach to symbiotic development. Compared to the commonly used “human in the loop” [196], which mainly emphasizes humans overseeing or making final decisions in AI systems, symbiotic agency theory goes further: it highlights mutual shaping between humans and AI. Humans guide AI development through ethical norms and clinical experience, while AI enhances human capabilities by expanding cognitive boundaries, forming a dynamic, mutually reinforcing relationship [11]. Policies should affirm human primacy in medical education, exemplified by reserving clinical empathy training exclusively for human instructors while limiting GAI to standardized case supplementation. An effective return to the essence of symbiotic agency means building collaborative mechanisms as shown in Figure 5: educators lead in setting teaching goals and ensuring ethical alignment (eg, reviewing GAI-generated cases to match real clinical logic); GAI supports personalized learning; students provide feedback to refine GAI tools; and policies clarify rights and responsibilities in this interaction. This human-centered approach ensures technological advancement aligns with pedagogical integrity and global equity imperatives.

Figure 5. Vision of human-machine symbiosis: a schematic diagram. GAI: generative artificial intelligence.



Limitations and Future Direction

This scoping review has several limitations that should be acknowledged. First, the rapidly evolving nature of GAI means our findings primarily reflect the landscape captured up to the search date; newer models and applications emerging subsequently may shift current patterns. Second, the inherent conceptual breadth and interdisciplinary nature of GAI pose challenges for exhaustive literature capture, potentially leading to omissions despite broad search parameters. Third, and most critically, while this study proposes 3 key conceptual frameworks (the RMA tripartite model, the hierarchical collaborative architecture, and the symbiotic agency principle) and argues for their feasibility based on synthesized evidence, it has not empirically tested their implementation or efficacy in authentic educational settings. Finally, reliance on published literature may underrepresent real-world implementation challenges and grassroots innovations.

Future research must bridge this critical gap by translating these frameworks into practice. Priority should be given to: (1) implementing and evaluating the RMA balancing strategies and the integrated system combining general and specialized medical GAI models in specific medical education contexts to assess their impact on learning outcomes and operational feasibility; (2) conducting longitudinal studies to track the dynamic evolution of GAI integration over time, observing its long-term empowerment effects on educational processes and outcomes; and (3) operationalizing the symbiotic agency framework to guide the design, deployment, and assessment of these interventions. This framework is essential for ensuring that human-AI collaboration in practice genuinely augments educator and learner agency, fosters critical competencies, and upholds pedagogical integrity, thereby realizing the envisioned synergistic educational ecosystem.

Conclusion

The application of GAI in medical education exhibits significant regional inequities, reflecting structural disparities in technological diffusion. Statistical findings from the model research reflect that researchers have certain preferences in its usage. The emergence of GAI has revitalized medical education, which is manifested in its promotion of the diversification of educational methods, the scientific evaluation of education assessment, and the dynamic optimization of education resources. However, these innovations are accompanied by current limitations and potential future challenges. By establishing the RMA tripartite model as a dynamic closed-loop system for educational optimization, proposing an integrated multimodel architecture to reconcile general and specialized

GAI capabilities, and advancing the symbiotic agency principle to safeguard human primacy, this study provides foundational frameworks for navigating GAI integration. These contributions collectively address critical gaps in conceptual standardization and collaborative design, while delineating actionable pathways for pedagogical innovation, equitable technology development, and governance reform, which ultimately steer the field toward responsible human-AI collaboration that enhances clinical education without compromising pedagogical integrity or global equity.

Acknowledgments

This study was financially supported by the Funding of Medical Science and Technology Research in Guangdong Province, China (A2023363), the Industry-University-Research Collaborative Education Program of Ministry of Education, China (230905518284433), and the Teaching Reform Research Project of Clinical Teaching Base in Guangdong Province, China (2023-30).

Data Availability

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

Authors' Contributions

Conceptualization: YL, ZL
Methodology: YL, ZL
Formal analysis: YL, ZL, ZY, NZ
Investigation: YL, ZL, ZY, NZ
Data curation: ZY, NZ
Writing – original draft: YL, ZL
Writing – review & editing: YC, ZC, XL
Visualization: ZY, NZ
Supervision: L Zhao, L Zhang
Project administration: L Zhao, L Zhang
Resources: YC, ZC, XL

Conflicts of Interest

None declared.

Multimedia Appendix 1
Technical features and application comparison of mainstream generative artificial intelligence (GAI) models.
[DOCX File, 18 KB - mededu_v11i1e71125_app1.docx]

Multimedia Appendix 2
Search strategy.
[DOC File, 58 KB - mededu_v11i1e71125_app2.doc]

Checklist 1
PRISMA-ScR checklist.
[DOCX File, 67 KB - mededu_v11i1e71125_app3.docx]

References

1. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. JMIR Med Educ 2023 Oct 20;9:e48785. [doi: 10.2196/48785] [Medline: 37862079]
2. Generative AI market (2025 - 2030). Grand View Research. URL: https://www.grandviewresearch.com/industry-analysis/generative-ai-market-report [accessed 2025-03-03]
3. Stretton B, Kovoor J, Arnold M, Bacchi S. ChatGPT-based learning: generative artificial intelligence in medical education. Med Sci Educ 2024 Feb;34(1):215-217. [doi: 10.1007/s40670-023-01934-5] [Medline: 38510403]
4. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023;6:1169595. [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
5. Karabacak M, Ozkara BB, Margetis K, Wintermark M, Bisdas S. The advent of generative language models in medical education. JMIR Med Educ 2023 Jun 6;9:e48163. [doi: 10.2196/48163] [Medline: 37279048]

6. Totlis T, Natsis K, Filos D, et al. The potential role of ChatGPT and artificial intelligence in anatomy education: a conversation with ChatGPT. Surg Radiol Anat 2023 Aug 16;45(10):1321-1329. [doi: 10.1007/s00276-023-03229-1]
7. Hanna JJ, Wakene AD, Lehmann CU, Medford RJ. Assessing racial and ethnic bias in text generation for healthcare-related tasks by ChatGPT1. medRxiv. 2023 Aug 28 p. 2023.08.28.23294730. [doi: 10.1101/2023.08.28.23294730] [Medline: 37693388]
8. Densen P. Challenges and opportunities facing medical education. Trans Am Clin Climatol Assoc 2011;122:48-58. [Medline: 21686208]
9. Xu T, Weng H, Liu F, et al. Current status of ChatGPT use in medical education: potentials, challenges, and strategies. J Med Internet Res 2024 Aug 28;26:e57896. [doi: 10.2196/57896] [Medline: 39196640]
10. Temsah O, Khan SA, Chaiah Y, et al. Overview of early ChatGPT's presence in medical literature: insights from a hybrid literature review by ChatGPT and human experts. Cureus 2023 Apr;15(4):e37281. [doi: 10.7759/cureus.37281] [Medline: 37038381]
11. Neff G, Nagy P. Agency in the digital age: using symbiotic agency to explain human–technology interaction. In: Papacharissi Z, editor. A Networked Self and Human Augmentics, Artificial Intelligence, Sentience, 1st edition: Routledge; 2018:97-107. [doi: 10.4324/9781315202082-8]
12. The 22 best generative AI tools for SMBs to stay competitive in 2025. WebFX. URL: https://www.webfx.com/blog/marketing/best-generative-ai-tools/ [accessed 2025-07-19]
13. Temsah MH, Alhuzaimi AN, Almansour M, et al. Art or artifact: evaluating the accuracy, appeal, and educational value of AI-generated imagery in DALL·E 3 for illustrating congenital heart diseases. J Med Syst 2024 May 23;48(1):54 [FREE Full text] [doi: 10.1007/s10916-024-02072-0] [Medline: 38780839]
14. Claude 2: reviews, prices & features. Appvizer. URL: https://www.appvizer.com/artificial-intelligence/llm/claude-2 [accessed 2025-07-19]
15. Global large language model (LLM) market research report 2024. : QYResearch; 2024 URL: https://www.qyresearch.com/reports/2212992/large-language-model--llm [accessed 2025-10-09]
16. OpenAI's o3 - AI model details. DocsBot AI. URL: https://docsbot.ai/models/o3 [accessed 2025-07-19]
17. OpeneVidence. AITop10. URL: https://aitop10.tools/zh/detail/openevidence [accessed 2025-07-19]
18. Sora Turbo: OpenAI's enhanced video generation model goes public. Neurohive. URL: https://neurohive.io/en/ai-apps/sora-turbo-openai-s-enhanced-video-generation-model-goes-public/ [accessed 2025-07-19]
19. AI tools for medical education and research. Macon & Joan Brock Virginia Health Sciences at Old Dominion University. URL: https://www.evms.edu/about_us/ai_resources/resources_and_ai_tools/ai_tools_for_medical_education_and_research/ [accessed 2025-07-26]
20. Cho J, Puspitasari FD, Zheng S, et al. Sora as an AGI world model? A complete survey on text-to-video generation. arXiv. Preprint posted online on Mar 8, 2024. [doi: 10.48550/ARXIV.2403.05131]
21. Hu H, Liang H, Wang H. Longitudinal study of the earliest pilot of tiered healthcare system reforms in China: will the new type of chronic disease management be effective? Soc Sci Med 2021 Sep;285:114284. [doi: 10.1016/j.socscimed.2021.114284]
22. Peek CJ, Allen M, Loth KA, et al. Harmonizing the tripartite mission in academic family medicine: a longitudinal case example. Ann Fam Med 2024;22(3):237-243. [doi: 10.1370/afm.3108] [Medline: 38806264]
23. Geenens R, De Schutter H. A tripartite model of federalism. Philos Soc Crit 2023 Sep;49(7):753-785. [doi: 10.1177/01914537211066850]
24. Windak A, Rochfort A, Jacquet J. The revised European definition of general practice/family medicine. a pivotal role of one health, planetary health and sustainable development goals. Eur J Gen Pract 2024 Dec;30(1):2306936. [doi: 10.1080/13814788.2024.2306936] [Medline: 38334099]
25. Human development report 2023-24. : United Nations Development Programme; 2024 Mar URL: https://hdr.undp.org/content/human-development-report-2023-24 [accessed 2024-12-05]
26. Grévisse C. LLM-based automatic short answer grading in undergraduate medical education. BMC Med Educ 2024 Sep 27;24(1):1060. [doi: 10.1186/s12909-024-06026-5] [Medline: 39334087]
27. Liu JQJ, Hui KTK, Al Zoubi F, et al. The great detectives: humans versus AI detectors in catching large language model-generated medical writing. Int J Educ Integr 2024 May 20;20(1):8. [doi: 10.1007/s40979-024-00155-6]
28. Li J, Zong H, Wu E, et al. Exploring the potential of artificial intelligence to enhance the writing of english academic papers by non-native english-speaking medical students - the educational application of ChatGPT. BMC Med Educ 2024 Jul 9;24(1). [doi: 10.1186/s12909-024-05738-y]
29. Li KC, Bu ZJ, Shahjalal M, et al. Performance of ChatGPT on Chinese master's degree entrance examination in clinical medicine. In: Grewal HS, editor. PLoS ONE 2024;19(4):e0301702. [doi: 10.1371/journal.pone.0301702] [Medline: 38573944]
30. Cherif H, Moussa C, Missaoui AM, Salouage I, Mokaddem S, Dhahri B. Appraisal of ChatGPT's aptitude for medical education: comparative analysis with third-year medical students in a pulmonology examination. JMIR Med Educ 2024 Jul 23;10:e52818. [doi: 10.2196/52818] [Medline: 39042876]
31. Ali K, Barhom N, Tamimi F, Duggal M. ChatGPT—A double‐edged sword for healthcare education? Implications for assessments of dental students. Eur J Dental Education 2024 Feb;28(1):206-211. [doi: 10.1111/eje.12937]

32. Panthier C, Gatinel D. Success of ChatGPT, an AI language model, in taking the French language version of the European Board of Ophthalmology examination: A novel approach to medical knowledge assessment. J Fr Ophtalmol 2023 Sep;46(7):706-711. [doi: 10.1016/j.jfo.2023.05.006] [Medline: 37537126]
33. Gandhi AP, Joesph FK, Rajagopal V, et al. Performance of ChatGPT on the India undergraduate community medicine examination: cross-sectional study. JMIR Form Res 2024 Mar 25;8:e49964. [doi: 10.2196/49964] [Medline: 38526538]
34. Yu P, Fang C, Liu X, et al. Performance of ChatGPT on the Chinese postgraduate examination for clinical medicine: survey study. JMIR Med Educ 2024 Feb 9;10:e48514. [doi: 10.2196/48514] [Medline: 38335017]
35. Morreel S, Verhoeven V, Mathysen D. Microsoft Bing outperforms five other generative artificial intelligence chatbots in the Antwerp University multiple choice medical license exam. In: Banerjee I, editor. PLOS Digit Health 2024 Feb;3(2):e0000349. [doi: 10.1371/journal.pdig.0000349] [Medline: 38354127]
36. Guerra GA, Hofmann H, Sobhani S, et al. GPT-4 artificial intelligence model outperforms ChatGPT, medical students, and neurosurgery residents on neurosurgery written board-like questions. World Neurosurg 2023 Nov;179:e160-e165. [doi: 10.1016/j.wneu.2023.08.042] [Medline: 37597659]
37. Huang RS, Lu KJQ, Meaney C, Kemppainen J, Punnett A, Leung FH. Assessment of resident and AI chatbot performance on the University of Toronto family medicine residency progress test: comparative study. JMIR Med Educ 2023 Sep 19;9:e50514. [doi: 10.2196/50514] [Medline: 37725411]
38. Watari T, Takagi S, Sakaguchi K, et al. Performance comparison of ChatGPT-4 and Japanese medical residents in the general medicine in-training examination: comparison study. JMIR Med Educ 2023 Dec 6;9:e52202. [doi: 10.2196/52202] [Medline: 38055323]
39. Terwilliger E, Bcharah G, Bcharah H, Bcharah E, Richardson C, Scheffler P. Advancing medical education: performance of generative artificial intelligence models on otolaryngology board preparation questions with image analysis insights. Cureus 2024 Jul;16(7):e64204. [doi: 10.7759/cureus.64204] [Medline: 39130878]
40. Revercomb L, Patel AM, Fu D, Filimonov A. Performance of novel GPT-4 in otolaryngology knowledge assessment. Indian J Otolaryngol Head Neck Surg 2024 Dec;76(6):6112-6114. [doi: 10.1007/s12070-024-04935-x] [Medline: 39559040]
41. Riedel M, Kaefinger K, Stuehrenberg A, et al. ChatGPT's performance in German OB/GYN exams – paving the way for AI-enhanced medical education and clinical practice. Front Med 2023;10. [doi: 10.3389/fmed.2023.1296615]
42. Patel EA, Fleischer L, Filip P, et al. Comparative performance of ChatGPT 3.5 and GPT4 on rhinology standardized board examination questions. OTO Open 2024;8(2):e164. [doi: 10.1002/oto2.164] [Medline: 38938507]
43. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. JMIR Med Educ 2023 Jun 29;9:e48002. [doi: 10.2196/48002] [Medline: 37384388]
44. Haddad F, Saade JS. Performance of ChatGPT on ophthalmology-related questions across various examination levels: observational study. JMIR Med Educ 2024 Jan 18;10:e50842. [doi: 10.2196/50842] [Medline: 38236632]
45. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
46. Anderson LW, Krathwohl DR. A Taxonomy For Learning, Teaching, And Assessing: A Revision Of Bloom's Taxonomy Of Educational Objectives: Addison Wesley Longman, Inc; 2001.
47. Yudovich MS, Makarova E, Hague CM, Raman JD. Performance of GPT-3.5 and GPT-4 on standardized urology knowledge assessment items in the United States: a descriptive study. J Educ Eval Health Prof 2024;21(17):17. [doi: 10.3352/jeehp.2024.21.17] [Medline: 38977032]
48. Bharatha A, Ojeh N, Fazle Rabbi AM, et al. Comparing the performance of ChatGPT-4 and medical students on MCQs at varied levels of Bloom's taxonomy. Adv Med Educ Pract 2024;15:393-400. [doi: 10.2147/AMEP.S457408] [Medline: 38751805]
49. Wong K, Fayngersh A, Traba C, Cennimo D, Kothari N, Chen S. Using ChatGPT in the development of clinical reasoning cases: a qualitative study. Cureus 2024 May;16(5):e61438. [doi: 10.7759/cureus.61438] [Medline: 38953081]
50. Shimizu I, Kasai H, Shikino K, et al. Developing medical education curriculum reform strategies to address the impact of generative AI: qualitative study. JMIR Med Educ 2023 Nov 30;9:e53466. [doi: 10.2196/53466] [Medline: 38032695]
51. Bakkum MJ, Hartjes MG, Piët JD, et al. Using artificial intelligence to create diverse and inclusive medical case vignettes for education. Brit J Clinical Pharma 2024 Mar;90(3):640-648. [doi: 10.1111/bcp.15977]
52. Smith A, Hachen S, Schleifer R, Bhugra D, Buadze A, Liebrenz M. Old dog, new tricks? Exploring the potential functionalities of ChatGPT in supporting educational methods in social psychiatry. Int J Soc Psychiatry 2023 Dec;69(8):1882-1889. [doi: 10.1177/00207640231178451]
53. Scherr R, Halaseh FF, Spina A, Andalib S, Rivera R. ChatGPT interactive medical simulations for early clinical education: case study. JMIR Med Educ 2023 Nov 10;9:e49877. [doi: 10.2196/49877] [Medline: 37948112]
54. Sardesai N, Russo P, Martin J, Sardesai A. Utilizing generative conversational artificial intelligence to create simulated patient encounters: a pilot study for anaesthesia training. Postgrad Med J 2024 Mar 18;100(1182):237-241. [doi: 10.1093/postmj/qgad137] [Medline: 38240054]

55. Magalhães Araujo S, Cruz-Correia R. Incorporating ChatGPT in medical informatics education: mixed methods study on student perceptions and experiential integration proposals. JMIR Med Educ 2024 Mar 20;10:e51151. [doi: 10.2196/51151] [Medline: 38506920]
56. Brennan L, Balakumar R, Bennett W. The role of ChatGPT in enhancing ENT surgical training – a trainees' perspective. J Laryngol Otol 2024 May;138(5):480-486. [doi: 10.1017/S0022215123001354]
57. Tabuchi H, Engelmann J, Maeda F, et al. Using artificial intelligence to improve human performance: efficient retinal disease detection training with synthetic images. Br J Ophthalmol 2024 Sep 20;108(10):1430-1435. [doi: 10.1136/bjo-2023-324923] [Medline: 38485215]
58. Seth I, Lim B, Cevik J, et al. Utilizing GPT-4 and generative artificial intelligence platforms for surgical education: an experimental study on skin ulcers. Eur J Plast Surg 2024 Jan 29;47(1):19. [doi: 10.1007/s00238-024-02162-9]
59. Fan BE, Chow M, Winkler S. Artificial intelligence-generated facial images for medical education. MedSciEduc 2023 Nov 14;34(1):5-7. [doi: 10.1007/s40670-023-01942-5]
60. Al-Worafi YM, Goh KW, Hermansyah A, Tan CS, Ming LC. The use of ChatGPT for education modules on integrated pharmacotherapy of infectious disease: educators' perspectives. JMIR Med Educ 2024 Jan 12;10:e47339. [doi: 10.2196/47339] [Medline: 38214967]
61. Robleto E, Habashi A, Kaplan MAB, et al. Medical students' perceptions of an artificial intelligence (AI) assisted diagnosing program. Med Teach 2024 Sep;46(9):1180-1186. [doi: 10.1080/0142159X.2024.2305369] [Medline: 38306667]
62. Kıyak YS, Kononowicz AA. Case-based MCQ generator: a custom ChatGPT based on published prompts in the literature for automatic item generation. Med Teach 2024 Aug 2;46(8):1018-1020. [doi: 10.1080/0142159X.2024.2314723]
63. Tong L, Wang J, Rapaka S, Garg PS. Can ChatGPT generate practice question explanations for medical students, a new faculty teaching tool? Med Teach 2025 Mar 4;47(3):560-564. [doi: 10.1080/0142159X.2024.2363486]
64. Artsi Y, Sorin V, Konen E, Glicksberg BS, Nadkarni G, Klang E. Large language models for generating medical examinations: systematic review. BMC Med Educ 2024 Mar 29;24(1):354. [doi: 10.1186/s12909-024-05239-y] [Medline: 38553693]
65. Kawahara T, Sumi Y. GPT-4/4V's performance on the Japanese National Medical Licensing Examination. Med Teach 2025 Mar;47(3):450-457. [doi: 10.1080/0142159X.2024.2342545] [Medline: 38648547]
66. Tran CG, Chang J, Sherman SK, De Andrade JP. Performance of ChatGPT on American Board of Surgery in-training examination preparation questions. J Surg Res 2024 Jul;299:329-335. [doi: 10.1016/j.jss.2024.04.060] [Medline: 38788470]
67. Botross M, Mohammadi SO, Montgomery K, Crawford C. Performance of Google's artificial intelligence chatbot "Bard" (now "Gemini") on ophthalmology board exam practice questions. Cureus 2024 Mar;16(3):e57348. [doi: 10.7759/cureus.57348] [Medline: 38690460]
68. Gan W, Ouyang J, Li H, et al. Integrating ChatGPT in orthopedic education for medical undergraduates: randomized controlled trial. J Med Internet Res 2024 Aug 20;26:e57037. [doi: 10.2196/57037] [Medline: 39163598]
69. Thomae AV, Witt CM, Barth J. Integration of ChatGPT into a course for medical students: explorative study on teaching scenarios, students' perception, and applications. JMIR Med Educ 2024 Aug 22;10:e50545. [doi: 10.2196/50545] [Medline: 39177012]
70. Favero TG. Using artificial intelligence platforms to support student learning in physiology. Adv Physiol Educ 2024 Jun 1;48(2):193-199. [doi: 10.1152/advan.00213.2023]
71. Ganjavi C, Eppler M, O'Brien D, et al. ChatGPT and large language models (LLMs) awareness and use. A prospective cross-sectional survey of U.S. medical students. PLOS Digit Health 2024 Sep;3(9):e0000596. [doi: 10.1371/journal.pdig.0000596] [Medline: 39236008]
72. Sallam M, Salim NA, Barakat M, Al-Tammemi AB. ChatGPT applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations. Narra J 2023 Apr;3(1):e103. [doi: 10.52225/narra.v3i1.103] [Medline: 38450035]
73. Arun G, Perumal V, Urias F, et al. ChatGPT versus a customized AI chatbot (Anatbuddy) for anatomy education: a comparative pilot study. Anatomical Sciences Ed 2024 Oct;17(7). [doi: 10.1002/ase.2502] [Medline: 39169464]
74. Deng A, Chen W, Dai J, et al. Current application of ChatGPT in undergraduate nuclear medicine education: Taking Chongqing Medical University as an example. Med Teach 2025 Jun 3;47(6):997-1003. [doi: 10.1080/0142159X.2024.2399673]
75. Garabet R, Mackey BP, Cross J, Weingarten M. ChatGPT-4 performance on USMLE step 1 style questions and its implications for medical education: a comparative study across systems and disciplines. MedSciEduc 2023 Dec 27;34(1):145-152. [doi: 10.1007/s40670-023-01956-z]
76. Saleem N, Mufti T, Sohail SS, Madsen D. ChatGPT as an innovative heutagogical tool in medical education. Cogent Education 2024 Dec 31;11(1):2332850. [doi: 10.1080/2331186X.2024.2332850]
77. Huang H, Lin HC. ChatGPT as a life coach for professional identity formation in medical education. Educational Technology & Society 2024;27(3):374-389 [FREE Full text]
78. Dhanvijay AKD, Pinjar MJ, Dhokane N, Sorte SR, Kumari A, Mondal H. Performance of large language models (ChatGPT, Bing Search, and Google Bard) in solving case vignettes in physiology. Cureus 2023 Aug;15(8):e42972. [doi: 10.7759/cureus.42972] [Medline: 37671207]

79. Wang T, Mainous AG 3rd, Stelter K, O'Neill TR, Newton WP. Performance evaluation of the generative pre-trained transformer (GPT-4) on the family medicine in-training examination. J Am Board Fam Med 2024 Oct 25;37(4):528-582. [doi: 10.3122/jabfm.2023.230433R1] [Medline: 39214695]
80. Abdullahi T, Singh R, Eickhoff C. Learning to make rare and complex diagnoses with generative AI assistance: qualitative study of popular large language models. JMIR Med Educ 2024 Feb 13;10:e51391. [doi: 10.2196/51391] [Medline: 38349725]
81. Guastafierro V, Corbitt DN, Bressan A, et al. Unveiling the risks of ChatGPT in diagnostic surgical pathology. Virchows Arch 2025 Apr;486(4):663-673. [doi: 10.1007/s00428-024-03918-1] [Medline: 39269615]
82. Sarangi PK, Irodi A, Panda S, Nayak DSK, Mondal H. Radiological differential diagnoses based on cardiovascular and thoracic imaging patterns: perspectives of four large language models. Indian J Radiol Imaging 2024 Apr;34(2):269-275. [doi: 10.1055/s-0043-1777289] [Medline: 38549881]
83. Shukla R, Mishra AK, Banerjee N, Verma A. The comparison of ChatGPT 3.5, Microsoft Bing, and Google Gemini for diagnosing cases of neuro-ophthalmology. Cureus 2024 Apr;16(4):e58232. [doi: 10.7759/cureus.58232] [Medline: 38745784]
84. Hadi A, Tran E, Nagarajan B, Kirpalani A. Evaluation of ChatGPT as a diagnostic tool for medical learners and clinicians. In: Ata F, editor. PLoS ONE 2024 Jul 31;19(7):e0307383. [doi: 10.1371/journal.pone.0307383]
85. Guthrie E, Levy D, Del Carmen G. The Operating and Anesthetic Reference Assistant (OARA): A fine-tuned large language model for resident teaching. Am J Surg 2024 Aug;234:28-34. [doi: 10.1016/j.amjsurg.2024.02.016] [Medline: 38365551]
86. Zhang Y, Hao Y. Traditional Chinese medicine knowledge graph construction based on large language models. Electronics (Basel) 2024 Jul;13(7):1395. [doi: 10.3390/electronics13071395]
87. Luke W, Seow Chong L, Ban KH, et al. Is ChatGPT 'ready' to be a learning tool for medical undergraduates and will it perform equally in different subjects? Comparative study of ChatGPT performance in tutorial and case-based learning questions in physiology and biochemistry. Med Teach 2024 Nov;46(11):1441-1447. [doi: 10.1080/0142159X.2024.2308779]
88. Oh N, Choi GS, Lee WY. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. Ann Surg Treat Res 2023 May;104(5):269-273. [doi: 10.4174/astr.2023.104.5.269] [Medline: 37179699]
89. Guillen-Grima F, Guillen-Aguinaga S, Guillen-Aguinaga L, et al. Evaluating the efficacy of ChatGPT in navigating the Spanish Medical Residency entrance examination (MIR): promising horizons for AI in clinical medicine. Clin Pract 2023 Nov 20;13(6):1460-1487. [doi: 10.3390/clinpract13060130] [Medline: 37987431]
90. Lai UH, Wu KS, Hsu TY, Kan JKC. Evaluating the performance of ChatGPT-4 on the United Kingdom Medical Licensing Assessment. Front Med 2023 Sep 19;10:1240915. [doi: 10.3389/fmed.2023.1240915]
91. Isleem UN, Zaidat B, Ren R, et al. Can generative artificial intelligence pass the orthopaedic board examination? J Orthop 2024 Jul;53:27-33. [doi: 10.1016/j.jor.2023.10.026]
92. Mackey BP, Garabet R, Maule L, Tadesse A, Cross J, Weingarten M. Evaluating ChatGPT-4 in medical education: an assessment of subject exam performance reveals limitations in clinical curriculum support for students. Discov Artif Intell 2024 May 16;4(1):38. [doi: 10.1007/s44163-024-00135-2]
93. Jaworski A, Jasiński D, Jaworski W, et al. Comparison of the performance of artificial intelligence versus medical professionals in the Polish Final Medical Examination. Cureus 2024 Aug;16(8):e66011. [doi: 10.7759/cureus.66011] [Medline: 39221376]
94. Abbas A, Rehman MS, Rehman SS. Comparing the performance of popular large language models on the National Board of Medical Examiners sample questions. Cureus 2024 Mar;16(3):e55991. [doi: 10.7759/cureus.55991] [Medline: 38606229]
95. Goodings AJ, Kajitani S, Chhor A, et al. Assessment of ChatGPT-4 in family medicine board examinations using advanced AI learning and analytical methods: observational study. JMIR Med Educ 2024 Oct 8;10:e56128. [doi: 10.2196/56128] [Medline: 39378442]
96. Jin HK, Lee HE, Kim E. Performance of ChatGPT-3.5 and GPT-4 in national licensing examinations for medicine, pharmacy, dentistry, and nursing: a systematic review and meta-analysis. BMC Med Educ 2024 Sep 16;24(1). [doi: 10.1186/s12909-024-05944-8]
97. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ (Chicago Ill) 2024 Nov;58(11):1276-1285. [doi: 10.1111/medu.15402]
98. Alkhaaldi SMI, Kassab CH, Dimassi Z, et al. Medical student experiences and perceptions of ChatGPT and artificial intelligence: cross-sectional study. JMIR Med Educ 2023 Dec 22;9:e51302. [doi: 10.2196/51302] [Medline: 38133911]
99. Hersh W, Fultz Hollis K. Results and implications for generative AI in a large introductory biomedical and health informatics course. NPJ Digit Med 2024 Sep 13;7(1):247. [doi: 10.1038/s41746-024-01251-0] [Medline: 39271955]
100. Altamimi I, Alhumimidi A, Alshehri S, et al. The scientific knowledge of three large language models in cardiology: multiple-choice questions examination-based performance. Annals of Medicine & Surgery 2024 May 3;86(6):3261-3266. [doi: 10.1097/MS9.0000000000002120]
101. Hou Y, Guo L, Luo F. Conflict of interest the authors declare that they have no conflict of interest. SSRN Journal 2022. [doi: 10.2139/ssrn.4258054]
102. Zong H, Li J, Wu E, Wu R, Lu J, Shen B. Performance of ChatGPT on Chinese national medical licensing examinations: a five-year examination evaluation study for physicians, pharmacists and nurses. BMC Med Educ 2024 Feb 14;24(1):143. [doi: 10.1186/s12909-024-05125-7] [Medline: 38355517]

103. Bongco EDA, Cua SKN, Hernandez M, Pascual JSG, Khu KJO. The performance of ChatGPT versus neurosurgery residents in neurosurgical board examination-like questions: a systematic review and meta-analysis. Neurosurg Rev 2024 Dec 7;47(1):892. [doi: 10.1007/s10143-024-03144-y] [Medline: 39643792]
104. Cuthbert R, Simpson AI. Artificial intelligence in orthopaedics: can Chat Generative Pre-trained Transformer (ChatGPT) pass Section 1 of the Fellowship of the Royal College of Surgeons (Trauma & Orthopaedics) examination? Postgrad Med J 2023 Sep 21;99(1176):1110-1114. [doi: 10.1093/postmj/qgad053] [Medline: 37410674]
105. Tangadulrat P, Sono S, Tangtrakulwanich B. Using ChatGPT for clinical practice and medical education: cross-sectional survey of medical students' and physicians' perceptions. JMIR Med Educ 2023 Dec 22;9:e50658. [doi: 10.2196/50658] [Medline: 38133908]
106. Nicikowski J, Szczepański M, Miedziaszczyk M, Kudliński B. The potential of ChatGPT in medicine: an example analysis of nephrology specialty exams in Poland. Clin Kidney J 2024 Aug;17(8):sfae193. [doi: 10.1093/ckj/sfae193] [Medline: 39099569]
107. Borchert RJ, Hickman CR, Pepys J, Sadler TJ. Performance of ChatGPT on the situational judgement test-a professional dilemmas-based examination for doctors in the United Kingdom. JMIR Med Educ 2023 Aug 7;9:e48978. [doi: 10.2196/48978] [Medline: 37548997]
108. Hudon A, Kiepura B, Pelletier M, Phan V. Using ChatGPT in psychiatry to design script concordance tests in undergraduate medical education: mixed methods study. JMIR Med Educ 2024 Apr 4;10:e54067. [doi: 10.2196/54067] [Medline: 38596832]
109. Agarwal M, Sharma P, Goswami A. Analysing the applicability of ChatGPT, Bard, and Bing to generate reasoning-based multiple-choice questions in medical physiology. Cureus 2023 Jun;15(6):e40977. [doi: 10.7759/cureus.40977] [Medline: 37519497]
110. Wu JH, Nishida T, Liu TYA. Accuracy of large language models in answering ophthalmology board-style questions: A meta-analysis. Asia Pac J Ophthalmol (Phila) 2024 Sep;13(5):100106. [doi: 10.1016/j.apjo.2024.100106]
111. Torres-Zegarra BC, Rios-Garcia W, Ñaña-Cordova AM, et al. Performance of ChatGPT, Bard, Claude, and Bing on the Peruvian National Licensing Medical Examination: a cross-sectional study. J Educ Eval Health Prof 2023;20:30. [doi: 10.3352/jeehp.2023.20.30] [Medline: 37981579]
112. Weng TL, Wang YM, Chang S, Chen TJ, Hwang SJ. ChatGPT failed Taiwan's Family Medicine Board Exam. J Chin Med Assoc 2023 Aug 1;86(8):762-766. [doi: 10.1097/JCMA.0000000000000946] [Medline: 37294147]
113. Yoon SH, Oh SK, Lim BG, Lee HJ. Performance of ChatGPT in the in-training examination for anesthesiology and pain medicine residents in South Korea: observational study. JMIR Med Educ 2024 Sep 16;10:e56859. [doi: 10.2196/56859] [Medline: 39284182]
114. Wang H, Wu W, Dou Z, He L, Yang L. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: Pave the way for medical AI. Int J Med Inform 2023 Sep;177:105173. [doi: 10.1016/j.ijmedinf.2023.105173]
115. Keshtkar A, Atighi F, Reihani H. Systematic review of ChatGPT accuracy and performance in Iran's medical licensing exams: A brief report. J Educ Health Promot 2024 Nov;13(1):421. [doi: 10.4103/jehp.jehp_1210_24]
116. Liu M, Okuhara T, Chang X, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. J Med Internet Res 2024 Jul 25;26:e60807. [doi: 10.2196/60807] [Medline: 39052324]
117. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. Neurosurgery 2023 Dec 1;93(6):1353-1365. [doi: 10.1227/neu.0000000000002632] [Medline: 37581444]
118. Elias ML, Burshtein J, Sharon VR. OpenAI's GPT‐4 performs to a high degree on board‐style dermatology questions. Int J Dermatology 2024 Jan;63(1):73-78. [doi: 10.1111/ijd.16913]
119. Sabri H, Saleh MHA, Hazrati P, et al. Performance of three artificial intelligence (AI)‐based large language models in standardized testing; implications for AI‐assisted dental education. J of Periodontal Research 2025 Feb;60(2):121-133. [doi: 10.1111/jre.13323]
120. Ilgaz HB, Çelik Z. The significance of artificial intelligence platforms in anatomy education: an experience with ChatGPT and Google Bard. Cureus 2023 Sep;15(9):e45301. [doi: 10.7759/cureus.45301] [Medline: 37846274]
121. Khorshidi H, Mohammadi A, Yousem DM, et al. Application of ChatGPT in multilingual medical education: How does ChatGPT fare in 2023's Iranian residency entrance examination. Informatics in Medicine Unlocked 2023;41:101314. [doi: 10.1016/j.imu.2023.101314]
122. Huang CH, Hsiao HJ, Yeh PC, Wu KC, Kao CH. Performance of ChatGPT on Stage 1 of the Taiwanese medical licensing exam. Digit HEALTH 2024;10:20552076241233144. [doi: 10.1177/20552076241233144] [Medline: 38371244]
123. Apornvirat S, Namboonlue C, Laohawetwanit T. Comparative analysis of ChatGPT and Bard in answering pathology examination questions requiring image interpretation. Am J Clin Pathol 2024 Sep 3;162(3):252-260. [doi: 10.1093/ajcp/aqae036]
124. Cross J, Robinson R, Devaraju S, et al. Transforming medical education: assessing the integration of ChatGPT into faculty workflows at a Caribbean medical school. Cureus 2023 Jul;15(7):e41399. [doi: 10.7759/cureus.41399] [Medline: 37426402]

125. Soulage CO, Van Coppenolle F, Guebre-Egziabher F. The conversational AI "ChatGPT" outperforms medical students on a physiology university examination. Adv Physiol Educ 2024 Dec 1;48(4):677-684. [doi: 10.1152/advan.00181.2023] [Medline: 38991037]
126. Gritti MN, AlTurki H, Farid P, Morgan CT. Progression of an artificial intelligence chatbot (ChatGPT) for pediatric cardiology educational knowledge assessment. Pediatr Cardiol 2024 Feb;45(2):309-313. [doi: 10.1007/s00246-023-03385-6] [Medline: 38170274]
127. Bartoli A, May AT, Al-Awadhi A, Schaller K. Probing artificial intelligence in neurosurgical training: ChatGPT takes a neurosurgical residents written exam. Brain Spine 2024;4:102715. [doi: 10.1016/j.bas.2023.102715] [Medline: 38163001]
128. Rasmussen ME, Akbarov K, Titovich E, et al. Potential of e-learning interventions and artificial intelligence-assisted contouring skills in radiotherapy: the ELAISA study. JCO Glob Oncol 2024 Aug;10(10):e2400173. [doi: 10.1200/GO.24.00173] [Medline: 39236283]
129. Mousavi M, Shafiee S, Harley JM, Cheung JCK, Abbasgholizadeh Rahimi S. Performance of generative pre-trained transformers (GPTs) in Certification Examination of the College of Family Physicians of Canada. Fam Med Com Health 2024 May;12(Suppl 1):e002626. [doi: 10.1136/fmch-2023-002626]
130. Temsah MH, Alhuzaimi AN, Almansour M, et al. Art or artifact: evaluating the accuracy, appeal, and educational value of AI-generated imagery in DALL·E 3 for illustrating congenital heart diseases. J Med Syst 2024 May 23;48(1):54. [doi: 10.1007/s10916-024-02072-0] [Medline: 38780839]
131. Fang Q, Reynaldi R, Araminta AS, et al. Artificial intelligence (AI)-driven dental education: exploring the role of chatbots in a clinical learning environment. J Prosthet Dent 2025 Oct;134(4):1296-1303. [doi: 10.1016/j.prosdent.2024.03.038] [Medline: 38644064]
132. Cheung BHH, Lau GKK, Wong GTC, et al. ChatGPT versus human in generating medical graduate exam multiple choice questions-A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). PLoS ONE 2023;18(8):e0290691. [doi: 10.1371/journal.pone.0290691] [Medline: 37643186]
133. Ignjatović A, Stevanović L. Efficacy and limitations of ChatGPT as a biostatistical problem-solving tool in medical education in Serbia: a descriptive study. J Educ Eval Health Prof 2023 Oct 16;20(28):28. [doi: 10.3352/jeehp.2023.20.28]
134. Agarwal M, Goswami A, Sharma P. Evaluating ChatGPT-3.5 and Claude-2 in answering and explaining conceptual medical physiology multiple-choice questions. Cureus 2023 Sep;15(9):e46222. [doi: 10.7759/cureus.46222] [Medline: 37908959]
135. Yanagita Y, Yokokawa D, Fukuzawa F, Uchida S, Uehara T, Ikusaka M. Expert assessment of ChatGPT's ability to generate illness scripts: an evaluative study. BMC Med Educ 2024 May 15;24(1):536. [doi: 10.1186/s12909-024-05534-8] [Medline: 38750546]
136. Sauder M, Tritsch T, Rajput V, Schwartz G, Shoja MM. Exploring generative artificial intelligence-assisted medical education: assessing case-based learning for medical students. Cureus 2024 Jan;16(1):e51961. [doi: 10.7759/cureus.51961] [Medline: 38333501]
137. Hanna RE, Smith LR, Mhaskar R, Hanna K. Performance of language models on the family medicine in-training exam. Fam Med 2024 Oct;56(9):555-560. [doi: 10.22454/FamMed.2024.233738] [Medline: 39207788]
138. Takahashi H, Shikino K, Kondo T, et al. Educational utility of clinical vignettes generated in Japanese by ChatGPT-4: mixed methods study. JMIR Med Educ 2024 Aug 13;10:e59133. [doi: 10.2196/59133] [Medline: 39137031]
139. Waikel RL, Othman AA, Patel T, et al. Recognition of genetic conditions after learning with images created using generative artificial intelligence. JAMA Netw Open 2024 Mar 4;7(3):e242609. [doi: 10.1001/jamanetworkopen.2024.2609] [Medline: 38488790]
140. Collins BR, Black EW, Rarey KE. Introducing AnatomyGPT: A customized artificial intelligence application for anatomical sciences education. Clin Anat 2024 Sep;37(6):661-669. [doi: 10.1002/ca.24178] [Medline: 38721869]
141. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. In: Dagan A, editor. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
142. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. Sci Rep 2023 Nov 22;13(1). [doi: 10.1038/s41598-023-46995-z]
143. Murphy Lonergan R, Curry J, Dhas K, Simmons BI. Stratified evaluation of GPT's question answering in surgery reveals artificial intelligence (AI) knowledge gaps. Cureus 2023 Nov;15(11):e48788. [doi: 10.7759/cureus.48788] [Medline: 38098921]
144. Wójcik S, Rulkiewicz A, Pruszczyk P, Lisik W, Poboży M, Domienik-Karłowicz J. Reshaping medical education: performance of ChatGPT on a PES medical examination. Cardiol J 2024;31(3):442-450. [doi: 10.5603/cj.97517] [Medline: 37830257]
145. Coşkun Ö, Kıyak YS, Budakoğlu I. ChatGPT to generate clinical vignettes for teaching and multiple-choice questions for assessment: A randomized controlled experiment. Med Teach 2025 Feb;47(2):268-274. [doi: 10.1080/0142159X.2024.2327477] [Medline: 38478902]
146. Knoedler L, Alfertshofer M, Knoedler S, et al. Pure wisdom or potemkin villages? A comparison of ChatGPT 3.5 and ChatGPT 4 on USMLE step 3 style questions: quantitative analysis. JMIR Med Educ 2024 Jan 5;10:e51148. [doi: 10.2196/51148] [Medline: 38180782]

147. Uribe SE, Maldupa I, Kavadella A, et al. Artificial intelligence chatbots and large language models in dental education: worldwide survey of educators. Eur J Dent Educ 2024 Nov;28(4):865-876. [doi: 10.1111/eje.13009] [Medline: 38586899]
148. Jarry Trujillo C, Vela Ulloa J, Escalona Vivas G, et al. Surgeons vs ChatGPT: assessment and feedback performance based on real surgical scenarios. J Surg Educ 2024 Jul;81(7):960-966. [doi: 10.1016/j.jsurg.2024.03.012] [Medline: 38749814]
149. Meo SA, Al-Khlaiwi T, AbuKhalaf AA, Meo AS, Klonoff DC. The scientific knowledge of Bard and ChatGPT in endocrinology, diabetes, and diabetes technology: multiple-choice questions examination-based performance. J Diabetes Sci Technol 2025 May;19(3):705-710. [doi: 10.1177/19322968231203987] [Medline: 37798960]
150. Shamim MS, Zaidi SJA, Rehman A. The revival of essay-type questions in medical education: harnessing artificial intelligence and machine learning. J Coll Physicians Surg Pak 2024 May;34(5):595-599. [doi: 10.29271/jcpsp.2024.05.595] [Medline: 38720222]
151. Meo SA, Alotaibi M, Meo MZS, Meo MOS, Hamid M. Medical knowledge of ChatGPT in public health, infectious diseases, COVID-19 pandemic, and vaccines: multiple choice questions examination based performance. Front Public Health 2024;12:1360597. [doi: 10.3389/fpubh.2024.1360597] [Medline: 38711764]
152. Ba H, zhang L, Yi Z. Enhancing clinical skills in pediatric trainees: a comparative study of ChatGPT-assisted and traditional teaching methods. BMC Med Educ 2024 May 22;24(1). [doi: 10.1186/s12909-024-05565-1]
153. Almazrou S, Alanezi F, Almutairi SA, et al. Enhancing medical students critical thinking skills through ChatGPT: An empirical study with medical students. Nutr Health 2025 Jul;31(3):1023-1033. [doi: 10.1177/02601060241273627]
154. Crawford LM, Hendzlik P, Lam J, et al. Digital ink and surgical dreams: perceptions of artificial intelligence–generated essays in residency applications. J Surg Res 2024 Sep;301:504-511. [doi: 10.1016/j.jss.2024.06.020] [Medline: 39042979]
155. Mosleh R, Jarrar Q, Jarrar Y, Tazkarji M, Hawash M. Medicine and pharmacy students' knowledge, attitudes, and practice regarding artificial intelligence programs: Jordan and West Bank of Palestine. Adv Med Educ Pract 2023;14:1391-1400. [doi: 10.2147/AMEP.S433255] [Medline: 38106923]
156. Western MJ, Smit ES, Gültzow T, et al. Bridging the digital health divide: a narrative review of the causes, implications, and solutions for digital health inequalities. Health Psychol Behav Med 2025;13(1):2493139. [doi: 10.1080/21642850.2025.2493139] [Medline: 40276490]
157. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023 May 4;6. [doi: 10.3389/frai.2023.1169595]
158. Liu L, Qu S, Zhao H, et al. Global trends and hotspots of ChatGPT in medical research: a bibliometric and visualized study. Front Med 2024 May 16;11. [doi: 10.3389/fmed.2024.1406842]
159. Khan N, Khan Z, Koubaa A, Khan MK, Salleh RB. Global insights and the impact of generative AI-ChatGPT on multidisciplinary: a systematic review and bibliometric analysis. Conn Sci 2024 Dec 31;36(1). [doi: 10.1080/09540091.2024.2353630]
160. 100+ eye-opening ChatGPT statistics: tracing the roots of generative AI to its global dominance. Master of Code. 2025 Jan. URL: https://masterofcode.com/blog/chatgpt-statistics [accessed 2025-07-26]
161. See BH, Gorard S, Lu B, Dong L, Siddiqui N. Is technology always helpful?: A critical review of the impact on learning outcomes of education technology in supporting formative assessment in schools. Res Pap Educ 2022 Nov 2;37(6):1064-1096. [doi: 10.1080/02671522.2021.1907778]
162. Nazi ZA, Peng W. Large language models in healthcare and medical domain: a review. Informatics (MDPI) ;11(3):57. [doi: 10.3390/informatics11030057]
163. Busch F, Hoffmann L, Rueger C, et al. Current applications and challenges in large language models for patient care: a systematic review. Commun Med 2025 Jan 21;5(1). [doi: 10.1038/s43856-024-00717-2]
164. Meyer JG, Urbanowicz RJ, Martin PCN, et al. ChatGPT and large language models in academia: opportunities and challenges. BioData Min 2023 Jul 13;16(1). [doi: 10.1186/s13040-023-00339-9]
165. Mao J, Chen B, Liu JC. Generative artificial intelligence in education and its implications for assessment. TechTrends 2024 Jan;68(1):58-66. [doi: 10.1007/s11528-023-00911-4]
166. Turner L, Hashimoto DA, Vasisht S, Schaye V. Demystifying AI: current state and future role in medical education assessment. Acad Med 2024 Apr 1;99(4S Suppl 1):S42-S47. [doi: 10.1097/ACM.0000000000005598] [Medline: 38166201]
167. Lakhtakia R, Otaki F, Alsuwaidi L, Zary N. Assessment as learning in medical education: feasibility and perceived impact of student-generated formative assessments. JMIR Med Educ 2022 Jul 22;8(3):e35820. [doi: 10.2196/35820] [Medline: 35867379]
168. Machkour M, El Jihaoui M, Lamalif L, Faris S, Mansouri K. Toward an adaptive learning assessment pathway. Front Educ 2025;10. [doi: 10.3389/feduc.2025.1498233]
169. Solis Trujillo BP, Velarde-Camaqui D, Gonzales Nuñez CA, Castillo Silva EV, Gonzalez Said de la Oliva MDP. The current landscape of formative assessment and feedback in graduate studies: a systematic literature review. Front Educ 2025 May 12;10. [doi: 10.3389/feduc.2025.1509983]
170. Wilson C, Scott B. Adaptive systems in education: a review and conceptual unification. IJILT 2017 Jan 3;34(1):2-19. [doi: 10.1108/IJILT-09-2016-0040]
171. Kolluru V, Mungara S, Chintakunta AN. Adaptive learning systems: harnessing AI for customized educational experiences. IJCSITY 2018 Aug 30;6(3):13-26 [FREE Full text] [doi: 10.5121/ijcsity.2018.6302]

172. Cross JL, Choma MA, Onofrey JA. Bias in medical AI: Implications for clinical decision-making. PLOS Digit Health 2024 Nov;3(11):e0000651. [doi: 10.1371/journal.pdig.0000651]
173. Sawan M. Balancing automation and empathy: how teachers can thrive with AI. Zenodo. Preprint posted online on May 18, 2025. [doi: 10.5281/zenodo.15456225]
174. Bond M, Khosravi H, De Laat M, et al. A meta systematic review of artificial intelligence in higher education: a call for increased ethics, collaboration, and rigour. Int J Educ Technol High Educ 2024 Jan 19;21(1). [doi: 10.1186/s41239-023-00436-z]
175. Resnik DB, Hosseini M. The ethics of using artificial intelligence in scientific research: new guidance needed for a new tool. AI Ethics 2025 Apr;5(2):1499-1521. [doi: 10.1007/s43681-024-00493-8]
176. Tong D, Jin B, Tao Y, Ren H, Atiquil Islam AYM, Bao L. Exploring the role of human-AI collaboration in solving scientific problems. Phys Rev Phys Educ Res 2025 May;21(1):010149. [doi: 10.1103/PhysRevPhysEducRes.21.010149]
177. Yu S, Lee SS, Hwang H. The ethics of using artificial intelligence in medical research. KMJ 2024 Dec;39(4):229-237. [doi: 10.7180/kmj.24.140]
178. Web-Based Medical Teaching Using a Multi-Agent System Applications and Innovations in Intelligent Systems XIII: Springer London:181-194. [doi: 10.1007/1-84628-224-1_14]
179. Wei H, Qiu J, Yu H, Yuan W. MEDCO: medical education copilots based on a multi-agent framework. arXiv. Preprint posted online on Aug 22, 2024. [doi: 10.48550/ARXIV.2408.12496]
180. Liu F, Zhou H, Gu B, et al. Application of large language models in medicine. Nat Rev Bioeng 2025;3(6):445-464 [FREE Full text] [doi: 10.1038/s44222-025-00279-5]
181. Zhang K, Meng X, Yan X, et al. Revolutionizing health care: the transformative impact of large language models in medicine. J Med Internet Res 2025 Jan 7;27:e59069. [doi: 10.2196/59069] [Medline: 39773666]
182. Hasanzadeh F, Josephson CB, Waters G, Adedinsewo D, Azizi Z, White JA. Bias recognition and mitigation strategies in artificial intelligence healthcare applications. NPJ Digit Med 2025 Mar 11;8(1):154. [doi: 10.1038/s41746-025-01503-7] [Medline: 40069303]
183. Li H, Li C, Wang J, et al. Review on security of federated learning and its application in healthcare. Future Generation Computer Systems 2023 Jul;144:271-290. [doi: 10.1016/j.future.2023.02.021]
184. Hu F, Qiu S, Yang X, Wu C, Nunes MB, Chen H. Privacy-preserving healthcare and medical data collaboration service system based on blockchain and federated learning. CMC 2024;80(2):2897-2915. [doi: 10.32604/cmc.2024.052570]
185. Ozer M. The Matthew Effect in Turkish Education System. BUJFED 2024 Nov 13. [doi: 10.14686/buefad.1359312]
186. Lucchi N. ChatGPT: a case study on copyright challenges for generative artificial intelligence systems. Eur j risk regul 2024 Sep;15(3):602-624. [doi: 10.1017/err.2023.59]
187. Mitra A, Mawson A. Neglected tropical diseases: epidemiology and global burden. TropicalMed 2017 Aug 5;2(3):36. [doi: 10.3390/tropicalmed2030036]
188. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019 Dec 3;5(2):e16048. [doi: 10.2196/16048] [Medline: 31793895]
189. Talib ZM, Kiguli-Malwadde E, Wohltjen H, et al. Transforming health professions' education through in-country collaboration: examining the consortia among African medical schools catalyzed by the Medical Education Partnership Initiative. Hum Resour Health 2015 Dec;13(1). [doi: 10.1186/1478-4491-13-1]
190. Ueda D, Kakinuma T, Fujita S, et al. Fairness of artificial intelligence in healthcare: review and recommendations. Jpn J Radiol 2024 Jan;42(1):3-15. [doi: 10.1007/s11604-023-01474-3]
191. Bui TTU, Tong TVA. The impact of AI writing tools on academic integrity: unveiling English-majored students' perceptions and practical solutions. acoj 2025 Jan 27;16(1):83-110 [FREE Full text] [doi: 10.54855/acoj.251615]
192. Yoo JH. Defining the boundaries of AI use in scientific writing: a comparative review of editorial policies. J Korean Med Sci 2025 Jun 16;40(23):e187. [doi: 10.3346/jkms.2025.40.e187] [Medline: 40524628]
193. Schwartzstein RM. Clinical reasoning and artificial intelligence: Can AI really think. Trans Am Clin Climatol Assoc 2024;134:133-145. [Medline: 39135584]
194. Kim Y, Jeong H, Chen S, et al. Medical hallucinations in foundation models and their impact on healthcare. arXiv. Preprint posted online on Feb 26, 2025. [doi: 10.48550/arXiv.2503.05777]
195. Alkhanbouli R, Matar Abdulla Almadhaani H, Alhosani F, Simsekler MCE. The role of explainable artificial intelligence in disease prediction: a systematic literature review and future research directions. BMC Med Inform Decis Mak 2025;25(1). [doi: 10.1186/s12911-025-02944-6]
196. Cohen IG, Babic B, Gerke S, Xia Q, Evgeniou T, Wertenbroch K. How AI can learn from the law: putting humans in the loop only on appeal. npj Digit Med 2023 Aug 25;6(1). [doi: 10.1038/s41746-023-00906-8]

Abbreviations

AI: artificial intelligence
GAI: generative artificial intelligence
HDI: Human Development Index

LLM: large language model
MCQ: multiple-choice question
RMA: resource-method-assessment
SAQ: short-answer question

Edited by B Lesselroth; submitted 10.01.25; peer-reviewed by B Meskó, C Wang, CN Hang, L Zhu, R Yin; revised version received 26.07.25; accepted 23.09.25; published 23.10.25.

Please cite as:
Lin Y, Luo Z, Ye Z, Zhong N, Zhao L, Zhang L, Li X, Chen Z, Chen Y
Applications, Challenges, and Prospects of Generative Artificial Intelligence Empowering Medical Education: Scoping Review
JMIR Med Educ 2025;11:e71125
URL: https://mededu.jmir.org/2025/1/e71125
doi:10.2196/71125

How Learning Styles Characterize Medical Students, Surgical Residents, Medical Staff, and General Surgery Teachers While Learning Surgery: Scoping Review

Gabriela Gouvea Silva[1], MSc; Marco Antonio Ribeiro Filho[1], MD; Carlos Dario da Silva Costa[1], MSc; Stela Regina Pedroso Vilela Torres de Carvalho[1], JD; Joao Daniel de Souza Menezes[1], MSc; Matheus Querino da Silva[1], MSc; William Donega Martinez[1], MSc; Bruno Cardoso Goncalves[1], MSc; Natália Almeida de Arnaldo Silva Rodriguez Castro[1], MSc; Luiz Vianney Cidrão Nunes[1], MD; Emerson Roberto Santos[1], MSc; Helena Landim Gonçalves Cristóvão[1], PhD; Alexandre Lins Werneck[1], PhD; Alex Bertolazzo Quitério[1], MD; Sonia Maria Maciel Lopes[1], MEd; Denise Vaz-Oliani[1,2], PhD; Fernando Facio[1], PhD; Patrícia da Silva Fucuta[3], PhD; Alba Regina de Abreu Lima[1], PhD; Vania M S Brienze[1], PhD; Heloisa Cristina Caldas[1], PhD; Julio Cesar Andre[1], PhD

[1]Center for Studies and Development of Health, Faculdade de Medicina de São José do Rio Preto, Avenida Brigadeiro Faria Lima, 5416, São José do Rio Preto, Brazil

[2]Reproductive Medicine Center of University Hospital Center Cova da Beira, University of Beira Interior, Covilhã, Portugal

[3]Epidemiology Department, Faculdades Integradas de Ciências, Educação, Cultura e Administração de Ceres (FACERES), São José do Rio Preto, Brazil

Corresponding Author:
Gabriela Gouvea Silva, MSc
Center for Studies and Development of Health, Faculdade de Medicina de São José do Rio Preto, Avenida Brigadeiro Faria Lima, 5416, São José do Rio Preto, Brazil

Abstract

Background: Learning style is a biologically and developmentally imposed configuration of personal characteristics that makes the same teaching method effective for some and ineffective for others. Studies support a relationship between learning style and career choice, resulting in learning style patterns observed in distinct types of residency programs, which can also be applied to general surgery, from medical school to the latest stages of training. The methodologies, populations, and contexts of the few studies pertinent to the matter are very different from one another, and a scoping review on this theme will unequivocally enhance and organize what is already known.

Objective: The goal of this study is to identify and map out data from studies that report on learning styles in medical students, surgical residents, medical staff, and surgical teachers.

Methods: The search strategy was performed on September 25, 2023, by a librarian and digital search strategy expert, through the descriptors “learning, style” and “surgery.” The databases consulted were Embase, SCOPUS, Web of Science, and PubMed through descriptors and their synonyms, according to MeSH (Medical Subject Headings). Of the 213 articles found, 135 articles remained after the exclusion of duplicates. The remaining 78 articles were analyzed by 3 of the researchers independently. A total of 27 articles were selected, and 2 articles were excluded because the full article was not found.

Results: A total of 25 articles were included in the review. A total of 96% (n=24) of the articles used cognitive theories as their theoretical basis. Regarding learning style instruments, 36% (n=9) articles used the visual, aural, read, and kinesthetic learning method instrument, and 40% (n=10) articles chose Kolb’s learning style inventory. The papers concentrate especially on the 2010s, and most of them are from North America (16/25, 64%) or Europe (6/25, 24%). The smallest study had 15 participants and the biggest had 1549 participants. The included studies primarily focused on surgical residents (21/25, 84%), with fewer targeting faculty and staff (9/25, 36%). The primary objectives of the studies were to investigate the relationship between learning styles and performance (15/25, 60%), gender differences (7/25, 28%), changes over time (4/25, 16%), and motivation (3/25, 12%).

Conclusions: This scoping review reveals a limited and geographically concentrated body of research on learning styles in surgery education, primarily focusing on surgical residents and using Kolb’s learning style inventory and visual, aural, read, and kinesthetic learning method instruments. Considerable gaps exist regarding geographical diversity and the study of medical staff and faculty. These findings underscore the need for future research with a broader scope to better inform educational strategies in surgery.

Trial Registration: OSF Registries 75ku4; https://osf.io/75ku4
International Registered Report Identifier (IRRID): RR2-10.2196/57229

(JMIR Med Educ 2025;11:e66766) doi:10.2196/66766

KEYWORDS

learning styles; general surgery; medical saff; surgical residents; students; medical faculty; medical school; cognitive theories; scoping review; PRISMA

Introduction

The concept of learning styles was first developed as a result of the interest in individual differences through the learning process, at the beginning of the 1960s [1]. According to Dunn, everyone has a peculiar learning style, like a signature. In this regard, tailoring teaching to different learning styles may help and improve results in education.

Several theories and inventories were created to assess and determine one's learning style, including the following: Kolb's model, Felder and Silverman's model, and Gregorc's model, especially in health care education. A debate about whether learning styles are fixed or flexible exists to this day, and to what extent the context can influence and determine them [2].

However, it is important to acknowledge the growing body of evidence that challenges the effectiveness and validity of learning styles theories. While these theories suggest that tailoring instruction to individual learning preferences can enhance learning outcomes, empirical studies have shown limited support for this approach [3]. Specifically, research indicates that many students do not study in ways that align with their self-reported learning styles, and that matching instruction to these styles does not necessarily improve performance [4].

For example, David Kolb describes learning as a process where knowledge is transformed through experience, and that acknowledgment is the combination of appropriation and transformation of experience. His theory is a holistic model called "experiential learning" and emphasizes experience in its core, differing from other theories [5]. Kolb's scheme hypothesizes that the learner has a concrete experience, upon which he reflects. Through reflection, it is possible to formulate abstract concepts and make appropriate generalizations, then consolidate the understanding by testing the implications of the knowledge in new situations. This then provides a concrete experience, and the cycle continues. Learners with different learning preferences will have different strengths and weaknesses in the quadrants of the (Kolb) cycle [6]. Based on that, he created the learning style inventory (LSI) to determine and assess individually the different learning styles, divided into converging, diverging, assimilating, and accommodating [7].

In medical education, it is particularly important to remember the heterogeneity of students. Some programs count on learners who have already completed a university degree; in others, the students come straight from secondary school, and many face a mixture of both. The broader concept of medical education includes postgraduate students and continuing professional development, too. Each of them will have variable individual constraints, experiences, and preferences [8].

The diversity in educational backgrounds, cultural influences, ethnic origins, and gender identities among contemporary surgical trainees presents unique challenges and opportunities for educators. While it is often assumed that these diverse backgrounds necessitate a personalized approach to learning, it is crucial to recognize that individual learning preferences may not always align with optimal learning strategies [9].

Instead, effective surgical education should focus on evidence-based instructional methods that cater to a wide range of learners, while also promoting critical thinking, adaptability, and lifelong learning skills.

Perry [10] noted that students change their learning approach as they progress through their college years. Students often begin with a "duality" approach, with a clear view between right and wrong, towards "multiplicity," where they recognize that context is important, and that there are various valuable sources of knowledge and experience.

Knowledge is the main domain of medical education, but the outcome depends strongly on other domains such as attitude, lifelong learning, empathy, communication, ethics, and professionalism. The clinical environment is challenging for both the student and the teacher, without mentioning the patient, who is at the center of the action. In this bigger context, it is vital to use different learning theories to promote effective learning [8].

Contemporary surgical trainees come from diverse educational, cultural, ethnic, and gender backgrounds [11], and are pressured to develop skills not only in the role as a medical expert, but also as a professional, scholar, health advocate, manager, collaborator, and communicator [12].

Educating surgeons is an ancient tradition that has existed since the development of surgery [13], and for centuries, surgical residency curricula have been guided primarily by tradition. The apprenticeship model has been one of the essential components of surgical training. It generally involves 3 steps: assisting at operations, performing operations with expert assistance, and operating without assistance.

Modern surgical education has been revolutionized by exponents such as Halsted: the historical model of apprenticeship was transformed into the current organized system that we call Residency [11].

The present day, however, requires the realization of more complex procedures, performed more regularly and in safer manners, demanding even more prepared professionals [14].

Studies report on a relationship between learning style and medical career choice, resulting in learning style patterns observed in distinct types of residency programs, which can also be applied to general surgery. Based on Kolb's LSI, students classified as accommodating and diverging frequently chose surgery as their career choice, whereas converging chose internal medicine, and assimilating chose academic medicine [15].

However, despite the theme's relevance, a preliminary search in the MEDLINE, Cochrane Database of Systematic Reviews, and Joanna Briggs Institute's evidence synthesis revealed no scoping review. Moreover, the methodologies, populations, and contexts of the few studies pertinent to the matter are very different from one another, and a scoping review on this theme would unequivocally enhance and organize what is already known.

Therefore, the aim of this scoping review is to identify, map, and synthesize the existing literature on learning styles in medical students, surgical residents, medical staff, and surgical teachers within the field of surgery. This review seeks to provide a comprehensive overview of the current state of knowledge, identify gaps in the literature, and inform future research and educational practices in surgical training.

Methods

Overview

The scoping review proposed here was carried out according to Arksey and O'Malley's [16] structure, using the first five stages: (1) identify the research question; (2) identify relevant studies; (3) select studies; (4) map out the data; and (5) collate, summarize, and report the results. Since this is preliminary research, it is likely that more studies on the theme will be finalized. Although the sixth stage of Arksey and O'Malley's [16] structure (consulting) will not be completed in this review, its results can inform this stage in a future study. The structure is congruent with The Joanna Briggs Institute's scoping review methodology. The protocol for this scoping review was duly developed in accordance with relevant reporting guidelines and has been publicly registered [17].

Inclusion Criteria

After a discussion involving the researchers, the eligibility criteria were defined.

Participants

Studies were made with medical students, surgical residents, medical staff in general surgery, and the general surgery's medical faculty. It was not obligatory to contain all the population's extracts. Studies were included if they involved at least one of these population groups; it was not required for a study to include all listed groups.

Concept

The included studies approached "learning styles" of the target population, regardless of the chosen instrument to define them.

Context

The eligible studies were those related to teaching surgery to the population in question, in any country.

Types of Sources

The following types of sources were included in the review: studies with qualitative and quantitative approaches, primary studies, systematic reviews, meta-analyses and meta-synthesis, books, and guidelines, published in indexed sources.

Research Strategy

The search strategy was performed on September 25, 2023, by a librarian and digital search strategy expert, through the descriptors "learning, style" and "surgery." There was no time frame restriction in the search. For the combination of descriptors, the Boolean operators "AND" and "OR" were considered. Words were reduced to their root to embrace variations in writing and broaden the search scope.

Databases consulted were Embase, SCOPUS, Web of Science, and PubMed through descriptors and their synonyms—according to the Medical Subject Headings—to every strategy item. These databases were selected because they are comprehensive and have a broad coverage of health publications. The databases' search resulted in a table made using a Microsoft Excel spreadsheet. In accordance with PRISMA-P (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Protocols) guidelines, the detailed search strategy used for one database (eg, PubMed) is provided in Multimedia Appendices 1 and 2 including specific search terms, Boolean operators, and any applied limits, to ensure reproducibility.

Data Selection and Extraction

The articles were evaluated by 3 independent researchers after discussions about inclusion and exclusion criteria. To align the eligibility criteria among the researchers, the titles and abstracts of 25 random articles were analyzed by 3 of the researchers. Disagreements regarding the inclusion or exclusion of the articles were discussed until a consensus was reached. There was a 100% agreement concerning the inclusion and exclusion of the articles. After the articles' selection, a form was used to extract data from the articles' full analysis, both quantitatively and qualitatively.

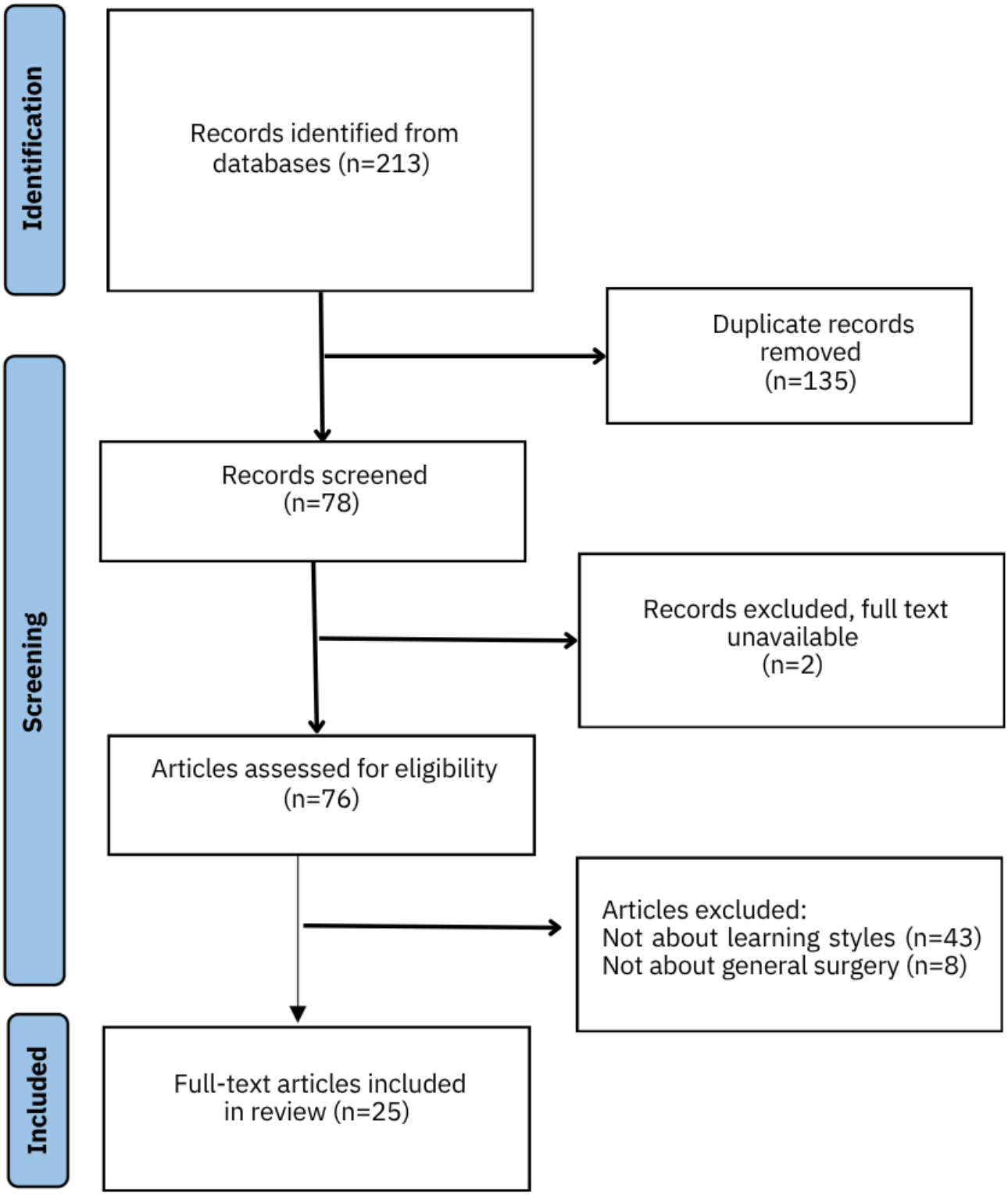
After the articles' selection, a form was used to extract data from the articles' full analysis, both quantitatively and qualitatively.

Results

Overview

Of the 213 articles found, 135 remained after the exclusion of duplicates. The remaining 78 articles were analyzed by 3 of the researchers independently, from which 27 articles were selected, and 2 articles were excluded because the full article was not found. A PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews; Checklist 1) flow diagram was produced (Figure 1). Thus, the analysis included a total of 25 articles.

Figure 1. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) flow diagram of learning styles in general surgery.



Learning Theory and Learning Style Instrument

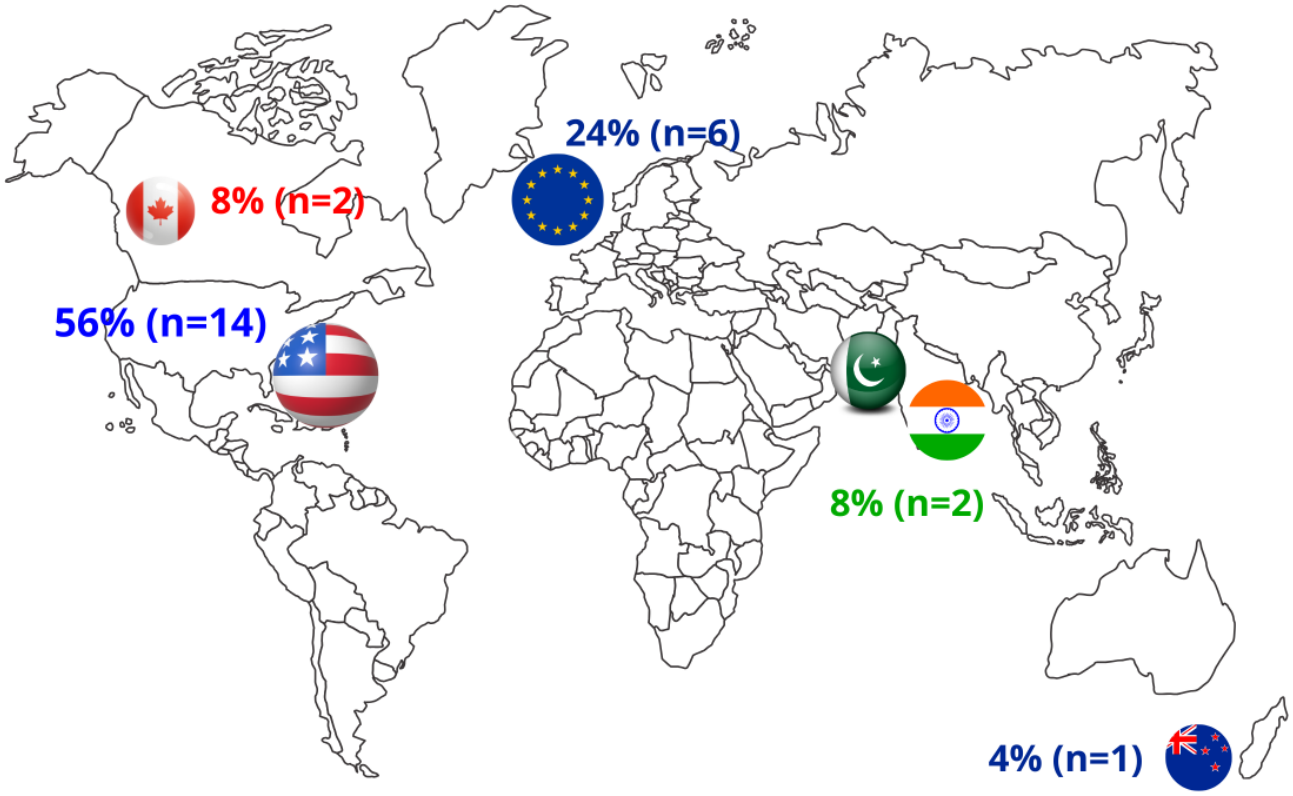
Regarding the theoretical basis, 96% (n=24) of the included articles used cognitive theories as a theoretical basis. Regarding the choice of learning style instrument: 36% (n=9) used the visual, aural, read, and kinesthetic learning method (VARK) instrument, and 40% (n=10) chose Kolb's LSI. One study opted for the Multiple Intelligences Developmental Assessment Scales [18] instrument; another used the Meyer-Briggs type indicator [19]. One study used the students' learning preferences questionnaire [20], and 2 others chose the 24 self-reported LSI questionnaire [21,22]. One study was conducted as a review, so it explained various learning styles approaches [23].

Mapping Studies Around the World

A majority of the studies were conducted in North America or Europe. Specifically, 14 studies were developed in the United States and 2 studies were from Canada, accounting for 64%

(n=16) of the studies with a northern American population. A total of 24% (n=6) of the studies took place in Europe, 4 studies in the United Kingdom, one study in Spain [24], and one study in the Netherlands [23]. Only one study was conducted in Oceania (New Zealand) [18], and the other 2 studies were from Asia (2/25, 8%), one from Pakistan [20] and the other from India [21]. No studies (0/25, 0%) were found regarding learning styles and surgery in Latin America or Africa. Figure 2 summarizes these findings.

Figure 2. Map of previous studies conducted around the globe regarding learning styles and surgery.



Interest in Learning Styles Across Time

Learning styles were not a common subject until the 1960s. Table 1 shows the number of articles on the theme in each decade. The year 2010 was the most interesting in the subject, with 15 published papers.

Table . Number of articles about learning styles in surgery through the decades.

Year	Articles, n
1970	1
2000	4
2010	15
2020	4

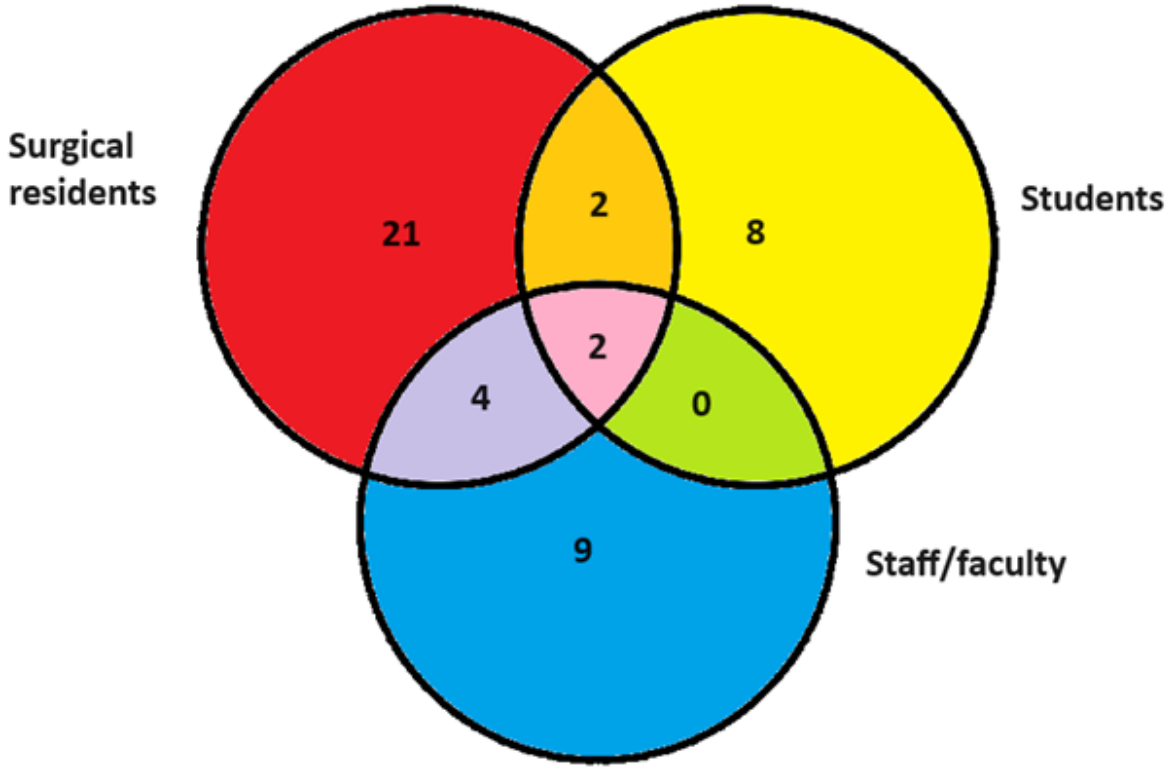
Characteristics

Most articles had a transversal design, accounting for 88% (n=22) of the total. One study (1/25, 4%) was a review, and 3 others (3/25, 12%) were cohorts. The number of participants was, on average, 114. The smallest study had 15 participants [25] and the biggest, 1549 participants [19].

The articles included students, surgical residents, and surgical staff/faculty. Figure 3 shows the number of articles with each of these populations. Specifically, 84% (n=21) of the papers included general surgery residents, while 36% (n=9) targeted faculty and staff.

The goals of these studies differed a bit. A total of 60% (n=15) of participants were interested in the relation between learning styles and performance, such as laparoscopic or robotic skills, academic achievement, and so on. Another 28% (n=7) of participants also looked over gender gaps and differences among female and male participants. Four studies (4/25, 16%) took an interest in changes in learning styles over time. Three studies (3/25, 12%) also evaluated motivation.

Figure 3. Number of articles with each population.



Discussion

Principal Findings

This scoping review identified 25 studies published between the 1960s and 2023 that explored learning styles in the context of surgery education. The majority of these studies were published in the 2010s and were predominantly conducted in North America and Europe. The most frequently used instruments to assess learning styles were Kolb's LSI and the VARK questionnaire. The included studies primarily focused on surgical residents, investigating the relationship between learning styles and various outcomes such as performance, gender differences, changes over time, and motivation.

There are several ways to face adult education. For example, there are multiple theories that can be superposed along with multiple ways of evaluating them according to the chosen theory. In health care, there are several specificities that make the educational process more complex, since there are a lot of abilities and competencies to be developed in a short period of time. On top of it, medical science evolution becomes faster every day, adding challenges to education and the preparation of professionals.

Surgery is commonly known as a delicate place for developing skills and training in high-risk procedures. When teaching surgery, it is important to account for medical students, residents, and continuous improvement for teachers and medical staff. These pieces of surgical educational chess are very different from each other, each one with special needs and goals. To understand how these population extracts capture and keep knowledge is vital to effective learning [26].

As highlighted in the results, Kolb's LSI and the VARK instrument were the most commonly used tools in the reviewed literature. Kolb's LSI, based on experiential learning theory, categorizes learners into 4 styles (diverging, assimilating, converging, and accommodating) based on their preferences for concrete experience, reflective observation, abstract conceptualization, and active experimentation [4,5]. The VARK questionnaire, on the other hand, focuses on sensory preferences for information intake and output, classifying learners as visual, auditory, read/write, or kinesthetic [16]. While both instruments aim to identify individual learning preferences, they stem from different theoretical frameworks and assess distinct aspects of the learning process. Kolb's model is more focused on how individuals process information through a cycle of experience and reflection, whereas VARK is centered on the modality through which information is best received and conveyed. The prevalence of these instruments in the literature may be attributed to their relative ease of administration and interpretation. However, it is important to note that the psychometric properties and theoretical underpinnings of some learning style instruments, including Kolb's LSI, have faced criticism in the broader educational psychology literature, raising questions about their validity and the implications of findings based solely on these tools [27-29]. Future research should consider these limitations and potentially explore alternative or complementary approaches to understanding individual differences in surgical learning.

The study field of learning styles and surgery is wide and still poorly explored. It is only possible to imagine the benefits that would emerge with the amplification of this understanding in teaching and learning surgery in the next decades.

Principal Results

Historically, there are not many studies on learning style in surgery, as pointed out by this scoping review. The papers concentrate especially on the year 2010, and most of them are in North America or Europe. Thanks to important cultural, financial, and political contexts, it is evident that more information on the participants from the rest of the world is needed.

The main instruments used in the papers were Kolb's LSI and the VARK instrument, both derived from cognitive theories. The ease of handling the instruments was probably an impact factor in their choice, on top of their proven effectiveness and reproducibility.

A total of 84% (n=21) of papers had general surgery residents as the target, while only 36% (n=9) targeted faculty and staff. The improvement of teaching goes by deep knowledge of teachers and their methods [30]. So, prospective surgeons should not be the only goal of surgical educational research. The ratio is to disarticulate the former master-learner vision, understanding that everyone involved in the learning process is a potential and continuous learner [16,24].

The main purpose of the papers was to relate learning styles and performance, motivation, gender, and changes over time.

The articles aiming for performance did it in a lot of diverse ways. One way to assess performance is to evaluate the scoring in admission tests, and research has concluded, many times, that a nonspecific learning style is being favored in those types of admissions [25-29]. Others took interest in psychomotor skills, measured by specific metrics [14,19,20,30]. There were also the ones that focused on academic achievement throughout college [31] or residency program [11,21], by overlooking the number of procedures [32], for example.

In surgery, gender is a substantial subject because the surgical environment is, in general, very masculine. There is still today a discrepancy between women in the early career (residents) and late career (teachers) [33-38], and differences in payment, rates of sexual harassment, rates of sexual abuse, rates of psychological abuse, and so on. The proportion of women in surgery, although it has been increasing in the past few decades [17,20,28,31,32,39]

Motivation is key to good learning, and many learning theories rely on it [40,41]. Without motivation, the learner lacks stimuli and fatally decreases their learning. There is no robust evidence in the literature to support the relationship between learning styles and degree of motivation [16,42-44].

Limitations

The main limitation of this scoping review is the fact that only English articles were searched, which diminishes the cultural range due to the language barrier, especially if we consider that surgical residency has very different types of models and resources around the world.

Comparison With Prior Work

No scoping review on learning styles in surgery was ever done.

Conclusions

This scoping review identified a limited number of studies exploring learning styles in surgery education, with the majority concentrated in the 2010s and predominantly conducted in North America and Europe. The review found that Kolb's LSI and the VARK instrument were the most frequently used tools, and that research has primarily focused on surgical residents, investigating relationships between learning styles and performance, gender, changes over time, and motivation. The scarcity of studies from other regions, particularly Latin America and Africa, and the limited focus on medical staff and faculty highlight considerable gaps in the current literature. Based on these findings, understanding the characteristics of the existing research landscape is key to informing future studies. Further research is needed to broaden the geographical scope, explore diverse populations within surgical education, and investigate the practical implications of learning style research for enhancing teaching and learning strategies in surgery.

Acknowledgments

The authors wish to thank Luiz Vianney Saldanha Cidrão Nunes, who, even through hard times, was ready to collaborate deeply in the success of the search strategy.

Authors' Contributions

GGS and JCA contributed to the conceptualization of the study. Data curation was performed by GGS, MARF, CDSC, and BCG. FSF was responsible for formal analysis. GGS, LVCN, and JCA conducted the investigation and developed the methodology. Project administration was carried out by DVO, FF, PSF, ARAL, VMSB, HCC, and JCA. Validation and visualization were undertaken by SRPVTC, JDSM, MQS, WDM, NAASRC, ERS, HLGC, ALL, ABQ, and SMML. The manuscript was written, reviewed, and edited by GGS and JCA.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed search strategy.

[DOCX File, 22 KB - mededu_v11i1e66766_app1.docx]

Multimedia Appendix 2
Search strategy used for one database (PubMed).
[DOCX File, 15 KB - mededu_v11i1e66766_app2.docx]

Checklist 1
PRISMA-ScR checklist.
[DOCX File, 87 KB - mededu_v11i1e66766_app3.docx]

References

1. Curry L. An organization of learning styles theory and constructs. 1983 Presented at: Annual Meeting of the American Educational Research Association; Apr 11-15, 1983; Montreal, Quebec.
2. Coffield F, Moseley D, Hall E, Ecclestone K. Learning styles and pedagogy in post 16 education: a critical and systematic review. : Learning and Skills Research Council; 2004. [Medline: 21197874]
3. Pashler H, Mcdaniel M, Rohrer D, Bjork R. Learning styles concepts and evidence. Psychol Sci Public Interest 2009;9(3):105-119 [FREE Full text] [doi: 10.1111/j.1539-6053.2009.01038.x]
4. Papadatou-Pastou M, Gritzali M, Barrable A. The learning styles educational neuromyth: lack of agreement between teachers' judgments, self-assessment, and students' intelligence. Front Educ 2018 Nov 29;3. [doi: 10.3389/feduc.2018.00105]
5. Sternberg RJ, Zhang L. Perspectives on Thinking, Learning, and Cognitive Styles: Taylor & Francis; 2001. URL: https://books.google.com.br/books?id=72xHSz5Z9CkC [accessed 2025-08-15]
6. Kolb D. Experiential Learning: Experience as the Source of Learning and Development: Prentice-Hall; 1984. URL: https://books.google.com.br/books?id=ufnuAAAAMAAJ [accessed 2025-08-15]
7. Kolb DA. Individual Learning Styles and the Learning Process: MIT; 1971.
8. Taylor DCM, Hamdy H. Adult learning theories: implications for learning and teaching in medical education: AMEE Guide No. 83. Med Teach 2013 Nov;35(11):e1561-e1572. [doi: 10.3109/0142159X.2013.828153] [Medline: 24004029]
9. Kirschner PA. Stop propagating the learning styles myth. Comput Educ 2017 Mar;106:166-171. [doi: 10.1016/j.compedu.2016.12.006]
10. Perry WG. Forms of Intellectual and Ethical Development in the College Years: A Scheme: Jossey-Bass Publishers; 1999. URL: https://books.google.com.br/books?id=EXLptwEACAAJ [accessed 2025-08-15]
11. Reznick RK, MacRae H. Teaching surgical skills — changes in the wind. N Engl J Med 2006 Dec 21;355(25):2664-2669. [doi: 10.1056/NEJMra054785]
12. Frank JR, Jabbour M, Fréchette D, Marks M, Valk N, Bourgeois G. The CanMEDS 2005 Physician Competency Framework Better Standards Better Physicians Better Care Framework: The Royal College of Physicians and Surgeons of Canada; 2005. URL: http://meds.queensu.ca/medicine/obgyn/pdf/CanMEDS2005.booklet.pdf [accessed 2025-08-15]
13. Jones WHS. The Hippocratic oath - Ludwig Edelstein: the Hippocratic oath. text, translation, and interpretation. pp. vii 64. Baltimore: Johns Hopkins Press, 1943. paper, $1.25. Classical Rev Cambridge University Press. Classical Rev 1945;59:14-15. [doi: 10.1017/S0009840X00087515]
14. Contessa J, Ciardiello KA, Perlman S. Surgery resident learning styles and academic achievement. Curr Surg 2005;62(3):344-347. [doi: 10.1016/j.cursur.2004.09.012] [Medline: 15890222]
15. Plovnick MS. Primary care career choices and medical student learning styles. J Med Educ 1975 Sep;50(9):849-855. [doi: 10.1097/00001888-197509000-00002] [Medline: 1171243]
16. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
17. Gouvea Silva G, Costa CDS, Gonçalves BC, et al. Learning styles of medical students, surgical residents, medical staff, and general surgery teachers when learning surgery: protocol for a scoping review. JMIR Res Protoc 2024;13:e57229. [doi: 10.2196/57229]
18. Windsor JA, Diener S, Zoha F. Learning style and laparoscopic experience in psychomotor skill performance using a virtual reality surgical simulator. Am J Surg 2008 Jun;195(6):837-842. [doi: 10.1016/j.amjsurg.2007.09.034] [Medline: 18417084]
19. Bell MA, Wales PS, Torbeck LJ, Kunzer JM, Thurston VC, Brokaw JJ. Do personality differences between teachers and learners impact students' evaluations of a surgery clerkship? J Surg Educ 2011 May;68(3):190-193. [doi: 10.1016/j.jsurg.2011.01.003]
20. Ahmad HN, Asif M. Medical student's learning habits: a mixed method study during clinical rotation in general surgery. J Pak Med Assoc 2018 Apr;68(4):600-605. [Medline: 29808051]
21. Bansal R, Mathew KA, Jith A, Narayanan D. A comparison of personality traits, learning style, and perceived stress among surgical and nonsurgical residents in a tertiary care hospital in India. Ind Psychiatry J 2021;30(2):329-334. [doi: 10.4103/ipj.ipj_93_21]

22. Preece RA, Cope AC. Are surgeons born or made? a comparison of personality traits and learning styles between surgical trainees and medical students. J Surg Educ 2016 Sep;73(5):768-773. [doi: 10.1016/j.jsurg.2016.03.017]
23. Dankelman J, Chmarra MK, Verdaasdonk EGG, Stassen LPS, Grimbergen CA. Fundamental aspects of learning minimally invasive surgical skills. Minim Invasive Ther Allied Technol 2005;14(4):247-256. [doi: 10.1080/13645700500272413] [Medline: 16754171]
24. Martín Parra JI, Toledo Martínez E, Martínez Pérez P, et al. Análisis de los estilos de aprendizaje en un curso de habilidades técnicas laparoscópicas. Implicaciones para el entrenamiento quirúrgico [analysis of learning styles in a laparoscopic technical skills course. Implications for surgical training]. Cirugía Española [Cir Esp] 2021 Dec;99(10):730-736. [doi: 10.1016/j.ciresp.2020.11.006]
25. Pang JHY, Goetz A, Hook L, Joshi ART, Leung PS. Self-awareness of learning styles among surgical trainees. J Am Coll Surg 2015;221(4):S56. [doi: 10.1016/j.jamcollsurg.2015.07.120]
26. Villet R. Teaching surgery in 2020. J Visc Surg 2020 Jun;157(3):S71-S72. [doi: 10.1016/j.jviscsurg.2020.03.003] [Medline: 32284242]
27. Metallidou P, Platsidou M. Kolb's learning style inventory-1985: validity issues and relations with metacognitive knowledge about problem-solving strategies. Learn Individ Differ 2008 Jan;18(1):114-119. [doi: 10.1016/j.lindif.2007.11.001]
28. Koob JJ, Funk J. Kolb's learning style inventory: issues of reliability and validity. Res Soc Work Pract 2002 Mar;12(2):293-308. [doi: 10.1177/104973150201200206]
29. Manolis C, Burns DJ, Assudani R, Chinta R. Assessing experiential learning styles: a methodological reconstruction and validation of the Kolb learning style inventory. Learn Individ Differ 2013 Feb;23:44-52. [doi: 10.1016/j.lindif.2012.10.009]
30. Castillo-Angeles M, Calvillo-Ortiz R, Barrows C, Chaikof EL, Kent TS. The learning environment in surgery clerkship: what are faculty perceptions? J Surg Educ 2020 Jan;77(1):61-68. [doi: 10.1016/j.jsurg.2019.07.003] [Medline: 31375466]
31. Linn BS, Cohen J, Wirch J, Pratt T, Zeppa R. The relationship of interest in surgery to learning styles, grades and residency choice. Soci Sci Med Part A: Med Psychol Med Sociol 1979 Jan;13(C):597-600. [doi: 10.1016/0271-7123(79)90102-0] [Medline: 538473]
32. Quillin RC, Pritts TA, Hanseman DJ, Edwards MJ, Davis BR. How residents learn predicts success in surgical residency. J Surg Educ 2013 Nov;70(6):725-730. [doi: 10.1016/j.jsurg.2013.09.016]
33. Kim RH, Gilbert T. Learning style preferences of surgical residency applicants. J Surg Res 2015 Sep;198(1):61-65. [doi: 10.1016/j.jss.2015.05.021] [Medline: 26070495]
34. Kim RH, Gilbert T, Ristig K. The effect of surgical resident learning style preferences on American board of surgery in-training examination scores. J Surg Educ 2015 Jul;72(4):726-731. [doi: 10.1016/j.jsurg.2014.12.009] [Medline: 25648283]
35. Kim RH, Gilbert T, Ristig K, Chu QD. Surgical resident learning styles: faculty and resident accuracy at identification of preferences and impact on ABSITE scores. J Surg Res 2013 Sep;184(1):31-36. [doi: 10.1016/j.jss.2013.04.050]
36. Mammen JMV, Fischer DR, Anderson A, et al. Learning styles vary among general surgery residents: analysis of 12 years of data. J Surg Educ 2007;64(6):386-389. [doi: 10.1016/j.jsurg.2007.08.005] [Medline: 18063274]
37. Retrosi G, Morris M, McGavock J. Does personal learning style predict the ability to learn laparoscopic surgery? a pilot study. J Laparoendosc Adv Surg Tech A 2019 Jan;29(1):98-102. [doi: 10.1089/lap.2018.0196] [Medline: 30052125]
38. Ballow D, Fang J, Kosarek C, Green T, Tarry W, Tarry S. Mp22-10 impact of matching educational materials to learning style on robotic surgical skills training. J Urol 2015 Apr;193(4S):e245. [doi: 10.1016/j.juro.2015.02.1022]
39. Dickinson KJ, Bass BL, Graviss EA, Nguyen DT, Pei KY. How learning preferences and teaching styles influence effectiveness of surgical educators. Am J Surg 2021 Feb;221(2):256-260. [doi: 10.1016/j.amjsurg.2020.08.028]
40. Guariente SMM, Guariente MDM, Moraes A. Perfil sociodemográfico e educacional do estudante ingressante no curso de graduação em medicina de 2004 a 2013: analise documental [socio-demographic profile and educational student newcomer course of graduation in medicine 2004 to 2013: documentary review]. Rev méd Minas Gerais 2020;30:e-30102 [FREE Full text] [doi: 10.5935/2238-3182.20200028]
41. Stephens EH, Heisler CA, Temkin SM, Miller P. The current status of women in surgery: how to affect the future. JAMA Surg 2020 Sep 1;155(9):876-885. [doi: 10.1001/jamasurg.2020.0312] [Medline: 32639556]
42. Lim WH, Wong C, Jain SR, et al. The unspoken reality of gender bias in surgery: a qualitative systematic review. PLoS ONE 2021;16(2):e0246420. [doi: 10.1371/journal.pone.0246420]
43. Motter SB, Brandão GR, Iaroseski J, et al. Women representation in academic and leadership positions in surgery in Brazil. Am J Surg 2022 Jan;223(1):71-75. [doi: 10.1016/j.amjsurg.2021.07.023] [Medline: 34315578]
44. Ferrari L, Mari V, Parini S, et al. Discrimination toward women in surgery: a systematic scoping review. Ann Surg 2022 Jul 1;276(1):1-8 [FREE Full text] [doi: 10.1097/SLA.0000000000005435] [Medline: 35275886]

Abbreviations

LSI: learning style inventory
MeSH: Medical Subject Headings
PRISMA-P: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Protocols

PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
VARK: visual, aural, read, and kinesthetic learning method

Edited by P Kanzow; submitted 22.09.24; peer-reviewed by B Niroomand, MA Jacobs; revised version received 02.07.25; accepted 03.07.25; published 05.09.25.

Please cite as:
Gouvea Silva G, Ribeiro Filho MA, da Silva Costa CD, Pedroso Vilela Torres de Carvalho SR, de Souza Menezes JD, Querino da Silva M, Donega Martinez W, Cardoso Goncalves B, Almeida de Arnaldo Silva Rodriguez Castro N, Vianney Cidrão Nunes L, Santos ER, Landim Gonçalves Cristóvão H, Lins Werneck A, Bertolazzo Quitério A, Maciel Lopes SM, Vaz-Oliani D, Facio F, da Silva Fucuta P, de Abreu Lima AR, Brienze VMS, Caldas HC, Andre JC
How Learning Styles Characterize Medical Students, Surgical Residents, Medical Staff, and General Surgery Teachers While Learning Surgery: Scoping Review
JMIR Med Educ 2025;11:e66766
URL: https://mededu.jmir.org/2025/1/e66766
doi:10.2196/66766

Technology Acceptance Model in Medical Education: Systematic Review

Jason Wen Yau Lee[1], PhD, MSc, BIT; Jenelle Yingni Tan[2]; Fernando Bello[1,3], PhD, BSc

[1]Duke-NUS Medical School, National University of Singapore, 8 College Road, Singapore, Singapore

[2]Faculty of Arts and Social Science, National University of Singapore, Singapore, Singapore

[3]Imperial College, London, United Kingdom

Corresponding Author:
Jason Wen Yau Lee, PhD, MSc, BIT
Duke-NUS Medical School, National University of Singapore, 8 College Road, Singapore, Singapore

Abstract

Background: With the growing use of technology in medical education, a framework is needed to evaluate learners' and educators' acceptance of these technologies. In this context, the Technology Acceptance Model (TAM) offers a valuable theoretical framework, providing insights into the determinants influencing users' acceptance and adoption of technology.

Objective: This review aims to systematically synthesize the body of research in medical education that uses the TAM.

Methods: An electronic literature search was conducted using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) approach in February 2024 on the Embase, MEDLINE, PsycINFO, PubMed, and Web of Science databases, yielding 680 articles. Upon elimination of duplicates and applying the exclusion criteria, a total of 39 articles were retained. To evaluate the quality of the study, the Medical Education Research Study Quality Instrument score was calculated for each analysis with a qualitative component.

Results: Studies using TAM in medical education began in 2010, with the model's application relatively rare up to 2016. Most of the studies were quantitative, operationalizing the TAM as a survey instrument, but it was also used as a research framework in qualitative data analysis. Structural equation modeling, descriptive analysis, and correlation analysis were the most common data analysis approaches in the studies. E-learning and mobile learning were the predominant learning interventions explored, but there were indications that novel learning technologies such as augmented reality, virtual reality, and 3D printing were being investigated.

Conclusions: The study's findings reveal an expanding scholarly engagement with using TAM in medical education. Although the TAM has been mostly used as a survey instrument, it can also be adapted as a qualitative research framework to analyze data. This systematic review provides a foundation for future research to understand the factors influencing users' acceptance of technology, especially in medical education.

(JMIR Med Educ 2025;11:e67873) doi:10.2196/67873

KEYWORDS

technology acceptance model; medical education; systematic review; TAM; learners; educators; technologies; technology; theoretical framework; technology adoption; electronic literature; qualitative; survey instrument; surveys; technology acceptance; learning interventions; e-learning; mobile learning

Introduction

Technology has changed how we learn and access knowledge, particularly with the introduction of digital devices and the internet. No longer are we constrained by time or space, and information is available anytime and anywhere. Today, learning can happen through massive open online courses such as Khan Academy [1], edX [2], and Coursera [3], or simply by viewing one of countless video tutorials online. Books can be supplemented or even replaced with multimedia resources that can provide learners with a richer learning experience. The way that knowledge is accessed has changed dramatically over the past 2 decades with the development of new technologies.

Medical education has traditionally relied on time-honored teaching methodologies. Cadaveric dissection has always been considered the gold standard for anatomy instruction [4], providing students with hands-on experience with human tissues and structures. However, as educational resources face constraints and medical knowledge expands, these traditional approaches have begun to be transformed through the use of technology. Teaching modalities such as virtual [5] and augmented reality [6] provide students with an immersive 3D learning experience. Three-dimensional printing technology [7]

has enabled the creation of anatomical models on demand that can be customized for specific learning outcomes [8], and e-learning resources [9] have democratized access to high-quality learning materials. In a study on rural posting clerkships, iPads equipped with mobile health information resources have positively influenced medical students' information-seeking behavior [10]. With the increasing use of technology in medical education, it is essential to understand how it is accepted for use in learning.

The Technology Acceptance Model (TAM) provides a framework to understand the factors influencing the decision to use new technologies in medical education [10-12]. The perceived ease of use is the extent to which a person believes the system will be free of effort. In contrast, perceived usefulness is the extent to which a person believes using the system would improve their productivity or job performance. However, one shortcoming of the TAM when applied to complex medical teaching environments is that it does not consider broader contextual factors, such as organizational culture, social influence, and other affective factors like attitudes and beliefs that may significantly impact the acceptance of educational technology.

To address this issue, the Technology Acceptance Model 2 (TAM2) was proposed by Venkatesh and Davis [13] as an improvement to the original model to include social influence and cognitive processes that may influence an individual's acceptance of technology. The purpose of developing the TAM2 was to include additional crucial factors influencing perceived usefulness and usage intention constructs to explain user behavior and acceptance. These factors include subjective norms, output quality, result demonstrability, and social factors, among others, that explain user behavior and acceptance. By integrating these factors, the TAM2 offers a more comprehensive framework for analyzing how individual and social variables influence beliefs, attitudes, and intentions to use the technology in medical education.

Despite the growing use of technology in medical education, understanding the factors that influence its adoption remains challenging for educators and institutions. Previous research has identified barriers to technology implementation such as technical difficulties [14], resistance to change, and varying acceptance by faculty and students [15]. The TAM has emerged as a valuable theoretical framework for examining these adoption challenges [10,16,17], yet its application within medical education contexts remains fragmented and inconsistently synthesized [16,18]. This knowledge gap may hinder evidence-based decision-making on the use of education technology that could enhance teaching and learning outcomes in medical education.

To address this limitation, this systematic review aims to synthesize the current research on the application of the TAM in medical education to provide insights into the factors influencing technology acceptance among medical professionals and students. The guiding questions that we aim to answer with this systematic review are as follows:

1. What is the state of TAM in medical education?
2. How has TAM been operationalized?
3. What education interventions are used in such studies?

Methods

This review was designed and is reported using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [19].

Search Strategy

A systematic search was conducted in February 2024 to identify original published articles on TAM and medical education from January 2003 to December 2023. We set the search criteria to focus on the last 20 years to capture only the most recent advancements in the field. An author (JYT) then systematically searched 5 databases accessible through the university library. A search of "all fields" with the keywords "TAM" or "Technology Acceptance Model" and "Medical Education" was used for the Embase, MEDLINE, PubMed, PsycINFO, and Web of Science databases.

Inclusion and Exclusion Criteria

Peer-reviewed articles were included if they used the TAM as a survey instrument in the study methodology or as a theoretical framework in medical education. This includes using the original TAM model proposed by Davis [20] or the TAM2 model proposed by Venkatesh and Davis [13]. We define medical education–related studies as training medical professionals, residents, and students pursuing their undergraduate, clerkship, postgraduate, or continuing medical education. If the study comprised a mix of medical students and students from other health care science professions (eg, nursing, pharmacy, emergency response), they were also included as part of the review.

Studies were excluded from our research if they were not related to medical education, such as research focused solely on nursing and allied health professions like pharmacy and physiotherapy, articles not written in English, articles published before 2003, and non–peer-reviewed documents, including theses or conference abstracts lacking comprehensive methodological details. When the cohort under study comprised a mixture of health professionals, including those who met the inclusion criteria, the entire cohort was included in the research analysis.

Final Study Selection

After retrieving the search results from the identified database, JYT removed the duplicates and uploaded the articles into a shared Microsoft Teams [21] folder. The shortlisted articles were entered into an Excel spreadsheet for screening by the authors. The final screening process involved JYT noting articles for inclusion or exclusion based on the title or abstract, which was verified independently by the primary author (JWYL). JYT also assigned a reason for exclusion for each excluded article. In cases of uncertainty, the articles in question were retained and screened together by both authors (JWYL and JYT). JYT extracted the full text of the retained articles and this was verified by JWYL for consistency.

Data Extraction and Analysis

After the shortlisted studies were identified, details of the studies were entered into the spreadsheet, including (1) general study information (eg, authors, title, and publication year), (2) participant-related information, (3) sample size, (4) application of the TAM framework, (5) study design, (6) statistical analysis used, (7) education intervention investigated, and (8) study quality (Medical Education Research Study Quality Instrument [MERSQI] score). Please see Table 1 for information gathered from the shortlisted studies.

Table . Articles included in the study.

Authors	Paper title	Publication year	Country	Study participants	Sample size	TAM[a] application	Study design	Statistical analysis	Education intervention	MERSQI[b] score
Wong G et al (2010) [22]	Internet-based medical education: a realist review of what works, for whom and in what circumstances	2010	United Kingdom	N/A[c] (systematic review)	N/A	Research framework	Qualitative	N/A	E-learning	N/A
McGowan BS et al (2012) [23]	Understanding the factors that influence the adoption and meaningful use of social media by physicians to share medical information	2012	United States	Health care professionals	485	Survey instrument	Quantitative	Correlation	E-learning	9.5
Knight JF (2013) [24]	Acceptability of video games technology for medical emergency training	2013	Denmark	Health care professionals	37	Survey instrument	Quantitative	Multiple regression	Serious game	12
Fang TY et al (2014) [25]	Evaluation of a haptics-based virtual reality temporal bone simulator for anatomy and surgery training	2014	Taiwan	Medical undergraduates and health care professionals	14	Survey instrument	Quantitative	*t* test	Haptic device	9
Briz-Ponce L and Garcia-Penalvo F (2015) [12]	An empirical assessment of a Technology Acceptance Model for apps in medical education	2015	Spain	Medical undergraduates and health care professionals	124	Survey instrument	Quantitative	SEM (CB)[d]	Mobile learning	10
Ryan JR et al (2015) [26]	Ventriculostomy simulation using patient-specific ventricular anatomy, 3D printing, and hydrogel casting	2015	United States	Medical undergraduates	10	Survey instrument	Quantitative	Descriptive	3D printing	7

Authors	Paper title	Publication year	Country	Study participants	Sample size	TAM[a] application	Study design	Statistical analysis	Education intervention	MERSQI[b] score
Huang HM et al (2016) [27]	Exploring learner acceptance of the use of virtual reality in medical education: a case study of desktop and projection-based display systems	2016	Taiwan	Medical undergraduates	230	Survey instrument	Quantitative	Correlation	Virtual reality	10.5
Briz-Ponce L et al (2017) [28]	Learning with mobile technologies — students' behavior	2017	Spain	Medical undergraduates	124	Survey instrument	Quantitative	SEM (PLS)[e]	E-learning	9
Tahamtan I et al (2017) [29]	Factors affecting smartphone adoption for accessing information in medical settings	2017	Iran	Medical undergraduates	112	Survey instrument	Mixed	SEM (CB)	Mobile learning	10
Yeom S et al (2017) [30]	Factors influencing undergraduate students' acceptance of a haptic interface for learning gross anatomy	2017	Australia	General undergraduates	89	Research framework	Quantitative	Descriptive	Haptic device	10
Basoglu N et al (2018) [31]	Exploring adoption of augmented reality smart glasses: applications in the medical industry	2018	Turkey	Medical undergraduates and health care professionals	71	Survey instrument	Quantitative	SEM (PLS)	Augmented reality	9

Authors	Paper title	Publication year	Country	Study participants	Sample size	TAM[a] application	Study design	Statistical analysis	Education intervention	MERSQI[b] score
Duch Christensen M et al (2018) [32]	Learners' perceptions during simulation-based training: an interview study comparing remote versus locally facilitated simulation-based training	2018	Denmark	Health care professionals	21	Research framework	Qualitative	N/A	Simulation-based training	N/A
Barteit S et al (2019) [17]	Technology acceptance and information system success of a mobile electronic platform for nonphysician clinical students in Zambia: prospective, nonrandomized intervention study	2019	Zambia	Medical undergraduates and health care professionals	109	Survey instrument	Quantitative	Correlation	E-learning	9
Chan KS and Zary N (2019)[33]	Applications and challenges of implementing artificial intelligence in medical education: integrative review	2019	United Arab Emirates	N/A (systematic review)	N/A	Research framework	Qualitative	N/A	Artificial intelligence in medical education	N/A
Johnson EM and Howard C (2019) [34]	A library mobile device deployment to enhance the medical student experience in a rural longitudinal integrated clerkship	2019	United States	Medical undergraduates	9	Survey instrument	Mixed	Descriptive	Mobile learning	9

Authors	Paper title	Publication year	Country	Study participants	Sample size	TAM[a] application	Study design	Statistical analysis	Education intervention	MERSQI[b] score
Abdekhoda M et al (2020) [35]	A conceptual model of flipped classroom adoption in medical higher education	2020	Iran	Medical undergraduates	110	Survey instrument	Quantitative	Correlation	Teaching approach	11
Kucuk S et al (2020) [36]	A model for medical students' behavioral intention to use mobile learning	2020	Turkey	Medical undergraduates	376	Survey instrument	Quantitative	SEM (CB)	Mobile learning	10
Lee CW et al (2020) [37]	User experience evaluation of the EPAs-based e-portfolio system and an analysis of its impact	2020	Taiwan	Health care professionals	20	Research framework	Qualitative	N/A	E-learning	N/A
Jeyakumar T et al (2021) [38]	Best practices for the implementation and sustainment of virtual health information system training: qualitative study	2021	Canada	Health care educators	18	Research framework	Qualitative	N/A	E-learning	N/A
Lee SS et al (2021) [39]	Mobile learning in clinical settings: unveiling the paradox	2021	Singapore	Health care professionals	171	Research framework	Mixed	Descriptive	Mobile learning	8.5
Zalat MM et al (2021) [40]	The experiences, challenges, and acceptance of e-learning as a tool for teaching during the COVID-19 pandemic among university medical staff	2021	Egypt	Health care professionals	346	Survey instrument	Quantitative	Descriptive	E-learning	8.5

Authors	Paper title	Publication year	Country	Study participants	Sample size	TAM[a] application	Study design	Statistical analysis	Education intervention	MERSQI[b] score
Almarzouqi A et al (2022) [41]	Prediction of user's intention to use metaverse system in medical education: a hybrid SEM-ML learning approach	2022	United Arab Emirates	General undergraduate and postgraduate	1858	Survey instrument	Quantitative	SEM (PLS)	E-learning	12
Bhardwaj M et al (2022) [42]	Perceptions and experience of medical students regarding e-learning during COVID-19 lockdown-a cross-sectional study	2022	India	Medical undergraduates	340	Research framework	Quantitative	Descriptive	E-learning	9
Bianchi I et al (2022) [43]	AnemiaAR: a serious game to support teaching of haematology	2022	Brazil	Medical undergraduates	14	Survey instrument	Quantitative	*U* test	Serious game	8
Chan E et al (2022) [44]	Medical teachers' experience of emergency remote teaching during the COVID-19 pandemic: a cross-institutional study.	2022	Hong Kong	Health care educators	139	Research framework	Quantitative	Correlation	E-learning	9.75
Do DH et al (2022) [10]	Drivers of iPad use by undergraduate medical students: the Technology Acceptance Model perspective	2022	Canada	Medical undergraduates	834	Survey instrument	Quantitative	SEM (PLS)	Mobile learning	10.5
Harmon DJ et al (2022) [45]	Development and assessment of an integrated anatomy mobile app	2022	United States	Medical undergraduates	195	Survey instrument	Quantitative	SEM (CB)	Mobile learning	10

Authors	Paper title	Publication year	Country	Study participants	Sample size	TAM[a] application	Study design	Statistical analysis	Education intervention	MERSQI[b] score
Komuhangi A et al (2022) [46]	Predictors for adoption of e-learning among health professional students during the COVID-19 lockdown in a private university in Uganda	2022	Uganda	Health science undergraduates	109	Survey instrument	Quantitative	Regression	E-learning	10
Lau V and Greer M (2022) [47]	Using technology adoption theories to maximize the uptake of e-learning in medical education	2022	United States	N/A (systematic review)	N/A	Research framework	Qualitative	N/A	E-learning	N/A
Bugli D et al (2023) [48]	Training the public health emergency response workforce: a mixed-methods approach to evaluating the virtual reality modality	2023	United States	Health care professionals	100	Survey instrument	Quantitative	Correlation	Virtual reality	9
Young Y et al (2023) [49]	Improving transitions between clinical placements	2023	United Kingdom	Medical undergraduates	19	Research framework	Qualitative	N/A	Website	N/A
Sallam M et al (2023) [50]	Assessing health students' attitudes and usage of ChatGPT in Jordan: validation study	2023	Jordan	General undergraduates	458	Survey instrument	Quantitative	Correlation	Artificial intelligence in medical education	11
Cabero-Almenara J et al (2023) [51]	Degree of acceptance of virtual reality by health sciences students	2023	Spain	Health science undergraduates	136	Survey instrument	Quantitative	Regression	Virtual reality	10

Authors	Paper title	Publication year	Country	Study participants	Sample size	TAM[a] application	Study design	Statistical analysis	Education intervention	MERSQI[b] score
Ndlovu K et al (2023) [52]	Evaluating the feasibility and acceptance of a mobile clinical decision support system in a resource-limited country: exploratory study	2023	Botswana	Health care professionals	28	Survey instrument	Mixed	Descriptive	Mobile learning	7
Lin CW et al (2023) [53]	Crowdsource authoring as a tool for enhancing the quality of competency assessments in healthcare professions	2023	Taiwan	Health care educators	50	Survey instrument	Quantitative	Correlation	E-learning	11
Rahadiani P et al (2023) [54]	Use of H5P interactive learning content in a self-paced MOOC [massive open online course] for learning activity preferences and acceptance in an Indonesian medical elective module	2023	Indonesia	Health science undergraduates	126	Survey Instrument	Quantitative	Correlation	E-learning	11
De Ruyck O et al (2024) [55]	A comparison of three feedback formats in an ePortfolio to support workplace learning in healthcare education: a mixed method study	2023	Belgium	Health care professionals	85	Survey instrument	Mixed	Correlation	E-learning	7

[a]TAM: Technology Acceptance Model.

[b]MERSQI: Medical Education Research Study Quality Instrument.

[c]N/A: not applicable.

[d]SEM (CB): covariance-based structural equation modeling.

[e]SEM (PLS): partial least squares structural equation modeling.

To assess the study quality of quantitative studies, the MERSQI was used to measure the methodological quality of the selected studies [56]. The MERSQI is an instrument that measures the quality of experimental, quasi-experimental, and observational studies. The MERSQI contains 6 domains (study design, sampling, type of data, validity evidence for the evaluation instrument, data analysis, and outcomes), with a study scoring a possible total of 18. The MERSQI was not designed for use in qualitative studies. Therefore, these studies will not be assessed using the MERSQI.

Results

Overview

A systematic literature retrieval and analysis was methodically executed across 5 authoritative databases (Embase, MEDLINE, PsycINFO, PubMed, and Web of Science), yielding 580 records. The PRISMA checklist informed the review protocol and is depicted in the flow diagram in Figure 1. Using automation tools to narrow the search criteria, 30 studies were removed, and 266 duplicate records were excluded. This resulted in 384 studies that were eligible for screening. Based on the abstract or title, 329 studies that did not meet the study inclusion criteria were eliminated. This left 55 studies for full-text retrieval. Both authors read through all shortlisted papers and excluded a further 18 papers, where 1 was a duplicate, 8 did not use the TAM in the study, and 9 were unrelated to medical education. Therefore, the total number of studies included was 37.

Table 2 presents the source database, publisher, and number of studies found in the search results.

Figure 1. PRISMA flowchart. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

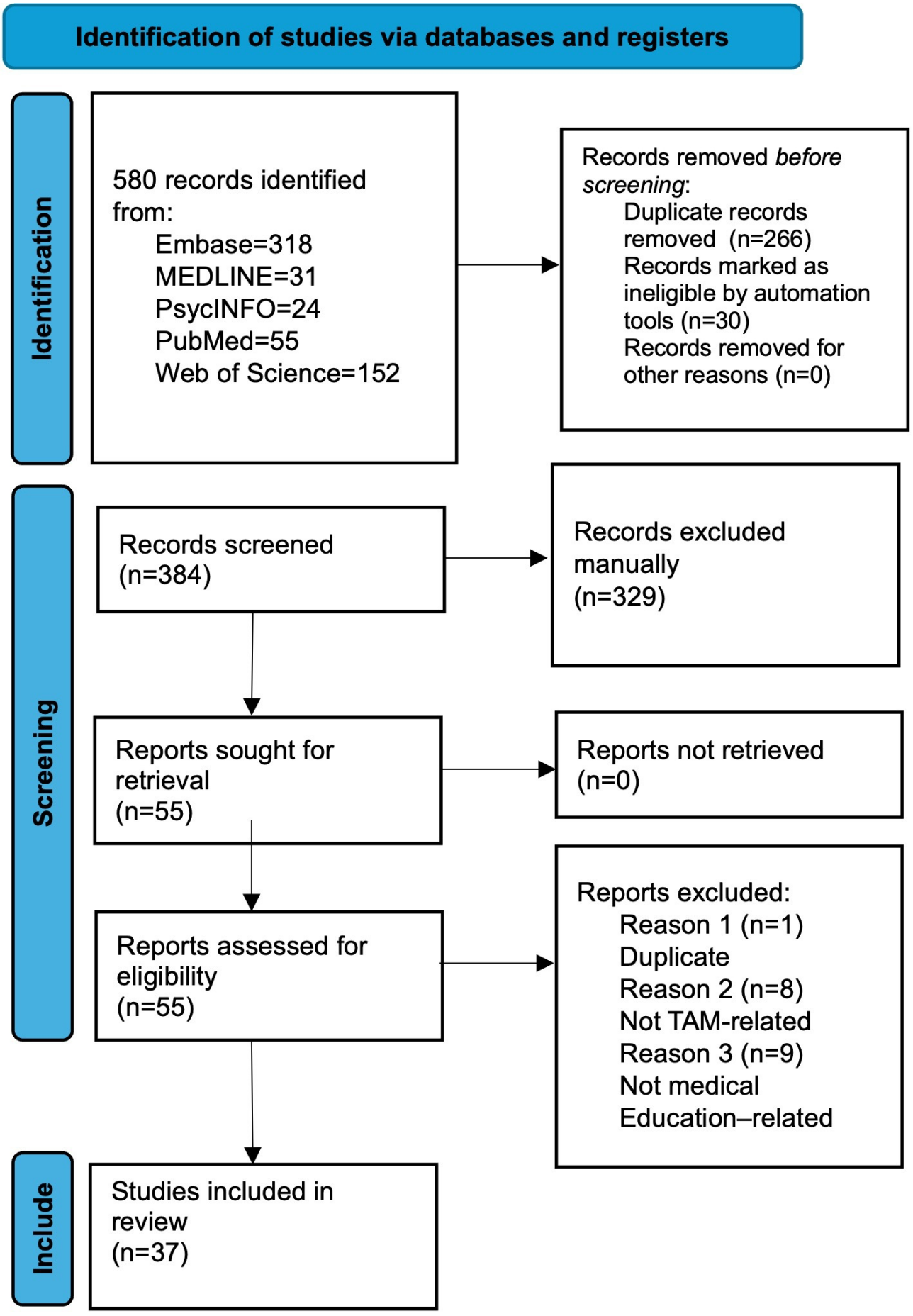


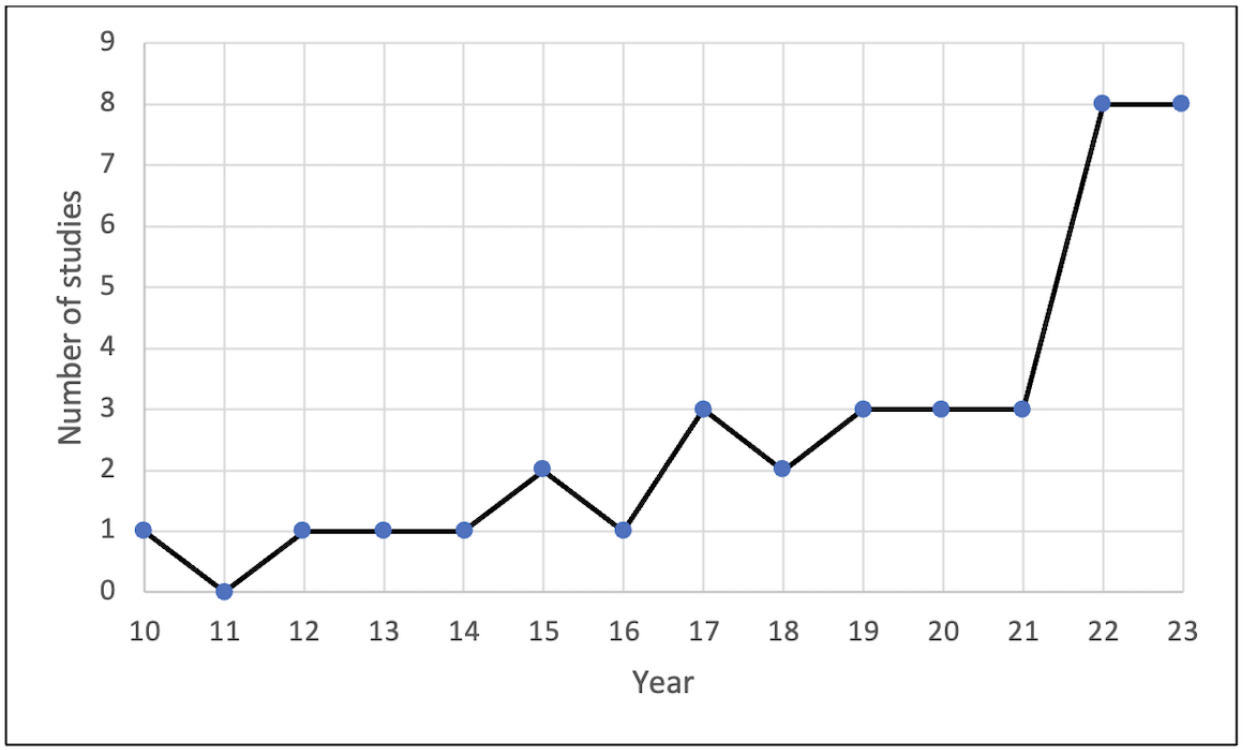
Table . Databases and search results (N=680).

Database	Vendor/publisher	Search results, n (%)
Embase	Elsevier	318 (46.3)
MEDLINE	OvidSP	31 (4.6)
PsycINFO	APA	24 (3.5)
PubMed	PubMed	155 (22.8)
Web of Science	Clarivate	152 (22.4)

Year of Publication

Despite the TAM being developed in the 1990s and our search spanning from 2003 to 2023, we only found a single study from 2010, which marked the earliest TAM usage in this review. The adoption of TAM in medical education remained relatively rare up to 2016. It was not until 2017 that a consistent uptick in the number of peer-reviewed publications using this model could be observed, with an average of 3 studies each year until 2022, when there was an almost 3-fold increase to 8 studies, which remained constant in 2023 (Figure 2). This surge in numbers was likely driven by the rapid integration of technologies from 2021 onward, which will be covered in the Discussion section.

Figure 2. Number of publications by year from 2010 to 2023.



Country of Study

The included studies were conducted in 21 countries, with no single region dominating the publications. Six of the studies were conducted in the United States [23,26,34,45,47,48], 4 in Taiwan [25,27,53], 3 in Spain [12,28,51]; 2 each in Canada [10,38], Denmark [24,32], Iran [29,35], Turkey [31,36], United Arab Emirates [33,41], and the United Kingdom [22,49]; and 1 each in Australia [30], Belgium [55], Botswana [52], Brazil [43], Egypt [40], Hong Kong [44], India [42], Indonesia [54], Jordan [50], Singapore [39], Uganda [46], and Zambia [17]. This diverse set of countries suggests that TAM has been applied globally across low-income and high-income nations, reflecting its adaptability to various education and technological contexts. Table 3 is a description of the countries by location.

Table . Location of study by country.

Country	Number of studies
Australia	1
Belgium	1
Botswana	1
Brazil	1
Canada	2
Denmark	2
Egypt	1
Hong Kong	1
India	1
Indonesia	1
Iran	2
Jordan	1
Singapore	1
Spain	3
Taiwan	4
Turkey	2
Uganda	1
United Arab Emirates	2
United Kingdom	2
United States	6
Zambia	1

Study Participants

The array of participants in the studies analyzed is quite diverse, reflecting the multifaceted nature of medical education. The review included 2 studies on general undergraduates of various disciplines [50], 1 on general undergraduate and postgraduate students of various disciplines [41], 3 on health care educators [38,44,53], and 9 on health care professionals [23,24,32,39,40,48,52,5537]. Three studies centered around health science undergraduates [46,51,54], 12 studies focused on undergraduate medical students [10,26-29,34-36,42,43,45,49], and 4 investigated undergraduate medical students and health care professionals [12,17,25,31]. Notably, 3 studies were systematic or scoping reviews [22,33,47], which, by their nature, did not involve direct study participants. Table 4 presents a summary of publications by study participants.

Table . Summary of publications by study participants.

Study participants	Publication count
General undergraduates	2
General undergraduate and postgraduate students	1
Health care educators	3
Health care professionals (doctors, nurses, pharmacists, residents)	9
Health science undergraduate students	3
Medical undergraduates	12
Medical undergraduates and health care professionals	4
Review articles (scoping or systematic review)	3

Application of TAM

The TAM served dual purposes in the surveyed studies. In 26 (70%) of the studies, it functioned as a survey instrument, quantitatively measuring the variables influencing user acceptance of and interaction with educational technology. The remaining 11 (30%) studies incorporated TAM as a foundational research framework, which involved thematic analysis of the

collected data or shaping the methodology for data collection. This 2-pronged application of the TAM highlights its adaptability and role in the empirical and theoretical examination of technology adoption in medical education. Table 5 is a summary of the application of TAM.

Table . Summary of the applications of the Technology Acceptance Model.

Application	Count
Research framework	11
Survey instrument	26

Study Design

The studies reviewed encompass quantitative, qualitative, and mixed methods research methodologies, each engaging the TAM differently. The quantitative studies operationalize the TAM through survey instruments, measuring variables such as perceived ease of use and perceived usefulness to explain the users' behavioral intentions and actual technology use. In contrast, qualitative studies contextualize the TAM within the broader theoretical landscape, using it to guide the thematic analysis of focus group discourse or to underpin systematic reviews that explore the factors influencing technology adoption. The mixed methods approach combines both, where survey data are analyzed quantitatively while concurrently using qualitative techniques such as semistructured interviews or textual analysis to capture the subtleties of user experience and perception. Most of the included studies (25/37) were quantitative, 7 were qualitative, and 5 adopted a mixed methods approach, as described in Table 6.

Table . Summary of study methodology.

Methodology	Count
Quantitative	25
Qualitative	7
Mixed method	5

Statistical Analysis

Correlation analysis was the predominant quantitative technique used in 10 studies to delineate the degree and direction of the linear relationship between the variables of interest. The next most used approach was structural equation modeling (SEM), with an equal number of studies (n=4) that used the covariance-based structural equation model and partial least squares structural equation model. Descriptive analysis was the third most frequently used method, implemented in 7 studies to succinctly summarize and describe the collected survey data. Three studies used regression analysis to predict the effect of the dependent variable based on the independent variable, and 2 other studies leveraged hypothesis testing, specifically the *t* test and *U* test, to conduct a comparative analysis of survey outcomes across different intervention groups. Seven studies were qualitative and therefore did not include statistical analysis. Table 7 summarizes the statistical approach taken by the reviewed studies, sorted by the complexity of the analysis.

Table . Statistical analysis approach of reviewed studies.

Statistical analysis approach	Count
Correlation analysis	10
SEM-CB	4
SEM-PLS	4
Descriptive analysis	7
Regression analysis	3
Hypothesis testing	2
Not applicable	7

Types of Education Intervention

The studies reviewed can be classified broadly into 2 categories of education interventions: education technologies and education methodologies. Under education technologies, 1 study examined 3D printing [26], 2 studies examined artificial intelligence [33,50], 3 studies focused on virtual reality [27,48,51], and 1 on augmented reality smart glasses [31], indicating an interest in integrating cutting-edge approaches into medical education. E-learning emerged as the most prevalent intervention, with 15 studies emphasizing digital learning [12,17,22,23,37,38,40-42,44,46,47,53-55]. In comparison, using mobile devices for learning was explored in 8 studies [10,28,29,34,36,39,45,52]. Two studies each investigated the use of serious games [24,43] and haptic devices [25,30]. One study evaluated the use of a website to improve transitions

between clinical placements [49]. Lastly, under the category of education methodologies, 1 study explored remote simulation training [34] and another explored flipped learning [45] in medical education. Table 8 summarizes the types of interventions investigated in the reviewed studies.

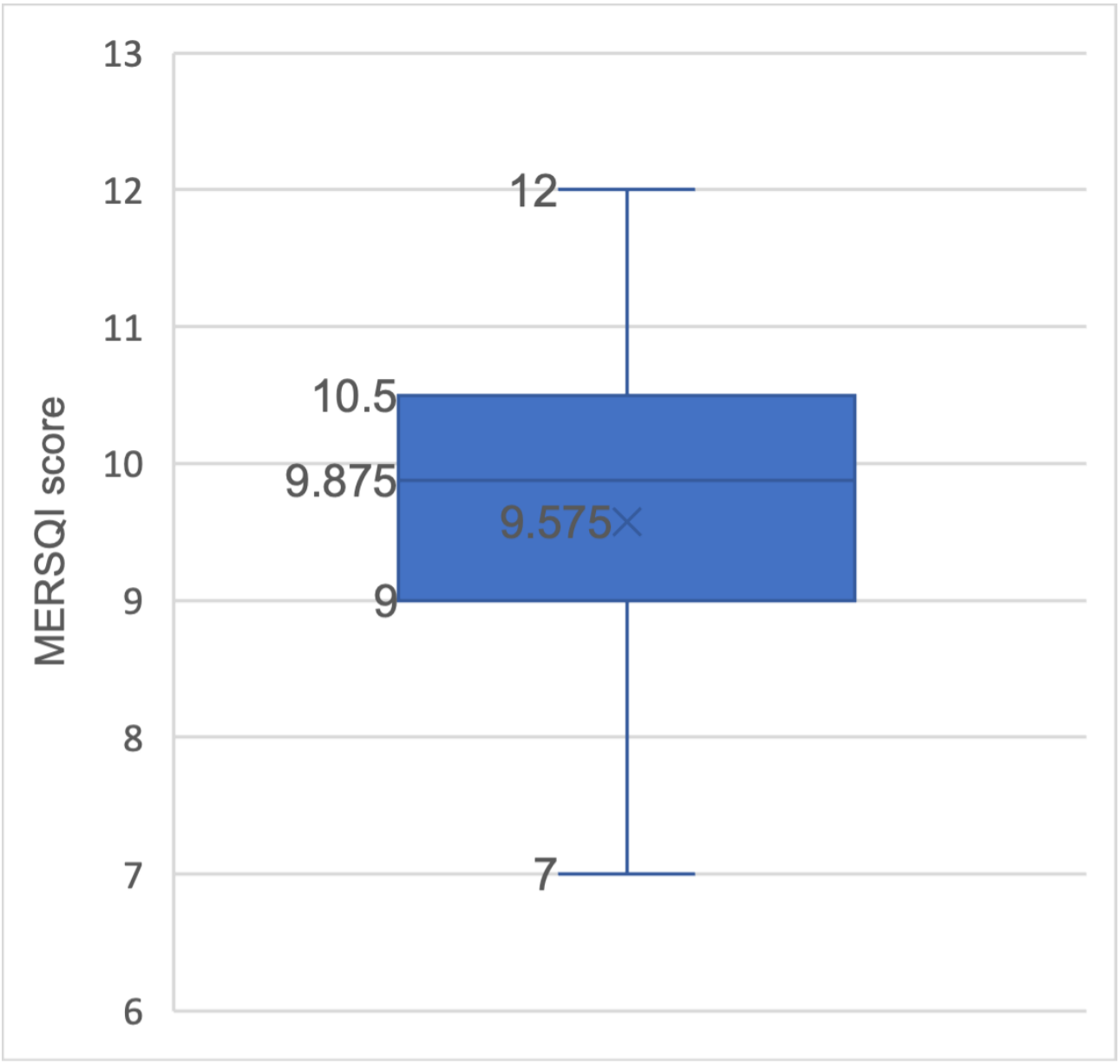
Table . Breakdown of the types of interventions investigated.

Intervention	Count
Education technologies	
3D printing	1
Artificial intelligence	2
Augmented/virtual reality	4
E-learning	15
Haptic device	2
Mobile device	8
Serious games	2
Website	1
Education methodologies	
Remote simulation training	1
Flipped learning	1

Study Quality

The MERSQI can be used to evaluate study quality in medical education research as it provides a validated, comprehensive framework for assessing methodological rigor across multiple dimensions. The MERSQI is a validated tool [57] consisting of 10 items across 6 domains: study design, sampling, data type, instrument validity, data analysis, and outcomes. Each domain can be scored up to 3, bringing the maximum score to 18. Thirty studies (25 quantitative and 5 mixed methods studies) were scored by the researchers using the MERSQI. The minimum score for the reviewed papers was 7, while the maximum was 12. The mean score was 9.58 (SD 1.31), with the mixed methods studies scoring generally below the mean score. Figure 3 is a boxplot diagram of the reviewed studies. Qualitative studies were not measured using the MERSQI. Table 1 displays a detailed summary of the MERSQI scores of the studies reviewed.

Figure 3. Boxplot diagram of the MERSQI score of reviewed studies. MERSQI: Medical Education Research Study Quality Instrument.



Discussion

Opportunities for TAM in Medical Education

Over the past two decades, technological progress has significantly shaped the education landscape. Traditional teaching approaches are enhanced with technology, making learning no longer bound by space or time. Yet, this systematic review found that the number of studies in medical education that use the TAM is notably infrequent when contrasted with other fields such as health informatics (134 studies) [16], higher education (104 studies) [58], mobile learning (87 studies) [59], and health profession education (142 studies) [18]. The review found a modest output of 1 study per year from 2010 to 2016. There was an uptick of 3 publications per year through 2021, followed by a large increase to 8 studies in 2022 and 2023, suggesting a growing interest in and recognition of TAM's relevance in medical education research.

This could be because the health care education field takes a conservative approach when adopting new digital initiatives. The curricula in medical education are highly structured and content-heavy, thus leaving little room for incorporating digital technologies. However, more recently, there have been calls for reforms within the curriculum [60], especially to integrate technology to enhance students' learning experience [61,62], which has been shown to affect student learning outcomes positively [63]. This can explain the steady increase in the number of studies that use TAM to understand user acceptance of learning interventions.

The low adoption rate of the TAM within medical education may be due to the focus on prioritizing satisfaction and basic usage statistics [64-66] when evaluating new technologies for learning. This approach overlooks the more nuanced dimensions that TAM examines, such as perceived usefulness and perceived ease of use. This emphasis on program-level satisfaction metrics fails to capture the complex psychological and organizational factors influencing technology acceptance in health care

educational environments. Such reliance on superficial evaluation matrices creates a significant gap between measuring program satisfaction and truly understanding the complex factors driving technology acceptance and sustained use in medical education.

The COVID-19 pandemic caused a global shift to digital platforms for learning. This created an urgent need to understand technology adoption in education and health care. The TAM became a framework for evaluating user acceptance of rapidly implemented technologies like e-learning platforms [26,34,56] and mobile learning [19-21]. The forced accelerated adoption of mobile and web-based learning highlighted the importance of TAM in assessing factors such as perceived usefulness and ease of use for remote teaching tools. The pandemic served as a global natural experiment in technology adoption, driving researchers to apply TAM across diverse contexts to address barriers to digital transitions. This surge in TAM applications demonstrated its adaptability in analyzing critical acceptance factors [32,36,42,46] during systemic disruptions, offering insights into user behavior that were essential for navigating the rapid technological transformations brought on by the crisis.

In this systematic review, each study with a quantitative element was appraised using the MERSQI, which is designed to assess the quality of published medical education research. Typically, a higher score is often associated with greater methodological rigor and would result in higher acceptance to quality journals [67,68]. With a mean MERSQI score of 9.6 (SD 1.17), the average score found within this review was higher than that found in a paper by Smith and Learman [57], yet it did not reach the benchmark of the high-quality score of 10.5 (SD 2.5) described by Reed et al [67]. Our analysis indicates that the substantial scores in this review are partly due to the inclusion of the TAM. Given that the TAM is a validated survey tool, its use—whether in its original or modified version—immediately contributes to a base score of five: 3 points for the tool's validity and 2 points for measuring behavioral outcomes. A study can accrue 4-5 points by using a methodologically robust and sound approach in the study design and reporting. Therefore, incorporating the TAM may potentially contribute to a higher quality of publication output.

Operationalizing the TAM

The TAM was originally developed as a theoretical framework based on the Theory of Planned Behavior [69], which can be operationalized as a survey based on the constructs within the model. This review found that the prevalent application of TAM in studies is through survey instruments, aligning with findings from other reviews [59]. Apart from the survey instrument, the TAM can be used qualitatively, such as adapting the constructs to guide the discourse in focus group discussions [32] or semistructured interviews [39].

Several different statistical analysis approaches were used to analyze the qualitative data. SEM stands out as one of the most comprehensive methods, adept at testing hypotheses concerning both observed and latent variables [70]; these studies generally had higher MERSQI scores [10,12,29,36,41,45]. Despite its robustness, SEM demands a thorough grasp of complex statistical concepts and a sufficiently large sample size to ensure the stability and accuracy of its estimates [71]. Correlation analysis offers a more straightforward approach to measuring the strength and direction of relationships between variables. Regression analysis further extends the analytical capability by providing predictive insights and facilitating the exploration of potential causal links between factors. Additionally, the TAM is frequently used in a descriptive capacity, offering an interpretive lens to dissect and articulate the intricacies of user interactions with technology, their attitudes, and the behavioral intentions that these factors precipitate.

This systematic review found that a large number of learning interventions were investigated for e-learning. This could be explained by the shift in higher education over the past 2 decades to web-based learning [72], accelerated by the COVID-19 pandemic, which necessitated and expedited the transition to web-based learning across various disciplines [73], including medical education. The integration of mobile technology into our everyday lives has naturally extended into the realm of education. This has prompted research on mobile devices for information access [15,23,32,47] and mobile apps for learning [12,45].

Additionally, this systematic review found studies delving into more innovative educational technologies beyond e-learning, such as virtual reality, augmented reality, serious games, and 3D printing. Virtual reality allows for students to practice their technical skills repeatedly in a risk-free setting, thus increasing their confidence and proficiency without jeopardizing patient safety [25,74]. Augmented reality overlays digital information onto physical or live environments, allowing students to understand complex anatomical structures [6] and to gain spatial awareness [75]. Another emerging tool that combines interactive gameplay with educational outcomes is serious games, which simulate real-world medical scenarios in a controlled and engaging environment [24,43]. 3D printing allows for the rapid development of customized models that can be used for teaching and learning [26]. Together, these technologies are changing the way learning is happening in medical education by providing immersive, interactive, and accessible tools to complement traditional teaching approaches. The TAM can serve as a valuable framework for understanding how these new and existing technologies are adopted and used for learning in medical education. By having a framework, educators and institutions can use the TAM to evaluate the integration of these technologies into their curricula and their potential for improving educational outcomes.

Limitations and Future Research

One limitation of this review is that it only encompasses data published up until 2023. Given the observed publication trend, it is plausible that subsequent studies using TAM in 2024 and beyond fall outside the scope of this review. Consequently, the conclusions drawn here are pertinent to the specified research period. Future studies should consider extending the review to include these additional years, thereby capturing a more comprehensive dataset, potentially offering a more current evaluation of TAM's application in the field.

Although this review contributes valuable insights into the use and application of TAM in education, the findings are primarily

within the context of medical education and exclude other health professions, including nursing and allied health professionals. Medical education is a highly specialized domain with unique challenges and practices that may not directly translate to the broader educational context outside of medicine. Furthermore, the complexity and heterogeneity within medical education, such as the variation in curricula and culture, may pose an additional challenge to generalizing findings even within the discipline.

Despite these limitations, the framework used in this review offers significant potential for broader application across other health care disciplines. Researchers could adapt this approach to explore TAM's adoption and effectiveness in nursing education, allied health training, or interdisciplinary health care programs. Expanding research beyond medical education would enhance the generalizability of findings and provide comparative insights into how TAM influences technology adoption across diverse health care professions. Additionally, extending TAM-based research to non–health care fields could further enrich our understanding of its applicability and utility in varied educational contexts.

Future researchers should consider adopting the additional constructs in TAM2 to better understand how social and cognitive factors influence technology acceptance beyond the perceived ease of use and perceived usefulness. For example, researchers can investigate how subjective norms or job relevance may influence the students' willingness to adopt new technologies and professional identities in an educational context.

Conclusions

This systematic review aimed to understand the use of the TAM in medical education over the past two decades, highlighting its utility as both a theoretical framework and survey instrument. This study reported on TAM's increasing popularity and versatility for measuring and understanding the learners' acceptance of the intervention. With the increasing integration of e-learning, digital learning, and other new learning modalities, it is critical that researchers can leverage technologies that learners will adopt. Such curriculum innovations are critical for maintaining educational continuity in the face of global health challenges by facilitating remote learning and continuous professional development. Consequently, these curricular reforms are expected to catalyze a significant surge in the adoption of digital technologies within medical education.

The growing importance of the TAM in understanding technology acceptance cannot be overstated, especially in medical education, where the use of artificial intelligence, virtual reality, and other adaptive learning platforms is increasingly popular. Educators and developers can use TAM as a theoretical framework to design curricula or interventions considering barriers to adoption, such as organizational support or the intervention's technical complexity. The TAM is relevant as an evaluation tool and can guide future innovations in medical education. Policymakers should consider using the insights gained using the TAM to develop strategies for the education system while meeting the challenges of cost, accessibility, and infrastructure development.

This review provides a comprehensive understanding of how the TAM has been applied within the field of medical education over the past 20 years. As the field continues to innovate, TAM will continue to play an important role in helping educators, policymakers, and researchers understand the dynamics of technology integration and the impact on teaching and student learning outcomes.

Acknowledgments

This work was supported by the National University of Singapore's Learning Improvement Teaching Enhancement Grant (TEG).

Conflicts of Interest

None declared.

Checklist 1
PRISMA checklist. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.
[PDF File, 62 KB - mededu_v11i1e67873_app1.pdf]

References

1. Khan Academy. URL: https://www.khanacademy.org/ [accessed 2024-04-18]
2. edX. URL: https://www.edx.org/ [accessed 2024-04-18]
3. Coursera. URL: https://www.coursera.org/ [accessed 2024-04-18]
4. Johnson EO, Charchanti AV, Troupis TG. Modernization of an anatomy class: from conceptualization to implementation. A case for integrated multimodal-multidisciplinary teaching. Anat Sci Educ 2012;5(6):354-366. [doi: 10.1002/ase.1296] [Medline: 22730175]
5. Birbara NS, Sammut C, Pather N. Virtual reality in anatomy: a pilot study evaluating different delivery modalities. Anat Sci Educ 2020 Jul;13(4):445-457. [doi: 10.1002/ase.1921] [Medline: 31587471]
6. Duncan-Vaidya EA, Stevenson EL. The effectiveness of an augmented reality head-mounted display in learning skull anatomy at a community college. Anat Sci Educ 2021 Mar;14(2):221-231. [doi: 10.1002/ase.1998] [Medline: 32583577]

7. Adams JW, Paxton L, Dawes K, Burlak K, Quayle M, McMenamin PG. 3D printed reproductions of orbital dissections: a novel mode of visualising anatomy for trainees in ophthalmology or optometry. Br J Ophthalmol 2015 Sep;99(9):1162-1167. [doi: 10.1136/bjophthalmol-2014-306189] [Medline: 25689987]
8. Lee JWY, Ong DW, Soh RCC, Rao JP, Bello F. Exploring student acceptance of learning technologies in anatomy education: a mixed-method approach. Clin Anat 2025 Apr;38(3):334-346. [doi: 10.1002/ca.24254] [Medline: 39673302]
9. Baptiste YM. Digital feast and physical famine: the altered ecosystem of anatomy education due to the Covid-19 pandemic. Anat Sci Educ 2021 Jul;14(4):399-407. [doi: 10.1002/ase.2098] [Medline: 33961346]
10. Do DH, Lakhal S, Bernier M, Bisson J, Bergeron L, St-Onge C. Drivers of iPad use by undergraduate medical students: the Technology Acceptance Model perspective. BMC Med Educ 2022 Feb 8;22(1):87. [doi: 10.1186/s12909-022-03152-w] [Medline: 35135525]
11. Baghcheghi N, Koohestani H, Karimy M, Alizadeh S. Factors affecting mobile learning adoption in healthcare professional students based on technology acceptance model. Acta fac medic Naissensis 2020;37(2):191-200. [doi: 10.5937/afmnai2002191B]
12. Briz-Ponce L, García-Peñalvo FJ. An empirical assessment of a Technology Acceptance Model for apps in medical education. J Med Syst 2015 Nov;39(11):176. [doi: 10.1007/s10916-015-0352-x] [Medline: 26411928]
13. Venkatesh V, Davis FD. A theoretical extension of the Technology Acceptance Model: four longitudinal field studies. Manage Sci 2000 Feb;46(2):186-204. [doi: 10.1287/mnsc.46.2.186.11926]
14. Papapanou M, Routsi E, Tsamakis K, et al. Medical education challenges and innovations during COVID-19 pandemic. Postgrad Med J 2022 May;98(1159):321-327. [doi: 10.1136/postgradmedj-2021-140032] [Medline: 33782202]
15. Gagnon MP, Ngangue P, Payne-Gagnon J, Desmartis M. m-Health adoption by healthcare professionals: a systematic review. J Am Med Inform Assoc 2016 Jan;23(1):212-220. [doi: 10.1093/jamia/ocv052] [Medline: 26078410]
16. Rahimi B, Nadri H, Lotfnezhad Afshar H, Timpka T. A systematic review of the Technology Acceptance Model in health informatics. Appl Clin Inform 2018 Jul;9(3):604-634. [doi: 10.1055/s-0038-1668091] [Medline: 30112741]
17. Barteit S, Neuhann F, Bärnighausen T, et al. Technology acceptance and information system success of a mobile electronic platform for nonphysician clinical students in Zambia: prospective, nonrandomized intervention study. J Med Internet Res 2019 Oct 9;21(10):e14748. [doi: 10.2196/14748] [Medline: 31599731]
18. AlQudah AA, Al-Emran M, Shaalan K. Technology acceptance in healthcare: a systematic review. Appl Sci (Basel) 2021 Jan;11(22):10537. [doi: 10.3390/app112210537]
19. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 2009 Jul 21;339(jul21 1):b2535. [doi: 10.1136/bmj.b2535] [Medline: 19622551]
20. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep;13(3):319. [doi: 10.2307/249008]
21. Microsoft Teams. Microsoft. 2023. URL: https://www.microsoft.com/en-us/microsoft-teams/group-chat-software [accessed 2025-06-19]
22. Wong G, Greenhalgh T, Pawson R. Internet-based medical education: a realist review of what works, for whom and in what circumstances. BMC Med Educ 2010 Feb 2;10(1):12. [doi: 10.1186/1472-6920-10-12] [Medline: 20122253]
23. McGowan BS, Wasko M, Vartabedian BS, Miller RS, Freiherr DD, Abdolrasulnia M. Understanding the factors that influence the adoption and meaningful use of social media by physicians to share medical information. J Med Internet Res 2012 Sep 24;14(5):e117. [doi: 10.2196/jmir.2138] [Medline: 23006336]
24. Knight JF. Acceptability of video games technology for medical emergency training. Int J Gaming Comput Mediat Simul 2013 Oct;5(4):86-99. [doi: 10.4018/ijgcms.2013100105]
25. Fang TY, Wang PC, Liu CH, Su MC, Yeh SC. Evaluation of a haptics-based virtual reality temporal bone simulator for anatomy and surgery training. Comput Methods Programs Biomed 2014 Feb;113(2):674-681. [doi: 10.1016/j.cmpb.2013.11.005] [Medline: 24280627]
26. Ryan JR, Chen T, Nakaji P, Frakes DH, Gonzalez LF. Ventriculostomy simulation using patient-specific ventricular anatomy, 3D printing, and hydrogel casting. World Neurosurg 2015 Nov;84(5):1333-1339. [doi: 10.1016/j.wneu.2015.06.016] [Medline: 26100167]
27. Huang HM, Liaw SS, Lai CM. Exploring learner acceptance of the use of virtual reality in medical education: a case study of desktop and projection-based display systems. Interactive Learning Environments 2016 Jan 2;24(1):3-19. [doi: 10.1080/10494820.2013.817436]
28. Briz-Ponce L, Pereira A, Carvalho L, Juanes-Méndez JA, García-Peñalvo FJ. Learning with mobile technologies – students' behavior. Comput Human Behav 2017 Jul;72:612-620. [doi: 10.1016/j.chb.2016.05.027]
29. Tahamtan I, Pajouhanfar S, Sedghi S, Azad M, Roudbari M. Factors affecting smartphone adoption for accessing information in medical settings. Health Info Libraries J 2017 Jun;34(2):134-145. [doi: 10.1111/hir.12174]
30. Yeom S, Choi-Lundberg DL, Fluck AE, Sale A. Factors influencing undergraduate students' acceptance of a haptic interface for learning gross anatomy. ITSE 2017 Apr 18;14(1):50-66. [doi: 10.1108/ITSE-02-2016-0006]
31. Basoglu N, Goken M, Dabic M, Ozdemir Gungor D, Daim TU. Exploring adoption of augmented reality smart glasses: applications in the medical industry. Front Eng 2018;0:0. [doi: 10.15302/J-FEM-2018056]

32. Duch Christensen M, Oestergaard D, Dieckmann P, Watterson L. Learners' perceptions during simulation-based training: an interview study comparing remote versus locally facilitated simulation-based training. Simul Healthc 2018 Oct;13(5):306-315. [doi: 10.1097/SIH.0000000000000300] [Medline: 29620704]
33. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. JMIR Med Educ 2019 Jun 15;5(1):e13930. [doi: 10.2196/13930] [Medline: 31199295]
34. Johnson EM, Howard C. A library mobile device deployment to enhance the medical student experience in a rural longitudinal integrated clerkship. J Med Libr Assoc 2019 Jan;107(1):30-42. [doi: 10.5195/jmla.2019.442] [Medline: 30598646]
35. Abdekhoda M, Maserat E, Ranjbaran F. A conceptual model of flipped classroom adoption in medical higher education. ITSE 2020 Mar 14;17(4):393-401. [doi: 10.1108/ITSE-09-2019-0058]
36. Kucuk S, Baydas Onlu O, Kapakin S. A model for medical students' behavioral intention to use mobile learning. J Med Educ Curric Dev 2020;7:2382120520973222. [doi: 10.1177/2382120520973222] [Medline: 33313399]
37. Lee CW, Chen GL, Lee YK. User experience evaluation of the EPAs-based e-portfolio system and an analysis of its impact. J Acute Med 2020 Sep 1;10(3):115-125. [doi: 10.6705/j.jacme.202009_10(3).0003] [Medline: 33209570]
38. Jeyakumar T, Ambata-Villanueva S, McClure S, Henderson C, Wiljer D. Best practices for the implementation and sustainment of virtual health information system training: qualitative study. JMIR Med Educ 2021 Oct 22;7(4):e30613. [doi: 10.2196/30613] [Medline: 34449402]
39. Lee SS, Tay SM, Balakrishnan A, Yeo SP, Samarasekera DD. Mobile learning in clinical settings: unveiling the paradox. Korean J Med Educ 2021 Dec;33(4):349-367. [doi: 10.3946/kjme.2021.204] [Medline: 34875152]
40. Zalat MM, Hamed MS, Bolbol SA. The experiences, challenges, and acceptance of e-learning as a tool for teaching during the COVID-19 pandemic among university medical staff. In: Hwang GJ, editor. PLoS One 2021;16(3):e0248758. [doi: 10.1371/journal.pone.0248758] [Medline: 33770079]
41. Almarzouqi A, Aburayya A, Salloum SA. Prediction of user's intention to use metaverse system in medical education: a hybrid SEM-ML learning approach. IEEE Access 2022;10:43421-43434. [doi: 10.1109/ACCESS.2022.3169285]
42. Bhardwaj M, Kashyap S, Aggarwal D, Bhawani R. Perceptions and experience of medical students regarding e-learning during COVID-19 lockdown- a cross-sectional study. JCDR 2022. [doi: 10.7860/JCDR/2022/54803.16051]
43. Bianchi I, Stefani CJM, Santiago P, Zanatta AL, Rieder R. AnemiaAR: a serious game to support teaching of haematology. J Vis Commun Med 2022 Jul;45(3):134-153. [doi: 10.1080/17453054.2021.2021798] [Medline: 35129054]
44. Chan E, Khong ML, Torda A, Tanner JA, Velan GM, Wong GTC. Medical teachers' experience of emergency remote teaching during the COVID-19 pandemic: a cross-institutional study. BMC Med Educ 2022 Apr 21;22(1):303. [doi: 10.1186/s12909-022-03367-x] [Medline: 35449047]
45. Harmon DJ, Burgoon JM, Kalmar EL. Development and assessment of an integrated anatomy mobile app. Clin Anat 2022 Jul;35(5):686-696. [doi: 10.1002/ca.23895] [Medline: 35452135]
46. Komuhangi A, Mpirirwe H, Robert L, Githinji FW, Nanyonga RC. Predictors for adoption of e-learning among health professional students during the COVID-19 lockdown in a private university in Uganda. BMC Med Educ 2022 Sep 10;22(1):671. [doi: 10.1186/s12909-022-03735-7] [Medline: 36088322]
47. Lau KHV, Greer DM. Using technology adoption theories to maximize the uptake of e-learning in medical education. Med Sci Educ 2022 Apr;32(2):545-552. [doi: 10.1007/s40670-022-01528-7] [Medline: 35261814]
48. Bugli D, Dick L, Wingate KC, et al. Training the public health emergency response workforce: a mixed-methods approach to evaluating the virtual reality modality. BMJ Open 2023 May 9;13(5):e063527. [doi: 10.1136/bmjopen-2022-063527] [Medline: 37160399]
49. Young Y, Leedham-Green K, Jensen-Martin J. Improving transitions between clinical placements. Clin Teach 2023 Aug;20(4):e13580. [doi: 10.1111/tct.13580] [Medline: 37146063]
50. Sallam M, Salim NA, Barakat M, et al. Assessing health students' attitudes and usage of ChatGPT in Jordan: validation study. JMIR Med Educ 2023 Sep 5;9(1):e48254. [doi: 10.2196/48254] [Medline: 37578934]
51. Cabero-Almenara J, Llorente-Cejudo C, Palacios-Rodríguez A, Gallego-Pérez Ó. Degree of acceptance of virtual reality by health sciences students. Int J Environ Res Public Health 2023 Apr 18;20(8):5571. [doi: 10.3390/ijerph20085571] [Medline: 37107853]
52. Ndlovu K, Stein N, Gaopelo R, et al. Evaluating the feasibility and acceptance of a mobile clinical decision support system in a resource-limited country: exploratory study. JMIR Form Res 2023 Oct 10;7:e48946. [doi: 10.2196/48946] [Medline: 37815861]
53. Lin CW, Clinciu DL, Salcedo D, Huang CW, Kang EYN, Li YCJ. Crowdsource authoring as a tool for enhancing the quality of competency assessments in healthcare professions. PLoS One 2023;18(11):e0278571. [doi: 10.1371/journal.pone.0278571] [Medline: 37917751]
54. Rahadiani P, Kekalih A, Krisnamurti DGB. Use of H5P interactive learning content in a self-paced MOOC for learning activity preferences and acceptance in an Indonesian medical elective module. African Journal of Science, Technology, Innovation and Development 2023 Nov 10;15(7):844-851. [doi: 10.1080/20421338.2023.2209482]
55. De Ruyck O, Embo M, Morton J, et al. A comparison of three feedback formats in an ePortfolio to support workplace learning in healthcare education: a mixed method study. Educ Inf Technol 2024 Jun;29(8):9667-9688. [doi: 10.1007/s10639-023-12062-3]

56. Reed DA, Cook DA, Beckman TJ, Levine RB, Kern DE, Wright SM. Association between funding and quality of published medical education research. JAMA 2007 Sep 5;298(9):1002-1009. [doi: 10.1001/jama.298.9.1002] [Medline: 17785645]
57. Smith RP, Learman LA. A plea for MERSQI: the Medical Education Research Study Quality Instrument. Obstet Gynecol 2017 Oct;130(4):686-690. [doi: 10.1097/AOG.0000000000002091] [Medline: 28885409]
58. Rosli MS, Saleh NS, Md. Ali A, Abu Bakar S, Mohd Tahir L. A systematic review of the Technology Acceptance Model for the sustainability of higher education during the COVID-19 pandemic and identified research gaps. Sustainability 2022 Sep 10;14(18):11389. [doi: 10.3390/su141811389]
59. Al-Emran M, Mezhuyev V, Kamaludin A. Technology Acceptance Model in M-learning context: a systematic review. Comput Educ 2018 Oct;125:389-412. [doi: 10.1016/j.compedu.2018.06.008]
60. Buja LM. Medical education today: all that glitters is not gold. BMC Med Educ 2019 Apr 16;19(1):110. [doi: 10.1186/s12909-019-1535-9] [Medline: 30991988]
61. Goh PS, Sandars J. A vision of the use of technology in medical education after the COVID-19 pandemic. MedEdPublish (2016) 2020;9:49. [doi: 10.15694/mep.2020.000049.1] [Medline: 38058893]
62. Kaul V, Gallo de Moraes A, Khateeb D, et al. Medical education during the COVID-19 pandemic. Chest 2021 May;159(5):1949-1960. [doi: 10.1016/j.chest.2020.12.026] [Medline: 33385380]
63. Vallée A, Blacher J, Cariou A, Sorbets E. Blended learning compared to traditional learning in medical education: systematic review and meta-analysis. J Med Internet Res 2020 Aug 10;22(8):e16504. [doi: 10.2196/16504] [Medline: 32773378]
64. Gray JA, DiLoreto M. The effects of student engagement, student satisfaction, and perceived learning in online learning environments. International Journal of Educational Leadership Preparation 2016 [FREE Full text]
65. Lim J. Impact of instructors' online teaching readiness on satisfaction in the emergency online teaching context. Educ Inf Technol (Dordr) 2023;28(4):4109-4126. [doi: 10.1007/s10639-022-11241-y] [Medline: 36247026]
66. Wu JH, Tennyson RD, Hsia TL. A study of student satisfaction in a blended e-learning system environment. Comput Educ 2010 Aug;55(1):155-164. [doi: 10.1016/j.compedu.2009.12.012]
67. Reed DA, Beckman TJ, Wright SM, Levine RB, Kern DE, Cook DA. Predictive validity evidence for medical education research study quality instrument scores: quality of submissions to JGIM's Medical Education Special Issue. J Gen Intern Med 2008 Jul;23(7):903-907. [doi: 10.1007/s11606-008-0664-3] [Medline: 18612715]
68. Reed DA, Beckman TJ, Wright SM. An assessment of the methodologic quality of medical education research studies published in The American Journal of Surgery. Am J Surg 2009 Sep;198(3):442-444. [doi: 10.1016/j.amjsurg.2009.01.024] [Medline: 19716888]
69. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991 Dec;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]
70. Burnette J, Williams L. Structural equation modeling (SEM): an introduction to basic techniques and advanced issues. In: Research in Organizations: Berrett-Koehler Publishers; 2005.
71. Hair JF, Risher JJ, Sarstedt M, Ringle CM. When to use and how to report the results of PLS-SEM. EBR 2019 Jan 14;31(1):2-24. [doi: 10.1108/EBR-11-2018-0203]
72. Lim CP. Trends in online learning and their implications for schools. Educ Technol 2002;42(6):43-48 [FREE Full text]
73. Kumar A, Kumar P, Palvia SCJ, Verma S. Online education worldwide: current status and emerging trends. Journal of Information Technology Case and Application Research 2017 Jan 2;19(1):3-9. [doi: 10.1080/15228053.2017.1294867]
74. Hogg ME, Tam V, Zenati M, et al. Mastery-based virtual reality robotic simulation curriculum: the first step toward operative robotic proficiency. J Surg Educ 2017;74(3):477-485. [doi: 10.1016/j.jsurg.2016.10.015] [Medline: 27884677]
75. Küçük S, Kapakin S, Göktaş Y. Learning anatomy via mobile augmented reality: effects on achievement and cognitive load. Anat Sci Educ 2016 Oct;9(5):411-421. [doi: 10.1002/ase.1603] [Medline: 26950521]

Abbreviations

MERSQI: Medical Education Research Study Quality Instrument
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SEM: structural equation modeling
TAM: Technology Acceptance Model

Edited by J Moen; submitted 23.10.24; peer-reviewed by K Mouloudj, O Egunlae; revised version received 26.03.25; accepted 12.05.25; published 16.07.25.

Please cite as:
Lee JWY, Tan JY, Bello F
Technology Acceptance Model in Medical Education: Systematic Review
JMIR Med Educ 2025;11:e67873
URL: https://mededu.jmir.org/2025/1/e67873
doi:10.2196/67873

Virtual Simulation Tools for Communication Skills Training in Health Care Professionals: Literature Review

Manuel Fernández-Alcántara[1,2*], PhD; Silvia Escribano[2,3*], PhD; Rocío Juliá-Sanchis[2,3*], PhD; Ana Castillo-López[3*]; Antonio Pérez-Manzano[4*], PhD; M Macur[5*], PhD; Sedina Kalender-Smajlović[5*], PhD; Sofía García-Sanjuán[2,3*], PhD; María José Cabañero-Martínez[2,3*], PhD

[1]Department of Health Psychology, Faculty of Health Sciences, University of Alicante, Alicante, Spain

[2]Institute of Health and Biomedical Research of Alicante, Alicante, Spain

[3]Department of Nursing, Faculty of Health Sciences, University of Alicante, Carretera San Vicente del Raspeig s/n, Alicante, Spain

[4]University of Murcia, Murcia, Spain

[5]Angela Boškin Faculty of Health Care, Spodnji Plavž 3, Jesenice, Slovenia

[*]all authors contributed equally

Corresponding Author:
Sofía García-Sanjuán, PhD
Institute of Health and Biomedical Research of Alicante, Alicante, Spain

Abstract

Background: Quality clinical care is supported by effective patient-centered communication. Health care professionals can improve their communication skills through simulation-based training, but our knowledge about virtual simulation and its effectiveness and use in training health professionals and students is still growing rapidly.

Objective: The objective of this study was to review the current academic literature to identify and evaluate the virtual simulation tools used to train communication skills in health care students and professionals.

Methods: This review was carried out in June 2023 by collecting data from the MEDLINE/PubMed and Web of Science electronic databases. Once applicable studies were identified, we recorded data related to type of technology used, learning objectives, degree of learning autonomy, outcomes, and other details.

Results: We found 35 articles that had developed and/or applied a virtual environment for training communication skills aimed at patients, in which 24 different learning tools were identified. Most had been developed to independently train communication skills in English, either generally or in the specific context of medical history (anamnesis) interviews. Many of these tools used a virtual patient that looked like a person and had the ability to vocally respond. Almost half of the tools analyzed allowed the person being trained to respond orally using natural language. Of note, not all these studies described the technology they had used in detail.

Conclusions: Many different learning tools with very heterogeneous characteristics are being used for the purposes of communication skills training. Continued research will still be required to develop virtual tools that include the most advanced features to achieve high-fidelity simulation training.

(JMIR Med Educ 2025;11:e63082) doi:10.2196/63082

KEYWORDS

communication skills; virtual patient; virtual simulation; health care professionals; virtual simulation tool; skill training; communication; heterogeneous; heterogeneous characteristics; virtual tool; patient-centered; patient-centered communication; implementation

Introduction

Effective patient-centered communication is one of the key components of quality clinical care [1]. Thus, it is vital that health care professionals adequately manage their communication skills. This involves mastering the transmission of information; listening and comprehensively understanding all the issues related to the health of each patient [2]; and responding appropriately to the physical and emotional needs of patients [3]. Better communication when supporting decision-making means that patients are better able to understand their situation, feel better informed, and are more active in the decision-making process [4]. Hence, acquiring good communication skills has been related to improved health outcomes, general patient satisfaction [5], better adherence to treatment plans [6], and positive effects on health care costs and length of hospital stay [7].

However, despite recognizing the importance of communication, health professionals are not always sufficiently skilled in this area [8]. Therefore, it is advisable that both health and educational institutions introduce different means of supporting the development of communication skills into their training plans as a priority objective. Furthermore, this training must also be implemented through effective educational strategies [9]. It has previously been shown that simulation-based learning is an effective means of acquiring communication skills [9]. Specifically, simulation with a standardized or simulated patient, which consists of using trained people to realistically portray a patient within learning contexts [10], is widely used to train communication skills [1].

Nonetheless, although the use of simulation methodologies has greatly advanced training in communication skills, its implementation also has limitations. For example, in terms of the human resources used in this type of training, it is particularly difficult to recruit actors able to simulate patients precisely and consistently in a completely standardized way [11,12]. Other difficulties include temporal–spatial issues because the availability of simulations with standardized patients is limited to a specific physical space and time [13]. A training alternative that could overcome these limitations is the use of standardized virtual patient programs that use computerized characters rather than real actors [14].

Indeed, compared to standardized patients, there are significant advantages to the use of virtual patients, including the need for fewer staff and resources once developed [15], unlimited availability, and the fact that they are highly customizable [14]. Additionally, these tools provide highly interactive, engaging, and more standardized experiences because educators can control their design, programming, delivery, and use [14]. It is also worth noting that these solutions can be personalized according to specific individual needs, given that they are not limited by time or space, so students can repeatedly engage in training in more clinical scenarios than is possible through traditional methods [15]. In addition, this technology also allows students to learn in a safe environment with low levels of risk and anxiety, which encourages them to gain greater personal awareness of their learning processes [16].

Virtual simulation has gained attention in recent years as a promising tool for training both undergraduate and graduate students, as well as health care professionals, in various competencies, including nontechnical skills. This growing interest is evident in an increasing number of studies focused on its potential applications in health care education [17]. However, despite this expanding body of research, it is advisable to continue researching with the aim of fully exploring and understanding which technological and technical skills are more suitable to train in virtual simulation [17]. Some reviews on virtual simulation and the learning of nontechnical skills such as communication are available [17-19]. For example, in their integrative review, Peddle et al [19] examined how interactions with virtual patients impacted nontechnical skills in general, without exclusively focusing on communication skills or technical and instructional design characteristics. Subsequently, both the systematic review by Lee et al [18] and the literature review by Battegazzorre et al [17] examined the technical characteristics of virtual learning applications aimed at improving communication skills. However, it is noteworthy that these reviews include studies published only up to December 2018 and May 2020, respectively, which highlights a gap in the literature regarding recent advancements in virtual simulation technologies.

The development of communication skills is fundamental for the effective clinical practice of health care professionals. However, the increasing diversity of virtual simulation tools and the rapid pace of technological innovation pose significant challenges to understanding which tools are most effective for training these skills. This raises the following key questions: what are the characteristics of the current virtual simulation tools used for training communication skills, and how effective are they in fostering realistic and immersive learning experiences? To address these questions, we conducted a systematic review of the virtual simulation tools available to train communication skills in health care professionals, analyzing their design, degree of immersion, and autonomy to identify their strengths and limitations.

Therefore, the objective of this study was to review the current academic literature to identify and evaluate the virtual simulation tools used to train communication skills in health care students and professionals and to assess their effectiveness and limitations in training health care personnel.

Methods

Design

We completed a literature review to identify virtual simulation tools designed to train communication skills in health care professionals, including students in training and practicing professionals. The inclusion criteria were studies that examined (1) virtual simulation tools and/or those based on artificial intelligence (AI), (2) tools used to train communication skills in health professionals, and (3) tools targeting training in communication skills and/or therapeutic relationships with patients. Studies were excluded if (1) the tools were designed to train interprofessional communication, (2) the objective was noneducational, and (3) the tool was designed to train patients in social and/or communication skills. This systematic review was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 [20] guidelines to ensure comprehensive and transparent reporting of the methodology and findings.

Search Strategy

The search for studies was conducted in June 2023 in the MEDLINE/PubMed and Web of Science electronic databases. As part of the search strategy, we consulted the PubMed thesaurus using the following Medical Subject Headings (MeSH) terms: “Artificial Intelligence,” “Machine learning,” “virtual reality,” and “social skills.” The natural language search terms included in the title and/or abstract fields were “artificial intelligence,” “machine learning,” “virtual reality,” “e-simulation,” “web-based simulation,” “virtual simulation,” “virtual patient,” “social skills,” “interpersonal skills,” “social ability,” “social competences,” and “communication skills.”

The complete search strategy was as follows: (((“Artificial Intelligence”[MeSH Terms] OR “Machine Learning”[MeSH Terms] OR “Artificial Intelligence”[Title/Abstract] OR “Machine Learning”[Title/Abstract])) OR ((“Virtual Reality”[MeSH Terms] OR “Virtual Reality”[Title/Abstract] OR “e-simulation”[Title/Abstract] OR “web-based simulation”[Title/Abstract] OR “virtual simulation”[Title/Abstract]) OR (“virtual patient”[Title/Abstract]))) AND ((“Social Skills”[MeSH Terms] OR “Social Skills”[Title/Abstract] OR “interpersonal skills”[Title/Abstract] OR (“social ability”[Title/Abstract] OR “social abilities”[Title/Abstract]) OR (“social competence”[Title/Abstract] OR “social competences”[Title/Abstract]) OR “communication skills”[Title/Abstract])).

No temporal restrictions were applied in any of these cases. Despite previous reviews focusing on similar topics [17-19], it was decided not to base the current review on them. This decision was due to differences in the search strategy used, which did not account for the wide range of synonyms associated with each term established for this review. Furthermore, it is important to note that Lee et al [18] focused their strategy exclusively on communication among medical students, while Peddle et al [19] directed their attention to all nontechnical skills, not just communication skills.

The eligibility of the studies was independently assessed by 2 of the authors (MJCM and RJS) and any discrepancies were resolved by another author (SE).

Data Extraction

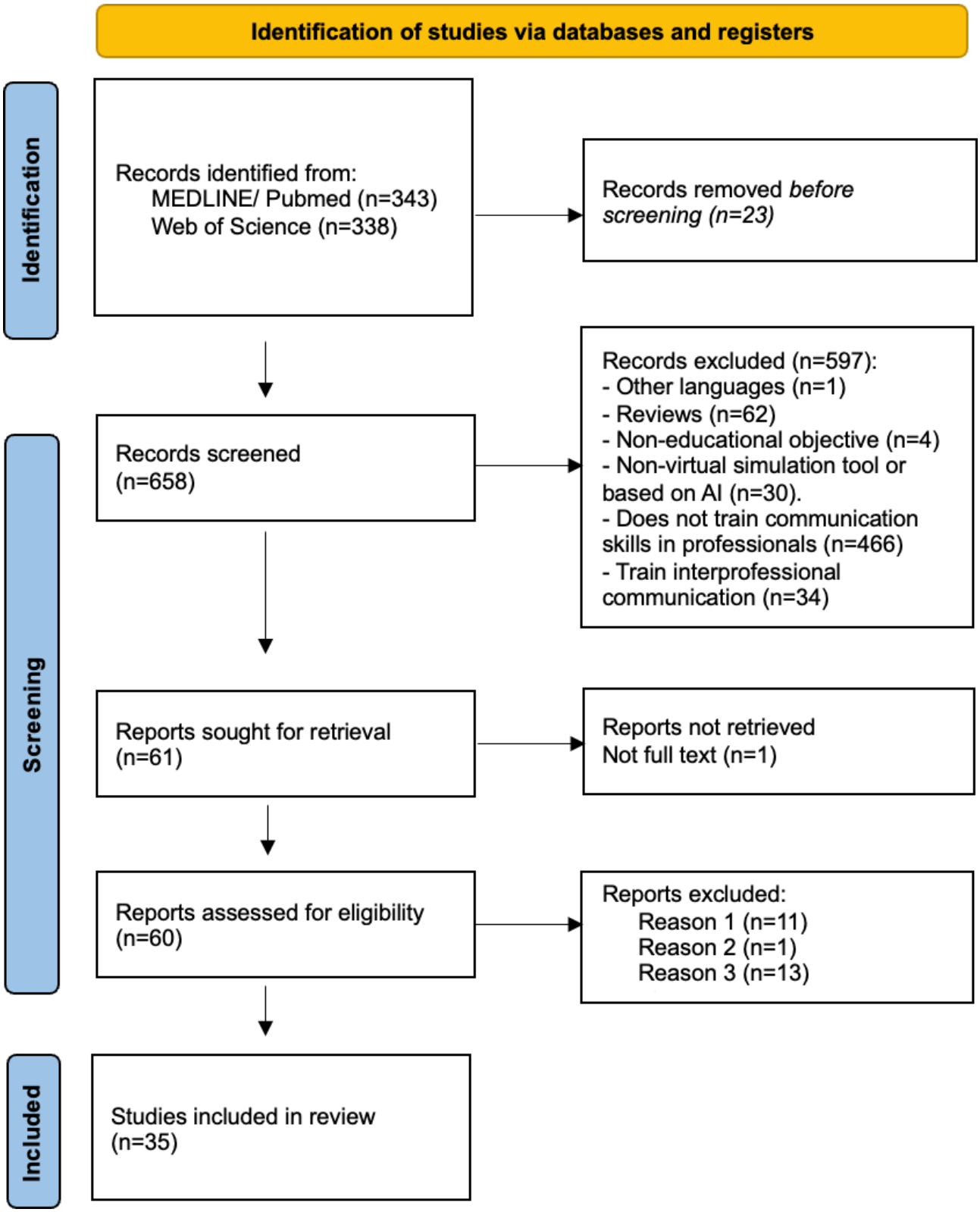
Data related to the characteristics of the studies (publication year, country, language, objective, and type) as well as data related to the outcome of the use of the digital/virtual training tool for improving communication skills in health care professionals were recorded. Specifically, we noted the tool name, training language, learning objective, degree of learning autonomy (fully autonomous vs instructor-mediated training), patient type (avatar/doll, virtual patient with a human-like appearance, real person, etc), type of answers given by the trainee (written or oral conversation), and type of technology used.

Results

Overview

The studies were manually screened and coded. Our search of PubMed and the Web of Science produced 681 records, of which 23 duplicates were eliminated. During the screening process, 2 of the authors independently analyzed 658 studies based on their titles and abstracts (Figure 1). After this initial screening, the full texts of 61 records were obtained for analysis. We requested the full texts of a further 2 articles from the corresponding authors by email and through ResearchGate; of these, we included 1 in this review. Of these 60 studies, 25 were excluded because they did not meet the inclusion criteria. Specifically, 11 articles had not directly trained clinical communication skills with patients (criterion 1), 1 had not studied virtual training (criterion 2), and 13 had not used a tool designed for training purposes (criterion 3). Therefore, a total of 35 articles were included in the review. Finally, one of the authors extracted the relevant data from these 35 studies and entered them into a database following the coding manual we had prepared for this purpose.

Figure 1. PRISMA flowchart. Reason 1: articles not directly related to training clinical communication skills with patients; reason 2: did not study virtual training; reason 3: did not use a tool designed for training purposes. AI: artificial intelligence; PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.



Characteristics of the Studies Included

A total of 35 articles were obtained that had developed and/or applied a virtual environment for training communication skills that would be directed toward patients; overall, 43% (n=15) were articles published in the United States and 17% (n=6) were from Australia, with the remaining articles having been published in Europe and Asia (Table 1). All the articles had been published in English and their objectives are shown in Table 1.

Table . Description of the studies (N=35).

Articles	Country	Language	Objective
Ali et al [21], 2020	United States	English	Describe the iterative participatory design of SOPHIE, an online virtual patient for "practice" based on feedback from sensitive conversations between patients and clinicians and discuss an initial qualitative evaluation of the system by professional end users.
Bánszki et al [22], 2018	Australia	English	Explore a novice clinical educator's experience in training essential communication and interpersonal skills using a virtual patient.
Bearman and Cesnik [23], 2001	Australia	English	Assess students' attitudes toward learning communication skills through a virtual patient; compare the acceptability of the 2 distinct types of virtual patient designs.
Bearman et al [24], 2001	Australia	English	Compare 2 types of virtual patients to understand how different virtual patient designs affect the student learning experience.
Bearman [25], 2003	Australia	English	Explore the students' experience with the virtual patient.
Borja-Hart et al [26], 2019	United States	English	Assess students' confidence and impressions when using their communication skills with a virtual patient and evaluate their competencies in the use of this technology.
Chae et al [27], 2023	Korea	English	The purpose of this study was to describe the development of SimCARE and evaluate the feasibility of its use in nursing education.
Courteille et al [28], 2014	Sweden	English	To investigate the dynamics and congruence of interpersonal behaviors and socioemotional interaction exhibited during the learning experience in a virtual patient, and to evaluate which interaction design features contribute most to behavioral and affective engagement in the medical student.
Deladisma et al [29], 2008	United States	English	Develop a virtual training environment system that can be accessed independently.
Dickerson et al [30], 2006	United States	English	Provide information about the advantages and disadvantages of using synthesized speech and evaluate the fidelity necessary for the training of communication skills.
Du et al [31], 2022	China	English	To evaluate the history-taking skills of nursing undergraduates using a virtual standardized patient, and to explore its independent influencing factors.
Guetterman et al [32], 2019	United States	English	To investigate the differential effects of a virtual patient–based simulation developed to train health care professionals in empathetic patient-provider and interprofessional communication.

Articles	Country	Language	Objective
Hwang et al [33], 2022	Taiwan, Japan	English	A virtual patient–based social learning approach is proposed to enhance nursing students' performance and clinical judgment in education programs.
Jacklin et al [34], 2018	United Kingdom	English	Create a virtual patient that simulates a primary care consultation, offering the opportunity to practice decision-making. A second objective was to involve patients in the design of a virtual patient simulation and inform the design process.
Jacklin et al [35], 2021	United Kingdom	English	This study aims to evaluate a virtual patient workshop for medical students aimed at developing the communication skills required for shared decision-making.
Kleinsmith et al [2], 2015	United States	English	Develop an understanding of whether students can respond empathically to expressions of concern from a virtual patient.
Lok [36], 2006	United States	English	Teach communication skills using virtual humans.
Maicher et al [37], 2019	United States	English	Describe a virtual standardized patient system that allows students to practice their history-taking skills and receive immediate feedback.
Mayor Silva et al [38], 2023	Spain	English	The objective was to develop a virtual reality simulator to improve communication skills and compare its results with a traditional workshop based on cases and theoretical content explained through video.
Nakagawa et al [39], 2022	Japan	English	The objective structured clinical examination is among validated approaches used to assess clinical competence through structured and practical evaluation.
Ochs et al [40], 2019	France	English	Evaluate the virtual reality training platform in which the user experience is analyzed based on the virtual environment.
Perez et al [41], 2022	United States	English	The purpose of this study was to explore the use of virtual simulation to experience difficult conversations and to evaluate differences in perceptions between nurse educator, family nurse practitioner, and nurse anesthesia students.
Plass et al [42], 2022	Germany	English	The purpose of this study is to evaluate the effectiveness of a brief virtual role-play motivational interviewing training program on motivational interviewing knowledge and skills in first-year undergraduate medical students, making use of both a pretest and a then-test (retrospective pre-test) to check for response shift in evaluating the educational intervention.

Articles	Country	Language	Objective
Quail et al [12], 2016	Australia	English	Investigate students' communication skills, knowledge, confidence, and empathy in simulated and traditional learning environments.
Real et al [43], 2017	United States	English	Develop an immersive virtual reality curriculum on addressing flu vaccine hesitancy using Kern's 6-step approach to curriculum design. The goal of the program was to teach best communication practices in cases of questions about the flu vaccine.
Real et al [44], 2017	United States	English	Create an immersive virtual reality curriculum to teach pediatric residents communication skills when discussing flu vaccination. Compare effectiveness with a control group.
Real et al [45], 2022	United States	English	Examined the acceptability and tolerability of the approach and the impact of deliberate practice using virtual reality simulations on clinicians' confidence related to shared decision-making communication skills.
Rouleau et al [46], 2022	Canada	English	This study aimed to assess the acceptability of a virtual patient simulation to improve nurses' relational skills in a continuing education context.
Sapkaroski et al [47], 2022	Australia	English	The aim of this study was to establish whether the mode of delivery, virtual reality simulated learning environments versus clinical role-play, could have a measurable effect on clinical empathic communication skills for magnetic resonance imaging scenarios.
Sezer and Sezer [48], 2019	Turkey	English	Design, develop, and evaluate a 3D virtual patient application that can move, has voice and lip synchronization, allows written communication, and is supported by a solid scenario to improve students' communication skills.
Şimşek Çetinkaya et al [49], 2022	Turkey	English	This study aimed to determine the effectiveness of 2 simulation types used for family planning consultation of midwifery students and to compare these methods.
Shorey et al [50], 2019	Singapore	English	Develop and evaluate the use of virtual patients to better prepare undergraduate nursing students to communicate with real-life patients, their families, and other health care professionals during their clinical stays.
Shorey et al [51], 2020	Singapore	English	To examine user attitudes and experiences and clinical facilitators' perspectives on student performance in the clinical environment following virtual patient training.

Articles	Country	Language	Objective
Shorey et al [52], 2023	United States	English	This study aimed to evaluate the effectiveness of this theory-based virtual intervention on nursing students' learning attitudes, communication self-efficacy, and clinical performance.
Stevens et al [53], 2006	United States	English	Create an interactive virtual clinical scenario of a patient with acute abdominal pain to teach medical students history-taking and communication techniques.

Features of the Virtual Tools

After reading the full text of the 35 articles, we identified 24 different learning tools that had been developed to train communication skills in students or health professionals (Table 2). Most of them (n=15; 62%) had provided training in English [2,21,22,24,26,28,29,32,34,37,41,43,46,47,52]. Regarding the learning objective of the virtual environment, 42% (n=10) aimed to train communication skills in the specific contexts of a clinical history and/or anamnesis interview [2,29,31,33,35,37,42,46,48,52], 42% (n=10) taught general communication skills [22,24,26-28,38,39,41,47,49], and 8% (n=2) covered giving bad news [21,40]. There was also a tool that had been specifically developed to train communication skills to address flu vaccination hesitancy [43-45]. Another tool that had been used to train communication skills focused on empathy is also worth highlighting [32].

Table . Virtual tools and their characteristics (n=24 tools).

Articles	Tool name	Language	Study purpose	Degree of learning autonomy	Patient type	Type of student responses during training	Type of technology used
Ali et al [21], 2020	SOPHIE	English	Train communication skills for the delivery of bad news. Aimed at health professionals.	Autonomous	Virtual patient with the appearance of a person. Responded with a voice. The entire transcript can be seen.	Oral conversation	Artificial intelligence
Bánszki et al [22], 2018; Quail et al [12], 2016	Not specified	English	Training communication skills. Aimed at health care students.	An instructor mediated the training.	Virtual patient with the appearance of a person. Responded with a voice.	Oral conversation	The instructor was in another room where they controlled everything and responded in the simulated interaction.
Bearman and Cesnik [23], 2001; Bearman et al [24], 2001; Bearman [25], 2003	Not specified	English	Training in communication skills. Aimed at medical students.	Autonomous	Real person speaking. Viewing of recorded videos.	Written. Choice of 3 or 4 written response options available after each video. The authors developed 2 types of responses to compare which was more effective: narrative (detailed communicative structures) or problem-solving (labels with possible actions).	A total of 154 recorded videos. The next video shown was adjusted depending on the response given. Therefore, the virtual patient became satisfied according to responses chosen by the student.
Borja-Hart et al [26], 2019	Used *Shadow Health* from Elsevier	English	Training in communications skills. Aimed at pharmacy students.	Autonomous	Virtual patient with the appearance of a person. Responded with a voice.	Natural language (written and spoken). Students could choose the interaction they would carry out: ask, empathize, or educate.	*Shadow Health* is simulation software that generates different scenarios. The article did not explain any more about the technology used.
Chae et al [27], 2023	SimCARE	Korean	Training in intercultural communication skills. Aimed at nursing students.	Autonomous	Virtual patient with the appearance of a person. Responded with a voice.	They selected a written response from among those on offer.	A virtual reality headset. The authors described the technology used to generate the 3D graphics (Unity 2019.4.0f1 game engine), avatars (DAZ 3D software), and avatar animation (iClone 7).

Articles	Tool name	Language	Study purpose	Degree of learning autonomy	Patient type	Type of student responses during training	Type of technology used
Courteille et al [28], 2014	Not specified	English and Swedish	Training in communication skills. Aimed at medical students.	Autonomous	Real person speaking. Viewing of recorded videos.	Written. Students replied in text written in natural language.	Interactive Simulation of Patients. A database with 200 videos for each case, allowing the simulator to respond according to the question posed by the student.
Deladisma et al [29], 2008; Dickerson et al [30], 2006; Lok [36], 2006; Stevens et al [53], 2006	Not specified	English	Training in communication skills and anamnesis techniques. Aimed at medical students.	Autonomous but with availability of additional resources. The technology that drives this interaction largely consisted of commodity hardware and software: 2 desktop computers, 2 cameras, a data projector, and a wireless microphone.	Virtual patient with the appearance of a person (an avatar called Diana) who spoke and produced natural gestures. The authors developed 2 types of communication for the avatar to study which one was more effective: real recorded communication or virtual communication.	Oral conversation. The students could speak using natural language. The software also detected various gestures.	The speech recognition worked using *Dragon Naturally Speaking* by Scansoft, which is a database developed with content organized in semantic categories to detect the communicative structures used by the students.
Du et al [31], 2022	University A Virtual Patient (UA-VP, 2021)	Chinese	Training in communication skills to carry out a nursing evaluation by following Gordon's Functional Patterns.	Autonomous	A virtual patient with the appearance of a person. Responded with text based on a predefined chat.	Written and oral conversation	Recognizes structures and offers feedback based on the uploaded chat scripts (as bullet points and not reflecting the most important part of the interaction). Used WeChat, a social media app.
Guetterman et al [32], 2019	Used MPathic-VR	English	Trained empathic communication skills. Aimed at medical students.	Autonomous	Virtual patient with the appearance of a person. Responded with a voice.	Oral conversation. It also detected gestures and movements.	Artificial intelligence.
Hwang et al [33], 2022	Not specified	Chinese	Trained students in diagnosis and treatment and has a specific medical history module which trains communication skills.	Autonomous	Virtual patient with the appearance of a person. Responded with voice and text.	Did not specify	Learning system designed as a decision tree.

Articles	Tool name	Language	Study purpose	Degree of learning autonomy	Patient type	Type of student responses during training	Type of technology used
Jacklin et al [34], 2018; Jacklin et al [35], 2021	Not specified	English	Training in communication skills for shared decision-making during clinical interviews. Aimed at medical and/or pharmacy students.	Autonomous	Virtual patient with the appearance of a person. Responded through a voice and with gestures.	Written text. Choice of 3 answer options.	A web-based virtual patient simulator.
Kleinsmith et al [2], 2015	Neurological Examination Rehearsal Virtual Environment	English	Trained communication skills for use during clinical interviews. Aimed at nursing students.	Autonomous	Virtual patient with the appearance of a person. A virtual patient responded with a voice and through text.	Written. The student inserted text written in natural language.	Virtual People Factory. A database used by the simulator to respond based on the student's question.
Maicher et al [37], 2019	Not specified	English	Trained skills for performing an anamnesis (to collect medical information). It does not address communicative listening strategies such as empathy. Aimed at medical students.	Autonomous	Virtual patient with the appearance of a person. Responded with voice and text.	Oral conversation. Text could also be written.	Artificial intelligence. The open-source natural language processing engine ChatScript is used for the conversion element. Unity gaming platform.
Mayor Silva et al [38], 2023	Not specified	Spanish	Training in communication skills. Aimed at nursing students.	An instructor mediated the evaluation.	Not specified	Not specified	A virtual reality headset.
Nakagawa et al [39], 2022	Not specified	Japanese	Trained communication skills such as desire suppression, expectation acceptance, facial expression, emotional communication, dominance, maintaining relationships, and dealing with disagreements. Aimed at pharmacy students.	Autonomous	A chatbot. Written and oral.	Oral conversation in natural language	Artificial intelligence. If the artificial intelligence did not detect the keywords, the conversation did not continue. There was no direct feedback.
Ochs et al [40], 2019	ACORFORMed	French	Training in the delivery of bad news. Aimed at medical practitioners (students and professionals).	Autonomous in some functions (eg, dialogue generator). In others (eg, categorizing the response and sending it to the simulator), the instructor mediated the learning.	Virtual patient with the appearance of a person. Responded with a voice.	Oral conversation	A virtual reality headset. The instructor categorized the response using a previously coded database and sent that information to the simulator.

Articles	Tool name	Language	Study purpose	Degree of learning autonomy	Patient type	Type of student responses during training	Type of technology used
Perez et al [41], 2022	Used the Mursion tool	English	Trained communication skills for use in difficult conversations. Aimed at nursing students.	Autonomous	Virtual patient with the appearance of a person. Responded with a voice.	Oral conversation in natural language.	Artificial intelligence (using the Mursion tool).
Plass et al [42], 2022	Used the Kognito Conversarion Platform	German	Training in person-centered communication skills for motivational interviewing. Aimed at medical students.	Autonomous	Virtual patient with the appearance of a person. Responded with a voice.	Select between different answer options.	Artificial intelligence (using the Kognito Conversation Platform).
Real et al [43], 2017; Real et al [45], 2022; Real et al [44], 2017	Not specified	English	Training in communication skills to inform patients about vaccination. Aimed at medical residents.	An instructor mediated the training.	Virtual patient with the appearance of a person. Responded through a voice and with gestures.	Oral conversation and natural language.	Unity gaming platform. A virtual reality headset.
Rouleau et al [46], 2022	Not specified	French, English	Training in nursing relational skills for use in motivational interviews.	Autonomous	Virtual patient with the appearance of a person. Responded with a voice.	Select between different answer options	Used the MedicActiV platform
Sapkaroski et al [47], 2022	Not specified	English	Training in communication skills. Aimed at medical students.	Autonomous	Virtual patient with the appearance of a person. Responded with voice and text.	Select from among answer options. This part of the case simulation was mandatory. It was also capable of natural language oral conversation and the ability to ask alternative questions was optional.	Clinical Education Training Solution virtual reality clinic software using the Oculus Rift CV1 virtual reality headset.
Sezer and Sezer [48], 2019	Not specified	Turkish	Training in basic communication skills for use in a medical interview. Aimed at training health care students.	Autonomous	Virtual patient with the appearance of a person. Responded with a voice and in writing.	Natural written text	Virtual People Factory for avatar and simulation generation. The scenario was created in Unity 3DTM. Different variations of the simulation interventions the students could apply at each stage were included and these answer combinations were compared to the closest preprogrammed scenario to give an answer.

Articles	Tool name	Language	Study purpose	Degree of learning autonomy	Patient type	Type of student responses during training	Type of technology used
Şimşek Çetinkaya et al [49], 2022	Not specified	Turkish	Training in communication skills for use in a family planning consultation. Aimed at midwifery students.	The instructor offered feedback after watching the simulation.	The patient type was not specified. Responded with a voice.	Oral conversation	Not specified
Shorey et al [50], 2019; Shorey et al [51], 2020; Shorey et al [52], 2023	Virtual Counselling Application using Artificial Intelligence	English	Trained basic communication skills for use in an interview. Aimed at nursing students.	Autonomous	Virtual patient with the appearance of a person. Responded with a voice and in writing.	Oral conversation in natural language	Artificial intelligence. Used the Dialogflow chatbot from Google Cloud to store and process natural language. The scenario was created in Unity 3D.

Several major virtual tools were identified in this review for training communication skills in health care professionals. SOPHIE [21] is a tool designed to train the delivery of bad news using a virtual patient that interacts through oral conversations, leveraging AI. Shadow Health [26] focuses on communication skills for pharmacy students, allowing both written and spoken interactions with a virtual patient. SimCARE [27] is a virtual reality–based tool aimed at nursing students, training intercultural communication skills through animated avatars. MPathic-VR [32] trains medical students in empathic communication, featuring virtual patients that respond with voice and detect nonverbal cues like gestures. ACORFORMed [40] trains medical practitioners in delivering bad news through virtual reality interactions with a virtual patient. Mursion [41] is designed for nursing students to practice difficult conversations using natural language processing for realistic interactions, while the Kognito Conversation Platform [42] supports motivational interviewing through person-centered communication training with virtual patients. VCAAI [50-52] trains basic communication skills in nursing interviews. These tools highlight the diversity of approaches in the use of virtual patients for communication training. Finally, 14 virtual tools did not specify their name.

Some (n=19, 79%) of the tools allowed students to train completely autonomously, whereas 21% (n=5) required an online instructor to mediate the training and respond during the interactions [22,39,40,44,49]. One of the tools could be defined as partially autonomous because a trained instructor had to perform some of the functions [40]. Regarding the patient type used for the training, the vast majority of the tools used virtual patients (n=19; 79%) with the appearance of a real person [2,21,22,26,29,31-33,35,37,40-42,44,46-48,51]. Of these, 95% (18/19) responded with a voice (18/24, 75%), except for the tool published by Du et al [46]. Two tools (8%) used videos recorded with real people [24,28].

Regarding the types of responses the user could give during the training, almost half of the tools analyzed (n=11, 45%) allowed the user to respond orally using natural language [21,22,26,29,31,32,37,39,41,44,49,51]. Shadow Health [26], for example, offers both written and spoken interactions, while SOPHIE [21] focuses solely on oral communication.

Discussion

This study reviews and analyzes the 24 virtual simulation tools available for training communication skills in health care professionals, assessing their characteristics, levels of immersion, and the autonomy they provide in learning processes. Although virtual simulation tools have shown significant growth in recent years, driven by technological advances, the review identified a high degree of heterogeneity in the approaches, technologies, and interaction methods used. This variety has made it challenging to standardize and effectively integrate these tools into consistent training plans. Most tools rely on virtual patients with a limited range of interaction capabilities, and very few offer fully immersive experiences that mimic real-world clinical communication. Furthermore, limited accessibility to tools in languages other than English, as well as a lack of high-fidelity technologies for simulating realistic, natural language–based conversations, continue to pose significant challenges. Considering these challenges, this review highlights several key findings regarding the applications of virtual environments to enhance communication skills training that will be detailed in the following paragraphs.

First, it is important to highlight the large number of different applications we identified that have been used to improve communication skills (either in basic or more specific situations) through virtual environments. Similarly, other reviews have also concluded that the use of virtual patients for clinical communication training has grown exponentially over the last decade [17,18], which has been driven by rapid technological advances [54], also providing further evidence of the benefits associated with this type of resource [18]. In fact, this work has included 13 new virtual simulation environments developed based on the published review by Battegazzorre et al [17].

Most of the applications we considered in this review used English, which could represent an obstacle for professionals and students who do not know this language. Indeed, only one of the tools identified used Spanish and in this case, it was also mediated by an instructor, thereby making it difficult for students to use it autonomously and independently [38] Therefore, there is still a long way to go to make these tools highly accessible at an international level. Regarding the more technical characteristics, we observed visible heterogeneity in the types of technologies used, including in the different types of patients used for training—for example, the use of chatbots, images, and/or recordings of real people and virtual patients. However, our results showed that almost all the applications we identified had designed virtual environments using virtual patients that looked like a person and could vocally respond to and receive oral responses to simulate a real conversation [21,22,26,29,32,37,39,41,44,49,51]. A key implementation across the tools was the use of natural language processing to simulate realistic conversations.

Training in simulation environments that assume an appropriate level of fidelity (a 3D term that includes physical/environmental, psychological, and conceptual elements) increases realism [55] and influences learning engagement [56]. For example, in their systematic review, Kaplonyi et al [1] reflected how simulations with the use of standardized patients are considered realistic environments and an effective means for learning communication skills. Indeed, the academic literature proposes that virtual patients can be used as a complementary alternative to working with standardized patients [57] and can represent patients in a realistic clinical environment [17] to effectively help students to acquire or improve their communication skills [18]. Nonetheless, it will be important for future lines of research to use standardized tests to evaluate the beneficial effects of training with this type of virtual tool before fully integrating them into training plans [18,54].

In terms of the fidelity of these tools, increasing the immersion of virtual simulations—defined as the psychological state of the perception of being inside or surrounded by something [58]—by using virtual patients with natural language processing and auditory and visual behavior [17,59] is positively related to better communication skills performance [17,19]. However, we must not forget that realism and authenticity, which are both relevant factors in design, are not only achieved through physical resemblance (physical fidelity) but also require other fidelity factors [19]. Hence, future research in this field should be designed to also consider conceptual fidelity (scenarios and cases consistent with reality) and psychological fidelity (the ability to provoke emotional responses like reality) in the design of virtual simulations [19], factors that were not considered in this review.

Nevertheless, we identified 2 tools that had specifically used recordings of real people in the clinical situations being trained, which could have generated a greater feeling of immersion among students because of the increased physical, auditory, and visual fidelity of these tools. However, in the interactions with the simulation developed by Bearman et al [23], users had to respond from a pool of pre-established options, limiting the immersion experience because the participant was unable to develop their own communication skills in the way they would have to when facing real situations. In a tool developed by Courteille et al [28], although the user had been allowed to issue a natural language response, this had to be done in writing, which also reduced the degree of reality and spontaneity one would expect from a real conversation. Therefore, highly immersive technologies must be designed to overcome these ongoing technological challenges, such as how to integrate effective natural language processing systems and natural conversation flows into these tools [60] and how to best capture nonverbal communication [17,18]. For example, in this review, we only identified 2 applications that could detect gestures and/or emotions [29,32].

Of note, most of the tools we identified were based on autonomous learning and therefore represented promising applications with potential great benefits such as high accessibility levels, the possibility of repeating the experience multiple times, and cost reduction once running [16,17]. In this sense, technological advances that can integrate systems that provide feedback to participants—such as AI and machine learning (ML)—without the need for an instructor/teacher to mediate the learning stand out in particular [60]. For example, compared to a previous literature review [18], we found more tools in which the feedback was provided by the virtual system itself. However, as discussed, despite cataloging the existence of various patient simulation tools with interesting characteristics, we did not identify any that simultaneously integrated the use of a real person (a standardized patient) with the objective of increasing the environmental fidelity to allow the user to train through an oral conversation using natural language and using complex technology, such as AI and ML, with the ability to detect, encode, and respond to complex communication structures [60].

Finally, it is important to note that there were several limitations to this review. First, we only consulted 2 medical databases—MEDLINE/PubMed and the Web of Science. Despite being a health science–specific database and a multidisciplinary database, respectively, having replicated the search in more technological databases may have provided some additional studies for consideration. Therefore, it is possible we did not recover all the relevant records on virtual simulation tools to train communication skills in health care professionals registered in the academic literature. Second, there is still inadequate standardization in academic and scientific fields regarding the term “virtual simulation” [16,55,61]. Thus, different terms in the academic literature are all used to refer to the concept of virtual simulation including “serious games,” “virtual worlds,” “virtual patients,” and “virtual reality,” [55] which may have also caused us to miss certain relevant records.

In conclusion, this review identified and analyzed the 24 main virtual tools described in the academic literature that have been used to date to train communication skills in the context of health sciences. The high heterogeneity in terms of their characteristics means that tools based on AI and ML that contribute to training both students and practicing health professionals with as high a fidelity as possible to real life remain to be developed. Although many tools offer a degree of realism, few incorporate advanced features like AI-driven

conversational flows or nonverbal cue detection, limiting the immersive experience. This highlights a need for further development to create more effective training environments. Addressing these gaps requires future innovations that integrate natural language processing and other advanced capabilities to enhance both the realism and educational value of virtual simulations.

Acknowledgments

This study is an Erasmus project funded by the European Union (Strategic Partnerships in Higher Education [KA203], in the Call for proposals Cooperation for Innovation and exchange of good practice 2020; grant agreement 2020--1-ES01-KA203-082566).

Conflicts of Interest

None declared.

Checklist 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File, 90 KB - mededu_v11i1e63082_app1.pdf]

References

1. Kaplonyi J, Bowles KA, Nestel D, et al. Understanding the impact of simulated patients on health care learners' communication skills: a systematic review. Med Educ 2017 Dec;51(12):1209-1219. [doi: 10.1111/medu.13387] [Medline: 28833360]
2. Kleinsmith A, Rivera-Gutierrez D, Finney G, Cendan J, Lok B. Understanding empathy training with virtual patients. Comput Human Behav 2015 Nov 1;52:151-158. [doi: 10.1016/j.chb.2015.05.033] [Medline: 26166942]
3. Stehr P, Reifegerste D, Rossmann C, Caspar K, Schulze A, Lindemann AK. Effective communication with caregivers to prevent unintentional injuries in children under seven years. A systematic review. Patient Educ Couns 2022 Aug;105(8):2721-2730. [doi: 10.1016/j.pec.2022.04.015] [Medline: 35537900]
4. Stacey D, Légaré F, Lewis K, et al. Decision aids for people facing health treatment or screening decisions. Cochrane Database Syst Rev 2017 Apr 12;4(4):CD001431. [doi: 10.1002/14651858.CD001431.pub5] [Medline: 28402085]
5. Boissy A, Windover AK, Bokar D, et al. Communication skills training for physicians improves patient satisfaction. J Gen Intern Med 2016 Jul;31(7):755-761. [doi: 10.1007/s11606-016-3597-2] [Medline: 26921153]
6. Ammentorp J, Graugaard LT, Lau ME, Andersen TP, Waidtløw K, Kofoed PE. Mandatory communication training of all employees with patient contact. Patient Educ Couns 2014 Jun;95(3):429-432. [doi: 10.1016/j.pec.2014.03.005] [Medline: 24666773]
7. Agarwal R, Sands DZ, Schneider JD. Quantifying the economic impact of communication inefficiencies in U.S. hospitals. J Healthc Manag 2010;55(4):265-281. [Medline: 20812527]
8. Synnot A, Bragge P, Lowe D, et al. Research priorities in health communication and participation: international survey of consumers and other stakeholders. BMJ Open 2018 May 8;8(5):e019481. [doi: 10.1136/bmjopen-2017-019481] [Medline: 29739780]
9. Gutiérrez-Puertas L, Márquez-Hernández VV, Gutiérrez-Puertas V, Granados-Gámez G, Aguilera-Manrique G. Educational interventions for nursing students to develop communication skills with patients: a systematic review. Int J Environ Res Public Health 2020 Mar 26;17(7):2241. [doi: 10.3390/ijerph17072241] [Medline: 32225038]
10. Lewis KL, Bohnert CA, Gammon WL, et al. The Association of Standardized Patient Educators (ASPE) Standards of Best Practice (SOBP). Adv Simul (Lond) 2017;2:10. [doi: 10.1186/s41077-017-0043-4] [Medline: 29450011]
11. Nestel D, Tabak D, Tierney T, et al. Key challenges in simulated patient programs: an international comparative case study. BMC Med Educ 2011 Sep 25;11:69. [doi: 10.1186/1472-6920-11-69] [Medline: 21943295]
12. Quail M, Brundage SB, Spitalnick J, Allen PJ, Beilby J. Student self-reported communication skills, knowledge and confidence across standardised patient, virtual and traditional clinical learning environments. BMC Med Educ 2016 Feb 27;16(72):73. [doi: 10.1186/s12909-016-0577-5] [Medline: 26919838]
13. Padilha JM, Machado PP, Ribeiro AL, Ramos JL. Clinical virtual simulation in nursing education. Clin Simul Nurs 2018 Feb;15:13-18. [doi: 10.1016/j.ecns.2017.09.005]
14. Yang H, Xiao X, Wu X, et al. Virtual standardized patients versus traditional academic training for improving clinical competence among traditional Chinese medicine students: prospective randomized controlled trial. J Med Internet Res 2023 Sep 20;25:e43763. [doi: 10.2196/43763] [Medline: 37728989]
15. Urresti-Gundlach M, Tolks D, Kiessling C, Wagner-Menghin M, Härtl A, Hege I. Do virtual patients prepare medical students for the real world? Development and application of a framework to compare a virtual patient collection with population data. BMC Med Educ 2017 Sep 22;17(1):174. [doi: 10.1186/s12909-017-1013-1] [Medline: 28938884]

16. Plotzky C, Lindwedel U, Sorber M, et al. Virtual reality simulations in nurse education: a systematic mapping review. Nurse Educ Today 2021 Jun;101:104868. [doi: 10.1016/j.nedt.2021.104868] [Medline: 33798987]
17. Battegazzorre E, Bottino A, Lamberti F. Training medical communication skills with virtual patients: literature review and directions for future research. In: Intelligent Technologies for Interactive Entertainment: Springer International Publishing; 2021:207-226. [doi: 10.1007/978-3-030-76426-5_14]
18. Lee J, Kim H, Kim KH, Jung D, Jowsey T, Webster CS. Effective virtual patient simulators for medical communication training: a systematic review. Med Educ 2020 Sep;54(9):786-795. [doi: 10.1111/medu.14152] [Medline: 32162355]
19. Peddle M, Bearman M, Nestel D. Virtual patients and nontechnical skills in undergraduate health professional education: an integrative review. Clin Simul Nurs 2016 Sep;12(9):400-410. [doi: 10.1016/j.ecns.2016.04.004]
20. Fuentes A. Reseña de sitio web: Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). Declaración PRISMA 2020 [Article in Spanish]. R Est Inv Psico y Educ 2022;9(2):323-327. [doi: 10.17979/reipe.2022.9.2.9368]
21. Ali MR, Sen T, Kane B, et al. Novel computational linguistic measures, dialogue system and the development of SOPHIE: standardized online patient for healthcare interaction education. IEEE Trans Affective Comput 2020;14(1):223-235. [doi: 10.1109/TAFFC.2021.3054717]
22. Banszki F, Beilby J, Quail M, Allen P, Brundage S, Spitalnick J. A clinical educator's experience using a virtual patient to teach communication and interpersonal skills. AJET 2018;34(3):60-73. [doi: 10.14742/ajet.3296]
23. Bearman M, Cesnik B. Comparing student attitudes to different models of the same virtual patient. Stud Health Technol Inform 2001;84(Pt 2):1004-1008. [Medline: 11604882]
24. Bearman M, Cesnik B, Liddell M. Random comparison of "virtual patient" models in the context of teaching clinical communication skills. Med Educ 2001 Sep;35(9):824-832. [doi: 10.1046/j.1365-2923.2001.00999.x] [Medline: 11555219]
25. Bearman M. Is virtual the same as real? Medical students' experiences of a virtual patient. Acad Med 2003 May;78(5):538-545. [doi: 10.1097/00001888-200305000-00021] [Medline: 12742794]
26. Borja-Hart NL, Spivey CA, George CM. Use of virtual patient software to assess student confidence and ability in communication skills and virtual patient impression: a mixed-methods approach. Curr Pharm Teach Learn 2019 Jul;11(7):710-718. [doi: 10.1016/j.cptl.2019.03.009] [Medline: 31227094]
27. Chae D, Kim J, Kim K, Ryu J, Asami K, Doorenbos AZ. An immersive virtual reality simulation for cross-cultural communication skills: development and feasibility. Clin Simul Nurs 2023 Apr;77:13-22. [doi: 10.1016/j.ecns.2023.01.005]
28. Courteille O, Josephson A, Larsson LO. Interpersonal behaviors and socioemotional interaction of medical students in a virtual clinical encounter. BMC Med Educ 2014 Apr 1;14(1):64. [doi: 10.1186/1472-6920-14-64] [Medline: 24685070]
29. Deladisma AM, Johnsen K, Raij A, et al. Medical student satisfaction using a virtual patient system to learn history-taking communication skills. Stud Health Technol Inform 2008;132:101-105. [Medline: 18391266]
30. Dickerson R, Johnsen K, Raij A, et al. Virtual patients: assessment of synthesized versus recorded speech. Stud Health Technol Inform 2006;119:114-119. [Medline: 16404028]
31. Du J, Zhu X, Wang J, et al. History-taking level and its influencing factors among nursing undergraduates based on the virtual standardized patient testing results: cross sectional study. Nurse Educ Today 2022 Apr;111:105312. [doi: 10.1016/j.nedt.2022.105312] [Medline: 35287063]
32. Guetterman TC, Sakakibara R, Baireddy S, et al. Medical students' experiences and outcomes using a virtual human simulation to improve communication skills: mixed methods study. J Med Internet Res 2019 Nov 27;21(11):e15459. [doi: 10.2196/15459] [Medline: 31774400]
33. Hwang GJ, Chang CY, Ogata H. The effectiveness of the virtual patient-based social learning approach in undergraduate nursing education: a quasi-experimental study. Nurse Educ Today 2022 Jan;108:105164. [doi: 10.1016/j.nedt.2021.105164] [Medline: 34627030]
34. Jacklin S, Maskrey N, Chapman S. Improving shared decision making between patients and clinicians: design and development of a virtual patient simulation tool. JMIR Med Educ 2018 Nov 6;4(2):e10088. [doi: 10.2196/10088] [Medline: 30401667]
35. Jacklin S, Maskrey N, Chapman S. Shared decision-making with a virtual patient in medical education: mixed methods evaluation study. JMIR Med Educ 2021 Jun 10;7(2):e22745. [doi: 10.2196/22745] [Medline: 34110299]
36. Lok B. Teaching communication skills with virtual humans. IEEE Comput Graph Appl 2006;26(3):10-13. [doi: 10.1109/mcg.2006.68] [Medline: 16711211]
37. Maicher KR, Zimmerman L, Wilcox B, et al. Using virtual standardized patients to accurately assess information gathering skills in medical students. Med Teach 2019 Sep;41(9):1053-1059. [doi: 10.1080/0142159X.2019.1616683] [Medline: 31230496]
38. Mayor Silva LI, Caballero de la Calle R, Cuevas-Budhart MA, Martin Martin JO, Blanco Rodriguez JM, Gómez Del Pulgar García Madrid M. Development of communication skills through virtual reality on nursing school students: clinical trial. Comput Inform Nurs 2023 Jan 1;41(1):24-30. [doi: 10.1097/CIN.0000000000000866] [Medline: 35363632]
39. Nakagawa N, Odanaka K, Ohara H, Kisara S. Communication training for pharmacy students with standard patients using artificial intelligence. Curr Pharm Teach Learn 2022 Jul;14(7):854-862. [doi: 10.1016/j.cptl.2022.06.021] [Medline: 35914846]

40. Ochs M, Mestre D, de Montcheuil G, et al. Training doctors' social skills to break bad news: evaluation of the impact of virtual environment displays on the sense of presence. J Multimodal User Interfaces 2019 Mar;13(1):41-51. [doi: 10.1007/s12193-018-0289-8]
41. Perez A, Gaehle K, Sobczak B, Stein K. Virtual simulation as a learning tool for teaching graduate nursing students to manage difficult conversations. Clin Simul Nurs 2022 Jan;62:66-72. [doi: 10.1016/j.ecns.2021.10.003]
42. Plass AM, Covic A, Lohrberg L, Albright G, Goldman R, Von Steinbüchel N. Effectiveness of a minimal virtual motivational interviewing training for first years medical students: differentiating between pre-test and then-test. Patient Educ Couns 2022 Jun;105(6):1457-1462. [doi: 10.1016/j.pec.2021.09.020] [Medline: 34598801]
43. Real FJ, DeBlasio D, Ollberding NJ, et al. Resident perspectives on communication training that utilizes immersive virtual reality. Educ Health (Abingdon) 2017;30(3):228-231. [doi: 10.4103/efh.EfH_9_17] [Medline: 29786025]
44. Real FJ, DeBlasio D, Beck AF, et al. A virtual reality curriculum for pediatric residents decreases rates of influenza vaccine refusal. Acad Pediatr 2017;17(4):431-435. [doi: 10.1016/j.acap.2017.01.010] [Medline: 28126612]
45. Real FJ, Hood AM, Davis D, et al. An immersive virtual reality curriculum for pediatric hematology clinicians on shared decision-making for hydroxyurea in sickle cell anemia. J Pediatr Hematol Oncol 2022 Apr 1;44(3):e799-e803. [doi: 10.1097/MPH.0000000000002289] [Medline: 35319512]
46. Rouleau G, Gagnon MP, Côté J, Richard L, Chicoine G, Pelletier J. Virtual patient simulation to improve nurses' relational skills in a continuing education context: a convergent mixed methods study. BMC Nurs 2022 Jan 4;21(1):1. [doi: 10.1186/s12912-021-00740-x] [Medline: 34983509]
47. Sapkaroski D, Mundy M, Dimmock MR. Immersive virtual reality simulated learning environment versus role-play for empathic clinical communication training. J Med Radiat Sci 2022 Mar;69(1):56-65. [doi: 10.1002/jmrs.555] [Medline: 34706398]
48. Sezer B, Sezer TA. Teaching communication skills with technology: creating a virtual patient for medical students. AJET 2019;35(5):183. [doi: 10.14742/ajet.4957]
49. Şimşek Çetinkaya Ş, Gümüş Çalış G, Kıbrıs Ş, Topal M. Effectiveness of virtual patient simulation versus peer simulation in family planning training in midwifery students: a comparative educational intervention. Interactive Learning Environments 2022 Aug 28;32(3):942-951. [doi: 10.1080/10494820.2022.2105897]
50. Shorey S, Ang E, Yap J, Ng ED, Lau ST, Chui CK. A virtual counseling application using artificial intelligence for communication skills training in nursing education: development study. J Med Internet Res 2019 Oct 29;21(10):e14658. [doi: 10.2196/14658] [Medline: 31663857]
51. Shorey S, Ang E, Ng ED, Yap J, Lau LST, Chui CK. Communication skills training using virtual reality: a descriptive qualitative study. Nurse Educ Today 2020 Nov;94:104592. [doi: 10.1016/j.nedt.2020.104592] [Medline: 32942248]
52. Shorey S, Ang ENK, Ng ED, et al. Evaluation of a theory-based virtual counseling application in nursing education. Comput Inform Nurs 2023 Jun 1;41(6):385-393. [doi: 10.1097/CIN.0000000000000999] [Medline: 36728150]
53. Stevens A, Hernandez J, Johnsen K, et al. The use of virtual patients to teach medical students history taking and communication skills. Am J Surg 2006 Jun;191(6):806-811. [doi: 10.1016/j.amjsurg.2006.03.002] [Medline: 16720154]
54. Mendez KJW, Piasecki RJ, Hudson K, et al. Virtual and augmented reality: implications for the future of nursing education. Nurse Educ Today 2020 Oct;93:104531. [doi: 10.1016/j.nedt.2020.104531] [Medline: 32711132]
55. Cant R, Cooper S, Sussex R, Bogossian F. What's in a name? Clarifying the nomenclature of virtual simulation. Clin Simul Nurs 2019 Feb;27:26-30. [doi: 10.1016/j.ecns.2018.11.003]
56. Watts PI, McDermott DS, Alinier G, et al. Healthcare Simulation Standards of Best PracticeTM simulation design. Clin Simul Nurs 2021 Sep;58:14-21. [doi: 10.1016/j.ecns.2021.08.009]
57. Maicher K, Danforth D, Price A, et al. Developing a conversational virtual standardized patient to enable students to practice history-taking skills. Sim Healthcare 2017;12(2):124-131. [doi: 10.1097/SIH.0000000000000195]
58. Witmer BG, Singer MJ. Measuring presence in virtual environments: a presence questionnaire. Presence (Camb) 1998 Jun;7(3):225-240. [doi: 10.1162/105474698565686]
59. Zielke MA, Zakhidov D, Hardee G, et al. Developing virtual patients with VR/AR for a natural user interface in medical teaching. In: 2017 IEEE 5th International Conference on Serious Games and Applications for Health (SeGAH): IEEE; 2017. [doi: 10.1109/SeGAH.2017.7939285]
60. Stamer T, Steinhäuser J, Flägel K. Artificial intelligence supporting the training of communication skills in the education of health care professions: scoping review. J Med Internet Res 2023 Jun 19;25(8):e43311. [doi: 10.2196/43311] [Medline: 37335593]
61. Foronda CL, Fernandez-Burgos M, Nadeau C, Kelley CN, Henry MN. Virtual simulation in nursing education: a systematic review spanning 1996 to 2018. Sim Healthcare 2020;15(1):46-54. [doi: 10.1097/SIH.0000000000000411]

Abbreviations

AI: artificial intelligence
MeSH: Medical Subject Headings
ML: machine learning

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by B Lesselroth; submitted 10.06.24; peer-reviewed by A Grilo, JD Ramos Pichardo; revised version received 29.10.24; accepted 02.01.25; published 06.05.25.

Please cite as:
Fernández-Alcántara M, Escribano S, Juliá-Sanchis R, Castillo-López A, Pérez-Manzano A, Macur M, Kalender-Smajlović S, García-Sanjuán S, Cabañero-Martínez MJ
Virtual Simulation Tools for Communication Skills Training in Health Care Professionals: Literature Review
JMIR Med Educ 2025;11:e63082
URL: https://mededu.jmir.org/2025/1/e63082
doi:10.2196/63082

Virtual Simulated Placements in Health Care Education: Scoping Review

Juliana Samson, BSc, Grad Cert, MSc; Marc Gilbey, BSc, MSc; Natasha Taylor, BSc, Grad Cert, MSc, MEd, EdD; Rosie Kneafsey, BSc, PhD

Coventry University, Richard Crossman Building, Priory Street, Coventry, United Kingdom

Corresponding Author:
Juliana Samson, BSc, Grad Cert, MSc
Coventry University, Richard Crossman Building, Priory Street, Coventry, United Kingdom

Abstract

Background: A virtual simulated placement (VSP) is a computer-based version of a practice placement. COVID-19 drove increased adoption of web-based technology in clinical education. Accordingly, the number of VSP publications increased from 2020. This review determines the scope of this literature to inform future research questions.

Objective: This study aimed to assess the range and types of evidence related to VSPs across the health care professions.

Methods: Studies that focussed on health care students participating in VSPs. Hybrid, augmented reality, and mixed reality placements were excluded. In total, 14 databases were searched, limited to English, and dated from January 1, 2020. Supplementary searches were employed, and an updated search was conducted on July 9, 2023. Themes were synthesized using the PAGER (patterns, advances, gaps, evidence for practice, and research recommendations) framework to highlight patterns, advances, gaps, evidence for practice, and research recommendations.

Results: In total, 28 papers were reviewed. All VSPs were designed in response to pandemic restrictions. Students were primarily from medicine and nursing. Few publications were from low and middle-income countries. There was limited stakeholder involvement in the VSP designs and a lack of robust research designs, consistent outcome measures, conceptual underpinnings, and immersive technologies. Despite this, promising trends for student experience, knowledge, communication, and critical thinking skills using VSPs have emerged.

Conclusions: This review maps the VSP evidence across health care education. Allied health and midwifery research require greater representation, and based on the highlighted gaps, other areas for future research are suggested.

Trial Registration: OSF Registries osf.io/ay5gh; https://osf.io/ay5gh/

(JMIR Med Educ 2025;11:e58794) doi:10.2196/58794

KEYWORDS

technology; students; learning; scoping review; simulation; healthcare education; virtual simulated placement; practice placement; clinical placement

Introduction

Background

Practice placements are important activities in the training of health care students. They promote the application of knowledge to a practical setting for developing the skills, attitudes, and behaviors expected of a health care professional [1-3]. Placements allow active involvement in care delivery under supervision, and the opportunity to receive feedback on student performance [4]. In other words, student learning on placement is contextualized to future practice.

Simulation-based placements present an alternative to traditional practice placements. In traditional placements, students enter a workplace and learn through observation and participation in actual clinical events. In contrast, health care simulation is a technique that produces a scenario designed to represent a real-life practice situation for experiential learning [5,6]. Compared with traditional placements, simulation can ensure that low-frequency and high-risk cases or situations receive sufficient practice in a safer space, without mistakes causing harm to real persons [7]. Thus, the advantage of simulation is the ability to control and direct case-based learning.

With advances in technology, simulation-based education is expanding into web-based environments, a trend accelerated during the pandemic. The increasing complexity of health care also requires an agile workforce of lifelong learners, capable of substituting skills across professions [8,9]. Consequently, health care training must keep pace with technology developments, and virtual simulations could support the training of these skills [10-12]. Furthermore, virtual simulations offer greater flexibility

and scalability compared with using standardized patients (people play acting the role of a service user) [13].

Problem Statement

As virtual simulated placements (VSPs) are an emerging field, mapping the literature across health care and analyzing gaps is recommended before more specific research questions are defined [14,15]. Therefore, our research team chose a scoping review method to conduct a broader search across medicine, nursing, midwifery, and allied health, for undergraduate and postgraduate students who undertook VSPs. Considering the importance of practice placement, the advantages of simulation-based learning, and recent advances in technology, this topic was relevant for the review. We define virtual simulations as computer-based activities according to the Healthcare Simulation Dictionary [16], and our aim was to determine the scope of the VSP literature to inform future research questions. Our objective was to assess the range and types of evidence related to VSPs, across the health care professions.

Review Questions

First, what is the scope of the literature relating to VSPs for health care students? Second, what outcomes are reported in relation to the students undertaking VSPs? Third, what are the patterns and gaps in the literature and the reported outcomes? Finally, what are the implications of the review findings for future directions in VSP research?

Methods

Overview

This study followed the stages detailed in a framework for scoping reviews [14]: (1) identify the research question; (2) identify relevant studies; (3) study selection; (4) charting the data; and (5) collating, summarizing, and reporting the results.

A preliminary search of MEDLINE, the Cochrane Database of Systematic Reviews, and Joanna Briggs Institute Evidence Synthesis was conducted on June 17, 2022 to locate any existing or underway reviews on the topic. One systematic review [13] was identified and focused on digital placements for undergraduate nursing and medical students. The review also included experiences such as telemedicine and on-screen role-play. While their search located 16 studies in April 2021, the increased trend toward implementing VSPs within undergraduate and postgraduate programs across the wider health professions justified this review.

An a priori protocol used the Joanna Briggs Institute template for scoping reviews [15] and was registered with the Open Science Framework (DOI 10.17605/OSF.IO/AY5GH) [17]. The PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews) checklist (Checklist 1) ensured methodological rigor when reporting this review [18].

Relevant Studies

The eligibility criteria are listed in Table 1 using the SPIDER (sample, phenomenon of interest, design, evaluation, research type) [19] and PCC (population, concept, and context) [15] formats:

Table . Eligibility criteria in population, concept, and context and sample, phenomenon of interest, design, evaluation, research type formats.

Item	Inclusion	Exclusion
S (sample) or population	Papers reporting on undergraduate and postgraduate health care students, from medicine, nursing, midwifery, and allied health	• Papers reporting on professions outside of the target group
PI (phenomenon of interest) or concept and context	Virtual simulation learning in a practice placement. Articles should stipulate that it is a placement, clerkship, elective, selective, practical, or practicum in the curriculum	• Onsite simulation • Augmented reality and mixed reality interventions • Contact with real or standardized patients, even if telecast to students or delivered in a virtual simulation suite • Hybrid or blended approaches (part online, part onsite) • Tutorials training isolated clinical skills and case studies • Theory-based education • Assessment of learning
D (design)	Studies with quantitative, qualitative, or mixed methods.	• Papers where no research methods were described
E (evaluation)	At least 1 student-centered outcome is included (eg, student satisfaction, confidence, self-efficacy, engagement, learning, knowledge, attitude, skills, or clinical performance)	• No student-centered outcomes recorded
R (research type)	Any primary research, including gray literature. In English language and published since January 1, 2020	• Reviews—although primary studies will be extracted from relevant reviews to determine their eligibility • Study protocols, expert opinion, discussion papers, letters, comments, editorials, and book chapters • Survey research (without a virtual simulated placement case)

The selection criteria were piloted by screening 50 randomly selected titles and abstracts, independently by 2 reviewers (JS and MG). This process generated 94% agreement (Cohen κ=0.6) and served to clarify the selection criteria. In discussion with a third reviewer (NT), the Health and Care Professions Council definition for allied health [20] was adopted in place of the National Health Service (NHS) criteria [21], since this definition includes practitioner psychologists—a population potentially well-suited to VSPs, with the emphasis on talking therapies.

Search Strategy

An initial limited search of MEDLINE and CINAHL was undertaken on June 28, 2022 to identify articles on the topic. The text words contained in the titles and abstracts of relevant articles and index terms were used to develop a full search strategy. This was checked by a health care research librarian and run on MEDLINE on August 3, 2022 (Multimedia Appendix 1). The search strategy was then adapted for each database. The databases searched included MEDLINE, CINAHL, Allied and Complementary Medicine Database, Cochrane Database, PsychINFO, Education Resources Information Center, SCOPUS, ScienceDirect, and Biomed Central. Gray literature sources include PubMed, Electronic Theses Online Service, ProQuest (dissertations), Google Scholar, and Institute of Electrical and Electronics Engineers Xplore. Searches were limited to English language and dated from January 1, 2020. The date limitation was justified given that VSP research has essentially emerged postpandemic.

Supplementary search strategies were employed using existing knowledge and networks, contacting relevant organizations, hand-searching journals, and checking the reference list of all included sources and relevant reviews. Advances in Simulation, British Medical Journal: Simulation and Technology Enhanced Learning (BMJ STEL) and Clinical Simulation in Nursing were hand-searched. These supplementary searches were conducted by one reviewer (JS) and checked by another (NT).

An updated database search was conducted on July 9, 2023. A second reviewer (MG) checked the title, abstract, and full-text selection decisions. Registries (Clinical Trials.gov, World Health Organization International Clinical Trials Registry Platform, and the Cochrane Database) were searched for additional papers [22]. Updated hand searches were performed in Advances in Simulation and Clinical Simulation in Nursing (BMJ STEL had since discontinued). A second reviewer (NT) checked these supplementary searches.

Source Selection

Following the database searches, all identified citations were uploaded into EndNote (Clarivate) [23], and duplicates were removed. Each potential duplicate was confirmed separately, rather than using batch automation to prevent the removal of

false positives [24]. Citations were exported to Rayyan and rechecked for any missed duplicates [25].

Once pilot screening was complete, the remaining titles and abstracts were screened independently by 2 reviewers (JS and MG) against the revised criteria, and potentially relevant sources were retrieved in full text. These were assessed in detail against the inclusion criteria by 2 independent reviewers (JS and MG), blinded in Rayyan. There was 83% agreement (Cohen κ=0.5) between reviewers. A 100% agreement was reached through discussion. Further details of the source selection, including a list of references excluded at full text screening are detailed in Multimedia Appendix 2.

Data Charting

A Microsoft Excel spreadsheet was used as a data charting tool to standardize obtaining information from the papers. Furthermore, 2 independent reviewers (JS and MG) conducted a pilot of 5 included papers to assess the utility of the information charted and generate emerging themes. Consensus was reached between reviewers (JS and MG) on the charting method, and modifications were made to the spreadsheet, to improve the quality of charted data (Multimedia Appendix 3). Following this, one reviewer (JS) charted the remaining data, which was checked by another (NT).

A table of included study characteristics was collated, and numerical analysis in Microsoft Excel was undertaken to provide descriptive statistics. The size of the dataset was manageable enough to organize findings across the PAGER (patterns, advances, gaps, evidence for practice, and research recommendations) domains [26], for synthesis, without the use of NVivo software (Lumivero; as was planned in the protocol).

Ethical Considerations

The Coventry University Ethical Approval process has been completed and the project has been confirmed and approved as low risk (project reference P139783). Date of approval is August 12, 2022.

Results

Overview

The search results and selection process are reported in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram (Figure 1).

The characteristics of the 28 included papers are summarized in Multimedia Appendix 4. Overall, VSPs were a combination of videoconferencing sessions with educators and peers, as well a variety of web-based material, including videos, reading, modules, and assignments. Most VSPs included some form of case-based learning that required problem-based activities to complete. Teaching methods ranged from didactic lecture-style sessions to peer learning and flipped classrooms. Session delivery featured more formal case conference-style sessions, as well as small group learning, and the use of online chat, polls, and quizzes.

The PAGER themes across the papers are summarized in Table 2. Key patterns and gaps are mapped across all included studies in Tables 3 and 4. The global distribution of publications is illustrated in Figure 2.

Patterns are mapped across all included studies in Table 3 and gaps are mapped in Table 4.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart (modified from) [27].

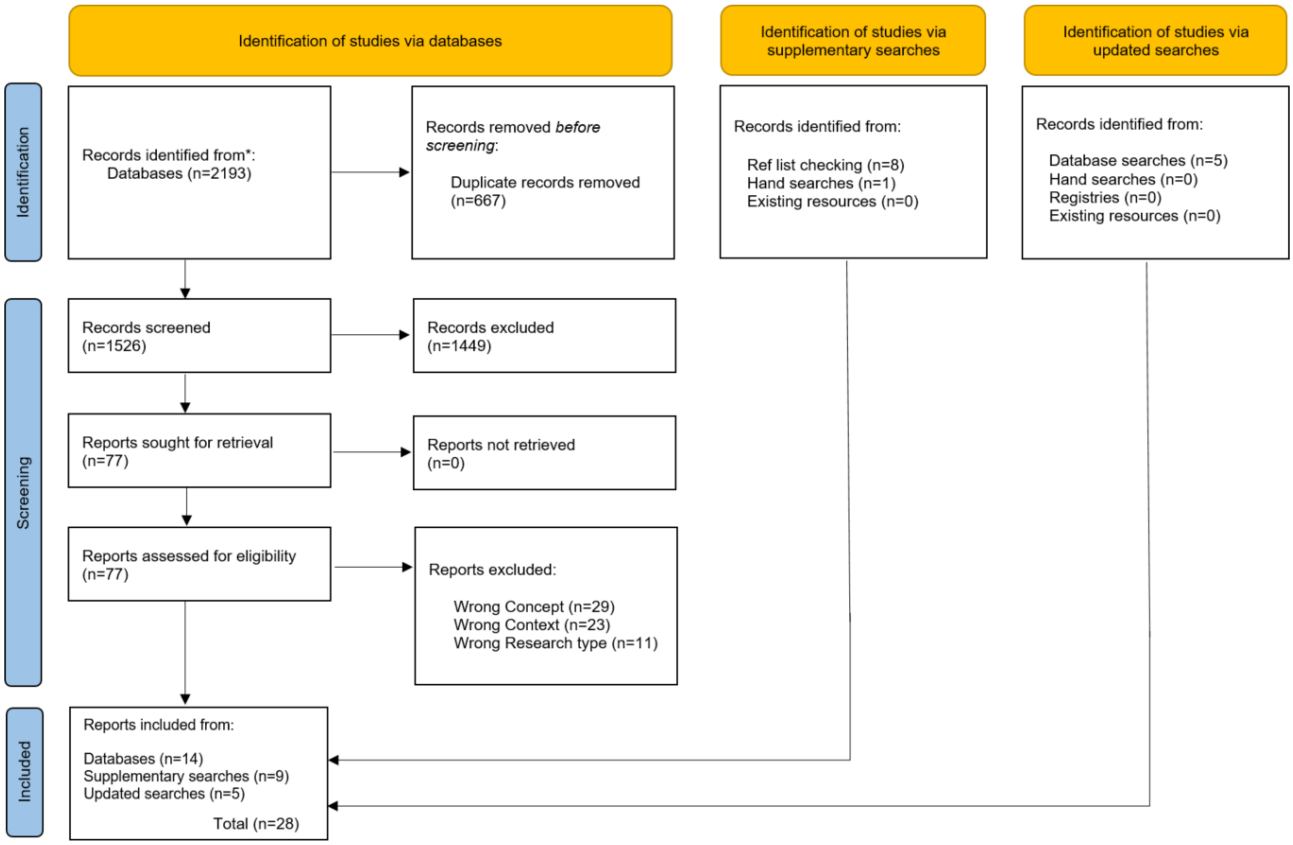


Table . PAGER (patterns, advances, gaps, evidence for practice, and research recommendations) framework themes summary.

Patterns	Advances	Gaps	Evidence for practice	Research recommendations
Publications from high-income countries (Figure 2)	Innovations occurred mostly in countries with resources to support VSP[a] development	Few publications from the LMICs[b]	VSPs can be delivered remotely and are scalable (useful for supporting training in the LMICs)	Sharing resources across countries and overcoming barriers such as internet connectivity or access to devices
Narrow profession focus	VSPs occurred within single profession silos. Populations were mostly medical or nursing	No IPE[c] and minimal allied health representation	Support for VSPs delivering on improved discipline specific skills	The development of IPE VSPs to train skills informed by allied health collaborations
Pandemic response	Rapid innovation to shift from in-person placement to VSPs in response to COVID-19 restrictions	Research planned under time pressure may explain the lack of robust experimental design and conceptual frameworks	Positive outcomes suggest that VSPs could be utilized beyond the pandemic response	With less time pressure, future research could consider conceptual frameworks, with more robust experimental designs
Stakeholder involvement in the VSP design	Most studies involved university faculty. Others also included clinicians	Few incorporated student input and consultation. No evidence of cocreation with service users	Design that involves student participation throughout the process better serves the end user needs	Participatory research designs should include all stakeholders, including students and service users (who ultimately benefit)
Use of generic platforms and screen-based delivery	Platforms such as Microsoft Teams, Zoom, and existing learning management systems were used to facilitate delivery	Limited use of bespoke software or VR[d]. No headsets, haptics, or conversational artificial intelligence systems	Student feedback frequently rated the live interaction with facilitators positively	Bespoke VR software, headsets, and haptic research may emerge as devices become more ubiquitous
A focus on case-based learning	VSPs were oriented toward clinical cases and knowledge, clinical reasoning, decision making, and communication	Practical skills training was rare. Few featured social determinants of health or community interventions	Evidence for improved knowledge, clinical thinking, and communication skills from VSP interventions	Hybrid is currently more suitable for practical skills but haptics may feature as technology improves. Community VSPs link well to IPE
Survey-based outcome measures	Most VSPs were evaluated through custom-designed surveys and student marks	Few validated outcome measure scales or standardized examinations	Evaluations were overall positive and test score improvements were equivalent to in-person cohorts	Validated outcome measures and standardized tests in future trials would provide more robust data for comparison

[a]VSP: virtual simulated placement.

[b]LMIC: low or middle-income country.

[c]IPE: interprofessional education.

[d]VR: virtual reality.

Table . Key patterns.

Citation	Patterns			
	High-income country	Medical or nursing profession	Pandemic response	Generic software
Alpert et al [28]	✓	✓	✓	✓
Bhashyam et al [29]	✓	✓	✓	✓
Creagh et al [30]	✓	✓	✓	✓
De Ponti et al [31]	✓	✓	✓	✓
Durfee et al [32]	✓	✓	✓	✓
Fehl et al [33]	✓	✓	✓	✓
Ganji et al [34]	×	×	✓	✓
Gomez et al [35]	✓	✓	✓	✓
He et al [36]	×	✓	✓	✓
Holmberg et al [37]	✓	✓	✓	✓
Joung et al [38]	✓	✓	✓	✓
Kasai et al [39]	✓	✓	✓	✓
Kubin et al [40]	✓	✓	✓	✓
Luo et al [41]	×	✓	✓	✓
Martin-Delgado et al [42]	✓	✓	✓	✓
Nguyen et al [43]	✓	✓	✓	✓
Rahm et al [44]	✓	✓	✓	✓
Redinger et al [45]	✓	✓	✓	✓
Samueli et al [46]	✓	✓	✓	✓
Smith et al [47]	✓	✓	✓	✓
Steehler et al [48]	✓	✓	✓	✓
Taylor et al [49]	✓	×	✓	✓
Villa et al [50]	✓	✓	✓	✓
Weston and Zauche [51]	✓	✓	✓	✓
White et al [52]	✓	✓	✓	✓
Wik et al [53]	✓	✓	✓	✓
Williams et al [54]	✓	✓	✓	✓
Zhou et al [55]	×	✓	✓	✓

Table . Key gaps.

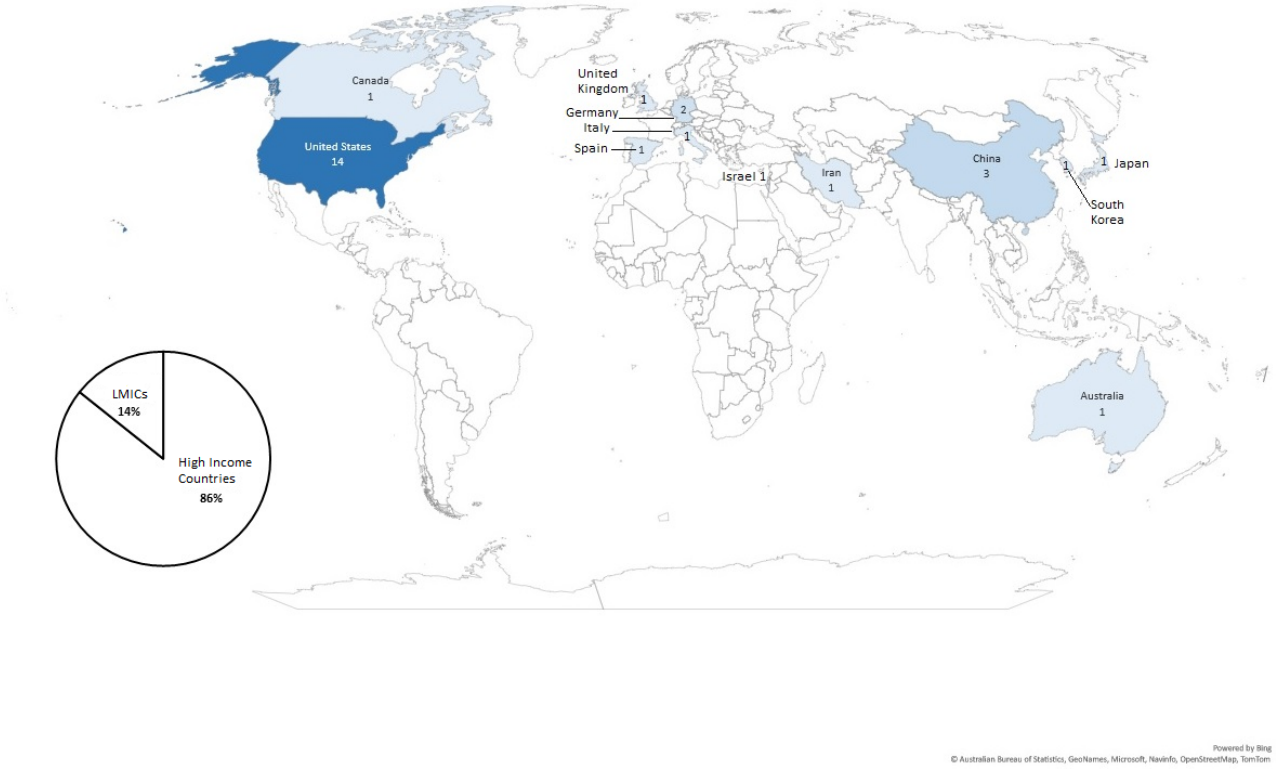
Citation	Gaps							
	Population		Experimental design				Software	Hardware
	IPE[a]	Allied health	Comparator group	Pre- and post-measures	Students involved in the design	Conceptual frameworks	Bespoke software	VR[b] equipment
Alpert et al [28]			✓					
Bhashyam et al [29]						✓		
Creagh et al [30]						✓	✓	
De Ponti et al [31]							✓	
Durfee et al [32]							✓	
Fehl et al [33]			✓			✓		
Ganji et al [34]				✓		✓		
Gomez et al [35]							✓	
He et al [36]								
Holmberg et al [37]				✓	✓			
Joung et al [38]							✓	
Kasai et al [39]				✓		✓		
Kubin et al [40]						✓	✓	
Luo et al [41]				✓		✓	✓	
Martin-Delgado et al [42]								
Nguyen et al [43]						✓		
Rahm et al [44]					✓		✓	
Redinger et al [45]			✓			✓	✓	
Samueli et al [46]						✓	✓	
Smith et al [47]							✓	
Steehler et al [48]				✓	✓			
Taylor et al [49]		✓				✓	✓	
Villa et al [50]				✓	✓	✓	✓	
Weston and Zauche [51]			✓				✓	
White et al [52]						✓	✓	

Citation	Gaps							
	Population		Experimental design				Software	Hardware
	IPE[a]	Allied health	Comparator group	Pre- and post-measures	Students involved in the design	Conceptual frameworks	Bespoke software	VR[b] equipment
Wik et al [53]							✓	
Williams et al [54]				✓		✓		
Zhou et al [55]			✓			✓	✓	

[a]IPE: interprofessional education.

[b]VR: virtual reality.

Figure 2. Country of origin of included papers. LMIC: low or middle-income country.



Countries of Origin

In total, 86% (24/28) of the included papers were published in high-income countries, as defined by the Organisation for Economic Co-operation and Development [56]. The VSP research was located primarily in the United States and the Northern Hemisphere.

Range of Professions

The literature was predominantly medical and nursing research, constituting 93% (26/28) of the included papers. The distribution by profession and breakdowns by specialty are illustrated in Multimedia Appendix 5. Diagnostic radiology rotations were the most prevalent VSPs in medicine and pediatrics in nursing. Where stated, learners were often in their latter stages of training, or undertaking these VSPs as postgraduates.

Pandemic Response

All the VSPs in the included papers were developed in response to COVID-19 restrictions, which aligns with the time span of the scoping search. The context at the time was that the pandemic necessitated that face-to-face (FTF) practice placements were often discontinued. VSPs were implemented to provide alternative placement hours, enabling students to progress toward professional registration and graduation.

Experimental Designs

The most basic study design was a single group, with a postintervention measure, featuring in 16 papers. In total, 7 papers compared measures pre- and postintervention [34,37,39,41,48,50,54]. Furthermore, 5 papers compared VSP outcomes with a previous cohort of students who completed FTF placements prepandemic [28,33,45,51,55].

Stakeholder Involvement

Practice partners (clinicians working in practice) were involved in the VSP course development with faculty in 8 studies [29,32,37,41,44,48,50,55] and students were involved in 4. Furthermore, 3 studies developed a needs assessment from student surveys [34,43,52]. None involved service users.

Conceptual Frameworks

Conceptual underpinnings include pedagogy, theoretical frameworks, and professional standards. Although no single paper covered all elements, underpinning concepts are evident across the literature, summarized in Multimedia Appendix 6. Pedagogies employed, focused on adult student learners, case-based activities, and experiential and web-based learning. The frameworks structured the VSP development, and the professional standards guided curriculum, simulation, and placement.

Software

All studies used generic software such as Zoom (Zoom Communications) or Microsoft Teams for screen-based communication, and many used existing learning management systems to host files and activities. Others adopted commercial software applications, allowing students to conduct a history by selecting from a menu of interview questions. None used conversational artificial intelligence (AI) systems (computer-generated conversation, assisted by AI). Some applications presented virtual reality (VR) patient avatars with which the student could direct a physical examination, although this was delivered via a screen [31,38,40,41,51] and 1 study provided an interactive community setting in screen-based VR [53]. All software resources are outlined in Multimedia Appendix 7.

Intended Learning Outcomes

The focus of most VSPs was clinical cases, through which knowledge, reasoning, decision-making, and communication skills (both verbal and written) were developed. Practical skills training was rarely practiced, with 1 study including home practice surgical kits as the exception [1]. Instead, skill learning was visualized through virtual patient encounters and instructional or walk-through procedure videos. The social determinants of health were the focus in 2 studies [50,53] and another facilitated students in teaching roles [42].

Outcomes

The most common outcome measures were custom-developed student evaluation questionnaires, followed by exam marks. Custom questionnaires provided positive feedback for student experience, satisfaction, and usability, although some technical issues and Zoom fatigue were cited [31,50]. In total, 3 papers reported a 100% pass rate on their VSPs [35,49,52], and 4 used a standardized exam to demonstrate comparable outcomes with FTF cohorts [45,51], or the national average [30,32].

Table 5 summarizes the outcomes of research that employed a repeated measures design or group comparisons.

Table . Outcomes from intra- and intergroup comparisons.

Study feature and outcomes	Papers
Measures compared pre- and post-VSP[a]	
Increase in self-rated competencies	Holmberg et al [37], Kasai et al [39], and Williams et al [54]
Increase in knowledge scores	Ganji et al [34], Steehler et al [48], and Villa et al [50]
Improvement in interview skills	Ganji et al [34]
Improvement in critical thinking ability	Luo et al [41]
Comparison between a VSP group and a previous cohort that attended a FTF[b] placement	
No significant difference in exam scores between groups (P>.05)	Redinger et al [45], Weston and Zauche [51], and Zhou et al [55]
Mixed outcomes from survey responses	Fehl et al [33] and Alpert et al [28]

[a]VSP: virtual simulated placement.

[b]FTF: face-to-face.

When measures were compared pre- and post-VSP, there was a trend of improvement in self-rated competencies, knowledge scores, and critical thinking skills. However, when the comparison is made with traditional FTF placements, the pattern is less clear. There were no differences in grades when post-VSP exam scores were compared with previous cohorts' who attended an FTF placement prepandemic. Student satisfaction was comparable in a study conducted in medical general practice, but professional exchange and learning scored higher in the VSP, while the attainment of new skills and attitudes scored higher in the FTF placement [33]. Furthermore, 1 paper compared students who participated in web-based readouts (the radiology equivalent of patient rounds) with students who attended workplace readouts prepandemic [28]. The educational value was comparable in survey results, although students on the VSP rated slightly higher for perceived interaction. That FTF students that were mostly observing on their placement might explain this finding. Conversely, FTF students had greater confidence in using the workstations, considered the case because the VSP students were unable to operate Picture Archiving Communication System workstations remotely.

Discussion

Principal Findings

This study mapped the literature describing VSPs across health care. All 28 papers were pandemic responses, primarily from medicine and nursing in high income countries . Selecting studies that conducted a web-based simulation, rather than employing a hybrid or blended approach may explain why all papers in this review were pandemic responses, and why the student populations were in their latter stages of training or postgraduates. COVID-19 necessitated a rapid shift to provide VSPs as a replacement for lost clinical hours to allow students to progress toward graduation [57]. However, these VSPs were often produced in a short time frame, under emergency situations, and may explain why few papers featured robust experimental designs and conceptual frameworks.

Replacing FTF placement hours with simulation is a contentious issue. Accordingly, a Delphi study considered the benefits and limitations of this approach [58]. Expert consensus across multiple professions agreed that between 11% - 30% of hours replaced with simulation would be acceptable, and this aligns with the current allocation set by the Nursing and Midwifery Medical Council [59]. VSPs in the curriculum may offset some pressure on workplace settings as they attempt to fulfill the NHS long-term plan to recruit and train more health care learners [11]. However, this does not diminish the importance of building further workplace placement capacity [58]. VSPs can be considered an additional pedagogy that offers a different, yet complimentary experience to traditional FTF placements.

Content and Technologies

In general, VSPs had a teleconferencing and a web-based learning component. The teleconferencing was commonly conducted with educators and peers over Zoom or Microsoft Teams, and the web-based learning activities included, but were not limited to videos, reading, modules, and assignments. There were a few examples of immersive learning with VR patient avatars, and these were delivered via a screen [31,38,40,41,51,53].

Disciplines that rely on image-based diagnoses may be more easily adapted to screen-based delivery, and consistent with this, diagnostic radiology, and pathology VSPs together constituted over 30% (6/19) of the medical papers in this review. In the development of this scoping review, we anticipated that psychology might be suited to VSPs due to the nature of talking-based therapies over physical skills, although it is possible that psychological presentations were considered too complex to portray accurately in computer-based simulations. With future developments in conversational AI systems and the growing acceptance of this technology, this situation may change. Similarly, professions that rely heavily on hands-on assessment, such as physiotherapy, may feature more in extended reality spaces with haptics, as further research and development into these technologies emerge. In the meantime, VSPs that require complex conversations are likely to include telecast or telemedicine simulations. Likewise, VSPs that teach advanced handling skills might adopt a hybrid or blended approach, thus combining the strengths of both web-based and FTF approaches.

Interprofessional Education

VSPs have the potential to break down silos between professions, by delivering interprofessional education (IPE) over a web-based platform. IPE is defined as 2 or more professions, "*learning with from and about one another to improve collaborative practice and quality of care*" [P4] [60]. The intended outcome is to improve mutual understanding, teamwork, and leadership among different professionals [61]. VSPs have advantages over FTF training in building asynchronous activities for flexibility in timetabling and hosting synchronous activities without geographical constraints [62]. Given the relevance of IPE to quality care and the fit with web-based technologies, IPE-VSPs may be an important area for future research.

VSP Design and Stakeholder Involvement

Elements of thoughtful VSP design are evident across several papers. Frameworks, such as ADDIE (analysis, design, development, implementation, and evaluation), ensure that there is structure to the process and stakeholder needs are met [30]. Existing curricula [54,55], or processes such as Kern's 6-step model for curricular development could be used [43,45,46,50,52]. If framed within existing standards [40,49], VSPs can align with specified learning outcomes. Principles in pedagogy, such as andragogy [29,30] and web-based learning [33,50], ensure that VSPs build features that engage students with experiential learning [30] and promote problem-solving [29,30,39] and active reflection [49]. The conceptual underpinnings documented across this body of literature could provide a blueprint for best practice in VSP design.

Stakeholder involvement is a key process to inform the design of a VSP. Service users could inform the content, which is especially important in computer-based simulations, yet no service user involvement was documented. Students are the end users of a VSP, yet they were involved in a minority of studies. When students were involved, surveys informed a needs assessment, or they were consulted early in the process. This is a tokenistic approach compared with cocreation, the preferred method of engaging with stakeholders. Cocreation involves a collective effort with all stakeholders to collaborate across the entire design, development, implementation, and testing phases [63]. One noteworthy research report provided an overview of VSP development within a nursing program, which included input from students, service users, and other universities throughout [64]. Their working group comprised of academics, clinicians, a service user, a carer involvement lead, and an education technology lead. Therefore, in addition to underpinning VSP design with the relevant conceptual frameworks (pedagogical principles, theoretical frameworks, and published standards), broad stakeholder cocreation is optimal.

Research Designs

The pattern of positive student evaluation, improvement from baseline measures post VSP, and equivalence in exam scores, compared with in-person cohorts, appears promising, although,

it should be remembered that the objective of a scoping review is to map the literature for patterns and gaps, rather than in-depth appraisal of the quality of the papers.

The findings compare with a systematic review that examined digital clinical education more broadly [13]. Stand-alone digital education was reported to be as effective as conventional learning for knowledge and practice, in nursing and medicine. However, there are some methodological concerns with this systematic review [13]. There was no a priori protocol, and the study lacked a pilot to test the methods. A librarian's involvement in verifying the search strategy was not reported, gray literature was not searched, and duplicate processes were absent for the study selection and data extraction stages.

There are several barriers to conducting a systematic review of VSPs across health care. First, there is insufficient research across midwifery and allied health [34,49]. Another consideration is that all student evaluations in this scoping review were custom-designed. Therefore, the inconsistency of outcome measures might prevent meaningful comparisons across papers. One study used previously researched scales for clinical thinking ability, academic self-efficacy, and student engagement, which demonstrated good reliability [41]. Some of the exams were standardized [30,32,45,51], but none compared the baseline marks of each group to determine whether there were differences at the outset. In all cases, VSP exam scores were compared with a previous cohort that attended placement FTF prepandemic, or the national average, rather than adopting a prospective design.

It is clear from the paucity of research outside nursing and medicine, the lack of prospective research designs and inconsistent, nonvalidated outcome measures, that research into VSPs is in its infancy. It is tempting to recommend greater consistency of outcome measures and more robust experimental designs to improve the evidence base. However, such approaches may not fit the study of complex educational interventions such as VSPs. More suitable approaches include quasi-experimental, qualitative, and evaluative designs to examine conceptual underpinnings, VSP cocreation, the mechanisms mediating learning responses, and individual case trends over time.

Strengths and Weaknesses

The strengths of this study relate to the methodology. A structured process for defining search terms was undertaken, and a librarian was consulted for the search strategy. A range of databases were searched across medical and technology specialties. Gray literature sources were searched, and an updated search included trial registries. An a priori protocol was registered, and a subset of data was piloted to determine the declared changes. Duplicate processes in study selection and data charting were employed, and existing guidelines were used to design the protocol, synthesize the findings, and report the paper.

One weakness is that many health care educators may have implemented VSPs without documenting their practices. As such, a scoping review of the literature will always underestimate the scale and depth of innovation in practice. Limiting the search to English language increased the risk of language bias. While the limited number of publications from low or middle income countries could reflect the language limitation, it is also likely that countries with greater resources were better positioned to make the rapid shift to web-based education and publish their research during a global health emergency. Web-based platforms are suited to sharing resources and overcoming geographical constraints to access expertise, and VSPs present an opportunity to address inequality in health care education moving forward.

Conclusion

This scoping review mapped the VSP evidence across health care, highlighting patterns and gaps in the evidence base. All papers documented pandemic responses, primarily in medicine and nursing in high income countries. There are notable gaps in the midwifery and allied health research. Although emerging trends for VSPs in this review demonstrate some positive outcomes, this review highlights the need for improvements in VSP design. These include cocreation with a wider range of stakeholders and underpinning by pedagogical principles, theoretical frameworks, and published standards. Research into student engagement using VR headsets, haptics, and conversational AI systems in VSPs, are areas for future research, as immersive technologies and their use cases develop. The pandemic has revealed an opportunity to augment placement capacity through VSPs. There is the potential for future VSPs to feature IPE, thus promoting joined-up care in health care graduates. There is also the opportunity for VSPs to improve local and global access to quality clinical education experiences.

Acknowledgments

We acknowledge our health librarian, Carlo Avillo, for assisting with the search strategy.

This review will contribute towards a PhD award for JS.

Conflicts of Interest

None declared.

Multimedia Appendix 1
MEDLINE search strategy.
[DOCX File, 21 KB - mededu_v11i1e58794_app1.docx]

Multimedia Appendix 2
Screening Decisions
[DOCX File, 84 KB - mededu_v11i1e58794_app2.docx]

Multimedia Appendix 3
Revised data charting tool.
[DOCX File, 23 KB - mededu_v11i1e58794_app3.docx]

Multimedia Appendix 4
Table of included study characteristics.
[DOCX File, 99 KB - mededu_v11i1e58794_app4.docx]

Multimedia Appendix 5
Papers by professions.
[DOCX File, 31 KB - mededu_v11i1e58794_app5.docx]

Multimedia Appendix 6
Conceptual frameworks.
[DOCX File, 25 KB - mededu_v11i1e58794_app6.docx]

Multimedia Appendix 7
Bespoke health care technology.
[DOCX File, 35 KB - mededu_v11i1e58794_app7.docx]

Checklist 1
PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[DOCX File, 30 KB - mededu_v11i1e58794_app8.docx]

References

1. Promoting excellence: standards for medical education and training. General Medical Council. 2015. URL: https://www.gmc-uk.org/education/standards.asp [accessed 2022-02-07]
2. Regulating health and care professionals. Health and Care Professions Council. 2017. URL: https://www.hcpc-uk.org [accessed 2022-04-07]
3. Nursing and Midwifery Council. Part 2: Standards for student supervision and assessment Part 1: Standards framework for nursing and midwifery education Part 2: Standards for student supervision and assessment Part 3: Programme standards. 2018. URL: https://www.nmc.org.uk/standards-for-education-and-training/standards-for-student-supervision-and-assessment/ [accessed 2022-04-07]
4. Burgess A, Mellis C. Feedback and assessment for clinical placements: achieving the right balance. Adv Med Educ Pract 2015;6(19):373-381. [doi: 10.2147/AMEP.S77890] [Medline: 26056511]
5. Leighton K, Kardong-Edgren S, Gilbert GE. Are traditional and simulated clinical environments meeting nursing students' learning needs? Clin Simul Nurs 2021 Oct;59(59):85-93. [doi: 10.1016/j.ecns.2021.06.003]
6. About simulation. Society for simulation in Healthcare. 2022. URL: https://www.ssih.org/ [accessed 2022-06-01]
7. Gaba DM. The future vision of simulation in health care. Quality and Safety in Health Care 2004 Oct 1;13(suppl_1):i2-i10. [doi: 10.1136/qshc.2004.009878]
8. Anderson M, O'Neill C, Macleod Clark J, et al. Securing a sustainable and fit-for-purpose UK health and care workforce. The Lancet 2021 May;397(10288):1992-2011. [doi: 10.1016/S0140-6736(21)00231-2]
9. NHS long term workforce plan. National Health Service England. 2023. URL: https://www.england.nhs.uk/publication/nhs-long-term-workforce-plan/ [accessed 2023-07-04]
10. Hobbs C, St John-Matthews J. Helping to ensure an essential supply of allied health professions (AHP) practice placements: challenges and solutions. National health service england. 2020. URL: https://www.hee.nhs.uk/our-work/allied-health-professions/increase-capacity/practice-placements-challenges-solutions [accessed 2022-04-04]
11. Health Education England. Harnessing digital technologies for workforce development, education and training: an overview. National Health Service England. 2022. URL: https://www.hee.nhs.uk/our-work/innovation-digital-transformation/harnessing-digital-technologies-workforce-development-education-training-overview [accessed 2022-12-05]
12. Health Education England. Preparing the healthcare workforce to deliver the digital future the topol review: an independent report on behalf of the secretary of state for health and social care. : National Health Service England; 2019 URL: https://topol.hee.nhs.uk/the-topol-review/ [accessed 2025-06-02]

13. Hao X, Peng X, Ding X, et al. Application of digital education in undergraduate nursing and medical interns during the COVID-19 pandemic: a systematic review. Nurse Educ Today 2022 Jan;108:105183. [doi: 10.1016/j.nedt.2021.105183] [Medline: 34741918]
14. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
15. Peters M, Godfrey C, McInerney P, Munn Z, Tricco AC, Khalil H. Chapter 11: scoping reviews. In: JBI Manual for Evidence Synthesis 2020. [doi: 10.46658/JBIRM-20-01]
16. Lioce L, Lopreiato J, Downing D, et al, editors. Healthcare Simulation Dictionary, 2nd edition: AHRQ Publication. [doi: 10.23970/simulationv2]
17. Samson J, Gilbey M, Taylor N, Kneafsey R. Virtual simulated placements in healthcare education: a scoping review. Med Educ (Chicago Ill) Preprint posted online on 2022. [doi: 10.1101/2023.10.12.23296932]
18. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850]
19. Cooke A, Smith D, Booth A. Beyond PICO: the SPIDER tool for qualitative evidence synthesis. Qual Health Res 2012 Oct;22(10):1435-1443. [doi: 10.1177/1049732312452938] [Medline: 22829486]
20. Professions and protected titles. Health & Care Professions Council. 2018. URL: https://www.hcpc-uk.org/about-us/who-we-regulate/the-professions/ [accessed 2022-09-01]
21. Allied health professions. National Health Service England. URL: https://www.england.nhs.uk/ahp/role/ [accessed 2022-06-01]
22. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 5;5(1):210. [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
23. Hunter KE, Webster AC, Page MJ, et al. Searching clinical trials registers: guide for systematic reviewers. BMJ 2022;26(377):e068791. [doi: 10.1136/bmj-2021-068791]
24. Clarivate. EndNote (EndNote X9). 2013. URL: https://www.endnote.com [accessed 2025-05-30]
25. McKeown S, Mir ZM. Considerations for conducting systematic reviews: evaluating the performance of different methods for de-duplicating references. Syst Rev 2021 Jan 23;10(1):38. [doi: 10.1186/s13643-021-01583-y] [Medline: 33485394]
26. Bradbury-Jones C, Aveyard H, Herber OR, Isham L, Taylor J, O'Malley L. Scoping reviews: the PAGER framework for improving the quality of reporting. Int J Soc Res Methodol 2022 Jul 4;25(4):457-470. [doi: 10.1080/13645579.2021.1899596]
27. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ ;2021(372):n71. [doi: 10.1136/bmj.n71]
28. Alpert JB, Young MG, Lala SV, McGuinness G. Medical student engagement and educational value of a remote clinical radiology learning environment: creation of virtual read-out sessions in response to the COVID-19 pandemic. Acad Radiol 2021 Jan;28(1):112-118. [doi: 10.1016/j.acra.2020.09.011] [Medline: 33268209]
29. Bhashyam AR, Dyer GSM. "Virtual" boot camp: orthopaedic intern education in the time of COVID-19 and beyond. J Am Acad Orthop Surg 2020 Sep 1;28(17):e735-e743. [doi: 10.5435/JAAOS-D-20-00559] [Medline: 32649439]
30. Creagh S, Pigg N, Gordillo C, Banks J. Virtual medical student radiology clerkships during the COVID-19 pandemic: distancing is not a barrier. Clin Imaging 2021 Dec;80(80):420-423. [doi: 10.1016/j.clinimag.2021.08.024] [Medline: 34537485]
31. De Ponti R, Marazzato J, Maresca AM, Rovera F, Carcano G, Ferrario MM. Pre-graduation medical training including virtual reality during COVID-19 pandemic: a report on students' perception. BMC Med Educ 2020 Sep 25;20(1):332. [doi: 10.1186/s12909-020-02245-8] [Medline: 32977781]
32. Durfee SM, Goldenson RP, Gill RR, Rincon SP, Flower E, Avery LL. Medical student education roadblock due to COVID-19: virtual radiology core clerkship to the rescue. Acad Radiol 2020 Oct;27(10):1461-1466. [doi: 10.1016/j.acra.2020.07.020] [Medline: 32747181]
33. Fehl M, Gehres V, Geier AK, et al. Medical students' adoption and evaluation of a completely digital general practice clerkship - cross-sectional survey and cohort comparison with face-to-face teaching. Med Educ Online 2022 Dec;27(1):2028334. [doi: 10.1080/10872981.2022.2028334] [Medline: 35107417]
34. Ganji J, Shirvani MA, Motahari-Tabari N, Tayebi T. Design, implementation and evaluation of a virtual clinical training protocol for midwifery internship in a gynecology course during COVID-19 pandemic: a semi-experimental study. Nurse Educ Today 2022 Apr;111(111):105293. [doi: 10.1016/j.nedt.2022.105293] [Medline: 35134637]
35. Gomez E, Azadi J, Magid D. Innovation born in isolation: rapid transformation of an in-person medical student radiology elective to a remote learning experience during the COVID-19 pandemic. Acad Radiol 2020 Sep;27(9):1285-1290. [doi: 10.1016/j.acra.2020.06.001] [Medline: 32565164]
36. He M, Tang XQ, Zhang HN, Luo YY, Tang ZC, Gao SG. Remote clinical training practice in the neurology internship during the COVID-19 pandemic. Med Educ Online 2021 Dec;26(1):1899642. [doi: 10.1080/10872981.2021.1899642] [Medline: 33685381]
37. Holmberg MH, Dela Cruz E, Longino A, Longino N, Çoruh B, Merel SE. Development of a single-institution virtual internal medicine subinternship with near-peer teaching in response to the COVID-19 pandemic. Acad Med 2021 Dec 1;96(12):1706-1710. [doi: 10.1097/ACM.0000000000004219] [Medline: 34192717]

38. Joung J, Kang KI. Can virtual simulation replace clinical practical training for psychiatric nursing? Issues Ment Health Nurs 2022 Aug;43(8):706-711. [doi: 10.1080/01612840.2022.2055684] [Medline: 35380910]
39. Kasai H, Shikino K, Saito G, et al. Alternative approaches for clinical clerkship during the COVID-19 pandemic: online simulated clinical practice for inpatients and outpatients-a mixed method. BMC Med Educ 2021 Mar 8;21(1):149. [doi: 10.1186/s12909-021-02586-y] [Medline: 33685442]
40. Kubin L, Fogg N, Trinka M. Transitioning child health clinical content from direct care to online instruction. J Nurs Educ 2021 Mar 1;60(3):177-179. [doi: 10.3928/01484834-20210222-11] [Medline: 33657238]
41. Luo Y, Geng C, Pei X, Chen X, Zou Z. The evaluation of the distance learning combining webinars and virtual simulations for senior nursing students during the COVID-19 period. Clin Simul Nurs 2021 Aug;57(57):31-40. [doi: 10.1016/j.ecns.2021.04.022] [Medline: 35915811]
42. Martin-Delgado L, Goni-Fuste B, Monforte-Royo C, de Juan M, Martin-Ferreres ML, Fuster P. A teaching role practicum during the COVID-19 for final year nursing students in Spain: A qualitative study. J Prof Nurs 2022;42(42):51-57. [doi: 10.1016/j.profnurs.2022.06.005] [Medline: 36150878]
43. Nguyen W, Fromer I, Remskar M, Zupfer E. Development and implementation of video-recorded simulation scenarios to facilitate case-based learning discussions for medical students' virtual anesthesiology clerkship. MedEdPORTAL 2023;19(19):11306. [doi: 10.15766/mep_2374-8265.11306] [Medline: 37025196]
44. Rahm AK, Töllner M, Hubert MO, et al. Effects of realistic e-learning cases on students' learning motivation during COVID-19. PLoS One 2021;16(4):e0249425. [doi: 10.1371/journal.pone.0249425] [Medline: 33882079]
45. Redinger KE, Greene JD. Virtual emergency medicine clerkship curriculum during the COVID-19 pandemic: development, application, and outcomes. West J Emerg Med 2021 Apr 28;22(3):792-798. [doi: 10.5811/westjem.2021.2.48430] [Medline: 34125062]
46. Samueli B, Sror N, Jotkowitz A, Taragin B. Remote pathology education during the COVID-19 era: crisis converted to opportunity. Ann Diagn Pathol 2020 Dec;49(49):151612. [doi: 10.1016/j.anndiagpath.2020.151612] [Medline: 32891922]
47. Smith JD, Jones PD. The COVID-19 e-lective: using innovation to manage disrupted medical student clinical placements. BMC Med Educ 2023 Feb 6;23(1):92. [doi: 10.1186/s12909-023-04067-w] [Medline: 36747169]
48. Steehler AJ, Pettitt-Schieber B, Studer MB, Mahendran G, Pettitt BJ, Henriquez OA. Implementation and evaluation of a virtual elective in otolaryngology in the time of COVID-19. Otolaryngol Head Neck Surg 2021 Mar;164(3):556-561. [doi: 10.1177/0194599820951150] [Medline: 32779955]
49. Taylor N, Wyres M, Green A, et al. Developing and piloting a simulated placement experience for students. Br J Nurs 2021 Jul 8;30(13):S19-S24. [doi: 10.12968/bjon.2021.30.13.S19] [Medline: 34251853]
50. Villa S, Janeway H, Preston-Suni K, et al. An emergency medicine virtual clerkship: made for COVID, here to stay. West J Emerg Med 2021 Dec 17;23(1):33-39. [doi: 10.5811/westjem.2021.11.54118] [Medline: 35060858]
51. Weston J, Zauche LH. Comparison of virtual simulation to clinical practice for prelicensure nursing students in pediatrics. Nurse Educ 2021;46(5):E95-E98. [doi: 10.1097/NNE.0000000000000946] [Medline: 33186190]
52. White MJ, Birkness JE, Salimian KJ, et al. Continuing undergraduate pathology medical education in the Coronavirus disease 2019 (COVID-19) global pandemic: the Johns Hopkins virtual surgical pathology clinical elective. Arch Pathol Lab Med 2021 Jul 1;145(7):814-820. [doi: 10.5858/arpa.2020-0652-SA] [Medline: 33740819]
53. Wik V, Barfield S, Cornwall M, Lajoie R. Finding the right balance: student perceptions of using virtual simulation as a community placement. Int J Nurs Educ Scholarsh 2022 Jan 1;19(1). [doi: 10.1515/ijnes-2021-0135] [Medline: 36103581]
54. Williams C, Familusi OO, Ziemba J, et al. Adapting to the educational challenges of a pandemic: development of a novel virtual urology subinternship during the time of COVID-19. Urology 2021 Feb;148(148):70-76. [doi: 10.1016/j.urology.2020.08.071] [Medline: 33045288]
55. Zhou T, Huang S, Cheng J, Xiao Y. The distance teaching practice of combined mode of massive open online course micro-video for interns in emergency department during the COVID-19 epidemic period. Telemed J E Health 2020 May;26(5):584-588. [doi: 10.1089/tmj.2020.0079] [Medline: 32271650]
56. Organisation for Economic Co-operation and Development. Guidance:Countries defined as developing by the OECD. URL: https://www.gov.uk/government/publications/countries-defined-as-developing-by-the-oecd [accessed 2022-11-15]
57. Goh PS, Sandars J. A vision of the use of technology in medical education after the COVID-19 pandemic. MedEdPublish (2016) 2020;9:49. [doi: 10.15694/mep.2020.000049.1] [Medline: 38058893]
58. Bridge P, Adeoye J, Edge CN, et al. Simulated placements as partial replacement of clinical training time: a Delphi consensus study. Clin Simul Nurs 2022 Jul;68:42-48. [doi: 10.1016/j.ecns.2022.04.009]
59. Current recovery programme standards. Nursing and Midwifery Council. 2022. URL: https://www.nmc.org.uk/globalassets/sitedocuments/education-standards/current-recovery-programme-standards.pdf [accessed 2022-10-17]
60. Barr H, Ford J, Gray R, et al. CAIPE (2017) interprofessional education guidelines. Centre for the Advancement of Interprofessional Education. 2017. URL: https://www.caipe.org/resources/publications/caipe-publications/caipe-2017-interprofessional-education-guidelines-barr-h-ford-j-gray-r-helme-m-hutchings-m-low-h-machin-reeves-s [accessed 2022-08-10]
61. CAIPE Strategy 2022-2027. Centre for the Advancement of Interprofessional Education. URL: https://www.caipe.org/strategy#1 [accessed 2022-11-15]

62. Fealy S, Jones D, Hutton A, et al. The integration of immersive virtual reality in tertiary nursing and midwifery education: a scoping review. Nurse Educ Today 2019 Aug;79(79):14-19. [doi: 10.1016/j.nedt.2019.05.002] [Medline: 31078869]
63. Sanders E, Sappers P. Convivial Toolbox: Generative Research at the Front End of Design: BIS; 2008.
64. Sanderson L, Choma L, Cappelli T, et al. Developing online simulated practice placements: a case study. Br J Nurs 2023 Jul 6;32(13):636-643. [doi: 10.12968/bjon.2023.32.13.636] [Medline: 37410679]

Abbreviations

ADDIE: analysis, design, development, implementation, and evaluation
AI: artificial intelligence
BMJ STEL: British Medical Journal: Simulation and Technology Enhanced Learning
FTF: face-to-face
IPE: interprofessional education
NHS: National Health Service
PAGER: patterns, advances, gaps, evidence for practice, research recommendations
PCC: population, concept, context
PRISMA: Preferred Reporting Items for Systematic reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews
SPIDER: sample, phenomenon of interest, design, evaluation, research type
VR: virtual reality
VSP: virtual simulated placement

Edited by B Lesselroth; submitted 25.03.24; peer-reviewed by A Duncan, C McCrorie, S Markham; revised version received 21.11.24; accepted 02.01.25; published 10.06.25.

Please cite as:
Samson J, Gilbey M, Taylor N, Kneafsey R
Virtual Simulated Placements in Health Care Education: Scoping Review
JMIR Med Educ 2025;11:e58794
URL: https://mededu.jmir.org/2025/1/e58794
doi:10.2196/58794

Multidisciplinary Oncology Education Among Postgraduate Trainees: Systematic Review

Houman Tahmasebi[1,2], MD, MSc; Gary Ko[2], MD; Christine M Lam[1], MSc; Idil Bilgen[3], MDc; Zachary Freeman[4], BSc; Rhea Varghese[1], BHSc; Emma Reel[5], MSW; Marina Englesakis[6], HBA, MLIS; Tulin D Cil[1,2,5], MD, MEd

[1]Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada

[2]Department of Surgery, University of Toronto, Toronto, ON, Canada

[3]School of Medicine, Koç University, Istanbul, Turkey

[4]Faculty of Science, Wilfrid Laurier University, Waterloo, ON, Canada

[5]Sprott Department of Surgery, Princess Margaret Cancer Centre, University Health Network, 6th floor, 700 University Ave, Toronto, ON, Canada

[6]Library and Information Services, University Health Network, Toronto, ON, Canada

Corresponding Author:
Tulin D Cil, MD, MEd
Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada

Abstract

Background: Understanding the roles and patient management approaches of the entire oncology team is imperative for effective communication and optimal cancer treatment. Currently, there is no standard residency or fellowship curriculum to ensure the delivery of fundamental knowledge and skills associated with oncology specialties with which trainees often collaborate.

Objective: This study is a systematic review that aims to evaluate the multidisciplinary oncology education in postgraduate medical training.

Methods: A systematic literature search was performed using MEDLINE, Embase, Cochrane Database of Systematic Reviews, Cochrane CENTRAL, APA PsycINFO, and Education Resources Information Center in July 2021. Updates were performed in February 2023 and October 2024. Original studies reporting the effectiveness of multidisciplinary oncology training among residents and fellows were included.

Results: A total of 6991 studies were screened and 24 were included. Fifteen studies analyzed gaps in existing multidisciplinary training of residents and fellows from numerous fields, including surgical, medical, and radiation oncology; geriatrics; palliative medicine; radiology; and pathology programs. Trainees reported limited teaching and knowledge of oncology outside of their respective fields and endorsed the need for further multidisciplinary oncology training. The remaining 9 studies assessed the effectiveness of educational interventions, including tumor boards, didactic sessions, clinical rotations, and case-based learning. Trainees reported significant improvements in multidisciplinary oncology knowledge and skills following the interventions.

Conclusions: These data suggest postgraduate medical trainees have limited formal multidisciplinary oncology training. Existing educational interventions show promising results in improving trainees' oncology knowledge and skills. There is a need for further research and the development of multidisciplinary oncology curricula for postgraduate medical training programs.

Trial Registration: PROSPERO CRD42022271308; https://www.crd.york.ac.uk/PROSPERO/view/CRD42022271308

(JMIR Med Educ 2025;11:e63655) doi:10.2196/63655

KEYWORDS

multidisciplinary; oncology; postgraduate medical education; resident; fellow; surgery; hematology; radiation oncology; geriatrics; palliative

Introduction

Cancer was the second leading cause of death in the United States in 2023 [1]. Cancer care often requires a team of physicians including surgical, medical, and radiation oncologists, as well as specialists in radiology and pathology [2]. Knowledge of collaborating oncologists' roles and appropriate multidisciplinary referrals may impact cancer treatment. There is evidence of improved adherence to standard treatment guidelines with multidisciplinary referrals for patients with prostate [3], lung cancer [4], and bladder cancer [5].

There is considerable potential to improve interdisciplinary communication between various oncologic specialists and to optimize psychosocial support for patient care. Therapies with different oncologists must be well coordinated and specifically selected based on the medical and social needs of each patient.

To achieve this, knowledge of other disciplines' roles, responsibilities, and treatment options is necessary for effective communication and optimal cancer care.

There is currently no standard curriculum for delivering multidisciplinary oncology education in residency and fellowship programs in the United States [6-10]. Mattes et al [11] identified that while many of the program requirements for oncology subspecialties emphasize the importance of providing multidisciplinary cancer care, how this occurs varies widely between subspecialties. Not all programs mandate multidisciplinary oncology rotations or experiential specialty training, and only a select few require attendance at multidisciplinary tumor board meetings (MTBM) [6-12]. Such a training gap may impact trainee education and, as a result, influence referral patterns and the timely access of patients to multimodal cancer therapies.

The objective of this study was to perform a systematic review of the literature to evaluate the multidisciplinary oncology education in postgraduate medical training (ie, interns, residents, and fellows). This study provides a review of literature analyzing the education of learners about the role of any collaborating physician specialty involved in oncology care, including but not limited to, medical oncology, radiation oncology, surgical oncology, and palliative care. These data summarize gaps in training programs identified across studies, the suggested educational interventions to bridge these gaps, and limitations in the literature within the field.

Methods

Research Design and Methodology

This systematic review was reported based on PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [13]. The protocol was registered and published by PROSPERO (ID: CRD42022271308).

Search Strategy

A search strategy was developed with the assistance of an information specialist using these and other related terms: "Residents or Fellows or Trainees or Medical Training" AND "Education or Training Programs" AND "Multidisciplinary" AND "Oncology." The following databases were searched from inception: MEDLINE, MEDLINE In-Process, Embase Classic + Embase, Cochrane Central Register of Controlled Trials, Cochrane Database of Systematic Reviews, and APA PsycINFO (all via the Ovid platform); and Education Resources Information Center via the EbscoHost platform. The search was initially performed on July 21, 2021, and updated twice (ie, on February 26, 2023, and October 9, 2024). Table S1 in Multimedia Appendix 1 shows the number of citations identified from each database. The search strategy and the number of citations identified via MEDLINE are included in Table S2 in Multimedia Appendix 1.

Eligibility Criteria

Eligibility criteria were developed prior to the search strategy. The scope of this study was to evaluate the multidisciplinary oncology education offered by residency and fellowship programs to postgraduate medical trainees. Thus, the first eligibility criterion was the inclusion of studies investigating postgraduate medical training (ie, interns, residents, and fellows). Studies about nonphysician specialties (eg, nursing, pharmacy, or dentistry), attending or staff physicians, or those involving solely Masters, PhD, or medical students were excluded. Studies were included if their focus was specific to oncology care. Selected studies focused on multidisciplinary aspects of medical education, which included knowledge of collaborating medical specialties and their roles in cancer care (eg, surgical trainees' knowledge of radiation or medical treatments). Trainees from all specialties were included, as long as the study was assessing the multidisciplinary oncology education of trainees, and therefore, these were not necessarily restricted to oncology residency or fellowship programs (eg, medical oncology, radiation oncology, surgical oncology). Only primary research papers and studies available in English (ie, both original and translations to English) were included. Thus, all reviews, case studies, opinion papers, abstract-only papers, conference literature, and short reports were excluded.

Study Selection

There were 2 stages of review: title and abstract screening, followed by full-text screening. A total of 6 reviewers (HT, GK, CML, IB, ZF, and RV) were involved, and studies were screened by a minimum of 2 independent reviewers at each stage. Discrepancies were resolved by a third reviewer. Both screening stages were performed on Covidence [14], a web-based systematic review organization software.

Data Extraction and Synthesis

Data extraction was performed on the selected studies. Studies were divided between 3 reviewers (HT, CML, and RV) who performed data extraction. Study design, study population, outcome measures, and main results were extracted from each study.

Quality Assessment

Selected studies were independently assessed for quality by 2 independent reviewers (CL, IB, and RV) using the Mixed Methods Appraisal Tool (MMAT) version 2018 [15]. Discrepancies were resolved through discussion with a third author (HT). The MMAT was chosen due to its ability to concomitantly assess multiple study types (ie, qualitative, quantitative randomized controlled trial, quantitative nonrandomized, quantitative descriptive, or mixed methods). Each study was evaluated on a set of 5 criteria depending on the study type. For survey studies, the risk of nonresponse bias was deemed to be high if the response rate was below 70%. Studies were assigned an overall quality score ranging from 0 to 5 stars based on the number of criteria that were met.

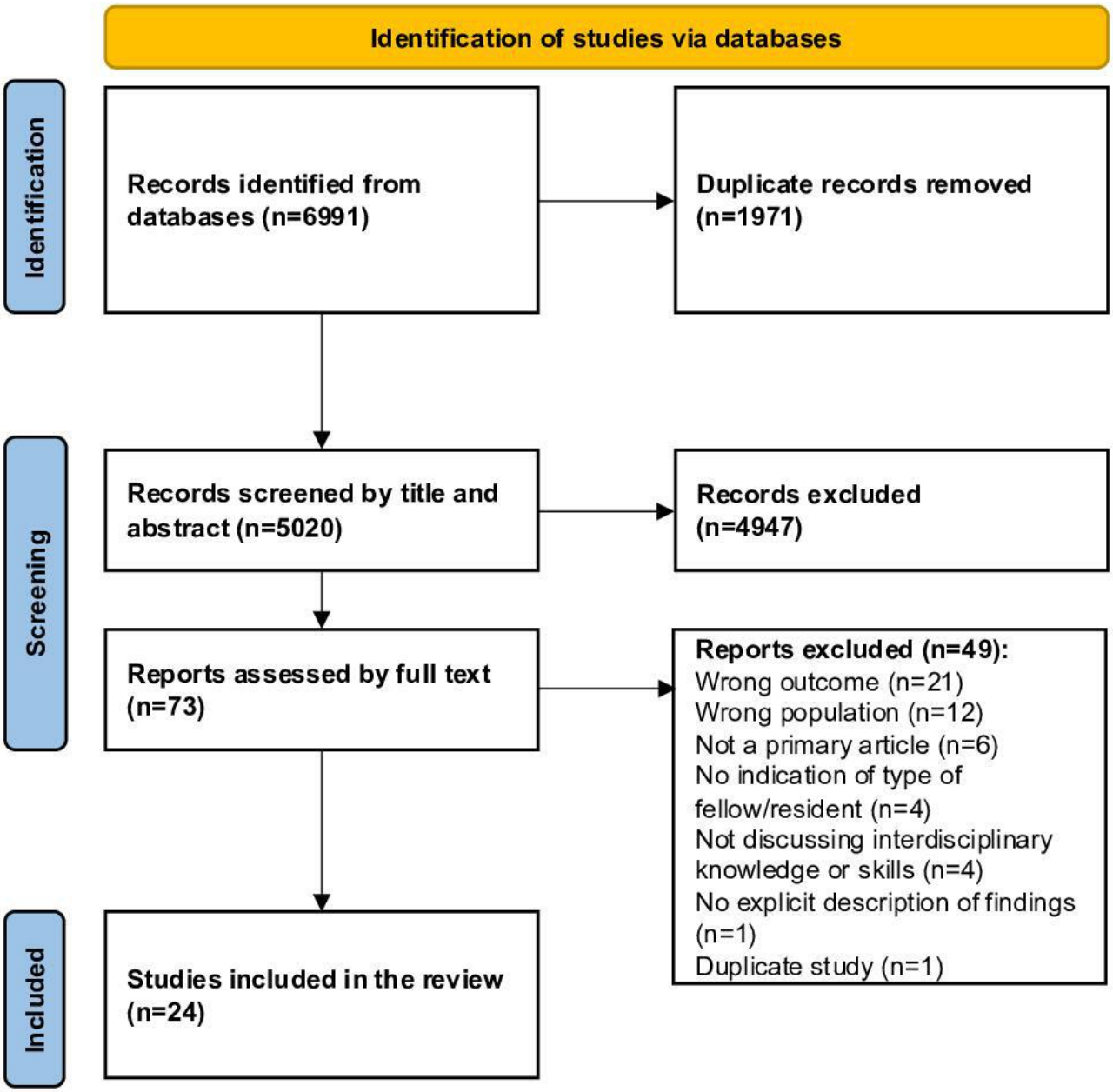
Results

Study Characteristics

The search strategy resulted in a total of 6991 studies. After removing duplicates between databases, 5020 unique studies were identified. A total of 73 studies remained after title and abstract screening. Full-text screening excluded 49 studies, and

24 studies were therefore included in the final analysis. The PRISMA flow diagram is demonstrated in Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram of the systematic review. Adapted from Page et al [13].



The remaining 24 studies were divided into 2 categories. Fifteen studies assessed the quality of existing postgraduate oncology training based on trainees' multidisciplinary knowledge. The remaining 9 assessed trainees' multidisciplinary knowledge following an educational intervention. For the latter category, all studies with educational interventions directed toward improving multidisciplinary oncology knowledge and skills among interns, residents, and fellows were included. These included studies that are part of the formal postgraduate medical training (eg, residency or fellowship program), as well as external initiatives for improving multidisciplinary oncology training. Studies involving educational interventions for medical students and staff or attending physicians were not included.

Existing Multidisciplinary Training

A summary of the 15 studies evaluating the impact of existing multidisciplinary oncology training is included in Table 1. These studies included surgical or surgical oncology fields [16-23], hematology or medical or hematology oncology [16,20-22,24,25], geriatrics or geriatric oncology [20,22,26], radiation oncology [16,20,21,23,27], palliative medicine [16,20], radiology [21], pathology [21], genetics [23], dermatology [23], pediatric specialties [28], and other medical fields (eg, internal medicine, nephrology, neurology) [22,23].

Table . Summary of studies evaluating the existing multidisciplinary education across postgraduate medical training programs.

Reference	Study design	Sample	Outcome measure	Main findings and conclusions
Akthar et al [16]	Electronic surveys completed by oncology trainees and program directors across the United States in 2013	• 557 hematology or medical oncology, surgical oncology, radiation oncology, and palliative medicine residents and fellows • 141 hematology or medical oncology, surgical oncology, radiation oncology, and palliative medicine program directors	Proportion of trainees who received formal education in oncology fields outside of their specialty	• Generally limited interdisciplinary oncology education: ≤70% of trainees reported receiving formal interdisciplinary education; highest rate of training in radiation oncology (70% of trainees) and lowest rate of training in geriatric oncology (19% of trainees) • Consistently lower rates of interdisciplinary oncology training reported by trainees compared to program directors (P<.01)
Brenner and De Donno [17]	Survey of postgraduate year 1 - 5 residents from 3 general surgery programs: Florida Atlantic University, The University of Iowa, and The University of Connecticut	• 135 general surgery residents	Proportion of residents who indicated receiving training in a specific multidisciplinary field	• Limited proportion of residents indicated receiving multidisciplinary training: radiation oncology (23%), chemotherapy (31%), and palliative medicine (53%) • Majority (82%) of residents endorsed further multidisciplinary training
David et al [25]	Survey of hematology residents (ie, postgraduate years 4 - 5) or fellows (ie, postgraduate year 6 - 7) across Canada as part of a cross-sectional study	• 29 hematology residents and 3 hematology fellows	Geriatric oncology curriculum needs assessment	• 56.3% did not receive geriatric oncology teaching • 96.9% endorsed the inclusion of geriatric training in hematology residency
Delaye et al [22]	Surveys completed by French residents and senior physicians regarding the field of onco-nephrology	• Residents (n=130) and senior physicians (n=98) from nephrology, oncology, hematology, surgery and geriatrics	Current practices in onco-nephrology, information resources, existing cooperation networks, and expectations about onco-nephrology	• Oncology residents rated their confidence in facing renal events as 5.5/10 • Nephrology residents rated their confidence in facing cancer events as 6.0/10 • 21% of residents had received onco-nephrology teaching, which was judged as insufficient
Eid et al [24]	Review of literature, expert consultation, review of fellows' rotation evaluations, and interviews with current and recently graduated fellows, as a means of needs assessment for the development of a geriatric oncology program at MD Anderson Cancer Center	• 9 current hematology-oncology fellows (years 1 - 3) at MD Anderson Cancer Center • 2 MD Anderson faculty members who recently graduated from a hematology-oncology fellowship	Geriatric oncology training program needs assessment	Top 3 identified needs for geriatric oncology programs, based on current educational gaps: • Geriatric assessment • Pharmacology knowledge • Psychosocial knowledge

Reference	Study design	Sample	Outcome measure	Main findings and conclusions
Givi et al [29]	Semistructured interviews with program directors and faculty in head and neck surgery across the United States and Canada over a 7-month period	• 58 participants including head and neck surgery program directors and faculty	Head and neck surgical oncology training needs assessment	• 38% endorsed increasing the number of head and neck surgery fellows' interactions with medical oncology, radiation oncology, and speech and language pathology • 85% view exposure to multidisciplinary teams as essential in training curricula
Le Nail and Samargandi [30]	Web-based questionnaire on various aspects of MTBMs[a] completed by French orthopedic oncology residents	• 27 orthopedic oncology residents	Residents' opinions on educational impact and areas of improvement for MTBMs	• 54% agreed that MTBM is an appropriate venue for teaching • 75% endorsed that MTBMs improved their knowledge of other specialties involved • 71% indicated opportunities to improve teaching during MTBM, the most popular suggestion being active participation of residents (voted by 46% of all residents)
Maggiore et al [26]	Web-based survey completed by program directors of geriatrics fellowship programs in the United States	• 67 geriatrics program directors	Proportion of program directors offering or endorsing future learning opportunities in the field of geriatric oncology	• Majority (81%) of program directors offered didactic teaching in the form of formal geriatric oncology lectures/seminars • Limited number of program directors offered clinical experience: 39% offered mandatory oncology clinical experience and 46% offered clinical electives • Majority (77%) endorsed oncology training as part of the geriatrics fellowship
Mäurer et al [23]	Web-based survey distributed to all junior oncology groups represented in Young Oncologists United in Germany regarding interdisciplinarity in oncology	• 294 participants including 268 physicians (staff and trainees) from internal medicine, gynecology, radiotherapy and radiation oncology, general surgery, genetics, neurosurgery, urology, neurology, and dermatology	Opinions on interdisciplinarity at clinic, educational, and research levels	• 63.1% assigned a high priority to interdisciplinary residency training • Only 18.3% had the opportunity to participate in rotations in other specialties beyond their curriculum • 71.4% were interested in participating in rotations in other specialties • 73.1% of those who completed interdisciplinary rotations benefited from them

Reference	Study design	Sample	Outcome measure	Main findings and conclusions
Morris et al [27]	Web-based survey completed by radiation oncology residents in Australia, New Zealand, and Singapore	• 61 radiation oncology residents	Proportion of residents who indicated receiving or endorsing future geriatric oncology training	• Majority (91.8%) did not receive any geriatric training • Limited number of residents (39.3%) comfortable managing complex geriatric issues • Majority (85.3%) endorsed additional geriatric training
Morris et al [20]	2-stage Delphi consensus with input from a panel of internationally recognized oncology experts, staff physicians, radiation oncology and clinical oncology trainees, allied health professionals, patients, and caregivers. Experts were from geriatrics, geriatric oncology, and radiation oncology. Staff physicians were from clinical/medical oncology, palliative care, and surgical oncology.	• A total of 103 and 54 individuals participated in rounds 1 and 2 of the modified Delphi consensus process, respectively • Majority were radiation oncologists (43%)	Establishing learning outcomes for a geriatric radiation oncology curriculum	• 33 learning outcomes identified in the areas of fundamental geriatric medicine concepts, epidemiology, geriatric screening, planning and delivery of radiation therapy, geriatric palliative care, surgery, systematic treatment, research, communication skills, and health advocacy
Park et al [18]	30 item self-efficacy survey completed by residents at Ohio State University Wexner Medical Center, in order to measure knowledge and skills in 6 breast cancer care aspects: genetics, surgery, medical oncology, radiation oncology, pathology, and radiology	• 31 general surgery residents	Residents' perceived capability (ie, self-efficacy score) in various domains of breast cancer care	• Highest self-efficacy in surgery (3.56/5) vs lowest in genetics (2.67/5), radiation oncology (2.67/5), and pathology (2.67/5) • Significant improvement of self-efficacy in surgery only (P=.002) with additional years in residency
Picca and Reed [28]	Semistructured interviews with faculty and trainees across pediatric oncology, radiology, pathology, surgical oncology, and palliative care	• 4 pediatric oncology fellows • 11 pediatric oncology, pathology, radiology, palliative care, and surgical oncology faculty physicians	Exploration of learning in tumor boards	• Trainees found tumor board presentation to be educational • Barriers to learning: competing clinical/administrative responsibilities • Facilitators to learning: learning-focused goals, faculty mentorship during presentation preparation, collaborative discussion, content tailored to learners and board exams, and supportive environment • Web-based tumor boards promoted accessibility and convenience but decreased learning due to limited engagement, discussion, and professional relationship development

Reference	Study design	Sample	Outcome measure	Main findings and conclusions
Walraven et al [21]	Semistructured interviews with Dutch residents and specialists in medical/surgical/radiation oncology, radiology, nuclear radiology, and pathology participating in MDTMs[b]	• 19 residents • 16 specialists	Residents' barriers and facilitators to participate in MDTMs	• 100% agreed that MDTMs play an important role in both education and patient care • Barriers: insufficient supervisor guidance, time constraints, meeting atmosphere and hierarchy, strict regulations, unfamiliarity, and resident's personal characteristics • Solutions: MDTM simulation training, and training courses on communication and meeting skills
Wilson et al [19]	Survey of applicants to Roswell Park Cancer Institute surgical oncology fellowship program	• 29 general surgery residents or recent general surgery graduates applying to surgical oncology fellowship	Proportion of applicants with breast surgery exposure and their comfort with medical and surgical management of breast cancer	• Majority (65%) had exposure to multidisciplinary breast cancer clinics, involving medical and surgical oncologists • Lower level of comfort (7.07/10) with breast cancer medical management compared to surgical management (7.34 - 9.10/10 depending on the type of surgery)

[a]MTBM: multidisciplinary tumor board meeting.

[b]MDTM: multidisciplinary team meetings.

Thirteen studies obtained opinions of trainees with respect to multidisciplinary oncology education within their training programs [16-25,27,28,30]. Morris et al [20] used a Delphi consensus process, and 4 studies directly interviewed trainees and faculty [21,24,28,29]. The remainder of the studies used surveys. Maggiore et al [26] surveyed geriatrics program directors, Givi et al [29] surveyed head and neck surgery program directors, and Akthar et al [16] surveyed program directors of pediatric and adult hematology oncology, surgical oncology, radiation oncology, and palliative medicine. Eid et al [24] used a combination of expert consultation, trainee interviews, review of trainee rotation evaluations, and literature review to assess their multidisciplinary educational needs.

While all studies analyzed the quality of existing multidisciplinary education, there were differences in the disciplines investigated across studies. Akthar et al [16], Delaye et al [22], Mäurer et al [23], Walraven et al [21], Picca and Reed [28], and Brenner and De Donno [17] focused on identifying broad gaps in multidisciplinary education including knowledge and skills of trainees in numerous fields, such as radiation, surgical, and medical oncology, radiology, pathology, geriatrics, palliative medicine, and other pediatric and medical fields. The remaining 8 studies focused on a more specific set of trainee skills. David et al [25], Eid et al [24], and Maggiore et al [26] assessed gaps in geriatric oncology education among hematology residents and fellows, hematology oncology fellows, and geriatrics fellows, respectively. Morris et al [20,27] assessed gaps in the radiation oncology training curriculum. Park et al [18] and Wilson et al [19] assessed the quality of general surgery residency training in breast cancer care. Le Nail and Samargandi [30] evaluated the quality of tumor boards for orthopedic oncology trainees. Finally, Givi et al [29] performed a needs assessment analysis of the head and neck surgery training curriculum.

13 studies assessed the strengths and weaknesses of oncology training programs [16-19,21-23,25-30]. Of these, 11 found that trainees had limited exposure to multidisciplinary oncology disciplines, barriers to attending multidisciplinary oncology meetings, and a low level of trainee comfort in multidisciplinary oncology knowledge [16-19,21-23,25-28]. Givi et al [29] found that 27% of interviewees indicated exposure to multidisciplinary care as a strength of the head and neck surgery training program, although 38% endorsed the need to improve fellows' multidisciplinary participation. In general, Akthar et al [16] found the least amount of multidisciplinary training in geriatric oncology, compared to palliative medicine, medical, radiation, and surgical oncology. Similarly, Morris et al [27] found that less than 10% of radiation oncology trainees received geriatrics training. Furthermore, less than half of geriatrics fellows were offered geriatric oncology rotations [26]. For multidisciplinary

breast cancer management, Park et al [18] found limited training in genetics, radiation oncology, and pathology among surgical residents, compared to rotations within surgery, radiology, and medical oncology. Brenner and De Donno [17] found that a small proportion of general surgery residents received training in the fields of radiation (23%) and medical oncology (31%), but over half (53%) received exposure to palliative care.

Additionally, 11 studies researched areas of improvement for multidisciplinary oncology education among the postgraduate programs via surveys, interviews, Delphi consensus, and literature search [17,20,21,23-30]. Maggiore et al [26] and Morris et al [27] found that 77% of geriatrics fellows and 85.3% of radiation oncology residents advocated for further geriatric oncology training. David et al [25] found that over 95% of hematology trainees endorsed geriatric training during residency. 82% of general surgery residents surveyed by Brenner and De Donno [17] agreed that additional multidisciplinary training is needed to optimize cancer care. Additionally, based on an educational needs assessment, Eid et al [24] found that the top 3 priorities for a geriatric oncology program included geriatric assessment, pharmacology, and psychosocial skills. MTBMs were found to enhance trainee experience and multidisciplinary oncology education [21,28,30]. However, some barriers to attending meetings included time constraints, clinical duties, and lack of active resident participation [21,28,30]. Residents and specialists interviewed by Walraven et al [21] suggested that the educational value of multidisciplinary team meetings could be improved through additional training such as multidisciplinary team meeting simulations and courses on effective communication and meeting skills.

Impact of Educational Interventions

A summary of the 9 studies analyzing the impact of educational interventions is included in Table 2. The majority included general surgery trainees [31-35]. Faculty and trainees from radiation oncology [32,35], medical oncology [12,35], respirology [12,36], thoracic surgery [12], gynecology [35], urology [37], and palliative medicine [38] were also included. All 9 studies demonstrated improvements in multidisciplinary oncology knowledge and skills postintervention.

Table . Summary of studies evaluating the impact of multidisciplinary educational interventions.[a]

Reference	Study design	Sample	Outcome measure	Main findings and conclusions
Cook et al [31]	Electronic surveys were sent to general surgery residents at the completion of 4-week rotations in MDB[b], USOS[c], and community-based TSR[d] at Oregon Health and Science University in 2010 - 2013. MDB included operative time, as well as half-days in pathology, radiology, medical oncology, and surgery clinic.	• Total sample size: 32 in MDB, 73 in USOS, and 51 in TSR • Operative logs of 29 residents in MDB, 11 in TSR, and 12 in USOS were obtained	• Trainee satisfaction based on surveys • Operative volume based on operative logs	• MDB rotation residents rated the opportunity to perform and learn procedures higher than those in USOS (P=.02) and TSR (P=.01) • 83% of MDB residents' operative experience included breast cancer operations, compared to 71% of USOS and 12% of TSR groups • MDB rotation residents rated higher on the quality of faculty teaching and educational materials than those on TSR (P=.03 and P=.04, respectively)
Khoshgoftar et al [37]	Short interviews were held with urology residents and faculty members regarding needs for holding web-based tumor boards prior to implementation of 20 monthly web-based tumor boards. Tumor boards were assessed through questionnaires postintervention, resident pretest and posttest scores for 5 consecutive tumor boards, and external evaluators from the faculty of urology.	• 35 urology residents • 25 urology faculty members • Panelists from pathology, radiation oncology, medical oncology, radiology, and nuclear medicine	• Needs assessment, satisfaction levels, pretest and posttest scores, recommendations from external evaluators	• Resident needs assessment was divided by level of importance and postgraduate years (ie, years 1 - 2 vs 3 - 4). An important limitation to participate was significant clinical responsibilities, particularly for lower year residents • High resident satisfaction rate (71% - 88%) based on various aspects of web-based tumor boards. The most important technical issue was the low bandwidth speed. • There was significant improvement in resident posttest scores in the majority of sessions
Mackay et al [36]	Respiratory and oncology trainees completed a 3-hour MDTM[e] simulation session and completed pre- and postsimulation questionnaires	• 19 oncology and respiratory trainees (specialty training years 3 - 7)	• Perceptions of current training programs, confidence presenting in MDTMs, use of the simulation, and impact on future clinical practice	• Trainees rated 4/10 for how well their program prepared them to present at MDTM • Trainee confidence in presenting in MDTMs increased from 5/10 to 7/10 postintervention (P<.01) • Trainees rated 9/10 for usefulness and 9/10 for likelihood the session will lead to changes in their practice

Reference	Study design	Sample	Outcome measure	Main findings and conclusions
Martin et al [38]	Fellows completed three 1-hour lectures in palliative radiotherapy, as well as pre- and postcourse questionnaires and objective knowledge assessment multiple-choice questions.	• 5 hospice and palliative medicine fellows at the University of California, San Diego	• Knowledge and confidence in palliative radiotherapy	• Postintervention improvement in trainee-reported confidence in discussion with patients about radiotherapy (0.009), managing its common side effects (*P*=.021), and identifying oncologic emergencies related to radiotherapy (*P*=.012) • Significant improvement in radiotherapy knowledge based on objective knowledge assessment questions (22% vs 86% pre- vs postintervention; *P*=.010) • Increased trainee-reported likelihood of collaboration with radiation oncologists postintervention (*P*=.014)
Mattes et al [12]	Faculty, fellows, and residents attended a didactic lecture on radiation therapy in lung cancer care. Knowledge was tested using multiple choice questions pre- and postintervention.	• A total of 121 faculty and trainees from pulmonology, thoracic surgery, and medical oncology • Pretest: 54 residents/fellows and 9 faculty participated • Posttest: 23 residents/fellows and 2 faculty participated	• Knowledge of radiation therapy in lung cancer treatment and comfort in appropriate referral to radiation oncology	• The majority had no didactic training (75%) or rotations (85.5%) in radiation oncology preintervention • Significant improvements in mean objective test scores postintervention (*P*<.001) • Postintervention, 100% of participants felt more knowledgeable in radiation therapy and 96% felt more comfortable making appropriate radiation oncology referrals
Meani et al [35]	Faculty and trainees completed a postintervention questionnaire following a multidisciplinary breast cancer course.	• A total of 42 participants in medical oncology, radiation oncology, gynecology, and general surgery • 11 heads of department/professors • 17 consultants/attending Physicians • 14 trainees: residents, medical fellows, PhD students, and postdoctoral fellows	• Opinions on the impact of the course	• Postintervention, 64% made changes in their clinical practice and 33% made institutional changes in breast cancer management • 95% reported increased knowledge of MDB cancer care
Sloan et al [32]		• 22 general surgery residents • 3 radiation oncology residents • 15 faculty at stations • 12 patients with breast cancer at stations	• Self-reported trainee improvement in breast cancer care–specific skills • Perception of faculty, patients, and residents of the overall quality of intervention	

Reference	Study design	Sample	Outcome measure	Main findings and conclusions
	Residents at the University of Kentucky received multidisciplinary instruction and completed 15 case-based stations about various domains of breast cancer care (ie, surgical oncology, medical oncology, radiology, radiation oncology, plastic surgery, and pathology). Surveys about the overall quality of intervention were completed by patients, faculty, and residents. Residents also completed pre- and postintervention surveys regarding specific breast cancer care-specific skills.			• Statistically significant trainee-reported improvement for all measured skills, including fine-needle aspiration, mammography interpretation, and treatment discussion with patients • Overall, intervention rated favorably by trainees, faculty, and patients
Sloan et al [33]	Residents at the University of Kentucky completed 12 case-based stations during a head and neck oncology workshop, designed by faculty from general surgery, speech pathology, dentistry, radiation therapy, otolaryngology, plastic and reconstructive surgery, pathology, anesthesiology, and cardiothoracic surgery. Surveys about the overall quality of intervention were completed by patients, faculty, and residents. Residents also completed pre- and postintervention surveys regarding head and neck-specific skills.	• 21 general surgery residents • 11 faculty at stations • 8 standardized patients at stations (including 6 patients with cancer)	• Self-reported trainee improvement in skills relevant to head and neck cancer care • Perception of faculty, patients, and residents of the overall quality of intervention	• Statistically significant trainee-reported improvement for most skills postintervention (P<.001) • Overall, intervention rated favorably by trainees, faculty, and patients • Residents generally endorsed having intervention minimum twice during residency
Sloan et al [34]	2 groups received multidisciplinary teaching in breast cancer care, including radiation oncology, radiology, surgery, and medical oncology, in the form of a 15-station workshop. The other 2 groups served as controls. 1 intervention and 1 control group were administered an 11-problem OSCE[f] assessment immediately postintervention and the remaining 2 groups were administered the same OSCE assessment 8 months later. Residents were assessed by faculty and standardized patients during OSCE assessments.	• 48 general surgery residents from the University of Kentucky, divided evenly into 4 groups • 15 faculty at stations • 12 standardized patients at stations (including 5 patients with cancer)	• Skills in diagnosis and management of breast cancer postintervention, assessed by faculty and standardized patients during OSCE assessments	• Improvement in skills of residents who attended the workshop, compared to the control group, both immediately and 8 months postintervention (P<.01) • Residents' skills diminished after 8 months, as evidence by the difference in skill set between the group tested immediately versus the one tested 8 months postintervention (P<.004)

[a]Patients who performed assessments included actual and simulated patients.

[b]MDB: multidisciplinary breast.

[c]USOS: university surgical oncology service.

[d]TSR: traditional surgical rotation.

[e]MDTM: multidisciplinary team meeting.

[f]OSCE: Objective Structured Clinical Examination.

The study by Cook et al [31] compared the impact of a multidisciplinary breast rotation to traditional oncology or community rotations using trainee self-evaluations. Martin et al [38] and Mattes et al [12] analyzed the effectiveness of didactic learning for palliative radiotherapy and lung cancer radiotherapy, respectively, using pre- and postcourse trainee evaluations. Meani et al [35] studied the impact of a multidisciplinary breast cancer course on the knowledge and practice of faculty and trainees using a questionnaire. Three studies by Sloan et al tested the quality of case-based instruction, involving workshops or Objective Structured Clinical Examination (OSCE) stations, where evaluations were completed by trainees, standardized patients, and faculty [32-34]. In the 2004 study by Sloan et al [34], faculty and standardized patient completed evaluations following the observation of trainees in OSCE stations. Patient ratings mainly included interpersonal skills, while faculty ratings included both the clinical and interpersonal skills of trainees. In the other 2 Sloan et al studies, faculty and standardized patients provided feedback on the overall quality of workshops, rather than a specific focus on trainee skills [32,33]. Many of the standardized patients were actual patients with cancer [32-34]. Two of the Sloan et al studies with breast cancer-specific stations focused on knowledge and skills in the following fields: surgical, medical, and radiation oncology; pathology; plastic surgery; and radiology [32,34]. A pilot study by the same group included a head and neck workshop in which stations were designed by faculty from general surgery, radiation oncology, cardiothoracic surgery, otolaryngology, plastic surgery, pathology, anesthesiology, speech pathology, and dentistry [33].

In 8 of these 9 studies, the benefit of educational interventions was noted by the trainees through self-assessment of knowledge or skills [12,31-33,35-38], while Sloan et al [34] demonstrated improvements in knowledge or skills, as assessed by faculty and patients following the observation of trainees in OSCE stations. In addition to reporting subjective benefits, Khoshgoftar et al [37], Mattes et al [12], and Martin et al [38] used objective assessments to demonstrate improvements in trainee knowledge postintervention. Interestingly, Sloan et al [34] showed that while the intervention benefited residents' knowledge and skill set in breast cancer management both immediately after and 8 months postintervention, it declined after 8 months. In the other 2 studies by this group [32,33], trainees, faculty, and patients rated the interventions highly.

Quality Assessment

A summary of the MMAT quality assessment is included in Table S3 in Multimedia Appendix 1. Five studies were categorized as nonrandomized, 4 as qualitative, 13 as quantitative descriptive, 1 as mixed methods, and 1 as randomized controlled. Studies were given a score out of 5, based on the number of MMAT criteria met. Two studies were given an overall MMAT quality rating of 3 stars, 14 studies were rated as 4 stars, and the remaining 8 were rated as 5 stars. Overall, all studies were deemed to be satisfactory by authors, based on MMAT quality assessment criteria.

Discussion

Principal Results

To our knowledge, this is the first systematic review of multidisciplinary oncology education in postgraduate medical training. These data summarize educational gaps and potential solutions to improve multidisciplinary education for future trainees. Of the 24 studies included in the final analysis, 15 obtained faculties' and trainees' opinions on deficiencies and areas of improvement for existing multidisciplinary oncology education [16-19,24,26,27]. They generally reported limited multidisciplinary oncology training or knowledge, barriers to multidisciplinary training, and advocated for further instruction in different areas. The remaining 9 studies studied the impact of educational interventions on trainees' oncology expertise [31-34,38]. Multidisciplinary rotations, tumor board meetings, didactic teaching, and case-based learning were found to be beneficial based on trainee self-assessments, written exams, and evaluations from faculty and patients following the observation of trainees in OSCE stations.

Filling the current gaps in multidisciplinary oncology education using the aforementioned educational interventions has the potential to improve multidisciplinary communication, appropriate referrals, and oncologic outcomes [3-5]. Studies by Mattes et al [12] and Martin et al [38] found that trainees were more likely to collaborate and make appropriate referrals to radiation oncologists after didactic teachings in lung cancer treatment and palliative radiotherapy, respectively. Several studies also found MTBMs to enhance trainee education [30,36,37]. In fact, the study by Mackay et al [36] found that tumor board simulation sessions significantly improved trainee's confidence in presenting in tumor board sessions. After all, improved communication and referral patterns are central to effective multidisciplinary collaboration among oncology specialists and ultimately improve the access of patients to evidence-based oncologic treatments.

Comparison With Prior Work

Geriatric oncology was consistently found to be an area in which trainees received limited training [16,26,27,39]. As cancer incidence increases in older adults, a population with a higher burden of comorbidities, trainees must gain sufficient knowledge and experience in geriatric oncology to optimize treatment [40]. These findings are echoed in a review by Morris et al [39] highlighting insufficient training and education in geriatric oncology among radiation oncology trainees across several different countries. This training should identify the specific needs of older patients and thereby result in a more informed and nuanced approach to this population's medical and psychosocial issues [24]. Development of these skills may be achieved through dedicated rotations or training in geriatric oncology.

Based on findings from this study, it is evident that the quality of multidisciplinary oncology education and training needs to be assessed and addressed. Implementation of benchmarks to ensure sufficient training across residency and fellowship programs commonly involved in cancer care would provide an educational quality metric [6-10]. This would encourage training

programs to develop and establish multidisciplinary oncology curricula. One approach to achieve this would be to ensure trainee participation in a variety of educational activities such as multidisciplinary case conferences, research, rotations, didactic teaching, and case-based learning led by faculty from other disciplines [11,31-34,38]. Furthermore, a review of each residency or fellowship program's curriculum by a multidisciplinary faculty committee may ensure sufficient trainee exposure to collaborating oncology areas.

Competency-based medical education is an outcome-based approach to evaluate medical trainees and ensure a high degree of graduate skill set [41]. This is often done via objective measures, such as entrustable professional activities (EPAs) and milestones. The development of standardized and program-specific EPAs, specifically for multidisciplinary oncology education, would provide training programs with a specific measure of their trainees' knowledge, skills, and progress in this area. Using EPAs would also identify areas of improvement for trainees early on in their training and would allow for additional support to improve multidisciplinary oncology competencies. Ultimately, these EPAs should mirror curriculum changes to ensure effective multidisciplinary oncology education. The benefits of using EPAs for geriatric oncology training are echoed by Eid et al [24]. They provide an example of an EPA to assess the appropriateness of chemotherapy for a geriatric patient, which includes the ability to perform a comprehensive geriatric assessment, having sufficient knowledge of chemotherapy toxicities and interactions, and assessment of suitability based on patients' comorbidities. This represents a geriatric oncology-specific EPA for medical or hematology oncology trainees. Oncology training programs may adopt similar EPAs to ensure a high quality of multidisciplinary oncology training within their residency and fellowship programs.

Despite its merits, there are potential barriers to the implementation of oncology training curricula. Several factors may prevent trainee participation in multidisciplinary education activities, including limited elective time, educational options, or available personnel. For instance, those training in the community or rural hospitals may not have access to many electives in other oncology fields. For the same reason, there may be limited available multidisciplinary faculty to either design effective oncology curricula or mentor trainees. Furthermore, many residency or fellowship programs may have strict curricula and elective requirements, and thus limit elective options for trainees. To overcome some of these challenges, studies have suggested the importance of web-based courses or teaching sessions to supplement their curriculum. As a result of the COVID-19 pandemic, web-based education has become an integral part of medical training that will likely remain used to various degrees in the future [42,43]. Data supports the effectiveness of web-based training, including web-based rotations or clinical training [44-46], tumor board meetings [28,37], surgical skills training [47], and didactic and case-based teaching [48-52].

Furthermore, local, state-wide or provincial, and national resources and programs could also be offered to trainees interested in further advancing their multidisciplinary oncology knowledge and skills outside their residency and fellowship programs. Certainly, didactic teaching [12,35,38], as well as workshops and OSCE-style evaluation sessions [32-34] are valuable in advancing trainee education in multidisciplinary oncology care. Depending on the topic, these teaching sessions could be offered in person, remotely via web-based applications, or as a prerecording to enhance trainee participation. As indicated by Mackay et al [36], tumor board simulation sessions contribute to significant improvements in trainee confidence and skills in participating in tumor boards. This is a novel educational intervention not traditionally offered by residency or fellowship programs. The addition of such resources and programs outside of the mainstream postgraduate training programs has the potential to supplement trainee education toward multidisciplinary oncology care.

Given the time constraint of residency and fellowship, it is not feasible for trainees to gain all relevant multidisciplinary knowledge and skills while also excelling in all core competencies relevant to their program. Every proposed intervention will have its own challenges to implement and needs to be balanced against other rotations within the curriculum. Yet, it is preferred that trainees obtain sufficient multidisciplinary knowledge during training rather than through experience during practice. It is crucial that training programs conduct an evaluation of any new educational intervention and prioritize selected interventions in their curricula based on outcomes and feedback.

Limitations

This study has limitations. Only 24 studies have analyzed the quality of multidisciplinary oncology education among postgraduate medical trainees. Furthermore, we limited our study to English-only and primary papers. It is possible that additional studies analyzing multidisciplinary oncology education in other languages or papers (eg, grey literature) exist that are missing from our results. Over a third of these studies were also published more than 5 years ago. Particularly, 3 of the intervention studies are by Sloan et al [32-34], published in 1997, 1999, and 2004, which could have had overlapping participants. This could limit the generalizability of the findings from these studies. There is a need for additional and more contemporary research assessing the needs of postgraduate medical trainees and the impact of newer educational interventions. It is particularly important to evaluate the use of technologies currently used in medical education such as web-based live teaching [43-47], clinical teaching tools such as case-based modules with built-in radiology software [53,54], and virtual reality surgical training [55-57]. Additionally, none of the studies on educational interventions were conducted with trainees in geriatric oncology. As previously discussed, this is an important aspect of oncology, though generally missing from oncology training curriculums. Thus, additional studies are needed within these fields. Furthermore, while a large proportion of studies solely focus on gaps in geriatric oncology education, this may not be generalizable to all multidisciplinary oncology education needs. Future research will be important in developing multidisciplinary oncology curricula for postgraduate trainees.

Conclusions

This systematic review demonstrated several gaps in the existing multidisciplinary oncology training of postgraduate medical trainees and the promising results of various educational interventions in bridging these gaps. Further studies investigating the needs of trainees at both local and national levels are needed to develop specific educational curricula and program requirements that focus on multidisciplinary oncology collaboration. Future research should also assess contemporary educational interventions to determine the most effective methods of attaining multidisciplinary oncology expertise among postgraduate medical trainees.

Authors' Contributions

The authors met all International Committee of Medical Journal Editors criteria for authorship. HT, GK, ER, ME, and TDC contributed to the study design. HT, GK, CML, IB, ZF, RV, and ME contributed to the processes of screening or data acquisition. HT, GK, CML, IB, ZF, ER, TDC, and RV participated in data analysis. All authors contributed to manuscript drafting and revision.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Research data, search strategy, and assessment results.
[DOCX File, 70 KB - mededu_v11i1e63655_app1.docx]

Checklist 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File, 33 KB - mededu_v11i1e63655_app2.docx]

References

1. Ahmad FB, Cisewski JA, Anderson RN. Mortality in the United States—provisional data, 2023. MMWR Morb Mortal Wkly Rep 2024 Aug 8;73(31):677-681. [doi: 10.15585/mmwr.mm7331a1] [Medline: 39116025]
2. Hong NJL, Wright FC, Gagliardi AR, Paszat LF. Examining the potential relationship between multidisciplinary cancer care and patient survival: an international literature review. J Surg Oncol 2010 Aug 1;102(2):125-134. [doi: 10.1002/jso.21589] [Medline: 20648582]
3. Aizer AA, Paly JJ, Michaelson MD, et al. Medical oncology consultation and minimization of overtreatment in men with low-risk prostate cancer. J Oncol Pract 2014 Mar;10(2):107-112. [doi: 10.1200/JOP.2013.000902] [Medline: 24399853]
4. Goulart BHL, Reyes CM, Fedorenko CR, et al. Referral and treatment patterns among patients with stages III and IV non-small-cell lung cancer. J Oncol Pract 2013 Jan;9(1):42-50. [doi: 10.1200/JOP.2012.000640] [Medline: 23633970]
5. Booth CM, Siemens DR, Peng Y, Mackillop WJ. Patterns of referral for perioperative chemotherapy among patients with muscle-invasive bladder cancer: a population-based study. Urol Oncol 2014 Nov;32(8):1200-1208. [doi: 10.1016/j.urolonc.2014.05.012] [Medline: 24968946]
6. ACGME common program requirements (residency). Accreditation Council for Graduate Medical Education. 2023 Jul 1. URL: https://www.acgme.org/globalassets/pfassets/programrequirements/cprresidency_2023.pdf [accessed 2024-12-24]
7. ACGME common program requirements (fellowship). Accreditation Council for Graduate Medical Education. 2022 Jul 1. URL: https://www.acgme.org/globalassets/pfassets/programrequirements/cprfellowship_2022v3.pdf [accessed 2024-12-24]
8. ACGME program requirements for graduate medical education in complex general surgical oncology. Accreditation Council for Graduate Medical Education. 2023 Jul 1. URL: https://www.acgme.org/globalassets/pfassets/programrequirements/446_complexgeneralsurgicaloncology_2023.pdf [accessed 2024-12-24]
9. ACGME program requirements for graduate medical education in hematology and medical oncology. Accreditation Council for Graduate Medical Education. 2024 Jul 1. URL: https://www.acgme.org/globalassets/pfassets/programrequirements/2024-prs/155_hematologyandmedicaloncology_2024.pdf [accessed 2024-12-24]
10. ACGME program requirements for graduate medical education in radiation oncology. Accreditation Council for Graduate Medical Education. 2023 Jul 1. URL: https://www.acgme.org/globalassets/pfassets/programrequirements/430_radiationoncology_2023.pdf [accessed 2024-12-24]
11. Mattes MD. Multidisciplinary oncology education: going beyond tumor board. J Am Coll Radiol 2016 Oct;13(10):1239-1241. [doi: 10.1016/j.jacr.2016.06.005] [Medline: 27474420]
12. Mattes MD, Ye JC, Peters GW, et al. Pilot study demonstrating the value of interdisciplinary education on the integration of radiation therapy in lung cancer management. J Cancer Educ 2023 Apr;38(2):590-595. [doi: 10.1007/s13187-022-02158-8] [Medline: 35357645]
13. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]

14. Covidence. URL: www.covidence.org [accessed 2025-05-16]
15. Hong QN, Fàbregues S, Bartlett G, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Educ Inf 2018;34(4):285-291. [doi: 10.3233/EFI-180221]
16. Akthar AS, Hellekson CD, Ganai S, et al. Interdisciplinary oncology education: a national survey of trainees and program directors in the United States. J Canc Educ 2018 Jun;33(3):622-626. [doi: 10.1007/s13187-016-1139-6]
17. Brenner BM, De Donno MA. Assessing gaps in surgical oncology training: results of a survey of general surgery residents. J Surg Educ 2020;77(4):749-756. [doi: 10.1016/j.jsurg.2020.01.011] [Medline: 32063511]
18. Park KU, Selby L, Chen XP, et al. Development of residents' self-efficacy in multidisciplinary management of breast cancer survey. J Surg Res 2020 Jul;251:275-280. [doi: 10.1016/j.jss.2020.02.014] [Medline: 32197183]
19. Wilson JP, Miller A, Edge SB. Breast education in general surgery residency. Am Surg 2012 Jan;78(1):42-45. [Medline: 22273306]
20. Morris L, Turner S, Thiruthaneeswaran N, et al. An international expert Delphi consensus to develop dedicated geriatric radiation oncology curriculum learning outcomes. Int J Radiat Oncol Biol Phys 2022 Aug 1;113(5):934-945. [doi: 10.1016/j.ijrobp.2022.04.030] [Medline: 35500796]
21. Walraven JEW, van der Meulen R, van der Hoeven JJM, et al. Preparing tomorrow's medical specialists for participating in oncological multidisciplinary team meetings: perceived barriers, facilitators and training needs. BMC Med Educ 2022 Jun 27;22(1):502. [doi: 10.1186/s12909-022-03570-w] [Medline: 35761247]
22. Delaye M, Try M, Rousseau A, et al. Onco-nephrology: physicians' expectations about a new subspecialty. J Cancer Educ 2023 Jun;38(3):878-884. [doi: 10.1007/s13187-022-02201-8] [Medline: 35840858]
23. Mäurer M, Staudacher J, Meyer R, et al. Importance of interdisciplinarity in modern oncology: results of a national intergroup survey of the Young Oncologists United (YOU). J Cancer Res Clin Oncol 2023 Sep;149(12):10075-10084. [doi: 10.1007/s00432-023-04937-2] [Medline: 37261525]
24. Eid A, Hughes C, Karuturi M, Reyes C, Yorio J, Holmes H. An interprofessionally developed geriatric oncology curriculum for hematology-oncology fellows. J Geriatr Oncol 2015 Mar;6(2):165-173. [doi: 10.1016/j.jgo.2014.11.003] [Medline: 25487037]
25. David V, Hsu T, Mithoowani S, Fraser G, Mian H. What do hematology residents know about caring for older adults with cancer? A national survey of Canadian hematology residents' knowledge and interests. J Geriatr Oncol 2022 Nov;13(8):1236-1240. [doi: 10.1016/j.jgo.2022.08.018] [Medline: 36050270]
26. Maggiore RJ, Callahan KE, Tooze JA, Parker IR, Hsu T, Klepin HD. Geriatrics fellowship training and the role of geriatricians in older adult cancer care: a survey of geriatrics fellowship directors. Gerontol Geriatr Educ 2018;39(2):170-182. [doi: 10.1080/02701960.2016.1247070] [Medline: 27749199]
27. Morris L, Thiruthaneeswaran N, Lehman M, Hasselburg G, Turner S. Are future radiation oncologists equipped with the knowledge to manage elderly patients with cancer? Int J Radiat Oncol Biol Phys 2017 Jul 15;98(4):743-747. [doi: 10.1016/j.ijrobp.2017.01.001] [Medline: 28258899]
28. Picca A, Reed S. Off to boarding school: exploring how physicians learn in tumor board. Pediatr Blood Cancer 2023 Nov;70(11):e30632. [doi: 10.1002/pbc.30632] [Medline: 37610271]
29. Givi B, Gordon AJ, Park YS, Lydiatt WM, Tekian A. Needs assessment in head and neck surgical oncology training: a qualitative study of expert opinions. Head Neck 2022 Nov;44(11):2528-2536. [doi: 10.1002/hed.27158] [Medline: 35920353]
30. Le Nail LR, Samargandi R. Teaching potential of multidisciplinary tumor board meetings for orthopedic residents: insights from a French sarcoma reference center. Cureus 2023 May;15(5):e39783. [doi: 10.7759/cureus.39783] [Medline: 37265907]
31. Cook MR, Graff-Baker AN, Moren AM, et al. A disease-specific hybrid rotation increases opportunities for deliberate practice. J Surg Educ 2016;73(1):1-6. [doi: 10.1016/j.jsurg.2015.09.005] [Medline: 26481268]
32. Sloan DA, Donnelly MB, Schwartz RW, et al. The multidisciplinary structured clinical instruction module as a vehicle for cancer education. Am J Surg 1997 Mar;173(3):220-225. [doi: 10.1016/s0002-9610(97)89596-7] [Medline: 9124631]
33. Sloan DA, Witzke DB, Plymale MA, et al. A multidisciplinary workshop to teach head and neck oncology to residents: results of a pilot study. J Cancer Educ 1999;14(4):228-232. [doi: 10.1080/08858199909528632]
34. Sloan DA, Plymale MA, Donnelly MB, Schwartz RW, Edwards MJ, Bland KI. Enhancing the clinical skills of surgical residents through structured cancer education. Ann Surg 2004 Apr;239(4):561-566. [doi: 10.1097/01.sla.0000118568.75888.04] [Medline: 15024318]
35. Meani F, Kovacs T, Wandschneider W, Costa A, Pagani O. Multidisciplinary blended learning to build a breast cancer specialist career: survey on the perspective of the first 2 cohorts of the ESO-ULM Certificate of Competence in Breast cancer (CCB). BMC Med Educ 2022 May 5;22(1):344. [doi: 10.1186/s12909-022-03414-7] [Medline: 35513883]
36. Mackay EC, Patel KR, Davidson C, et al. Simulation as an effective means of preparing trainees for active participation in MDT meetings. Future Healthc J 2024 Mar;11(1):100017. [doi: 10.1016/j.fhj.2024.100017] [Medline: 38646046]
37. Khoshgoftar Z, Sodeifian F, Allameh F. Improving the educational gap with implementing of teaching scholarship in virtual multidisciplinary tumor boards. Int J Cancer Manag 2023;In Press(In Press). [doi: 10.5812/ijcm-137490]
38. Martin EJ, Nalawade VV, Murphy JD, Jones JA. Incorporating palliative radiotherapy education into hospice and palliative medicine fellowship training: a feasibility study. Ann Palliat Med 2019 Sep;8(4):436-441. [doi: 10.21037/apm.2019.04.02]

39. Morris L, Turner S, Thiruthaneeswaran N, Agar M. Improving the education of radiation oncology professionals in geriatric oncology: where are we and where should we be? Semin Radiat Oncol 2022 Apr;32(2):109-114. [doi: 10.1016/j.semradonc.2021.11.008] [Medline: 35307112]
40. Balducci L, Ershler WB. Cancer and ageing: a nexus at several levels. Nat Rev Cancer 2005 Aug;5(8):655-662. [doi: 10.1038/nrc1675] [Medline: 16056261]
41. Harris P, Bhanji F, Topps M, et al. Evolving concepts of assessment in a competency-based world. Med Teach 2017 Jun;39(6):603-608. [doi: 10.1080/0142159X.2017.1315071] [Medline: 28598736]
42. Dedeilia A, Sotiropoulos MG, Hanrahan JG, Janga D, Dedeilias P, Sideris M. Medical and surgical education challenges and innovations in the COVID-19 era: a systematic review. In Vivo 2020;34(3 suppl):1603-1611. [doi: 10.21873/invivo.11950]
43. Wendt S, Abdullah Z, Barrett S, et al. A virtual COVID-19 ophthalmology rotation. Surv Ophthalmol 2021;66(2):354-361. [doi: 10.1016/j.survophthal.2020.10.001] [Medline: 33058927]
44. Chandra S, Laoteppitaks C, Mingioni N, Papanagnou D. Zooming-out COVID-19: virtual clinical experiences in an emergency medicine clerkship. Med Educ 2020 Dec;54(12):1182-1183. [doi: 10.1111/medu.14266] [Medline: 32502282]
45. Haws BE, Mannava S, Schuster BK, DiGiovanni BF. Implementation and evaluation of a formal virtual medical student away rotation in orthopaedic surgery during the COVID-19 pandemic. Foot Ankle Orthopaedic 2022 Jan;7(1):2473011421S00229. [doi: 10.1177/2473011421S00229]
46. Villa S, Janeway H, Preston-Suni K, et al. An emergency medicine virtual clerkship: made for COVID, here to stay. West J Emerg Med 2021 Dec 17;23(1):33-39. [doi: 10.5811/westjem.2021.11.54118] [Medline: 35060858]
47. Harrell Shreckengost CS, Reitz A, Ludi E, Rojas Aban R, Jáuregui Paravicini L, Serrot F. Lessons learned during the COVID-19 pandemic using virtual basic laparoscopic training in Santa Cruz de la Sierra, Bolivia: effects on confidence, knowledge, and skill. Surg Endosc 2022 Dec;36(12):9379-9389. [doi: 10.1007/s00464-022-09215-9] [Medline: 35419639]
48. Wilson HC, Lim TR, Axelrod DM, et al. A multimedia paediatric cardiology assessment tool for medical students and general paediatric trainees: development and validation. Cardiol Young 2023 Mar;33(3):444-448. [doi: 10.1017/S1047951122001123] [Medline: 35411842]
49. Nozari A, Mukerji S, Lok LL, et al. Perception of web-based didactic activities during the COVID-19 pandemic among anesthesia residents: pilot questionnaire study. JMIR Med Educ 2022 Mar 31;8(1):e31080. [doi: 10.2196/31080] [Medline: 35275840]
50. Haring RS, Rydberg LK, Mallow MK, Kortebein P, Verduzco-Gutierrez M. Development and implementation of an international virtual didactic series for physical medicine and rehabilitation graduate medical education during COVID-19. Am J Phys Med Rehabil 2022 Feb 1;101(2):160-163. [doi: 10.1097/PHM.0000000000001926] [Medline: 35026777]
51. Murdock HM, Penner JC, Le S, Nematollahi S. Virtual morning report during COVID-19: a novel model for case-based teaching conferences. Med Educ 2020 Sep;54(9):851-852. [doi: 10.1111/medu.14226] [Medline: 32403168]
52. Shih KC, Chan JCH, Chen JY, Lai JSM. Ophthalmic clinical skills teaching in the time of COVID-19: a crisis and opportunity. Med Educ 2020 Jul;54(7):663-664. [doi: 10.1111/medu.14189] [Medline: 32324929]
53. El-Ali A, Kamal F, Cabral CL, Squires JH. Comparison of traditional and web-based medical student teaching by radiology residents. J Am Coll Radiol 2019 Apr;16(4 Pt A):492-495. [doi: 10.1016/j.jacr.2018.09.048] [Medline: 30449521]
54. Sugi MD, Kennedy TA, Shah V, Hartung MP. Bridging the gap: interactive, case-based learning in radiology education. Abdom Radiol 2021 Dec;46(12):5503-5508. [doi: 10.1007/s00261-021-03147-z]
55. Bernardo A. Virtual reality and simulation in neurosurgical training. World Neurosurg 2017 Oct;106:1015-1029. [doi: 10.1016/j.wneu.2017.06.140] [Medline: 28985656]
56. Bakshi SK, Lin SR, Ting DSW, Chiang MF, Chodosh J. The era of artificial intelligence and virtual reality: transforming surgical education in ophthalmology. Br J Ophthalmol 2021 Oct;105(10):1325-1328. [doi: 10.1136/bjophthalmol-2020-316845]
57. Goh GS, Lohre R, Parvizi J, Goel DP. Virtual and augmented reality for surgical training and simulation in knee arthroplasty. Arch Orthop Trauma Surg 2021 Dec;141(12):2303-2312. [doi: 10.1007/s00402-021-04037-1] [Medline: 34264380]

Abbreviations:

EPA: entrustable professional activity
MMAT: Mixed Methods Appraisal Tool
MTBM: multidisciplinary tumor board meeting
OSCE: Objective Structured Clinical Examination
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by B Lesselroth; submitted 25.06.24; peer-reviewed by N Waheed, R Bansal, S Ganesh; revised version received 26.01.25; accepted 19.03.25; published 26.05.25.

Please cite as:
Tahmasebi H, Ko G, Lam CM, Bilgen I, Freeman Z, Varghese R, Reel E, Englesakis M, Cil TD
Multidisciplinary Oncology Education Among Postgraduate Trainees: Systematic Review
JMIR Med Educ 2025;11:e63655
URL: https://mededu.jmir.org/2025/1/e63655
doi:10.2196/63655

Bridging Gaps in Telemedicine Education in Romania to Support Future Health Care: Scoping Review

Mircea Adrian Focsa[1], MD, PhD; Virgil Rotaru[2], MPhys; Octavian Andronic[3], MD, PhD; Marius Marginean[4], MD, PhD; Sorin Florescu[5], MD, PhD

[1]Medical Informatics and Biostatistics Department, Victor Babe University of Medicine and Pharmacy Timi oara, Timisoara, Romania

[2]Victor Babe University of Medicine and Pharmacy Timi oara, Eftimie Murgu 2 Sqr, Timi oara, Romania

[3]General Surgery Department and Innovation and eHealth Center, Carol Davila University of Medicine and Pharmacy, Bucharest, Romania

[4]Public Health Authority, Brasov, Romania

[5]Orthopedics-traumatology II, Victor Babe University of Medicine and Pharmacy Timi oara, Timisoara, Romania

Corresponding Author:
Virgil Rotaru, MPhys
Victor Babe University of Medicine and Pharmacy Timi oara, Eftimie Murgu 2 Sqr, Timi oara, Romania

Abstract

Background: Telemedicine is a key element of modern health care, providing remote medical consultations and bridging the gap between patients and health care providers. Despite legislative advancements and pilot programs, the integration of telemedicine education in Romania remains limited. Addressing these educational gaps is essential for preparing current and future medical professionals to effectively use telemedicine technologies.

Objective: This study aimed to evaluate the current state of telemedicine education for medical professionals in Romania, focusing on the integration of diagnostic and therapeutic capabilities into medical curricula, identifying the challenges and opportunities, and providing recommendations for improving telemedicine education.

Methods: A scoping review was conducted following Arksey and O'Malley's framework. Peer-reviewed articles from 2019 to 2023 were identified using databases such as PubMed and Scopus. Additional gray literature was reviewed to provide a comprehensive understanding of telemedicine education in Romania. Data were thematically analyzed to extract key findings and recommendations.

Results: The review identified significant progress in the legislative and infrastructural aspects of telemedicine in Romania, but highlighted gaps in integrating telemedicine education into curricula for medical professionals and other health care practitioners directly involved in telemedicine practices. While some universities have included telemedicine components, dedicated telemedicine courses and hands-on training remain insufficient. Barriers include a lack of infrastructure, digital literacy, and practical exposure to telemedicine technologies.

Conclusions: For telemedicine to be effectively integrated into Romania's health care system, medical education must be adapted to include comprehensive telemedicine training. Recommendations include enhancing digital literacy, fostering public-private partnerships, and incorporating telemedicine into undergraduate and continuous professional education programs. These efforts are essential for improving healthcare access and quality through telemedicine.

(JMIR Med Educ 2025;11:e66458) doi:10.2196/66458

KEYWORDS

telemedicine; digital health; healthcare education; micro-credentials; scoping review; health education; Romania; future healthcare; telehealth; healthcare providers; technologies; digital literacy; healthcare system; quality care

Introduction

Telemedicine in Romania has undergone significant development over the years, marked by early pilot projects, legislative advancements, and a growing recognition of its potential to transform health care delivery. Given the rapid advancements in digital health globally, understanding and improving telemedicine education in Romania will not only enhance national health care delivery but also offer insights for similar health care systems transitioning to more digital practices.

For the purposes of this study and in accordance with Romanian legislation, telemedicine is defined as the use of digital technologies to deliver medical acts such as diagnosis, treatment, and therapy performed exclusively by licensed medical professionals. This differs from telehealth, which encompasses broader health-related services, including health education, prevention, and administrative activities.

While telemedicine inherently involves medical acts performed by licensed clinicians, it relies on a collaborative team, including nonclinical professionals, to ensure efficient and comprehensive care delivery.

Several interdependent factors, including educational competencies, regulatory policies, technological infrastructure, and institutional readiness shape telemedicine education. This study adopts a theoretical framework that integrates competency-based medical education, digital health policies, and workforce development strategies to evaluate the current state of telemedicine training in Romania. The framework is informed by internationally recognized guidelines, such as the World Health Organization (WHO) Digital Health Competency Framework (DHCF), the International Society for Telemedicine & eHealth (ISfTeH) guidelines, and the European Health Telematics Association (EHTEL) recommendations, which outline essential skills and knowledge areas for telemedicine practitioners.

Within this framework, telemedicine competencies are influenced by both regulatory structures and digital readiness, which shape how educational programs can be effectively implemented. While Romania has made legislative progress in supporting telemedicine, educational curricula remain inconsistent, lacking standardized competencies and hands-on training opportunities. Furthermore, limited technological infrastructure and digital literacy among both professionals and patients present additional challenges. By assessing these dimensions, this study identifies the current gaps in telemedicine education and proposes targeted recommendations to improve training programs, ensuring alignment with international best practices.

The journey began in 2001 when the Romanian Space Agency (ROSA) launched the Demonstrative Pilot of Telemedicine. This project, a significant milestone in the country's health care history, focused on diagnostic, clinical, and educational applications, serving as a pioneering effort to explore the capabilities of telemedicine in the country. In 2003, further progress was made in establishing the Romanian Association for Telemedicine and Space Applications for Health (ATASS), which aimed to promote and develop telemedicine technologies [1].

A significant legislative milestone came in 2018 when telemedicine was formally incorporated into Romanian law through Government Emergency Ordinance number 8/2018, which amended Health Reform Law 95/2006. This legislative change aimed to address the chronic shortage of medical personnel, particularly in remote areas, and ensure more equitable access to healthcare services.

The COVID-19 pandemic in 2020 acted as a catalyst for telemedicine's rapid adoption. In response to the crisis, the Romanian government took swift and decisive action to establish a regulatory framework for telemedicine services. Government Decision 252/2020 and subsequent ordinances laid the groundwork for telemedicine during states of emergency and beyond. These regulations facilitated various telemedicine services, including teleconsultation, tele-expertise, teleradiology, and telemonitoring. By 2022, the regulatory framework had further evolved with Government Decision 1133/2022 [2], which approved comprehensive implementation norms for telemedicine. This decision standardized procedures for scheduling remote appointments, protecting data privacy, and setting up payment mechanisms through the National Health care Insurance House. These measures ensured that telemedicine services could be provided seamlessly and securely, enhancing their integration into the healthcare system.

In the last 2 years, significant financial investments have supported the expansion of telemedicine in Romania. The recovery and resilience plan allocated substantial funds, including approximately €100 million for telemedicine support and €300-€400 million for hospital digitalization. These investments underscored the government's commitment to advancing telemedicine as a critical component of health care delivery.

Despite these significant financial investments, the low levels of health and digital literacy among Romanian citizens present substantial barriers to the successful adoption and utilization of these technologies. Digital literacy for the public refers to the ability to access, understand, and use digital technologies for obtaining health information and services. For health care professionals, digital literacy extends to proficiency in using digital tools for clinical care, such as eHealth records, telemedicine platforms, and data privacy protocols.

Eurostat statistics [3] indicate that only 40% of Romanians use the internet to search for health information, significantly below the European Union average of 55%. This discrepancy highlights the role of education in digital engagement, with just 17% of individuals with low educational attainment using the internet for health purposes, compared with 41% of those with medium education and 66.5% of highly educated individuals.

The first cross-sectional study on health literacy in Romania [4] underscores the challenges faced by the population in processing health information. Approximately 21.6% of respondents found it difficult to protect themselves from illness based on health information provided by the media. Moreover, 7.5% of participants demonstrated inadequate health literacy, and 33.2% had problematic health literacy, leaving a majority (59.2%) with sufficient health literacy. Key determinants of health literacy included age, gender, education, and self-reported health status, while the residential area did not appear to influence health literacy levels. These findings underscore the considerable gaps in both health and digital literacy among the Romanian population.

Moreover, significant gaps still need to be addressed in integrating telemedicine education effectively within medical curricula, particularly in ensuring that current and future medical professionals are adequately prepared to leverage these technologies in practice. Most medical universities and medical schools have started incorporating telemedicine into their curricula, mainly as part of the medical informatics discipline, aiming to familiarize future health care professionals with digital health tools and focusing on both telemedicine's technical and ethical aspects. Professional development opportunities have also expanded, with continuous education programs incorporating modules on telemedicine. Online training

platforms and workshops have become vital resources for health care providers, helping them stay updated with the latest advancements and best practices in telemedicine. Beyond serving as a means of continuing education, these platforms often provide essential initial training for health care professionals new to telehealth, equipping them with foundational knowledge and skills.

Telehealth competency, encompassing knowledge, skills, and attitudes essential for effectively delivering care via telemedicine, is increasingly recognized as a critical aspect of modern health care practice. However, in Romania, health care professionals often lack structured and standardized training in telemedicine, resulting in gaps in areas such as conducting teleconsultations, ensuring data security, and communicating effectively with patients in virtual settings. These gaps highlight the need for targeted educational interventions to prepare health care professionals for the demands of telemedicine.

The scope of this study is confined to telemedicine, which involves clinical activities performed by physicians and other licensed medical professionals, ensuring a clear distinction from the broader concept of telehealth.

The aim of this scoping review is to evaluate the current state of telemedicine education in Romania, identify the challenges and opportunities associated with its integration into medical curricula, and provide recommendations for improving telemedicine education. Specifically, this study maps existing telemedicine education initiatives, assesses barriers to implementation, and proposes strategies to enhance training programs for health care professionals.

Methods

Study Design

This study used a scoping review methodology to comprehensively explore the current landscape of telemedicine education in Romania. The study specifically evaluates educational approaches for medical professionals performing telemedicine, addressing clinical activities such as diagnosis, therapy, and patient management. Telemedicine education evaluated in this study encompasses competencies applicable to multiple specialties, including primary care, chronic disease management, and specialized services such as telemonitoring and telerehabilitation. A scoping review was chosen for its ability to map key concepts, types of evidence, and gaps in research related to a defined area or field of interest, particularly in complex or under-reviewed topics. This approach is especially suitable for telemedicine education in Romania, given its rapidly evolving nature and the need to synthesize diverse sources of information

Research Question

The primary research question guiding this scoping review was: “What is the current state of telemedicine education in Romania, and what are the key challenges and opportunities for its integration into medical curricula?”. This question was formulated to encompass the broad scope of telemedicine education, including formal educational programs, professional development initiatives, and digital literacy efforts.

Literature Search Strategy

A systematic search of peer-reviewed journal papers was conducted across multiple databases, including PubMed, Scopus, Web of Science, and Google Scholar. The search covered articles published between January 2019 and December 2023. The following keywords and their combinations were used: telemedicine, telehealth, digital health, medical education, telemedicine education, Romania, eHealth, and digital literacy.

Additional filters were applied to include only papers available in English or Romanian, with a focus on education, telemedicine implementation, and health care policy in Romania. Papers were limited to those that addressed telemedicine within the context of health care education, the use of digital tools in clinical training, and barriers or facilitators to telemedicine adoption.

In addition to peer-reviewed papers, relevant gray literature was included in the scoping review. Sources of gray literature encompassed reports from governmental and nongovernmental organizations, policy briefs, and institutional documents related to telemedicine education in Romania. These documents were identified through searches of online repositories and institutional websites, and they provided critical context on legislative developments, pilot projects, and educational initiatives that may not have been extensively covered in academic databases.

Study Selection

The study selection process involved 2 independent researchers (MF and VR) who screened titles, abstracts, and full texts to ensure alignment with the inclusion and exclusion criteria. Any disagreements were resolved through discussion, and when consensus could not be reached, a third researcher reviewed the articles to make the final decision. The process was facilitated using Rayyan platform, allowing efficient tracking and documentation of the selection process.

The initial search yielded 105 journal papers, screened for relevance based on titles and abstracts. Articles that focused on telemedicine in clinical practice without addressing educational aspects were excluded. After this preliminary screening, 35 papers remained for full-text review. A further screening based on inclusion and exclusion criteria led to the identification of 19 papers deemed highly relevant to the study’s aims.

Eligibility Criteria

The inclusion and exclusion criteria are listed in Textbox 1.

Textbox 1. Inclusion and exclusion criteria

Inclusion criteria

- Studies published between 2019 and 2023.
- Studies addressing telemedicine education or digital literacy training in medical or health care fields.
- Studies conducted in or relevant to the Romanian context.

Exclusion criteria

- Articles focusing exclusively on telemedicine in clinical practice without reference to education.
- Studies not available in English or Romanian.

Data Extraction

For each of the 19 selected peer-reviewed papers, data were extracted using a standardized data charting form, which included the following variables: (1) author(s), publication year, and country of origin; (2) study design (eg, qualitative, quantitative, and mixed methods); (3) the focus of the study (eg, telemedicine education, barriers to digital health adoption, and telemedicine curriculum development); (4) key findings relevant to telemedicine education, digital literacy, and health care training; and (5) recommendations for telemedicine integration into medical education.

Data Analysis

Two independent researchers conducted a thematic analysis to identify recurring patterns and themes within the selected literature. An inductive approach allowed themes to emerge naturally from the data. The analysis involved coding and categorizing data using manual methods, followed by a review of the themes to ensure consistency and relevance. The themes were grouped into broad categories, such as the current status of telemedicine education in Romania, barriers to telemedicine education, opportunities for development, and digital literacy challenges.

Ethical Considerations

Since this study was based on a review of publicly available literature, no ethical approval was required. However, all articles were reviewed with a commitment to academic integrity and transparency, and any potential conflicts of interest were disclosed.

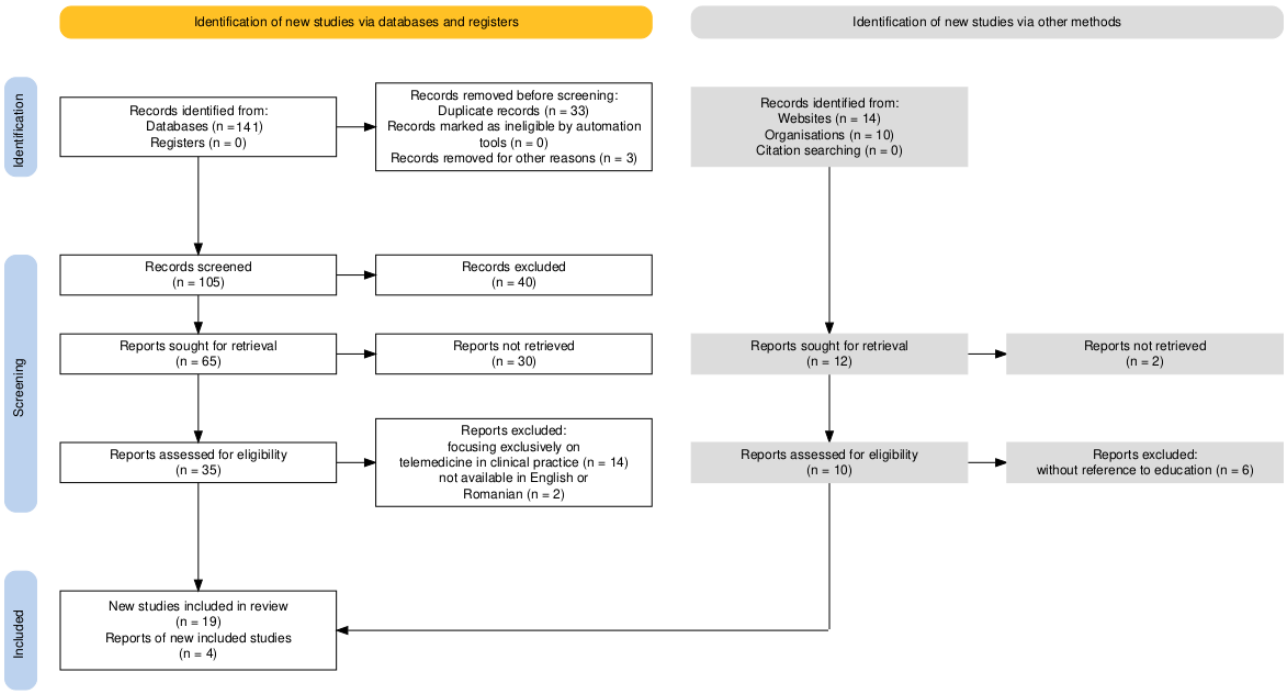
Results

Included Studies

The initial search yielded 105 journal papers, screened for relevance based on titles and abstracts. Articles that focused on telemedicine in clinical practice without addressing educational aspects were excluded. After this preliminary screening, 35 papers remained for full-text review. A further screening based on inclusion and exclusion criteria led to the identification of 19 papers deemed highly relevant to the study's aims (Figure 1 and Checklist 1) .

The results presented are based on themes identified through inductive thematic analysis, which highlighted key areas such as the integration of telemedicine into curricula, barriers to adoption, and digital literacy as a critical enabler for telemedicine education.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart for identification and selection of studies.



Current State of Telemedicine Education in Romania

The data collected and synthesized in Table 1 offers a comprehensive overview of recent research in telemedicine, eHealth, and artificial intelligence (AI)-based health care interventions in Romania. It includes various study designs such as cross-sectional, prospective, case-control, and system architecture studies and spans diverse health care topics, including telemedicine-driven rehabilitation, the impact of virtual communities on telemedicine adoption, and the integration of AI and Internet of Things (IoT) technologies in health care. The analysis covers key aspects of telemedicine's role during and post–COVID-19 and its relevance for both chronic diseases and mental health management. The studies underscore the positive reception of telemedicine in various medical fields, from diabetes management to cardiac rehabilitation, with the technology being especially vital during the COVID-19 pandemic. However, challenges such as digital literacy, accessibility, and infrastructure remain.

While Table 1 provides a summary of the reviewed studies, the specific components of telemedicine education and digital literacy training are discussed in greater detail within the following subsection, to align with the thematic approach of this scoping review.

Table . Overview of recent research in telemedicine, eHealth, and artificial intelligence–based health care interventions in Romania [5-23].

PMID	The focus	Key findings	Recommendations
35879930 [5]	Openness of medical students to telemedicine	Moderate-to-high acceptance, need for practical exposure	Integrate telemedicine into medical curricula, enhance digital literacy
35935629 [6]	Evaluating telemedicine benefits for cardiovascular patients during COVID-19.	Telemedicine facilitated patient management, including medication adjustments, but barriers like data security and reimbursement need addressing.	Improve telemedicine frameworks and regulations for cardiovascular care beyond the pandemic.
37297692 [7]	Perceptions of telemedicine among health care professionals	Positive outlook, concerns over digital literacy	Increase digital education for health care professionals
37510969 [8]	Telemedicine-driven pulmonary rehabilitation for post–COVID-19	Significant improvement in physical and mental health	Integrate telerehabilitation into postacute COVID-19 management
36141685 [9]	AI[a] in eHealth, telemedicine, and remote monitoring	AI advances in eHealth and telemedicine	Enhance integration of AI into eHealth for better health care services
36141899 [10]	Home-based robotic cardiac telerehabilitation system	RoboTeleRehab system is feasible, positive feedback	Further testing on cardiac patients, integration into rehabilitation programs
33923842 [11]	Blockchain-enabled framework for mHealth[b] systems	Improves security, transparency, and immutability	Implement blockchain for secure patient data management in mHealth
35062129 [12]	AI in primary care and telemedicine	AI aids in primary care diagnosis, treatment, and decision-making	Enhance AI systems to support primary care and telemedicine workflows
36141297 [13]	Impact of virtual communities on telemedicine usage	Virtual communities influence patient satisfaction and usage intention	Use virtual communities to promote telemedicine services
35954526 [14]	Adoption of eHealth and mHealth for mental health	Accessibility and data security are critical for adoption	Improve accessibility and data security in digital mental health tools
36556027 [15]	Management of dilated cardiomyopathy during COVID-19 using telemedicine	Telemedicine maintains clinical stability in patients with home monitoring	Use multiparametric home monitoring to manage dilated cardiomyopathy during crises
33953605 [16]	Perception of Romanian family doctors on telemedicine during COVID-19	Positive view, but tele-diagnostic challenges and time constraints	Training for family doctors and continued telemedicine reimbursement
34898981 [17]	Assessing patient adherence to telemedicine in diabetes	Developed reliable and valid tool for telemedicine adherence in diabetes care	Use the instrument to optimize telemedicine platforms based on patient needs
35150518 [18]	Attitudes toward eHealth during the COVID-19 pandemic	Negative attitudes in Greece due to forced usage	Address technical skills gaps and improve ease of access
37372846 [19]	Telemonitoring for cardiovascular disease during and post-COVID-19	Telemedicine improved cardiovascular prevention during the pandemic	Universal access to home telemonitoring for high-risk cardiovascular patients
36900948 [20]	Virtual assistant in cardiac rehabilitation	Similar results for virtual versus in-person rehabilitation	Optimize virtual assistants for cardiac rehab
33800728 [21]	IoT[c]-based biometric monitoring system for elders	Scalable solution for monitoring and cognitive assessment	Expand to predictive analysis for cognitive and physiological data
35979169 [22]	Telerehabilitation for Parkinson disease	Improved walking performance using telerehabilitation	Expand telerehabilitation for other neurorehabilitation settings
31407668 [23]	Use of telemedicine for rare diseases in Romania	Telemedicine improves access to rare disease care	Increase telemedicine use for remote consultations in rare diseases

[a]AI: artificial intelligence.

[b]mHealth: mobile health.

[c]IoT: Internet of Things.

University Programs and Curricula

In Romania, prominent medical universities in Timisoara, Cluj-Napoca, Iasi, and Bucharest have incorporated telemedicine and digital health into their educational programs. These changes were particularly accelerated following the COVID-19 pandemic, which highlighted the need for digital health competencies among health care professionals. New curricula encompass various aspects, such as telemedicine applications, electronic health records (EHRs), and data management strategies. Despite these inclusions, the development of dedicated disciplines and laboratory classes specific to telemedicine remains ongoing.

While telemedicine education in Romania has introduced theoretical foundations and basic digital health tools, it remains limited in scope and practical application. Key gaps include the lack of hands-on training, comprehensive and stand-alone telemedicine courses, and interprofessional education initiatives. Furthermore, advanced topics such as AI and patient engagement strategies remain underexplored, emphasiing the need for structured and innovative approaches to telemedicine education.

On October 14, 2023, the Center for Innovation and e-Health at the University of Medicine and Pharmacy “Carol Davila” in Bucharest hosted a specialized course titled “Telemedicine - Current Information and Skills.” [24]. This initiative covers fundamental theoretical concepts of telemedicine, including applicability and legal frameworks, alongside practical skills for using associated technologies.

Many health care professionals participate in continuing medical education (CME) programs to stay updated on telemedicine. These programs focus on skills for telemedicine platforms, data privacy laws, and improving remote patient interaction. The Digital Innovation Zone Association [25], affiliated with the North-East Regional Development Agency, offers comprehensive 1- to 2-month programs delivered in a hybrid online and on-site format.

A significant milestone in telemedicine education within the country was marked by the launch of the Erasmus+ project “TEAM: Supporting Innovation in Telemedicine Education with Cross-European Collaboration” in November 2023. The University of Medicine and Pharmacy “Victor Babes” in Timisoara is a principal participant in this consortium, which also includes partners from Belgium, Slovenia, Croatia, Greece, and Ukraine. The project aims to develop adaptable learning pathways that conform to international best practices, focusing on surmounting challenges such as limited digital literacy and fostering cross-sectoral cooperation. Targeting higher education students in health care and IT, educators, and institutions, the project also extends its reach to health care and IT professionals and policymakers. One of the primary outcomes of this initiative is the establishment of flexible microcredentials in Telemedicine designed to enhance proficiency in digital health among students and professionals alike.

Professional Development Workshops and Seminars

Universities and health care institutions in Romania proactively conduct workshops and seminars to offer practical training in telemedicine technologies. These educational sessions frequently use case studies and practical simulations to enrich the learning experience. They cover a spectrum of topics, from fundamental telemedicine principles to advanced applications such as the integration of AI tools in clinical settings.

Across Romania, hospitals and medical associations also organize workshops and seminars aimed at practicing health care professionals. These sessions foster proficiency in telemedicine platforms, digital communication skills, and data security. Participants, including doctors, nurses, and ancillary staff, benefit from the self-paced learning environment these workshops provide. They typically include hands-on sessions where attendees can interact with telemedicine software, acquire best practices for conducting virtual consultations, and gain insights into the legal and ethical dimensions of remote health care.

A notable initiative was the establishment of the ROHEALTH cluster in 2015, which brought together various entities within the health and bioeconomy sectors to enhance their competitive edge. This cluster supports an online platform offering diverse courses and webinars, including specialized offerings such as “eTELEDOC, emergency telemedicine” [26]. These resources aim to bolster the competencies of health care professionals in the evolving landscape of telemedicine.

Digital Literacy Programs

Several nongovernmental organizations offer training programs and workshops to enhance digital literacy among health care providers. Initiated in 2022 by PALMED, the Patronage of Private Medical Service Providers, the project titled “Digital Skills for Employees - Support for SMEs in the Health Sector to Assimilate Technologies and Develop Telemedicine Services (TELMed)” is a notable endeavor under the Human Capital Operational Program for 2014–2020 [27]. This project specifically aims to augment the digital competencies of personnel across 35 small and medium-sized enterprises in the health care sector, including hospitals, clinics, offices, and laboratories. By enhancing these skills, the initiative not only supports the adaptation and expansion of telemedicine activities but also prepares the workforce for advancements related to Industry 4.0 and smart specialization areas.

Challenges in Telemedicine Education

The challenges identified in this section were derived from thematic analysis of the 19 peer-reviewed papers and relevant gray literature, including institutional reports and policy briefs. These sources highlighted common barriers such as limited digital literacy among health care professionals, the lack of standardized curricula, and inadequate infrastructure for telemedicine training.

The pandemic propelled telemedicine into prominence, shedding light on Romanian family doctors’ diverse experiences and perceptions [16]. Over a quarter of general practitioners reported that remotely addressing patients’ health care needs was more manageable, demonstrating adaptability to telemedicine modalities. Nevertheless, challenges such as the time-intensive nature of teleconsultations, diagnostic uncertainties, and patients’ difficulties with technology have surfaced. These issues highlight the critical need for specialized training programs in

telemedicine for both health care professionals and patients to mitigate disruptions in health care delivery effectively.

Furthermore, the moderate-to-high acceptance of telemedicine among Romanian medical students emphasizes the necessity of integrating telemedicine education early in their medical training. Telemedicine fundamentally transforms the patient-physician relationship, requiring physicians to develop new communication skills, known as "webside manner." The methodological norms for telemedicine services in Romania also emphasize the importance of well-trained professionals who can navigate legal, ethical, business, and practical challenges. Training should start at the undergraduate level and continue through all professional stages for medical staff in all specialties.

While the primary objective of this paper is to explore telemedicine education for health care professionals, patient experiences with telemedicine offer indirect yet critical insights. These experiences highlight areas where health care providers may require additional training, such as building virtual rapport, managing technological issues, and addressing patient concerns about telemedicine efficacy and privacy. Integrating these considerations into training programs can better align telemedicine education with real-world practice.

In 2021, a study [17] evaluating the desirability, acceptability, and adherence to telemedicine among diabetes patients underscores the need for educational programs. Such initiatives targeting patients are essential to foster a positive perception and readiness to use telemedicine services, particularly in chronic conditions like diabetes, where continuous care and monitoring are essential. Patients, particularly those managing chronic diseases, require thorough education to use virtual technology effectively, including understanding how to operate software and hardware, such as mobile communication devices and other digital interfaces essential for remote health care.

Discussion

Principal Findings

This scoping review identified significant gaps in telemedicine education in Romania, despite recent legislative and infrastructural advancements supporting telemedicine adoption. While some medical universities have incorporated telemedicine content into their curricula, there remains a lack of structured, hands-on training and dedicated telemedicine courses. The study also highlighted key barriers, including limited digital literacy among health care professionals, insufficient policy support for mandatory telemedicine training, and a lack of standardized competency frameworks. These findings underscore the need for formalized telemedicine education initiatives, integrated into both undergraduate and CME programs, to enhance digital health preparedness among health care providers.

The findings of this scoping review highlight the significant progress and persistent challenges in telemedicine education in Romania. This discussion will interpret these results, explore their implications, and propose strategies for advancing telemedicine education in the country. To advance telemedicine in Romania, several strategic directions need to be pursued.

Medical schools must develop comprehensive programs that include hands-on training with telemedicine platforms and technologies, such as conducting teleconsultations or using telemonitoring devices in simulated or real clinical environments. The launch of specialised courses and international collaborations, such as the Erasmus+ TEAM project, demonstrates a commitment to enhancing telemedicine education.

In Romania, health care professionals are required by law to participate in CME programs to retain their licenses. While telemedicine-specific education is not yet a mandatory component of these requirements, its growing integration into CME programs highlights the recognition of telemedicine as an essential skill set for modern medical practice. This approach aligns with recommendations from the EHTEL, which emphasizes the importance of ongoing training in digital health for all health care providers [28]. The involvement of industry clusters like ROHEALTH in providing specialized webinars indicates a promising collaboration between academia and industry.

However, the review also suggests that these efforts may not be sufficient to meet the rapidly evolving needs of the health care system. There appears to be a need for more structured, comprehensive, and widely accessible professional development programs in telemedicine. While beneficial for updating specific skills, self-directed CME is often insufficient in providing a holistic understanding of telehealth practice. Without structured education, health care professionals may lack critical knowledge of professional telehealth standards, guidelines, and best practices, placing them at risk of learning by trial and error. Comprehensive, formalized training programs are essential to equip professionals with the competencies needed to deliver high-quality, safe, and effective telehealth care.

The identified gaps in digital literacy among both health care providers and patients represent a significant barrier to the effective implementation of telemedicine. The initiatives aimed at enhancing digital skills, such as PALMED's "Digital Skills for Employees" project, are steps in the right direction. These efforts align with European Union–wide initiatives like the Digital Education Action Plan (2021 - 2027) [29], which emphasizes the importance of digital skills across all sectors, including health care. However, these efforts need to be scaled up and integrated more systematically into both medical education and public health initiatives. The review also highlights the need for patient education in telemedicine, particularly for managing chronic conditions. In addition, robust policy support, increased public awareness, and education are crucial for the effective implementation of telemedicine, ultimately improving health care access and outcomes across Romania.

To fully realize the benefits of telemedicine, regulatory bodies in Romania should consider introducing mandatory education requirements for telemedicine practice. While technological proficiency is an important component of telemedicine education, health care professionals must also develop a broader range of competencies to provide effective telehealth care. These include clinical decision-making, patient communication, and

understanding the ethical and legal dimensions of telemedicine. Incorporating recognized telemedicine competency frameworks into educational programs can ensure comprehensive preparation for future telehealth practitioners.

Telemedicine is pivotal in advancing integrated care by fostering coordination among health care professionals and enabling patient-centered approaches. To fully realize its potential, telemedicine education must include interprofessional education and training for collaborative care, equipping health care teams with the skills to work cohesively in delivering seamless and effective telehealth services [30].

The results highlight a diverse range of technologies relevant to telemedicine education, including mobile health platforms, IoT devices, EHR systems, and continuous monitoring technologies. Although only a minority of reviewed studies explicitly addressed AI, its inclusion underscores the importance of preparing professionals for future technological advancements. While AI is not yet pervasive in all areas of health care, equipping professionals with foundational knowledge will prepare them for its growing integration into clinical practice.

Creating a thorough telemedicine education program necessitates teamwork between health care professionals, educators, technology specialists, and policymakers. Collaborating can help ensure the curriculum meets health care system needs, includes cutting-edge telemedicine technologies, and prepares providers and patients for future health care delivery.

Implications of Findings

To close the skills gaps, it is essential to develop educational programs that address the points mentioned below.

Improve Digital Literacy

Training must concentrate on enhancing the digital skills of health care professionals and patients so they can successfully navigate eHealth and mobile health platforms. While improving, digital literacy among Romanian health care professionals remains largely dependent on voluntary initiatives such as webinars, workshops, and seminars. This fragmented approach highlights the need for structured and comprehensive education programs to ensure that professionals are fully equipped to leverage telemedicine technologies effectively.

Integrates AI Technologies

Educational initiatives must cover foundational AI and machine learning (ML) concepts relevant to health care applications, including data privacy, ethical considerations, and interpreting AI-generated insights.

Develop New Skills

Training for telemedicine should include soft skills such as remote patient interaction, digital communication etiquette, and managing online patient relationships to adapt to remote health care dynamics. It also includes educating patients on how to use telemedicine services, getting ready for virtual visits, and handling their health data online.

Promote Trust in the Efficacy of Telemedicine

The research suggests the importance of increasing trust among health care providers and patients regarding telemedicine's effectiveness. For that purpose, telemedicine training should cover not only the technical aspects but also the clinical relevance and influence on patient outcomes, especially in managing chronic conditions.

Deal With Ethical and Privacy Concerns

Telemedicine training should incorporate adherence to recognized telehealth and telemedicine standards and guidelines, such as GDPR (General Data Protection regulations) and those created by the International Society for Telemedicine and eHealth [31]. Training should address ethical issues, privacy concerns, legal regulations, patient privacy protection strategies, and ethical guidelines for virtual patient interactions, ensuring alignment with global best practices.

Limitations and Future Directions

This scoping review, while comprehensive, has certain limitations. The focus on English and Romanian language publications may have excluded relevant studies in other European languages. In addition, the rapid evolution of telemedicine, particularly in response to the COVID-19 pandemic, means that some recent developments may not be fully captured in the published literature.

Future research should focus on (1) longitudinal studies assessing the long-term impact of telemedicine education on health care delivery and outcomes in Romania; (2) comparative analyses of telemedicine education approaches across different European countries, particularly comparing Eastern and Western European contexts; and (3) in-depth qualitative studies exploring the experiences and perspectives of medical students, health care providers, and patients regarding telemedicine education and implementation in the Romanian and broader international context.

Conclusion

Despite significant advancements, telemedicine in Romania still faces challenges. Infrastructure deficiencies, digital literacy gaps, and regulatory hurdles remain significant obstacles. However, ongoing investments in infrastructure, education, and regulatory frameworks are expected to address these issues, paving the way for broader adoption of telemedicine and improved health care access across the country.

The integration of telemedicine into medical education in Romania is crucial for the future of health care delivery. By addressing the current challenges and learning from successful global models, Romania can enhance its telemedicine capabilities and ensure that health care providers are well-prepared to leverage telemedicine technologies to improve health care access and quality. Moving forward, efforts should focus on enhancing digital health literacy, optimizing telemedicine systems, and expanding the use of AI and IoT for more integrated health care.

The future of telemedicine in Romania looks promising. As the country continues to invest in telemedicine education and infrastructure, health care providers will be better prepared to

leverage digital health technologies, ultimately enhancing the quality and accessibility of health care services for all Romanians.

Acknowledgments

We would like to acknowledge Victor Babes University of Medicine and Pharmacy Timisoara for their support in covering the costs of publication for this research paper. Part of the activities described in this manuscript were supported by funding from the EU Erasmus+ program under the TEAM project.

Conflicts of Interest

None declared.

Checklist 1
PRISMA-ScR checklist.
[PDF File, 349 KB - mededu_v11i1e66458_app1.pdf]

References

1. Panait L, Doarn CR, Saftoiu A, Popovici C, Valeanu V, Merrell RC. A review of telemedicine in Romania. J Telemed Telecare 2004;10(1):1-5. [doi: 10.1258/135763304322764103] [Medline: 15006207]
2. Romanian Government. Decision Nr. 1133 from 14 September 2022. URL: https://legislatie.just.ro/Public/DetaliiDocument/259367 [accessed 2023-12-15]
3. Eurostat. Individuals – internet activities: seeking health information. URL: https://ec.europa.eu/eurostat/databrowser/view/ISOC_CI_AC_I__custom_4264326/default/table?lang=en [accessed 2023-12-15]
4. Coman MA, Forray AI, Van den Broucke S, Chereches RM. Measuring health literacy in Romania: validation of the HLS-EU-Q16 survey questionnaire. Int J Public Health 2022;67:1604272. [doi: 10.3389/ijph.2022.1604272] [Medline: 35185446]
5. Cretu S, Gorzko AM, Salavastru CM. Openness toward the use of telemedicine among medical students in Romania: A cross-sectional study. JAAD Int 2022 Sep;8:139-141. [doi: 10.1016/j.jdin.2022.06.011] [Medline: 35879930]
6. Ghilencea LN, Chiru MR, Stolcova M, et al. Telemedicine: benefits for cardiovascular patients in the COVID-19 Era. Front Cardiovasc Med 2022;9:868635. [doi: 10.3389/fcvm.2022.868635] [Medline: 35935629]
7. Andronic O, Petrescu GED, Artamonov AR, et al. Healthcare professionals' specialists' perception of telemedicine in romania-a quantitative study of beliefs, practices, and expectations. Healthcare (Basel) 2023 May 25;11(11):1552. [doi: 10.3390/healthcare11111552] [Medline: 37297692]
8. Pescaru CC, Crisan AF, Marc M, et al. A systematic review of telemedicine-driven pulmonary rehabilitation after the acute phase of COVID-19. J Clin Med 2023 Jul 24;12(14):4854. [doi: 10.3390/jcm12144854] [Medline: 37510969]
9. Pap IA, Oniga S. A review of converging technologies in eHealth pertaining to artificial intelligence. Int J Environ Res Public Health 2022 Sep 10;19(18):11413. [doi: 10.3390/ijerph191811413] [Medline: 36141685]
10. Mocan B, Mocan M, Fulea M, Murar M, Feier H. Home-based robotic upper limbs cardiac telerehabilitation system. Int J Environ Res Public Health 2022 Sep 15;19(18):11628. [doi: 10.3390/ijerph191811628] [Medline: 36141899]
11. Taralunga DD, Florea BC. A blockchain-enabled framework for mHealth systems. Sensors (Basel) 2021 Apr 16;21(8):2828. [doi: 10.3390/s21082828] [Medline: 33923842]
12. Turcian D, Stoicu-Tivadar V. Artificial intelligence in primary care: an overview. Stud Health Technol Inform 2022 Jan 14;289:208-211. [doi: 10.3233/SHTI210896] [Medline: 35062129]
13. Priescu I, Oncioiu I. Measuring the impact of virtual communities on the intention to use telemedicine services. Healthcare (Basel) 2022 Sep 4;10(9):1685. [doi: 10.3390/healthcare10091685] [Medline: 36141297]
14. Tudor AIM, Nichifor E, Litră AV, Chi u IB, Brătucu TO, Brătucu G. Challenges in the adoption of eHealth and mHealth for adult mental health management-evidence from Romania. Int J Environ Res Public Health 2022 Jul 27;19(15):9172. [doi: 10.3390/ijerph19159172] [Medline: 35954526]
15. Iliuta L, Andronesi AG, Panaitescu E, Rac-Albu ME, Scafa-Udri te A, Moldovan H. Challenges for management of dilated cardiomyopathy during COVID-19 pandemic-a telemedicine application. J Clin Med 2022 Dec 14;11(24):7411. [doi: 10.3390/jcm11247411] [Medline: 36556027]
16. Florea M, Lazea C, Gaga R, et al. Lights and shadows of the perception of the use of telemedicine by Romanian family doctors during the COVID-19 pandemic. Int J Gen Med 2021;14:1575-1587. [doi: 10.2147/IJGM.S309519] [Medline: 33953605]
17. Patrascu R, Albai A, Braha A, et al. Instrument for assessing patients' desirability, acceptability, and adherence to telemedicine in diabetes: development, validity, and reliability. Patient Prefer Adherence 2021;15:2705-2713. [doi: 10.2147/PPA.S343869] [Medline: 34898981]

18. Giannouli V, Stoyanova S, Drugas M, Ivanova D. Attitudes towards eHealth during the COVID-19 pandemic: untangling the gordian Knot in Greece, Bulgaria and Romania in healthcare professionals and students? Psychiatr Danub 2021 Dec;33(Suppl 13):415-419. [Medline: 35150518]
19. Iliu ă L, Andronesi AG, Rac-Albu M, et al. Challenges in caring for people with cardiovascular disease through and beyond the COVID-19 pandemic: the sdvantages of universal access to home telemonitoring. Healthcare (Basel) 2023 Jun 12;11(12):1727. [doi: 10.3390/healthcare11121727] [Medline: 37372846]
20. Lăcraru AE, Busnatu S, Pană MA, et al. Assessing the efficacy of a virtual assistant in the remote cardiac rehabilitation of heart failure and ischemic heart disease patients: case-control study of Romanian adult patients. Int J Environ Res Public Health 2023 Feb 22;20(5):3937. [doi: 10.3390/ijerph20053937] [Medline: 36900948]
21. Vizitiu C, Bîră C, Dinculescu A, Nistorescu A, Marin M. Exhaustive description of the system architecture and prototype implementation of an IoT-based eHealth biometric monitoring system for elders in independent living. Sensors (Basel) 2021 Mar 6;21(5):1837. [doi: 10.3390/s21051837] [Medline: 33800728]
22. Anghelescu A. Telerehabilitation: a practical remote alternative for coaching and monitoring physical kinetic therapy in patients with mild and moderate disabling Parkinson's disease during the COVID-19 Pandemic. Parkinsons Dis 2022;2022:4370712. [doi: 10.1155/2022/4370712] [Medline: 35979169]
23. Giunti G, Guisado-Fernandez E, Belani H, Lacalle-Remigio JR. Mapping the access of future doctors to health information technologies training in the European union: cross-sectional descriptive study. J Med Internet Res 2019 Aug 12;21(8):e14086. [doi: 10.2196/14086] [Medline: 31407668]
24. Center of Innovation and e-Health – UMFCD. Telemedicine - current information and skills. 2023. URL: https://cieh.umfcd.ro/2023/09/22/Telemedicina-info [accessed 2025-05-01]
25. Digital Innovation Zone. Telemedicine specialist curriculum. 2024. URL: https://digital-innovation.zone/cultura-si-competente-digitale/telemedicine-specialist/ [accessed 2025-05-01]
26. RoHEALTH webinar. 2024. URL: https://rohealth.ro/detalii-curs/37 [accessed 2025-05-01]
27. Telemed. 2024. URL: https://www.telemed.org.ro/comunicat-de-presa-lansare-proiect.html [accessed 2025-05-01]
28. EHTEL European health telematics association. Vision and mission. 2025. URL: https://ehtel.eu/about/vision-and-mission.html [accessed 2025-05-01]
29. European Commission. Digital education action plan. URL: https://education.ec.europa.eu/focus-topics/digital-education/action-plan [accessed 2025-05-01]
30. Interprofessional education collaborative. IPEC core competencies for interprofessional collaborative practice: version 3. : Interprofessional Education Collaborative; 2023.
31. ISfTeH education program. URL: https://www.isfteh.org/education/isfteh_education_program [accessed 2025-01-25]

Abbreviations

AI: artificial intelligence
ATASS: Association for Telemedicine and Space Applications for Health
CME: Continuing Medical Education
DHCF: Digital Health Competency Framework
EHR: electronic health record
EHTEL: European Health Telematics Association
GDPR: General Data Protection regulations
IoT: Internet of Things
ISfTeH: International Society for Telemedicine & eHealth
ROSA: Romanian Space Agency
WHO: World Health Organization

Edited by B Lesselroth; submitted 23.09.24; peer-reviewed by K Drude, T Vagg; revised version received 01.02.25; accepted 25.02.25; published 14.05.25.

Please cite as:
Focsa MA, Rotaru V, Andronic O, Marginean M, Florescu S
Bridging Gaps in Telemedicine Education in Romania to Support Future Health Care: Scoping Review
JMIR Med Educ 2025;11:e66458
URL: https://mededu.jmir.org/2025/1/e66458
doi:10.2196/66458

AI in the Health Sector: Systematic Review of Key Skills for Future Health Professionals

Javier Gazquez-Garcia[1], MSc; Carlos Luis Sánchez-Bocanegra[2], PhD; Jose Luis Sevillano[3], PhD

[1]Servicio Andaluz de Salud, Distrito Sanitario Almeria, Almeria, Spain

[2]Faculty of Health Sciences, Universidad Oberta de Catalunya (UOC), Barcelona, Spain

[3]Universidad de Sevilla, ETS Ingenieria Informatica, Avda Reina Mercedes s/n, Sevilla, Spain

Corresponding Author:
Jose Luis Sevillano, PhD
Universidad de Sevilla, ETS Ingenieria Informatica, Avda Reina Mercedes s/n, Sevilla, Spain

Abstract

Background: Technological advancements have significantly reshaped health care, introducing digital solutions that enhance diagnostics and patient care. Artificial intelligence (AI) stands out, offering unprecedented capabilities in data analysis, diagnostic support, and personalized medicine. However, effectively integrating AI into health care necessitates specialized competencies among professionals, an area still in its infancy in terms of comprehensive literature and formalized training programs.

Objective: This systematic review aims to consolidate the essential skills and knowledge health care professionals need to integrate AI into their clinical practice effectively, according to the published literature.

Methods: We conducted a systematic review, across databases PubMed, Scopus, and Web of Science, of peer-reviewed literature that directly explored the required skills for health care professionals to integrate AI into their practice, published in English or Spanish from 2018 onward. Studies that did not refer to specific skills or training in digital health were not included, discarding those that did not directly contribute to understanding the competencies necessary to integrate AI into health care practice. Bias in the examined works was evaluated following Cochrane's domain-based recommendations.

Results: The initial database search yielded a total of 2457 articles. After deleting duplicates and screening titles and abstracts, 37 articles were selected for full-text review. Out of these, only 7 met all the inclusion criteria for this systematic review. The review identified a diverse range of skills and competencies, that we categorized into 14 key areas classified based on their frequency of appearance in the selected studies, including AI fundamentals, data analytics and management, and ethical considerations.

Conclusions: Despite the broadening of search criteria to capture the evolving nature of AI in health care, the review underscores a significant gap in focused studies on the required competencies. Moreover, the review highlights the critical role of regulatory bodies such as the US Food and Drug Administration in facilitating the adoption of AI technologies by establishing trust and standardizing algorithms. Key areas were identified for developing competencies among health care professionals for the implementation of AI, including: AI fundamentals knowledge (more focused on assessing the accuracy, reliability, and validity of AI algorithms than on more technical abilities such as programming or mathematics), data analysis skills (including data acquisition, cleaning, visualization, management, and governance), and ethical and legal considerations. In an AI-enhanced health care landscape, the ability to humanize patient care through effective communication is paramount. This balance ensures that while AI streamlines tasks and potentially increases patient interaction time, health care professionals maintain a focus on compassionate care, thereby leveraging AI to enhance, rather than detract from, the patient experience.

(JMIR Med Educ 2025;11:e58161) doi:10.2196/58161

KEYWORDS

artificial intelligence; healthcare competencies; systematic review; healthcare education; AI regulation

Introduction

Technological advancements have transformed health care, improving diagnostics and patient care through solutions such as diagnostic support systems and telemedicine. These technologies reduce costs and enhance care quality, but their adoption faces challenges, particularly in terms of infrastructure, training, and education [1,2]. To address these challenges, the European Commission's DigComp framework [3] outlines key competencies needed to adapt to new technologies, with a particularly challenging issue being the integration of artificial intelligence (AI) into health care.

AI systems use complex algorithms to analyze large datasets, identify patterns, and improve decision-making [4]. AI

classification is multifaceted. The European Commission distinguishes between AI software, such as virtual assistants, and AI embedded in physical devices, such as robots [3]. Alternatively, Russell and Norvig's [5] taxonomy assesses AI based on cognitive and behavioral capabilities, differentiating systems that emulate human or rational thought and action.

Each AI approach is designed to refine its problem-solving abilities, whether by mimicking human behavior or optimizing logical decisions [4]. AI improves clinical decision-making by addressing human limitations in data processing, supporting evidence-based practices through technologies such as machine learning and deep learning [6]. Various AI applications, such as image processing and convolutional neural networks, enhance clinical outcomes [6].

While human interaction remains central to health care, AI mitigates cognitive biases and provides faster, more precise outcomes [7]. For example, AI can process millions of medical images far faster than a human radiologist, improving accuracy through continual learning [6].

The integration of AI into health care is not just a technological advancement but a profound transformation of the operational, cultural, and ethical frameworks of health care organizations. This shift requires professionals to develop specialized knowledge in AI disciplines, such as machine learning, deep learning, and natural language processing, to use these tools effectively and ethically [8].

AI's integration demands expertise and regulatory oversight to ensure ethical use, bias mitigation, and privacy protection [9]. The US Food and Drug Administration is exploring regulatory frameworks for AI-based algorithms in medicine, though a definitive process has yet to be established [10]. A recent study in *JAMA Ophthalmology* [11,12] highlighted the risks of AI misuse, including sophisticated models such as ChatGPT. Professionals are essential in identifying data falsifications that may not be evident to untrained individuals [12].

Addressing these challenges requires a concerted effort to enhance AI literacy and redesign clinical processes, fostering synergy between human judgment and AI-augmented decision-making [13,14]. This transition demands rigorous training in technical, procedural, and collaborative skills to ensure health care professionals can effectively integrate AI into practice, enabling them to adapt to the evolving technological landscape and improve clinical practice [15].

Numerous studies have explored the integration of AI in medical education, demonstrating its potential to enhance practical skills and personalize the learning experience for students. However, most of these reviews focus on the application of AI to improve medical education [16], which examines the use of various AI methods to enhance training in different medical domains.

Some studies [17] propose that training health care professionals in AI requires specialized roles, such as health information management professionals, to manage and adapt AI technologies in clinical settings. This approach ensures that AI integration occurs safely and efficiently, considering data quality, ethical, and legal aspects.

A review of how training programs for health care professionals deal with AI shows both the relatively low number of programs available and their significant limitations [18]. The authors recommend future curricula be designed with AI-related core competencies in mind.

This review aims to identify and highlight the critical competencies necessary to guide the development of educational programs designed to optimize the use of AI in clinical settings, thus addressing a growing need at the intersection of medicine and technology. By analyzing and synthesizing the existing literature on AI training for health care professionals, our goal is to provide a comprehensive framework that informs continuous education and specialized training in AI, ensuring the safe and effective implementation of these advanced technologies in daily clinical practice.

The integration of AI into health care presents several critical challenges. This study aims to consolidate and articulate the specific skills and knowledge required for health care professionals to effectively implement AI in routine clinical practice. Our research question is as follows: "In healthcare professionals, what specific skills and competencies are necessary for the effective implementation and use of AI technologies in daily clinical practice compared to their current skill set?" This question focuses on identifying and defining the critical competencies required for health care professionals to effectively use AI, providing a solid foundation for the development of educational and training programs in this emerging field.

Methods

Study Design

In conducting a systematic review between November and December 2023, we adhered to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [19] (Multimedia Appendix 1 and Checklist 1).

Data Sources and Search Strategy

Comprehensive searches were performed across databases including PubMed, Scopus, and Web of Science, using Health Science Descriptors and Medical Subject Headings descriptors pertinent to AI in health care and associated skills. The search strategy was refined using Boolean and truncation operators. The search queries included combinations such as ("Artificial Intelligence" OR "AI") AND "Healthcare Professionals" AND ("Skills" OR "Competencies" OR "Education"), ("Artificial Intelligence" OR "AI") AND ("data analysis" OR "ethical considerations").

Inclusion and Exclusion Criteria

We established specific selection criteria, prioritizing peer-reviewed original articles, review articles, editorials, and commentaries that directly explored the required skills for health care professionals to integrate AI into their practice. Regarding the inclusion criteria, studies were considered if they: were published in English or Spanish, taking advantage of the linguistic accessibility for the research team in order to reflect the possible specific characteristics of the Spanish-speaking

community; published from 2018 onward, to ensure the relevance and timeliness of the research given the rapid advancements in AI technologies in recent years; focused on the necessary skills for the effective use of AI by health care professionals, encompassing both technical and management competencies; and included aspects related to training in digital health, highlighting the importance of specific training in the use of emerging technologies.

Conversely, studies were excluded if they did not meet these criteria: studies written in languages other than English or Spanish, as they could not be accurately analyzed by the research team; studies published before 2018, to focus on the most recent trends in AI in health care; studies not specifically related to the skills or needs of health care professionals for the use of AI tools, excluding research that did not directly address this focus; and studies that did not refer to specific skills or training in digital health, discarding those that did not directly contribute to understanding the competencies necessary to integrate AI into health care practice.

This selection methodology was designed to identify studies that provided significant evidence on the key competencies health care professionals need to develop for effectively integrating AI into their clinical practice, thus, ensuring that the systematic review focused on research offering practical and applicable insights.

Data Extraction and Study Quality Assessment

The search strategy was developed iteratively to optimize the retrieval of relevant studies. An initial search used the aforementioned databases and descriptors. Titles and abstracts underwent rigorous review for relevance by 2 independent researchers (JG-G and CLS-B), with duplicates removed and articles not meeting inclusion criteria or fitting exclusion criteria discarded. Mendeley (Elsevier Ltd) served as the reference management tool, not involved in the data extraction process.

Subsequently, full-text articles were retrieved for an in-depth content review based on the inclusion criteria specified earlier. However, 3 articles were not retrievable despite repeated efforts. Two articles were inaccessible due to subscription restrictions, and attempts to obtain them via interlibrary loans were unsuccessful. The third article had a broken link, and the authors were unresponsive after multiple contact attempts. These articles have been documented in the “reports not retrieved” section of the PRISMA flow diagram in the Results section.

To assure methodological integrity, discrepancies were resolved through consultation with a third researcher (JLS), reaching consensus on study admissibility. The screening process was manual, without the use of automation tools, and each study was carefully evaluated. Mendeley was used again for reference management, without affecting the data extraction process. The PRISMA flow diagram is provided in the Results section.

Data were extracted and synthesized from eligible studies, encompassing details such as authors, publication year, study design, identified skills, and methodological quality assessment. Tables served as the primary method for tabulating and visually presenting the results and their synthesis.

The data search spanned all pertinent dimensions of these outcomes, including variables relevant to the analyzed studies. These variables covered aspects such as humanization, social skills, and participants’ AI usage experience, offering a comprehensive perspective on the competencies necessary for integrating AI into health care practices.

The GRADE (Grading of Recommendations Assessment, Development, and Evaluation) framework [20] was applied to assess the evidence quality of the included works, considering the quality of studies, result consistency, imprecision, potential bias, and other pertinent factors. Mixed methods or qualitative studies were appraised using the mixed methods appraisal tool [21].

Bias in the examined works was evaluated following Cochrane’s domain-based recommendations [22], considering five types of bias, each with its domains. Bias risk was independently assessed by at least 2 reviewers, with discrepancies resolved via discussion or consultation with a third reviewer when necessary.

Synthesis Method

The research question guided the synthesis groupings, focusing the analysis on the skills and competencies necessary for the effective implementation of AI technologies by health care professionals. No standardization metric was applied. The synthesis method involved extracting relevant sections of the studies related to the identified skills and competencies. Overall, the risk of bias assessment of the studies showed no critical results. Consequently, the included studies were synthesized with equal weight.

The categorization of skills and competencies was based on a qualitative synthesis approach, where 2 independent reviewers extracted and coded data from each study. The categories were developed iteratively by grouping competencies that were conceptually aligned. The final domains were determined based on the frequency with which competencies appeared across the studies, with higher-frequency competencies categorized as key domains. For example, “AI fundamentals” and “ethical and legal considerations” were identified as critical due to their frequent mention in the selected studies, whereas others such as “data governance” and “programming” appeared less often but were still considered relevant.

Study design was not restricted, so a meta-analysis was not performed due to the heterogeneity of the studies and the differing amounts and qualities of information reported on skills and competencies. The reported effectiveness of the competencies was synthesized based on each study’s findings. The identified competencies are presented in the results section. Tables and figures aggregate information about this study’s characteristics and focal areas of this review (skills and competencies).

This review aimed to identify and delineate the skills and competencies essential for health care professionals to use AI in their clinical practices effectively. The objective encompassed various outcome domains, not limited to interpreting results from machine learning models, managing AI biases, ethical

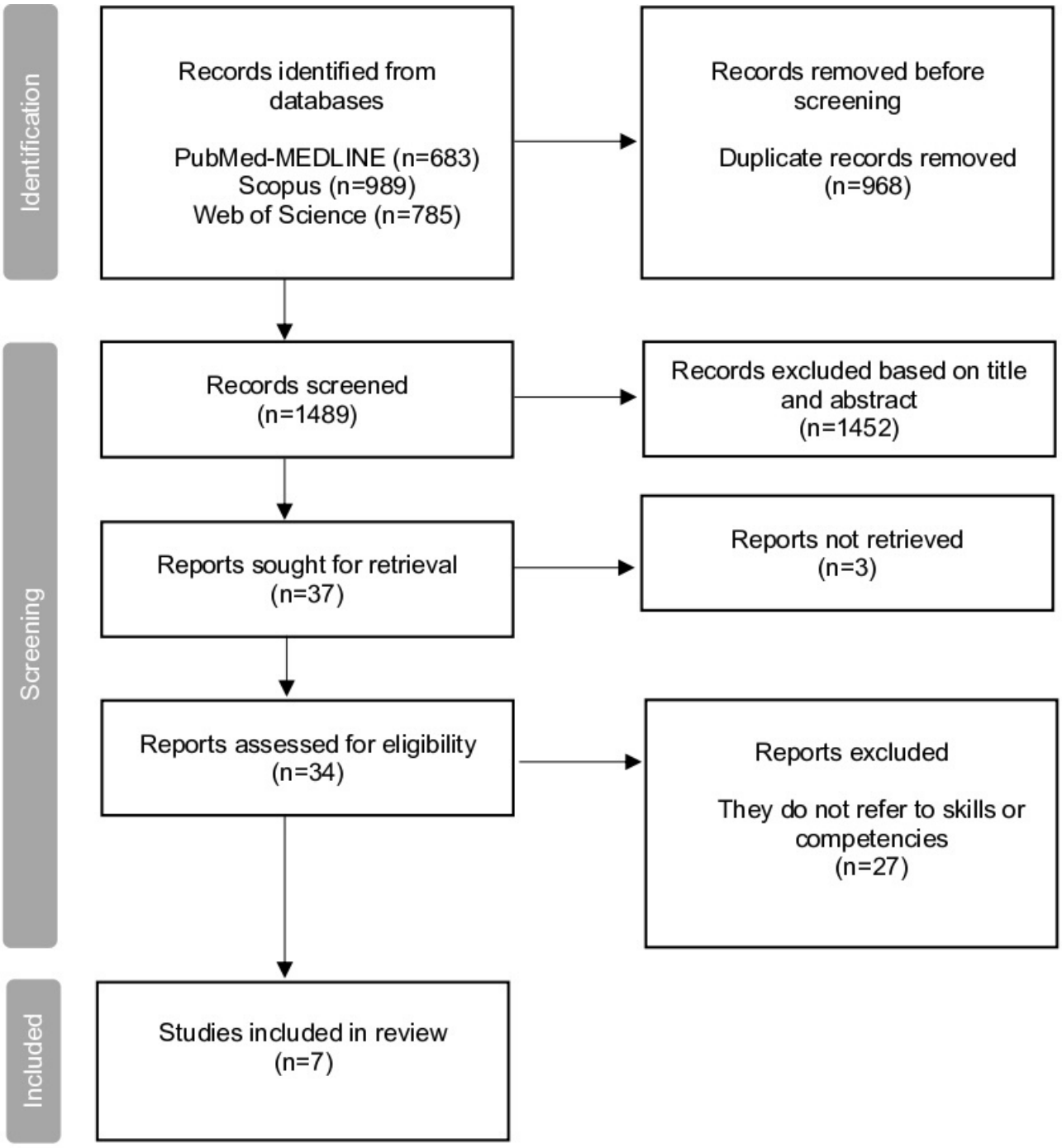
considerations in AI-assisted decision-making, and the technical skills required for effective AI tool use.

Results

Overview

The initial database searches yielded a total of 2457 articles (Figure 1). After deleting duplicates and screening titles and abstracts, 37 articles were selected for full-text review. Out of these, only 7 met all the inclusion criteria for this systematic review [23-29]. Each selected article specifically concentrated on the competencies essential for the integration of AI into routine clinical practice. A substantial number of the excluded articles (n=28) made reference to, but did not directly engage with, the competencies in question. Table S1 in Multimedia Appendix 1 delineates the characteristics of these studies and the principal competencies identified therein.

Figure 1. PRISMA flow diagram. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.



The review identified a diverse range of skills and competencies necessary for the effective implementation of AI in health care. These competencies were categorized into several key areas based on their frequency of appearance in the selected studies. The identified skills, definitions, and the number of studies that reported each are summarized in Table S2 in Multimedia Appendix 1.

Identified Skills

The analysis of the selected studies highlights several key competencies deemed essential for the effective integration of AI into routine clinical practice.

AI fundamentals were identified as an essential competency by 86% of the studies [23-27,29]. This indicates a strong consensus on the importance of foundational AI knowledge for health care professionals. The studies emphasize that a solid understanding of AI principles is crucial for effectively implementing and using AI technologies in clinical settings. This foundational knowledge forms the bedrock upon which more advanced competencies are built, ensuring that professionals can confidently engage with AI tools and integrate them into their practice.

Ethical and legal considerations were emphasized by 71% of the studies [23-26,29]. The studies underscored the importance of having a solid knowledge base in various aspects of ethics, including patient privacy, data security, biases in algorithms, and transparency and explainability.

Data analysis and management skills were highlighted by 43% of the studies [24,26,28]. These studies emphasize several secondary skills crucial for effective data handling. For instance, McCoy et al [26] stress the importance of data acquisition, cleaning, and visualization as foundational steps in preparing data for AI applications. Singh et al [24] underline the need for robust data management practices to ensure the integrity and reliability of AI outputs. Wiljer and Hakim [28] highlight the necessity of developing capabilities in data governance, which includes the secure storage and regulatory compliance of health care data.

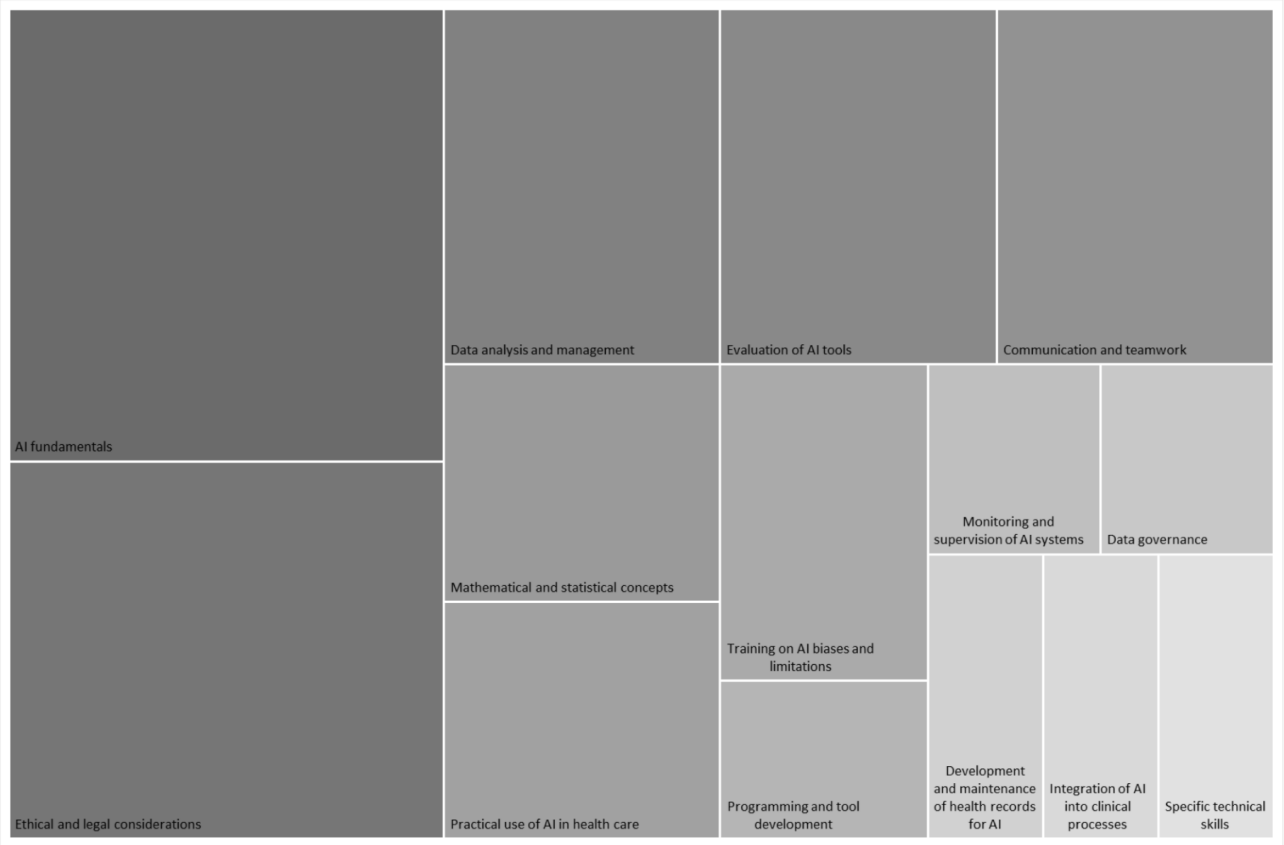
Communication and teamwork were identified as important competencies by 43% of the studies [23,25,27]. These studies underscore the critical need for effective communication skills to convey complex AI-related information to both colleagues and patients. Çalışkan et al [23] emphasize the role of interdisciplinary teamwork in the development and implementation of AI applications, highlighting the necessity for seamless collaboration between health care professionals and AI experts. Liaw et al [25] point out the importance of clear and empathetic communication with patients regarding the use of AI in their care, ensuring transparency and maintaining trust. Sujan et al [27] stress the need for health care teams to work cohesively to monitor and supervise AI systems, ensuring their safe and ethical use.

Evaluation of AI tools was mentioned by 43% of the studies [25,26,29]. These studies highlight several essential secondary skills necessary for the rigorous assessment of AI technologies. Liaw et al [25] emphasize the importance of understanding evidence-based evaluation methods to critically assess the performance and utility of AI tools in clinical practice. McCoy et al [26] discuss the need for health care professionals to be proficient in performing critical evaluations, including assessing the accuracy, reliability, and validity of AI algorithms. Sapci and Sapci [29] stress the significance of ongoing evaluation and monitoring of AI tools to ensure they consistently meet clinical standards and enhance patient outcomes.

Some investigations highlight more technical abilities, such as programming [26], or a deeper mathematical acumen [24,28]. These skills, while crucial, were not as frequently cited as the previously mentioned competencies. Conversely, a capability deemed highly significant across the literature is the aptitude for evaluating AI tools to ascertain their quality and rationalize their application, in addition to scrutinizing potential biases and limitations [24,25].

Figure 2 presents a mosaic plot displaying the distribution of identified competencies based on the percentage of studies that reported each skill. The size of each tile corresponds to the proportion of studies mentioning that competency, providing a visual representation of the relative importance of each skill across the reviewed literature.

Figure 2. Distribution of identified competencies. AI: artificial intelligence.



Discussion

General Aspects

This review underscores the critical need for health care professionals to acquire competencies related to the effective use of AI in clinical practice. Five key areas of competence were identified as essential for AI integration in clinical practice: AI fundamentals, ethical and legal considerations, data analysis and management, communication and teamwork, and the evaluation of AI tools. These competencies are essential to ensure the safe and reliable integration of AI technologies into routine health care workflows. Furthermore, ethical and legal considerations, such as data security and transparency, are crucial for building trust in AI-driven decisions (Table S2 in Multimedia Appendix 1).

In line with our findings, recent studies have emphasized the growing need to define and standardize AI competencies within clinical settings. A 2022 exploratory review [30] highlighted this gap, calling for more structured educational programs to support the integration of AI into health care practice. Additionally, a survey among nursing staff [31] indicated that most professionals had acquired their AI skills independently, further reinforcing the need for formalized training curricula.

Skills

AI Fundamentals

The consensus across the literature highlights the essential nature of AI fundamentals [23-27,29]. This foundational knowledge is crucial for health care professionals to effectively implement and use AI technologies, providing the bedrock upon which more advanced competencies are built. AI fundamentals encompass a basic understanding of machine learning, neural networks, and deep learning. While health care professionals are not expected to have deep expertise in these technical areas, a solid grasp of how AI models function—how they are trained, how they process data, and how they support clinical decision-making—is necessary. This understanding allows professionals to effectively integrate AI into their workflows and ensure the appropriate use of AI-driven tools in patient care.

A key competency is the ability to critically assess AI algorithms, identify biases, and interpret the results. Studies, such as those by Faes et al [32], provide guidelines to help clinicians evaluate AI outputs without needing advanced computational skills, allowing them to focus on patient care. For example, clinicians should recognize when AI models may be limited by biased data or poor generalizability. Similar to using ultrasound technology—where clinicians do not fully understand the physics but can accurately interpret images—the same principle applies to AI. Health care professionals must develop the skills to interpret AI-generated information and effectively communicate it to patients, fostering trust and transparency in AI-driven care [33].

Ethical and Legal Considerations

Ethical and legal considerations were emphasized in 71% of the reviewed studies [23-26,29], highlighting the importance of health care professionals understanding and adhering to the ethical principles and legal frameworks governing the use of AI in clinical practice. These considerations include ensuring

patient privacy, data security, and compliance with relevant regulations, which are critical for maintaining the trust and safety of patients.

Stöger et al [34] emphasized the crucial role of managing high-quality data and the risks associated with its mismanagement, underscoring both the ethical and legal implications for health care professionals using AI systems. In this context, health care providers must be familiar with the laws and regulations governing AI to ensure that the technology is implemented in compliance with legal and ethical standards.

In the European Union, the recent approval of the AI Act introduces comprehensive regulations [35] aimed at safeguarding fundamental rights and democracy by controlling AI systems based on their potential risks and impact. The AI Act establishes stringent requirements for high-risk AI systems, including mandatory impact assessments and prohibitions on certain AI practices that threaten fundamental rights. For health care professionals, this underscores the need for a thorough understanding of these legal frameworks to ensure that AI is used responsibly and safely in clinical practice.

Moreover, the opacity of machine learning algorithms, often referred to as “black boxes” due to their complex and abstract problem-solving methods, represents a significant challenge in ensuring accountability and transparency in AI-driven health care solutions [36]. Although efforts to make these algorithms more interpretable are still developing [37], health care professionals must rely on regulated AI tools that meet established standards for safety and efficacy. Regulatory bodies such as the US Food and Drug Administration have already approved specific AI tools for medical applications [38], providing a level of assurance that these technologies meet rigorous safety and legal standards. This regulatory oversight is critical for fostering trust in AI among clinicians and patients, while also mitigating liability concerns for health care providers.

Data Analysis and Management

Data analysis and management were identified as key competencies in 43% of the reviewed studies [24,26,28], emphasizing the critical role these skills play in the effective implementation of AI in health care. This domain encompasses the ability to handle large volumes of health care data, including data acquisition, cleaning, visualization, and governance. Health care professionals need to be proficient in organizing and managing these datasets to ensure the integrity and reliability of AI outputs.

A core competency in this area includes the preparation of clean and structured datasets for AI models, as highlighted by McCoy et al [26], who stress the importance of data preparation steps such as cleaning and visualization to optimize AI performance. Singh et al [24] underline the need for robust data management practices to guarantee that the data used by AI systems is accurate, reliable, and compliant with health care regulations.

Wiljer and Hakim [28] further highlight the significance of data governance, which includes not only managing data but also ensuring its secure storage and adherence to regulatory requirements. Proficient data handling is crucial, as health care professionals are often the primary data generators. Their ability to manage and interpret large datasets ensures that AI methodologies are applied effectively in clinical practice, supporting both the accuracy of AI predictions and the quality of patient care [39].

Evaluate AI Tools

The ability to evaluate AI tools emphasises the need for health care professionals to critically assess the performance, accuracy, and reliability of AI technologies in clinical practice. Effective evaluation is closely tied to a foundational understanding of AI principles, which allows professionals to determine whether AI tools are suitable for their specific clinical settings.

A key competency in this domain involves assessing AI tools with the same rigor applied to new drugs, diagnostic tests, or treatment protocols. AI-based tools must undergo extensive testing for accuracy, generalizability, efficacy, and fairness to ensure they meet clinical standards [25]. This evaluation process is crucial for ensuring that AI technologies deliver reliable outcomes and can be integrated safely into health care workflows.

Additionally, health care professionals must develop skills for the ongoing evaluation and monitoring of AI tools to detect any performance shifts or biases that may arise over time [40]. Given the continuous learning nature of AI systems, regular reassessment is necessary to maintain their reliability across diverse patient populations and clinical scenarios.

Communication and Teamwork

Communication and teamwork were identified as key competencies in 43% of the reviewed studies [23,25,27], focusing on the dual importance of effectively conveying AI-related insights to patients and facilitating collaboration among health care professionals. Health care providers must clearly explain AI-generated results, addressing patient concerns with empathy, ensuring transparency, and maintaining trust in the use of AI technologies in clinical care.

Equally important is the interdisciplinary collaboration required for integrating AI into clinical workflows. Health care professionals need to work closely with AI specialists, data scientists, and other colleagues to ensure the safe and effective use of AI tools. Studies such as those by Çalışkan et al [23] stress the necessity of seamless teamwork between health care providers and AI experts, while Liaw et al [25] underscore the need for clear and compassionate communication with patients, ensuring they understand how AI is applied in their care and the potential implications.

This competency involves not only understanding AI outputs but also contextualizing them for patients in a way that fosters trust and humanizes AI-assisted care. At the same time, health care teams must collaborate cohesively to monitor and manage AI systems. As AI streamlines routine tasks, health care providers will have more time for patient interaction, making effective and empathetic communication even more critical [25]. Ensuring that both patients and professional teams feel supported and understood is essential for the ethical and successful integration of AI in health care practice.

Competencies Based on Identified Skills

Based on the previously identified skills, and following an approach similar to that outlined by the Association of American Medical Colleges in 2021 [41], we propose a set of competencies that health care professionals should acquire to effectively integrate AI into clinical practice. These competencies, derived from the literature reviewed, encompass the key domains and provide a framework to guide educational programs and ongoing training in AI-driven tools in health care settings.

Table S3 in Multimedia Appendix 1 outlines the competencies for each skill, offering a structured approach to developing AI proficiency. This framework ensures that professionals not only grasp AI fundamentals but also apply these technologies ethically, manage health care data effectively, rigorously evaluate AI tools, and communicate insights to both patients and interdisciplinary teams. However, these competencies should be considered only a first proposal; this list may be modified depending on the specific profile of the training program, on foreseeable technological developments, as well as on the evaluation of the academic results of the new curricula.

AI in Health Care Training

Recent data from Rock Health, a venture capital firm specializing in digital health, have underscored an exponential increase in investments directed toward digital health enterprises and related technologies [42]. This trend distinctly signals an evolving health care landscape, increasingly reliant on novel technologies, thereby accentuating the imperative for health care professionals to proficiently integrate such advancements into their clinical practices [43].

Despite widespread agreement on the necessity for comprehensive AI training from the outset of medical education, there is a lack of consensus on the specific content and approach of such training [15]. The discussion around this issue is abundant, yet concrete resolutions are rare. The findings of this review aim to provide an approach to the possible competencies required, offering a structured framework for developing comprehensive AI training programs for health care professionals.

Integrating AI education into health care curricula presents several challenges due to the variability and lack of standardization of required competencies. To address this need, we propose general guidelines for the development of essential AI competencies for health care professionals. These guidelines provide a reference framework that educators and course coordinators can use to design training programs, assess student progress, and establish performance criteria. For example, health care professionals should acquire the ability to assess the quality of algorithms and their interpretations, as well as identify and mitigate potential biases. This approach facilitates the integration of AI into clinical practice and provides a clear guide for curriculum development and performance evaluation.

Augmenting the education and training of health care professionals is posited to elevate their confidence in using these tools. Although concerns persist regarding AI's potential to supplant human roles, a more discerning view proposes that AI will primarily alleviate the burden of mundane tasks. This reallocation of time and resources is anticipated to enhance patient interactions and elevate the quality of health care services provided [39].

Several authors argue that the existing academic infrastructure is ill-equipped to incorporate AI education, citing time constraints and a lack of teaching expertise as significant obstacles. An alternative proposed involves the use of specific AI tools not only for clinical applications but also to elucidate the underlying algorithms, focusing on their practical use and ethical implications [44].

A solution for integrating AI education into health care curricula, addressing the shortage of instructors with expertise in clinical AI applications, involves leveraging established training programs from other institutions [45]. Programs by Stanford University [46] and Harvard University [47] serve as examples, providing access to high-quality educational content. These programs offer a comprehensive curriculum that covers essential AI concepts, practical applications, and ethical considerations, enabling health care professionals to gain a deep understanding of AI technologies.

The integration of AI into clinical practice is expected to augment, not replace, the roles of health care professionals. It calls for a workforce proficient in digital health and communication, capable of leveraging AI's benefits while recognizing its limitations and ethical considerations [48]. This paradigm shift offers an opportunity to enhance patient care, delegating computational tasks to AI and focusing on the human aspects of health care delivery [37].

Limitations

This systematic review encounters several challenges, primarily due to the limited availability of literature specifically addressing the competencies required for integrating AI into health care practice. The scarcity of targeted studies can be attributed to the nascent and rapidly evolving nature of AI applications in health care.

To mitigate this issue, the search criteria were broadened to include general terms related to AI in health care and competencies required by health care professionals. While necessary, this broadening may have introduced studies that do not exclusively focus on AI competencies, potentially affecting the homogeneity of the findings. The review also faced potential language bias, as it primarily focused on literature in English and Spanish. Pertinent studies in other languages might have been excluded. Despite rigorous and independent review processes, the selection could still be influenced by subjective interpretation of the inclusion and exclusion criteria.

Limiting the review to studies published after 2018 aimed to capture the most recent advancements, but this restriction might omit emerging research. Despite these efforts, there remains a clear knowledge gap in the specific skills required for effective AI use in health care. The literature predominantly addresses AI applications and benefits, but lacks detailed research on the precise skills health care professionals need. Current research is often fragmented and varies significantly in scope and depth. Few studies offer comprehensive models or curricula for AI competency training in health care. This inconsistency highlights

the need for more robust research to identify essential AI skills and explore effective methods for integrating these skills into health care education and practice.

Conclusions and Future Works

This systematic review has identified essential competencies for health care professionals to effectively integrate AI into clinical practice. The analysis reveals a consensus on the importance of five key areas: AI fundamentals, ethical and legal considerations, data analysis and management, communication and teamwork, and evaluation of AI tools. These competencies are crucial for leveraging AI technologies to enhance patient care and health care delivery. A consensus within the scholarly discourse suggests the necessity for health care professionals to attain proficiency in these domains to ensure the judicious application of AI tools, thereby accruing benefits for both patients and the health care ecosystem.

AI fundamentals form the backbone of necessary knowledge, enabling health care professionals to understand and use AI technologies effectively. Ethical and legal considerations ensure that AI applications adhere to patient privacy, data security, and transparency standards, maintaining trust and compliance within health care settings. Data analysis and management skills are vital for handling large datasets, ensuring accurate AI outputs, and supporting informed clinical decisions.

Communication and teamwork are also critical, facilitating the clear conveyance of AI-related information among health care professionals and to patients, thereby promoting transparency and trust. The ability to evaluate AI tools is essential for assessing the performance and reliability of AI technologies, ensuring they meet clinical standards and deliver safe, effective patient care.

Augmenting the education and training of health care professionals is posited to elevate their confidence in using these tools. Although concerns persist regarding AI's potential to supplant human roles, a more discerning view proposes that AI will primarily alleviate the burden of mundane tasks. This reallocation of time and resources is anticipated to enhance patient interactions and elevate the quality of health care services provided.

In this context, the importance of communication skills becomes increasingly paramount. The introduction of AI tools is expected to afford health care professionals additional time per patient encounter, potentially heightening patient satisfaction and care quality.

The ambition extends beyond merely acquiring proficiency in disciplines ancillary to traditional health care paradigms. Considering the already intricate and comprehensive nature of health care education, particularly in medicine, the emphasis is placed on fostering an in-depth comprehension of AI's functionalities, inherent biases, pragmatic utility, and cost-effectiveness compared to abstaining from AI applications.

The integration of AI into health care is indispensable for advancing patient care but requires a concerted effort to develop and standardize competencies among health care professionals. Regulatory oversight and enhanced educational frameworks are essential for overcoming existing barriers and leveraging AI's full potential in clinical settings.

Despite the progress highlighted in this review, significant gaps remain in the literature, particularly concerning the specific educational frameworks and training programs needed to develop these competencies. Most existing research focuses on the potential applications and benefits of AI, with less emphasis on the precise skills required for effective implementation.

Future research should prioritize the development and validation of standardized AI competency frameworks tailored for health care professionals. These frameworks should cover technical skills, ethical and legal aspects, and data management practices. Collaborative efforts between academic institutions, health care organizations, and AI experts can create comprehensive training programs to address these competencies.

Additionally, longitudinal studies are necessary to evaluate the long-term effectiveness of AI training programs. Research should explore how health care professionals apply their AI training in clinical settings, assessing the impact on patient outcomes, clinical decision-making, and health care efficiency.

Acknowledgments

This work has been partially supported by the Telefonica Chair on "Intelligence in Networks" of the Universidad de Sevilla.

Data Availability

All data generated or analyzed during this study are included in this published article. This encompasses detailed descriptions of the databases consulted, the search criteria used, the selection process for the included studies, and the analytical methods applied. Specifically, this paper delineates the comprehensive search strategy, including the exact search terms, the databases accessed (PubMed, Scopus, and Web of Science), and the filters used (such as publication date ranges and language restrictions). Additionally, the criteria for study selection—both inclusion and exclusion criteria—are explicitly outlined to ensure reproducibility and transparency of the review process. The methodologies used for data extraction and analysis are also described, providing insight into how the findings were synthesized and interpreted. Through this approach, we aim to ensure that our systematic review process is fully transparent, enabling other researchers to replicate this study or to conduct further analysis based on the procedures and datasets detailed within this article.

Authors' Contributions

All authors contributed significantly to this literature review. Specifically, each author participated in one or more of the following: conceptualization and design of the review, data acquisition, analysis, interpretation, or critically revising this work. All authors approved the final paper for publication and agree to be accountable for all aspects of this work, ensuring the integrity and accuracy of their contribution.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional tables.
[DOCX File, 109 KB - mededu_v11i1e58161_app1.docx]

Checklist 1
PRISMA checklist.
[DOCX File, 31 KB - mededu_v11i1e58161_app2.docx]

References

1. do Nascimento IJB, Abdulazeem HM, Vasanthan LT, et al. The global effect of digital health technologies on health workers' competencies and health workplace: an umbrella review of systematic reviews and lexical-based and sentence-based meta-analysis. Lancet Digit Health 2023 Aug;5(8):e534-e544. [doi: 10.1016/S2589-7500(23)00092-4] [Medline: 37507197]
2. do Nascimento IJB, Abdulazeem H, Vasanthan LT, et al. Barriers and facilitators to utilizing digital health technologies by healthcare professionals. NPJ Digit Med 2023 Sep 18;6(1):161. [doi: 10.1038/s41746-023-00899-4] [Medline: 37723240]
3. Vuorikari R, Kluzer S, Punie Y. DigComp 22: The Digital Competence Framework for Citizens - With New Examples of Knowledge, Skills and Attitudes: Publications Office of the European Union; 2022. [doi: 10.2760/490274]
4. Qué es la inteligencia artificial. Gobierno de España, Plan de Recuperación Transformación y Resiliencia. URL: https://planderecuperacion.gob.es/noticias/que-es-inteligencia-artificial-ia-prtr [accessed 2025-01-14]
5. Russell S, Norvig P. Artificial Intelligence: A Modern Approach, 4th edition: Pearson – prentice hall.
6. Mintz Y, Brodie R. Introduction to artificial intelligence in medicine. Minim Invasive Ther Allied Technol 2019 Apr;28(2):73-81. [doi: 10.1080/13645706.2019.1575882] [Medline: 30810430]
7. Liu X, Zhang W, Zhang Q, et al. Development and validation of a machine learning-augmented algorithm for diabetes screening in community and primary care settings: a population-based study. Front Endocrinol (Lausanne) 2022;13:1043919. [doi: 10.3389/fendo.2022.1043919] [Medline: 36518245]
8. Kim HS, Kim DJ, Yoon KH. Medical big data is not yet available: why we need realism rather than exaggeration. Endocrinol Metab (Seoul) 2019 Dec;34(4):349-354. [doi: 10.3803/EnM.2019.34.4.349] [Medline: 31884734]
9. Arencibia MG, Cardero DM. Dilemas éticos en el escenario de la inteligencia artificial. Econom y Socied 2020;25(57):1-18. [doi: 10.15359/eys.25-57.5]
10. Harvey HB, Gowda V. How the FDA regulates AI. Acad Radiol 2020 Jan;27(1):58-61. [doi: 10.1016/j.acra.2019.09.017] [Medline: 31818387]
11. Taloni A, Scorcia V, Giannaccare G. Large language model advanced data analysis abuse to create a fake data set in medical research. JAMA Ophthalmol 2023 Dec 1;141(12):1174-1175. [doi: 10.1001/jamaophthalmol.2023.5162] [Medline: 37943569]
12. Naddaf M. ChatGPT generates fake data set to support scientific hypothesis. Nature New Biol 2023 Nov 30;623(7989):895-896. [doi: 10.1038/d41586-023-03635-w]
13. Lauricella LL, Pêgo-Fernandes PM. Databases, big data and artificial intelligence: what healthcare professionals need to know about them. Sao Paulo Med J 2022;140(6):737-738. [doi: 10.1590/1516-3180.2022.140611082022]
14. Kamble SS, Gunasekaran A, Goswami M, Manda J. A systematic perspective on the applications of big data analytics in healthcare management. Int J Healthc Manag 2019 Jul 3;12(3):226-240. [doi: 10.1080/20479700.2018.1531606]
15. Grunhut J, Wyatt AT, Marques O. Educating future physicians in artificial intelligence (AI): an integrative review and proposed changes. J Med Educ Curric Dev 2021;8:23821205211036836. [doi: 10.1177/23821205211036836] [Medline: 34778562]
16. Nagi F, Salih R, Alzubaidi M, et al. Applications of artificial intelligence (AI) in medical education: a scoping review. Stud Health Technol Inform 2023 Jun 29;305:648-651. [doi: 10.3233/SHTI230581] [Medline: 37387115]
17. Stanfill MH, Marc DT. Health information management: implications of artificial intelligence on healthcare data and information management. Yearb Med Inform 2019 Aug;28(1):56-64. [doi: 10.1055/s-0039-1677913] [Medline: 31419816]
18. Charow R, Jeyakumar T, Younus S, et al. Artificial intelligence education programs for health care professionals: scoping review. JMIR Med Educ 2021 Dec 13;7(4):e31043. [doi: 10.2196/31043] [Medline: 34898458]

19. Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097. [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
20. Aguayo-Albasini JL, Flores-Pastor B, Soria-Aledo V. Sistema GRADE: clasificación de la calidad de la evidencia y graduación de la fuerza de la recomendación. Cir Esp 2014 Feb;92(2):82-88. [doi: 10.1016/j.ciresp.2013.08.002]
21. Hong QN, Fàbregues S, Bartlett G, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Educ Inf 2018 Dec 18;34(4):285-291. [doi: 10.3233/EFI-180221]
22. Higgins JPT, Thomas J, Chandler J, et al. Cochrane Handbook for Systematic Reviews of Interventions Version 64: Cochrane; 2023. URL: https://training.cochrane.org/handbook [accessed 2025-01-14]
23. Çalışkan SA, Demir K, Karaca O. Artificial intelligence in medical education curriculum: an e-Delphi study for competencies. PLoS ONE 2022;17(7):e0271872. [doi: 10.1371/journal.pone.0271872] [Medline: 35862401]
24. Singh RP, Hom GL, Abramoff MD, Campbell JP, Chiang MF, AAO Task Force on Artificial Intelligence. Current challenges and barriers to real-world artificial intelligence adoption for the healthcare system, provider, and the patient. Transl Vis Sci Technol 2020 Aug;9(2):45. [doi: 10.1167/tvst.9.2.45] [Medline: 32879755]
25. Liaw W, Kueper JK, Lin S, Bazemore A, Kakadiaris I. Competencies for the use of artificial intelligence in primary care. Ann Fam Med 2022;20(6):559-563. [doi: 10.1370/afm.2887] [Medline: 36443071]
26. McCoy LG, Nagaraj S, Morgado F, Harish V, Das S, Celi LA. What do medical students actually need to know about artificial intelligence? NPJ Digit Med 2020;3(1):86. [doi: 10.1038/s41746-020-0294-7] [Medline: 32577533]
27. Sujan MA, White S, Habli I, Reynolds N. Stakeholder perceptions of the safety and assurance of artificial intelligence in healthcare. Saf Sci 2022 Nov;155(8):105870. [doi: 10.1016/j.ssci.2022.105870]
28. Wiljer D, Hakim Z. Developing an artificial intelligence-enabled health care practice: rewiring health care professions for better care. J Med Imaging Radiat Sci 2019 Dec;50(4 Suppl 2):S8-S14. [doi: 10.1016/j.jmir.2019.09.010] [Medline: 31791914]
29. Sapci AH, Sapci HA. Artificial intelligence education and tools for medical and health informatics students: systematic review. JMIR Med Educ 2020 Jun 30;6(1):e19285. [doi: 10.2196/19285] [Medline: 32602844]
30. Garvey KV, Craig KJT, Russell R, Novak LL, Moore D, Miller BM. Considering clinician competencies for the implementation of artificial intelligence-based tools in health care: findings from a scoping review. JMIR Med Inform 2022 Nov 16;10(11):e37478. [doi: 10.2196/37478] [Medline: 36318697]
31. Abuzaid MM, Elshami W, Fadden SM. Integration of artificial intelligence into nursing practice. Health Technol (Berl) 2022;12(6):1109-1115. [doi: 10.1007/s12553-022-00697-0] [Medline: 36117522]
32. Faes L, Liu X, Wagner SK, et al. A clinician's guide to artificial intelligence: how to critically appraise machine learning studies. Transl Vis Sci Technol 2020 Feb 12;9(2):7. [doi: 10.1167/tvst.9.2.7] [Medline: 32704413]
33. Nagy M, Radakovich N, Nazha A. Why machine learning should be taught in medical schools. Med Sci Educ 2022 Apr;32(2):529-532. [doi: 10.1007/s40670-022-01502-3] [Medline: 35528308]
34. Stöger K, Schneeberger D, Kieseberg P, Holzinger A. Legal aspects of data cleansing in medical AI. Comput Law Secur Rev 2021 Sep;42:105587. [doi: 10.1016/j.clsr.2021.105587]
35. European Commission. Official Journal of the European Union; Regulation (EU) 2023/206 on artificial intelligence. EU Artificial Intelligence Act. 2024. URL: https://artificialintelligenceact.eu/the-act/ [accessed 2025-01-14]
36. Lysaght T, Lim HY, Xafis V, Ngiam KY. AI-assisted decision-making in healthcare. ABR 2019 Sep;11(3):299-314. [doi: 10.1007/s41649-019-00096-0]
37. Savage N. Breaking into the black box of artificial intelligence. Nature New Biol 2022 Mar 29. [doi: 10.1038/d41586-022-00858-1] [Medline: 35352042]
38. Milam ME, Koo CW. The current status and future of FDA-approved artificial intelligence tools in chest radiology in the United States. Clin Radiol 2023 Feb;78(2):115-122. [doi: 10.1016/j.crad.2022.08.135] [Medline: 36180271]
39. Khanra S, Dhir A, Islam A, Mäntymäki M. Big data analytics in healthcare: a systematic literature review. Ent Inf Syst 2020 Aug 8;14(7):878-912. [doi: 10.1080/17517575.2020.1812005]
40. James CA, Wheelock KM, Woolliscroft JO. Machine learning: the next paradigm shift in medical education. Acad Med 2021 Jul 1;96(7):954-957. [doi: 10.1097/ACM.0000000000003943] [Medline: 33496428]
41. AAMC. AAMC new and emerging areas in medicine series; telehealth competencies across the learning continuum. Arizona Telemedine Program.: AAMC; 2021. URL: https://telemedicine.arizona.edu/sites/default/files/training/2022/Feb/5%20-%20AAMC-2021-telehealth-competencies%20across%20the%20learning%20continum.pdf [accessed 2025-01-14]
42. Zweig M, Desiliva J. Q1 2021 funding report: digital health is all grown up. Rock Health. 2021. URL: https://rockhealth.com/insights/q1-2021-funding-report-digital-health-is-all-grown-up/ [accessed 2025-01-14]
43. Marwaha JS, Landman AB, Brat GA, Dunn T, Gordon WJ. Deploying digital health tools within large, complex health systems: key considerations for adoption and implementation. NPJ Digit Med 2022 Jan 27;5(1):13. [doi: 10.1038/s41746-022-00557-1] [Medline: 35087160]
44. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. JMIR Med Educ 2019 Jun 15;5(1):e13930. [doi: 10.2196/13930] [Medline: 31199295]

45. Grunhut J, Marques O, Wyatt ATM. Needs, challenges, and applications of artificial intelligence in medical education curriculum. JMIR Med Educ 2022 Jun 7;8(2):e35587. [doi: 10.2196/35587] [Medline: 35671077]
46. Stanford center for professional development. Stanford University. URL: https://online.stanford.edu/ [accessed 2025-01-14]
47. Artificial intelligence in medicine program. Harvard University. URL: https://aim.hms.harvard.edu/ [accessed 2025-01-14]
48. Heckman GA, Hirdes JP, McKelvie RS. The role of physicians in the era of big data. Can J Cardiol 2020 Jan;36(1):19-21. [doi: 10.1016/j.cjca.2019.09.018] [Medline: 31787436]

Abbreviations

AI: artificial intelligence
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by A Hasan Sapci, B Lesselroth; submitted 07.03.24; peer-reviewed by B Matejko, K Varikara, Z Zandesh; revised version received 04.10.24; accepted 02.01.25; published 05.02.25.

Please cite as:
Gazquez-Garcia J, Sánchez-Bocanegra CL, Sevillano JL
AI in the Health Sector: Systematic Review of Key Skills for Future Health Professionals
JMIR Med Educ 2025;11:e58161
URL: https://mededu.jmir.org/2025/1/e58161
doi:10.2196/58161

Review

Motivation Theories and Constructs in Experimental Studies of Online Instruction: Systematic Review and Directed Content Analysis

Adam Gavarkovs[1], PhD; Erin Miller[2], PhD; Jaimie Coleman[3], MPT; Tharsiga Gunasegaran[4], HBSc; Rashmi A Kusurkar[5], MD, PhD; Kulamakan Kulasegaram[6], PhD; Melanie Anderson[7], MLiS; Ryan Brydges[8], PhD

[1]Division of Continuing Professional Development, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada

[2]School of Physical Therapy, Faculty of Health Sciences, Western University, London, ON, Canada

[3]School of Physical Therapy, University of Toronto, Toronto, ON, Canada

[4]University of Toronto, Toronto, ON, Canada

[5]Amsterdam UMC location Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

[6]Department of Family and Community Medicine, Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada

[7]University Health Network, Toronto, ON, Canada

[8]Department of Medicine, Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada

Corresponding Author:
Adam Gavarkovs, PhD
Division of Continuing Professional Development
Faculty of Medicine
University of British Columbia
555 West 12th Avenue
Suite 200
Vancouver, BC, V5Z 3X7
Canada
Phone: 1 6046753777
Email: adam.g@ubc.ca

Abstract

Background: The motivational design of online instruction is critical in influencing learners' motivation. Given the multifaceted and situated nature of motivation, educators need access to a range of evidence-based motivational design strategies that target different motivational constructs (eg, interest or confidence).

Objective: This systematic review and directed content analysis aimed to catalog the motivational constructs targeted in experimental studies of online motivational design strategies in health professions education. Identifying which motivational constructs have been most frequently targeted by design strategies—and which remain under-studied—can offer valuable insights into potential areas for future research.

Methods: Medline, Embase, Emcare, PsycINFO, ERIC, and Web of Science were searched from 1990 to August 2022. Studies were included if they compared online instructional design strategies intending to support a motivational construct (eg, interest) or motivation in general among learners in licensed health professions. Two team members independently screened and coded the studies, focusing on the motivational theories that researchers used and the motivational constructs targeted by their design strategies. Motivational constructs were coded into the following categories: intrinsic value beliefs, extrinsic value beliefs, competence and control beliefs, social connectedness, autonomy, and goals.

Results: From 10,584 records, 46 studies were included. Half of the studies (n=23) tested strategies aimed at making instruction more interesting, enjoyable, and fun (n=23), while fewer studies tested strategies aimed at influencing extrinsic value beliefs (n=9), competence and control beliefs (n=6), social connectedness (n=4), or autonomy (n=2). A focus on intrinsic value beliefs was particularly evident in studies not informed by a theory of motivation.

Conclusions: Most research in health professions education has focused on motivating learners by making online instruction more interesting, enjoyable, and fun. We recommend that future research expand this focus to include other motivational constructs,

such as relevance, confidence, and autonomy. Investigating design strategies that influence these constructs would help generate a broader toolkit of strategies for educators to support learners' motivation in online settings.

Trial Registration: PROSPERO CRD42022359521; https://www.crd.york.ac.uk/PROSPERO/view/CRD42022359521

(JMIR Med Educ 2025;11:e64179) doi:10.2196/64179

KEYWORDS

motivation; internet; systematic review; experimental studies; online instruction; educator; learner; researcher; health professional; education; tool-kit; autonomy

Introduction

The internet has become a preferred modality for health professions education (HPE) in the postpandemic landscape [1]. A recent global survey found that 60% of health professionals preferred blended learning, while 32% preferred fully online learning [2]. Online instruction can ameliorate barriers due to geography, scheduling, and cost that make in-person learning infeasible for many health professionals and trainees [3]. However, one challenge of online learning is keeping learners motivated. Motivation—the energetic force that instigates and sustains behavior [4]—is key to success when learning online [5,6]. A lack of face-to-face interaction and the metacognitive demands associated with learning online can lead to feelings of isolation, frustration, and diminished motivation [7,8]. To address these challenges and keep learners motivated, educators must build motivational support into online instruction through a process known as motivational design [9].

Motivational design is defined by Keller [9] as "the process of arranging resources and procedures to bring about changes in people's motivation." This process involves selecting, adapting, and applying motivational design strategies, which are resources and procedures that facilitate the motivational processes underpinning learning. For example, Colonnello et al [10] enhanced medical students' motivation by supplementing surgical videos with emotionally salient patient information. Other studies have demonstrated that other motivational design strategies, such as using narration in online modules, can impact learner motivation [11,12].

Motivational design strategies work by influencing various motivational constructs—cognitive factors that shape learners' moment-to-moment motivation [4]. Broad categories of motivational constructs include goals ("What am I aiming to do?"), competence beliefs ("Can I do it?"), value beliefs ("Do I want to do it? Why?"), and attributional beliefs ("Why did it happen this way?") [13]. For example, an educator might use a strategy to make learning seem more relevant, increase learners' interest, or boost their confidence that they can learn the material.

Theories of motivation emphasize that learners' motivation is influenced by several motivational constructs, any one of which may be the cause of poor motivation during online learning [4]. For example, medical students completing an online module on a basic science topic may be confident in their ability to learn but struggle to see the value in the material beyond their next examination. Conversely, students completing a virtual examination with a standardized patient may see the value in what they are learning but not feel confident in their ability to succeed. In the first case, an educator could use a strategy that targets learners' value beliefs (eg, a prompt to reflect on the clinical relevance of the material [14]), while in the second, an educator could use a strategy that targets' learners competence beliefs (eg, providing a demonstration that learners can observe beforehand [9]). Given the multifaceted and situated nature of motivation, educators need access to a range of evidence-based motivational design strategies that target different motivational constructs, such as strategies for enhancing confidence or perceived value [15].

Researchers can support educators by providing evidence on the effectiveness of different motivational design strategies [16]. However, we do not have a good understanding of which motivational constructs are most frequently targeted in research on online motivational design. For example, are researchers disproportionately focused on testing ways to make online instruction more interesting or enjoyable? An expanding literature on serious games and gamification in HPE suggests this may be the case, as games are often framed as a strategy to enhance interest [17-24]. While enhancing interest is important, if researchers focus too narrowly on this construct at the expense of others (eg, confidence), then educators may not receive the full range of design strategies needed to support learner motivation [4]. To inform future research, it is important to identify which motivational constructs have been most emphasized and which remain under-studied.

To address this gap, our review aims to catalog the motivational constructs targeted in studies of online motivational design strategies. This is a novel objective, as no previous reviews have organized the instructional design literature based on the motivational constructs that strategies aim to influence. By identifying which constructs have received the most attention, we aim to guide future literature syntheses on the most effective design strategies for supporting these constructs. Additionally, by identifying under-studied constructs, we aim to guide areas for future primary research. Ultimately, our review is intended as a resource for researchers interested in conducting future studies on motivational design for online instruction. Stimulating ongoing research in this area will ensure that educators have access to evidence-based guidance to design more motivating online instruction.

We hypothesize that there are two reasons why certain motivational constructs may be underrepresented in research on online motivational design strategies: (1) studies are not informed by a theory of motivation or model of motivational design, or (2) studies are informed by such theories but choose

not to focus on specific constructs. To disentangle these explanations, we posed two research questions: (1) Which theories or models of motivation, if any, inform experimental comparison studies of motivational design strategies for online instruction? (2) Within experimental comparison studies of motivational design strategies for online instruction, which motivational constructs, if any, have been targeted?

Methods

We conducted a systematic review and directed content analysis focused on experimental comparison studies in HPE [25]. Experimental comparison studies, which compare 1 version of online instruction to another, are uniquely positioned to generate empirical evidence for the causal effects of motivational design strategies [25-28]. Motivational design is, at its core, a process of making predictions about the causal effects of motivational design strategies ("If I use this strategy, will it cause my learners to be more motivated?"). Since experimental comparison studies are best suited for making causal claims, we consider them a necessary source of evidence for educators and serve as the focus for our review. Bajpai et al [29] adopted a similar position in their recent review of learning theories in randomized trials of digital instruction in HPE.

Given our focus on experimental comparison studies, we identified a systematic review as the most appropriate review methodology [30]. We registered (PROSPERO CRD42022359521) and published a review protocol [31], and report our findings in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 updated guidelines [32], with a few exceptions. We omit items 12 (effect measures), 14 (reporting bias assessment), 15 (certainty assessment), 19 (results of individual studies), 21 (risk of bias due to missing results), and 22 (certainty of evidence), as we did not intend to appraise nor synthesize the outcomes of included studies. Further details on our methods can be found in our published protocol [31]. To increase the clarity and brevity of reporting, this paper omits data related to a few research questions listed in our published protocol. Additional data regarding these questions is available upon request.

Eligibility Criteria

Study Characteristics

We included individual and cluster randomized controlled trials and quasi-experimental studies published in English from 1990 to August 2, 2022 (for databases) and September 15, 2022 (for registries). Our date range aligns with prior reviews of digital education in HPE [33]. We included protocols for planned or ongoing studies but excluded conference abstracts and unpublished studies. Studies were not excluded based on quality or risk of bias as we did not aim to synthesize the results of studies. However, we appraised the risk of bias to provide readers with additional context regarding the quality of studies.

Participants

We included studies focusing on learners in the health professions regardless of training status (see protocol for list of health professions), either exclusively or when mixed with other learners (eg, psychology students).

Interventions

We included studies comparing online instructional designs (or that could have been delivered online, such as CD-ROM instruction), which targeted a motivational construct (eg, interest) or motivation more generally. By "targeting" motivation, we mean that researchers stated that their instructional design aimed to enhance learner motivation to engage with instruction. Several studies demonstrated a cursory treatment of motivation, for example, by discussing the impact of design strategies on constructs (eg, interest) without grounding the construct in a theoretical framework. We decided to include these studies because they contribute to our understanding of the foci among researchers interested in this area of HPE. Studies comparing online instruction against paper-based or face-to-face instruction were excluded.

Outcomes

We included studies that assessed any learner outcome.

Search Strategy and Selection Process

Database Searching

Strategies were developed for Ovid Medline, Embase, Emcare PsycINFO, EBSCO ERIC, and Web of Science Core Collection (Social Sciences Citation Index; Arts & Humanities Citation Index; Book Citation Index-Social Sciences & Humanities; Conference Proceedings Citation Index-Science; Emerging Sources Citation Index; Science Citation Index; Book Citation Index-Social Sciences & Humanities; and Conference Proceedings Citation Index-Social Science & Humanities) by a health sciences librarian (MA) in collaboration with the review team (Multimedia Appendix 1). Appropriate subject headings and keywords for motivation, online instruction, and HPE focused on the licensed professions were used for each database. The results were limited to those published from 1990 to the date of the searches. The searches were run on August 2, 2022, and the 14,736 results were uploaded to Covidence for screening.

Registry Searching

For the Open Science Framework Registries, we developed 12 searches, comprised of different combinations of the highest yielding terms in our database searches (Multimedia Appendix 2). The searches yielded between 7277 and 16,018 hits for each combination of terms. AG manually screened the first 10 pages of results (10 results per page) for each search (1200 studies screened in total) and uploaded 19 potentially relevant studies to Covidence.

Hand and Reference Searching

AG manually screened several published literature reviews on online instruction in HPE [18-23,34-39] and the references of included studies and uploaded 161 potentially relevant studies to Covidence.

Screening

After removing duplicates, we screened 10,584 records. Two team members independently screened abstracts and, as necessary, the paper's full text. Before independent screening, all 6 team members who participated in the screening process practiced screening the same 30 abstracts, and then discussed and refined the inclusion criteria. AG also developed a decision tool to support full-text screening. As screening progressed, AG periodically reviewed conflicts for any systematic issues and further refined the inclusion and exclusion criteria. Two senior team members (EM or RB) not involved in the initial decision resolved all conflicts. We included 61 studies in the data extraction phase. During the extraction phase we excluded an additional 15 studies. In 12 cases, the papers were excluded because they did not discuss the potential motivational effects of a strategy in the introduction or did not state an objective to assess the effects of a strategy on motivation. Therefore, we concluded that these were not motivational design strategies [40-51]. This yielded 46 studies included in our review.

Data Collection and Synthesis Methods

Overview

The data items we extracted can be found in Multimedia Appendix 3. We conducted a directed content analysis during the extraction process [52], coding each study deductively regarding the motivational theories used and the motivational constructs targeted. We piloted and refined the extraction process in Covidence with a few included studies. AG trained team members to extract and code data. Two team members independently extracted data from each study. Conflicts were resolved through discussion, with an experienced team member (ie, currently in, or having completed, a PhD program) not involved in the initial decision leading to resolution.

Theories of Motivation (Aligned With Research Question 1)

We developed an a priori list of 6 prominent theories of motivation and 1 model of motivational design to deductively guide our coding. We defined theories as "prominent" based on meeting one of the following criteria: (1) they were included in a 2020 special issue of *Contemporary Educational Psychology* titled "Prominent Motivation Theories: The Past, Present, and Future" [53-57], or (2) they have been the subject of an AMEE Guide in *Medical Teacher* [58,59]. We also added Keller's ARCS model of motivational design, which we assumed would be cited in HPE studies [24]. Brief descriptions of these theories can be found in Table 1. Beyond this initial list, we considered any theory aiming to explain the energetic basis and direction of learners' engagement to be a theory of motivation [60]. We also coded whether these theories informed 4 key aspects of the research process: the research questions, the design of the experimental conditions, the selection of methods and measures, and the interpretation of results [61].

Table 1. Overview of and reported use of established theories of motivation and models of motivational design.

Theory or model	Description	Frequency used, n (%)	References
SDT[a] [55]	Ryan and Deci's SDT differentiates between types of motivation depending on learners' reasons for engaging in learning, such as feeling pressured to satisfy external demands (external regulation), feeling pressured to quell feelings of guilt or shame (introjected regulation), identifying with the value of an activity (identified regulation), or finding the activity inherently interesting (intrinsic motivation). SDT also emphasizes the influence of the social environment on learners' motivation, as mediated by the satisfaction of feelings of autonomy (ie, being in control of one's actions), competence (ie, feeling efficacious in one's actions), and relatedness (ie, feeling connected to others).	8 (17)	[10,11,62-67]
ARCS[b] model [9]	Keller's ARCS model states that, for learners to become and remain motivated to learn, their attention must be captured via feelings of curiosity, they must perceive instruction to be relevant to their current needs and long-term goals, they must feel confident that they can succeed, and they must feel satisfied with the intrinsic and extrinsic consequences of engaging with instruction.	6 (13)	[5,68-72]
SCT[c] [56]	Bandura's SCT emphasizes the primary role of learners' self-efficacy beliefs (ie, that they can execute courses of action needed to attain particular outcomes) and outcome expectancies (ie, that courses of action will lead to particular outcomes) in motivating their learning goal pursuit.	3 (7)	[64,73,74]
CVT[d] [75]	Pekrun's CVT posits that the achievement emotions that learners experience (as well as their self-regulation and learning) are most proximally a function of the subjective control and value beliefs they ascribe to actions and outcomes for an activity. Subjective control beliefs are based on action-control expectations (ie, expectations that actions can be performed) and action-outcome expectations (ie, expectations that particular actions will lead to certain outcomes). Subjective value beliefs are based on the perceived intrinsic and extrinsic value of engaging in the activity and attaining resultant outcomes.	2 (4)	[10,63]
EVT[e] [53]	Eccles and Wigfield's EVT (now called situated expectancy-value theory) posits that learners' motivation is most proximally a function of their expectations of success and the subjective value they ascribe to an activity. Subjective value is composed of interest value (ie, the interest or enjoyment an activity brings), utility value (ie, an activity's usefulness for attaining other valued goals), attainment value (ie, an activity's importance in confirming a salient aspect of one's identity), and cost (ie, the drawbacks of completing an activity).	1 (2)	[76]
Other theories or models	Theory of narrative engagement [77,78]; 4-phase model of interest development [11]; engagement modes model [73]; information and communication acceptance model [79]; social interdependence theory [80]; Guthrie and Wigfield engagement model [81]	N/A[f]	See description
None mentioned	N/A	24 (52)	[82-105]

[a]SDT: self-determination theory.

[b]ARCS: attention, relevance, confidence, and satisfaction.

[c]SCT: social cognitive theory.

[d]CVT: control-value theory.

[e]EVT: expectancy-value theory

[f]N/A: not applicable.

Motivational Constructs (Aligned With Research Question 2)

We used our list of theories and previous research [13] to create a priori categories of motivational constructs to deductively guide our coding. During the coding process, our categorization scheme changed slightly from that documented in our protocol [31], as we determined that a more parsimonious categorization scheme involved aggregating more constructs into fewer categories (Multimedia Appendix 4). Our list included the following categories of motivational constructs: intrinsic value beliefs (eg, interest), extrinsic value beliefs (eg, instrumentality), competence and control beliefs (eg, self-efficacy), social connectedness (eg, relatedness), autonomy, and goals. Intrinsic value refers to the value derived from the experience of completing an activity (eg, interest or enjoyment), whereas extrinsic value refers to the value derived from attaining outcomes external to an activity (eg, progress toward future goals) [53,55].

Study Risk of Bias Assessment

We rated each study's risk of bias across 9 dimensions contained within the Cochrane Collaboration's Effective Practice and Organization of Care risk of bias tool: random sequence generation, allocation concealment, similar baseline outcome measurements, similar baseline characteristics, incomplete

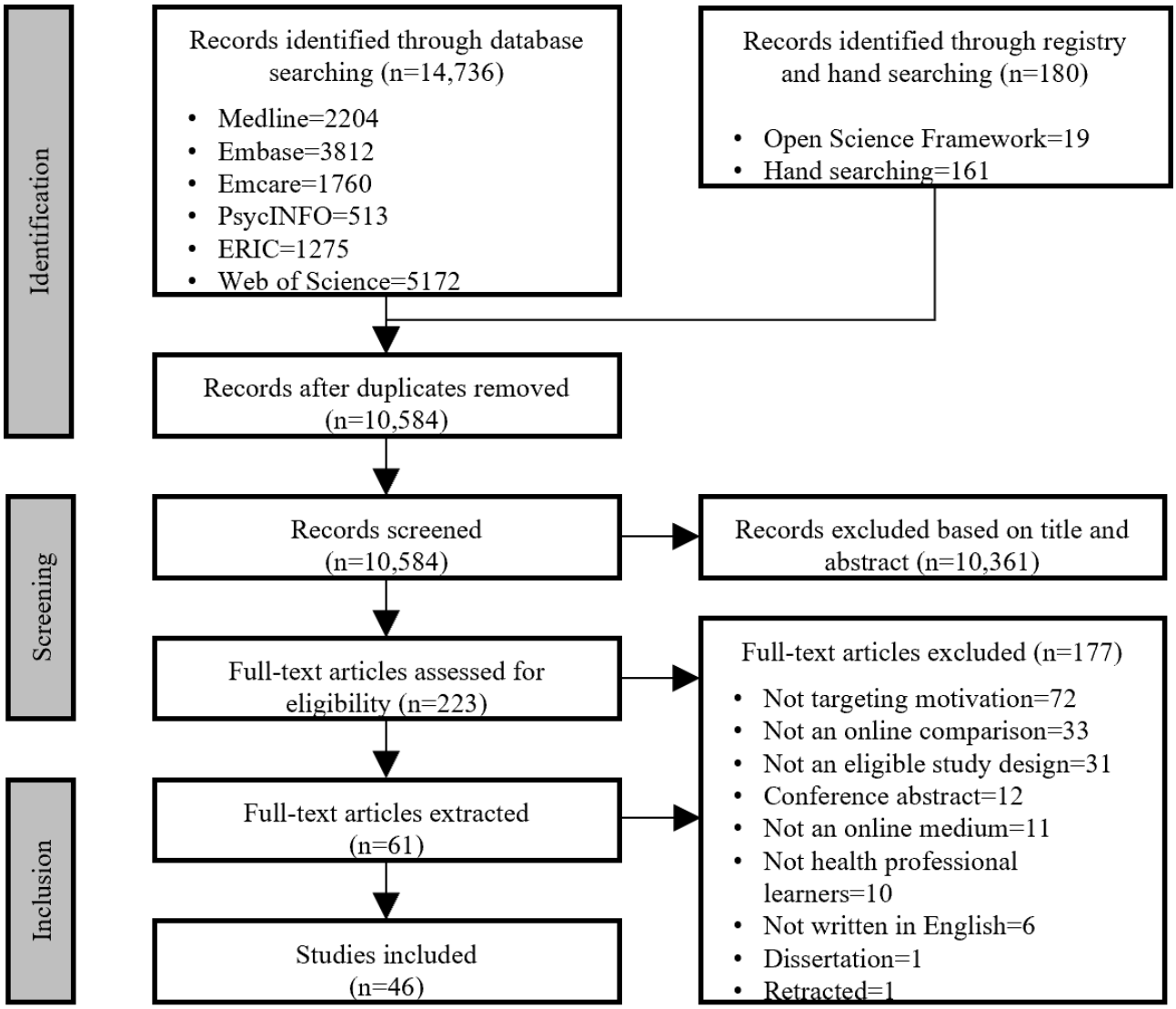
outcome data, blinded outcome measurement, protection against contamination, selective outcome reporting, and other risks of bias [30]. This tool has been used in similar systematic reviews of online instruction in HPE [19,36]. Team members reported particular difficulty in identifying "other risks of bias," and we observed that raters frequently documented different sources of bias (or no bias) within this broad category. Accordingly, we decided to exclude this dimension.

Results

Characteristics of Included Studies

The characteristics of the included studies are presented in Multimedia Appendix 5. Most studies were conducted with trainees (n=40), primarily medical students (n=17) and nursing students (n=11). Study designs were predominantly randomized parallel-group trials (n=27), followed by quasi-experimental trials (n=12), randomized cross-over trials (n=4), and cluster randomized trials (n=3). The risks of bias for each study are presented in Multimedia Appendix 6. Although 74% (34/46) of the included studies were identified as randomized trials, only 30% (14/46) were rated as low risk of bias for random sequence generation, and 33% (15/46) were rated as low risk of bias for allocation concealment. For other dimensions of bias, low risk was observed in 35% (16/46) of studies for baseline outcome measurements, 37% (17/46) for baseline characteristics, 50% (23/46) for blinded outcome measurements, 50% (23/46) for contamination, 57% (26/46) for missing outcome data, and 80% (37/46) for selective outcome reporting. The PRISMA flowchart for our review is presented in Figure 1, and the PRISMA checklist can be found in Multimedia Appendix 7.

Figure 1. PRISMA flowchart. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.



Which Theories or Models of Motivation Inform Existing Experimental Studies of Motivational Design Strategies?

Table 1 presents the number of studies that were informed by a theory of motivation or model of motivational design. SDT and the ARCS model were the most commonly used theories, while 24 studies did not cite any theory. Five studies cited more than 1 theory of motivation. Among the 22 studies that used at least 1 theory, we judged the theory as informing the research questions in 20 (91%) studies, informing the experimental conditions in 15 (68%) studies, informing methods and measures

in 17 (77%) studies, and informing the interpretation of results in 17 (77%) studies. Nine studies used theory to inform all 4 aspects of their research process [5,10,11,62,63,68,69,77,80].

Which Motivational Constructs Have Studies Targeted With Their Motivational Design Strategies?

Studies investigated motivational design strategies that targeted intrinsic value beliefs in 23 of the 46 (50%) studies, extrinsic value beliefs in 9 (20%) studies, competence and control beliefs in 6 (13%) studies, social connectedness in 4 (9%) studies, and autonomy in 2 (4%) studies. Ten (22%) studies targeted more than 1 construct; of these, 5 (11%) were informed by the ARCS model. Sixteen (35%) studies did not report targeting any specific motivational construct, instead aiming to enhance motivation in general.

While intrinsic value beliefs were the most commonly targeted construct, researchers drawing on a prominent theory or model (as listed in Table 1) tended to be more pluralistic in their foci. Specifically, studies that used a motivation theory or model targeted intrinsic value beliefs (n=11) at a similar level to extrinsic value beliefs (n=9) and, to a lesser extent, competence and control beliefs (n=6). By contrast, studies that did not use a theory or model focused solely on intrinsic value beliefs (n=10) compared to extrinsic value beliefs (n=0) and competence and control beliefs (n=0).

Discussion

Key Findings and Implications for Future Research

In this systematic review, we analyzed experimental comparison studies of online motivational design strategies in HPE. We aimed to identify which motivational constructs have been most frequently targeted in these studies and which remain understudied, offering insights into potential areas for future research.

A significant finding was that nearly one-third of the studies in our review did not specify which motivational constructs their design strategy was targeting, instead broadly aiming to enhance motivation. We argue that such research is of limited value to educators. Motivational design expertise relies on educators understanding how strategies work, specifically what constructs they influence and under what conditions they are most effective [106,107]. Studies that do not clarify which constructs a design strategy influences, either conceptually or empirically, cannot provide educators with the information needed to build expertise [16]. Therefore, we recommend that researchers explicitly define the motivational constructs their strategies aim to influence and test their impact on those constructs. This recommendation can be supported through the greater use of motivational theories, which were cited in fewer than half of the studies in our review. This lack of theory use is consistent with other reviews, such as those by Maheu-Cadotte et al [19] and Bajpai et al [29], who found similarly low levels of theory use in their reviews of serious games and digital education in HPE. Motivational theory should be used to inform the research questions, the design strategy, the outcome measures, and the interpretation of results. Excellent examples of theory use are present in our sample [5,11,80].

Among the studies that did specify targeted constructs, most focused on intrinsic value beliefs (eg, interest or enjoyment), compared to extrinsic value beliefs, competence and control beliefs, social connectedness, and autonomy. Accordingly, research in this area is disproportionately focused on ways to make online instruction more interesting and enjoyable. Given the volume of studies on design strategies targeting intrinsic value beliefs, we recommend that future research synthesize existing findings to identify the most effective strategies for enhancing interest and enjoyment and outline areas for future research.

A disproportionate focus on enhancing intrinsic value beliefs aligns with an increased uptake of SDT in HPE, as documented in our studies and other reviews [24,108]. SDT emphasizes the role of intrinsic motivation—which is grounded in feelings of interest and enjoyment—in effective learning [55]. However, we found that studies using SDT were often pluralistic in the constructs they targeted, suggesting a more nuanced approach than studies without a theoretical basis. A theoretical perspective, whether based on SDT or another theory, may help researchers avoid equating motivation solely with enjoyment and interest, thus neglecting other facets of motivation, such as confidence and relatedness, despite evidence suggesting that these constructs may be particularly at risk when learning online [7,8]. Supporting this perspective, we found that studies informed by the ARCS model—which explicitly states the importance of supporting learners' attention, relevance, confidence, and satisfaction—were most likely to report targeting multiple motivational constructs. We recommend that studies test design strategies targeting a broader range of motivational constructs to expand the set of design strategies that educators can choose from (eg, confidence-enhancing strategies or relatedness-enhancing strategies). For example, though serious games are often framed as ways to enhance interest and enjoyment, they may also be configured to support feelings of practical relevance or boost confidence [24]. Researchers could build on the serious games literature by investigating ways to design serious games to support feelings of extrinsic value, confidence, social connectedness, and autonomy.

We encourage researchers to study ways of motivating learners in established online modalities (eg, asynchronous modules or webinars) and by using emerging technologies such as virtual reality and artificial intelligence. For example, artificial intelligence chatbots have the potential to provide personalized coaching and feedback during learning [109,110]. Providing such support and scaffolding instruction in a learner's zone of proximal development may foster a sense of autonomy and confidence. As research on the motivational design of emerging online modalities is still in its infancy, future studies could investigate how to design emerging technology-enabled instruction to optimize learner motivation.

The risk of bias was a concern across many of the included studies. To ensure that future research can make more defendable claims regarding the effects of design strategies, researchers should clearly specify procedures for random sequence generation and allocation concealment, which are often missing from published papers. They should also capture

relevant variables at baseline, blind assessors to condition, and attempt to limit attrition and contamination [27].

Limitations

Several limitations are worth noting. We did not include any synonyms for the word "motivation" (eg, "engagement" or "satisfaction") or motivational constructs (eg, "value," "relevance," or "confidence") in our search terms because we believed these terms would greatly increase the number of nonrelevant studies in our search results. We assumed that studies using synonyms for "motivation" or referencing motivational constructs would also use the word "motivation" and thus would be retrieved in our searches. Consequently, we may have missed some otherwise eligible studies that exclusively referenced concepts that are related to, or treated as synonymous to, motivation (eg, engagement) or motivational constructs (eg, confidence). We also chose to exclude studies written in a language other than English, which may have resulted in missed studies.

We decided to focus our review on experimental studies because they provide a critical source of evidence regarding the effectiveness of design strategies. We acknowledge that many different kinds of studies can generate evidence to support educators' motivational design efforts when producing online learning [31,111]. For example, qualitative studies can help us understand how learners make meaning of instructional designs in context [112], and single-group studies can investigate the factors influencing engagement with motivational design strategies [113]. It may be that studies leveraging nonexperimental designs demonstrate a different distribution of foci regarding motivational constructs. We recommend that a breadth of methodologies, including but not limited to experimental comparison studies, be used to investigate novel motivational design strategies in the future.

Finally, our review focused on online instruction in HPE, and it is unclear whether the trends we observed apply to other types of HPE, such as in-person simulation. While the trend toward enhancing interest and enjoyment may also be present in other HPE contexts—such as through the gamification of in-person instruction [114-116]—we cannot make definitive claims about the generalizability of our results to other types of HPE. Conducting similar reviews in other areas of HPE may be a focus of future research.

Conclusions

A key challenge for educators when teaching online involves keeping learners motivated. To address this challenge, educators need access to motivational design strategies that target a range of motivational constructs. The existing research provides an important starting point, but there is much work to be done. Researchers can use our findings to guide future primary and secondary research that generates a more robust evidence base for educators wishing to motivate their learners.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Database search strategies.
[DOCX File , 53 KB - mededu_v11i1e64179_app1.docx]

Multimedia Appendix 2
Registry search strategy.
[DOCX File , 18 KB - mededu_v11i1e64179_app2.docx]

Multimedia Appendix 3
Data items.
[DOCX File , 13 KB - mededu_v11i1e64179_app3.docx]

Multimedia Appendix 4
Categories of motivational constructs.
[DOCX File , 14 KB - mededu_v11i1e64179_app4.docx]

Multimedia Appendix 5
Characteristics of included studies.
[DOCX File , 41 KB - mededu_v11i1e64179_app5.docx]

Multimedia Appendix 6
Risk of bias ratings for included studies.
[DOCX File , 23 KB - mededu_v11i1e64179_app6.docx]

Multimedia Appendix 7
PRISMA 2020 checklist. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.
[PDF File (Adobe PDF File), 138 KB - mededu_v11i1e64179_app7.pdf]

References

1. Heldt JP, Agrawal A, Loeb R, Richards MC, Castillo EG, DeBonis K. We're not sure we like it but we still want more: trainee and faculty perceptions of remote learning during the COVID-19 pandemic. Acad Psychiatry 2021;45(5):598-602 [FREE Full text] [doi: 10.1007/s40596-021-01403-4] [Medline: 33594628]
2. Cassidy D, Edwards G, Bruen C, Kelly H, Arnett R, Illing J. Are we ever going back? Exploring the views of health professionals on postpandemic continuing professional development modalities. J Contin Educ Health Prof 2023;43(3):172-180 [FREE Full text] [doi: 10.1097/CEH.0000000000000482] [Medline: 36877815]
3. Cook DA. The value of online learning and MRI: finding a niche for expensive technologies. Med Teach 2014;36(11):965-972. [doi: 10.3109/0142159X.2014.917284] [Medline: 25072533]
4. Cook DA, Artino AR. Motivation to learn: an overview of contemporary theories. Med Educ 2016;50(10):997-1014 [FREE Full text] [doi: 10.1111/medu.13074] [Medline: 27628718]
5. Cook DA, Beckman TJ, Thomas KG, Thompson WG. Measuring motivational characteristics of courses: applying Keller's instructional materials motivation survey to a web-based course. Acad Med 2009;84(11):1505-1509. [doi: 10.1097/ACM.0b013e3181baf56d] [Medline: 19858805]
6. Song HS, Kalet AL, Plass JL. Interplay of prior knowledge, self‐regulation and motivation in complex multimedia learning environments. J Comput Assist Learn 2016;32(1):31-50. [doi: 10.1111/jcal.12117]
7. Butz NT, Stupnisky RH. A mixed methods study of graduate students' self-determined motivation in synchronous hybrid learning environments. Internet High Educ 2016;28:85-95. [doi: 10.1016/j.iheduc.2015.10.003]
8. Scheiter K. The learner control principle in multimedia learning. In: Mayer RE, Fiorella L, editors. The Cambridge Handbook of Multimedia Learning. Cambridge, England: Cambridge University Press; 2021:418-429.
9. Keller JM. Motivational design for learning and performance. In: The ARCS Model Approach. Boston, MA: Springer; 2010:1-345.
10. Colonnello V, Mattarozzi K, Agostini A, Russo PM. Emotionally salient patient information enhances the educational value of surgical videos. Adv Health Sci Educ Theory Pract 2020;25(4):799-808. [doi: 10.1007/s10459-020-09957-y] [Medline: 31960188]
11. Dousay TA. Effects of redundancy and modality on the situational interest of adult learners in multimedia learning. Educ Technol Research Dev 2016;64(6):1251-1271. [doi: 10.1007/s11423-016-9456-3]
12. Bland T, Guo M, Dousay TA. Multimedia design for learner interest and achievement: a visual guide to pharmacology. BMC Med Educ 2024;24(1):113 [FREE Full text] [doi: 10.1186/s12909-024-05077-y] [Medline: 38317141]
13. Pintrich PR. A motivational science perspective on the role of student motivation in learning and teaching contexts. J Educ Psychol 2003;95(4):667-686. [doi: 10.1037/0022-0663.95.4.667]
14. Gavarkovs A, Crukley J, Miller E, Kusurkar R, Kulasegaram K, Brydges R. Effectiveness of life goal framing to motivate medical students during online learning: a randomized controlled trial. Perspect Med Educ 2023;12(1):444-454 [FREE Full text] [doi: 10.5334/pme.1017] [Medline: 37901885]
15. Gavarkovs AG, Glista D, O'Hagan R, Moodie S. Applying the purpose, autonomy, confidence, engrossment model of motivational design to support motivation for continuing professional development. J Contin Educ Health Prof 2025. [doi: 10.1097/CEH.0000000000000595] [Medline: 39907433]
16. Gavarkovs AG, Kusurkar RA, Kulasegaram K, Brydges R. Going beyond the comparison: toward experimental instructional design research with impact. Adv Health Sci Educ Theory Pract 2024. [doi: 10.1007/s10459-024-10365-9] [Medline: 39196469]
17. Nagengast B, Marsh HW, Scalas LF, Xu MK, Hau K, Trautwein U. Who took the "x" out of expectancy-value theory? A psychological mystery, a substantive-methodological synergy, and a cross-national generalization. Psychol Sci 2011;22(8):1058-1066. [doi: 10.1177/0956797611415540] [Medline: 21750248]
18. Gentry SV, Gauthier A, L'Estrade Ehrstrom B, Wortley D, Lilienthal A, Tudor Car L, et al. Serious gaming and gamification education in health professions: systematic review. J Med Internet Res 2019;21(3):e12994 [FREE Full text] [doi: 10.2196/12994] [Medline: 30920375]
19. Maheu-Cadotte M, Cossette S, Dubé V, Fontaine G, Lavallée A, Lavoie P, et al. Efficacy of serious games in healthcare professions education: a systematic review and meta-analysis. Simul Healthc 2021;16(3):199-212. [doi: 10.1097/SIH.0000000000000512] [Medline: 33196609]
20. Min A, Min H, Kim S. Effectiveness of serious games in nurse education: a systematic review. Nurse Educ Today 2022;108:105178. [doi: 10.1016/j.nedt.2021.105178] [Medline: 34717098]
21. Silva RDOS, Pereira AM, Araújo DCSAD, Rocha KSS, Serafini MR, de Lyra Jr DP. Effect of digital serious games related to patient care in pharmacy education: a systematic review. Simul Gaming 2021;52(5):554-584. [doi: 10.1177/1046878120988895]

22. Sipiyaruk K, Hatzipanagos S, Reynolds PA, Gallagher JE. Serious games and the COVID-19 pandemic in dental education: an integrative review of the literature. Computers 2021;10(4):42. [doi: 10.3390/computers10040042]
23. Wang R, DeMaria S, Goldberg A, Katz D. A systematic review of serious games in training health care professionals. Simul Healthc 2016;11(1):41-51. [doi: 10.1097/SIH.0000000000000118] [Medline: 26536340]
24. Krath J, Schürmann L, von Korflesch HF. Revealing the theoretical basis of gamification: a systematic review and analysis of theory in research on gamification, serious games and game-based learning. Comput Human Behav 2021;125:106963. [doi: 10.1016/j.chb.2021.106963]
25. Cook DA. The research we still are not doing: an agenda for the study of computer-based learning. Acad Med 2005;80(6):541-548. [doi: 10.1097/00001888-200506000-00005] [Medline: 15917356]
26. Cook DA, Beckman TJ. Reflections on experimental research in medical education. Adv Health Sci Educ 2010;15(3):455-464. [doi: 10.1007/s10459-008-9117-3] [Medline: 18427941]
27. Shadish WR, Cook TD, Campbell TD. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston: Houghton Mifflin; 2001.
28. Mayer RE. How to assess whether an instructional intervention has an effect on learning. Educ Psychol Rev 2023;35(2):64. [doi: 10.1007/s10648-023-09783-9]
29. Bajpai S, Semwal M, Bajpai R, Car J, Ho AHY. Health professions' digital education: review of learning theories in randomized controlled trials by the digital health education collaboration. J Med Internet Res 2019;21(3):e12912 [FREE Full text] [doi: 10.2196/12912] [Medline: 30860483]
30. Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M, et al. Cochrane handbook for systematic reviews of interventions version 6.4. Cochrane. The Cochrane Collaboration. 2023. URL: https://training.cochrane.org/handbook/current [accessed 2025-03-21]
31. Gavarkovs A, Kusurkar RA, Kulasegaram K, Crukley J, Miller E, Anderson M, et al. Motivational design for web-based instruction in health professions education: protocol for a systematic review and directed content analysis. JMIR Res Protoc 2022;11(11):e42681 [FREE Full text] [doi: 10.2196/42681] [Medline: 36350706]
32. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Syst Rev 2021;10(1):89 [FREE Full text] [doi: 10.1186/s13643-021-01626-4] [Medline: 33781348]
33. Car J, Carlstedt-Duke J, Tudor Car L, Posadzki P, Whiting P, Zary N, Digital Health Education Collaboration. Digital education in health professions: the need for overarching evidence synthesis. J Med Internet Res 2019;21(2):e12913 [FREE Full text] [doi: 10.2196/12913] [Medline: 30762583]
34. Arruzza E, Chau M. A scoping review of randomised controlled trials to assess the value of gamification in the higher education of health science students. J Med Imaging Radiat Sci 2021;52(1):137-146. [doi: 10.1016/j.jmir.2020.10.003] [Medline: 33153931]
35. Cook DA, Levinson AJ, Garside S, Dupras DM, Erwin PJ, Montori VM. Instructional design variations in internet-based learning for health professions education: a systematic review and meta-analysis. Acad Med 2010;85(5):909-922. [doi: 10.1097/ACM.0b013e3181d6c319] [Medline: 20520049]
36. Fontaine G, Cossette S, Maheu-Cadotte M, Mailhot T, Deschênes MF, Mathieu-Dupuis G, et al. Efficacy of adaptive e-learning for health professionals and students: a systematic review and meta-analysis. BMJ Open 2019;9(8):e025252 [FREE Full text] [doi: 10.1136/bmjopen-2018-025252] [Medline: 31467045]
37. Szeto MD, Strock D, Anderson J, Sivesind TE, Vorwald VM, Rietcheck HR, et al. Gamification and game-based strategies for dermatology education: narrative review. JMIR Dermatol 2021;4(2):e30325 [FREE Full text] [doi: 10.2196/30325] [Medline: 37632819]
38. Tudor Car L, Kyaw BM, Teo A, Fox TE, Vimalesvaran S, Apfelbacher C, et al. Outcomes, measurement instruments, and their validity evidence in randomized controlled trials on virtual, augmented, and mixed reality in undergraduate medical education: systematic mapping review. JMIR Serious Games 2022;10(2):e29594 [FREE Full text] [doi: 10.2196/29594] [Medline: 35416789]
39. Xu Y, Lau Y, Cheng LJ, Lau ST. Learning experiences of game-based educational intervention in nursing students: a systematic mixed-studies review. Nurse Educ Today 2021;107:105139. [doi: 10.1016/j.nedt.2021.105139] [Medline: 34563963]
40. Becker EA, Godwin EM. Methods to improve teaching interdisciplinary teamwork through computer conferencing. J Allied Health 2005;34(3):169-176. [Medline: 16252680]
41. Crowley RS, Legowski E, Medvedeva O, Tseytlin E, Roh E, Jukic D. Evaluation of an intelligent tutoring system in pathology: effects of external representation on performance gains, metacognition, and acceptance. J Am Med Inform Assoc 2007;14(2):182-190 [FREE Full text] [doi: 10.1197/jamia.M2241] [Medline: 17213494]
42. DeBate RD, Severson HH, Cragun D, Bleck J, Gau J, Merrell L, et al. Randomized trial of two e-learning programs for oral health students on secondary prevention of eating disorders. J Dent Educ 2014;78(1):5-15. [Medline: 24385519]
43. Gauthier A, Corrin M, Jenkinson J. Exploring the influence of game design on learning and voluntary use in an online vascular anatomy study aid. Comput Educ 2015;87:24-34. [doi: 10.1016/j.compedu.2015.03.017]

44. Harned MS, Dimeff LA, Woodcock EA, Kelly T, Zavertnik J, Contreras I, et al. Exposing clinicians to exposure: a randomized controlled dissemination trial of exposure therapy for anxiety disorders. Behav Ther 2014;45(6):731-744 [FREE Full text] [doi: 10.1016/j.beth.2014.04.005] [Medline: 25311284]
45. Hege I, Dietl A, Kiesewetter J, Schelling J, Kiesewetter I. How to tell a patient's story? Influence of the case narrative design on the clinical reasoning process in virtual patients. Med Teach 2018;40(7):736-742 [FREE Full text] [doi: 10.1080/0142159X.2018.1441985] [Medline: 29490538]
46. Hege I, Ropp V, Adler M, Radon K, Mäsch G, Lyon H, et al. Experiences with different integration strategies of case-based e-learning. Med Teach 2007;29(8):791-797. [doi: 10.1080/01421590701589193] [Medline: 18236274]
47. Hendriks WJAJ, Bakker N, Pluk H, de Brouwer A, Wieringa B, Cambi A, et al. Certainty-based marking in a formative assessment improves student course appreciation but not summative examination scores. BMC Med Educ 2019;19(1):178 [FREE Full text] [doi: 10.1186/s12909-019-1610-2] [Medline: 31151456]
48. Kalet AL, Song HS, Sarpel U, Schwartz R, Brenner J, Ark TK, et al. Just enough, but not too much interactivity leads to better clinical skills performance after a computer assisted learning module. Med Teach 2012;34(10):833-839. [doi: 10.3109/0142159X.2012.706727] [Medline: 22917265]
49. Noll C, von Jan U, Raap U, Albrecht U. Mobile augmented reality as a feature for self-oriented, blended learning in medicine: randomized controlled trial. JMIR mHealth uHealth 2017;5(9):e139 [FREE Full text] [doi: 10.2196/mhealth.7943] [Medline: 28912113]
50. Janda MS, Botticelli AT, Mattheos N, Nebel D, Wagner A, Nattestad A, et al. Computer-mediated instructional video: a randomised controlled trial comparing a sequential and a segmented instructional video in surgical hand wash. Eur J Dent Educ 2005;9(2):53-58. [doi: 10.1111/j.1600-0579.2004.00366.x] [Medline: 15811151]
51. Van Es SL, Kumar RK, Pryor WM, Salisbury EL, Velan GM. Cytopathology whole slide images and adaptive tutorials for senior medical students: a randomized crossover trial. Diagn Pathol 2016;11:1 [FREE Full text] [doi: 10.1186/s13000-016-0452-z] [Medline: 26746436]
52. Hsieh H, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
53. Eccles JS, Wigfield A. From expectancy-value theory to situated expectancy-value theory: a developmental, social cognitive, and sociocultural perspective on motivation. Contemp Educ Psychol 2020;61(4):101859. [doi: 10.1016/j.cedpsych.2020.101859]
54. Graham S. An attributional theory of motivation. Contemp Educ Psychol 2020;61:101861. [doi: 10.1016/j.cedpsych.2020.101861]
55. Ryan RM, Deci EL. Intrinsic and extrinsic motivation from a self-determination theory perspective: definitions, theory, practices, and future directions. Contemp Educ Psychol 2020;61:101860. [doi: 10.1016/j.cedpsych.2020.101860]
56. Schunk DH, DiBenedetto MK. Motivation and social cognitive theory. Contemp Educ Psychol 2020;60:101832. [doi: 10.1016/j.cedpsych.2019.101832]
57. Urdan T, Kaplan A. The origins, evolution, and future directions of achievement goal theory. Contemp Educ Psychol 2020;61:101862. [doi: 10.1016/j.cedpsych.2020.101862]
58. Artino AR, Holmboe ES, Durning SJ. Control-value theory: using achievement emotions to improve understanding of motivation, learning, and performance in medical education: AMEE Guide No. 64. Med Teach 2012;34(3):e148-e160. [doi: 10.3109/0142159X.2012.651515] [Medline: 22364472]
59. Ten Cate TJ, Kusurkar RA, Williams GC. How self-determination theory can assist our understanding of the teaching and learning processes in medical education. AMEE guide No. 59. Med Teach 2011;33(12):961-973. [doi: 10.3109/0142159X.2011.595435] [Medline: 22225433]
60. Schunk DH, Pintrich PR, Meece JL. Motivation in Education: Theory, Research, and Applications. 3rd Edition. Upper Saddle River, NJ: Pearson Education Internat; 2010.
61. Cheung JJH, Apramian T, Brydges R. Starting your research project: from problem to theory to question. In: Nestel D, Hui J, Kunkler K, Scerbo M, Calhoun A, editors. Healthcare Simulation Research. Cham, Switzerland: Springer; 2019.
62. Liu C, Lim R, Taylor S, Calvo RA. Students' behavioural engagement in reviewing their tele-consultation feedback within an online clinical communication skills platform. Comput Hum Behav 2019;94:35-44. [doi: 10.1016/j.chb.2019.01.002]
63. Wang M, Wu B, Kirschner PA, Michael Spector J. Using cognitive mapping to foster deeper learning with complex problems in a computer-based environment. Comput Hum Behav 2018;87:450-458. [doi: 10.1016/j.chb.2018.01.024]
64. Zwart DP, Goei SL, Van Luit JEH, Noroozi O. Nursing students' satisfaction with the instructional design of a computer-based virtual learning environment for mathematical medication learning. Interact Learn Environ 2022;31(10):7392-7407. [doi: 10.1080/10494820.2022.2071946]
65. Haftador AM, Shirazi F, Mohebbi Z. Online class or flipped-jigsaw learning? Which one promotes academic motivation during the COVID-19 pandemic? BMC Med Educ 2021;21(1):499 [FREE Full text] [doi: 10.1186/s12909-021-02929-9] [Medline: 34548075]
66. Mahnken AH, Baumann M, Meister M, Schmitt V, Fischer MR. Blended learning in radiology: is self-determined learning really more effective? Eur J Radiol 2011;78(3):384-387. [doi: 10.1016/j.ejrad.2010.12.059] [Medline: 21288674]

67. Rudolphi-Solero T, Lorenzo-Alvarez R, Ruiz-Gomez MJ, Sendra-Portero F. Impact of compulsory participation of medical students in a multiuser online game to learn radiological anatomy and radiological signs within the virtual world Second Life. Anat Sci Educ 2022;15(5):863-876. [doi: 10.1002/ase.2134] [Medline: 34449983]
68. Drees C, Ghebremedhin E, Hansen M. Development of an interactive e-learning software "Histologie für Mediziner" for medical histology courses and its overall impact on learning outcomes and motivation. GMS J Med Educ 2020;37(3):Doc35 [FREE Full text] [doi: 10.3205/zma001328] [Medline: 32566737]
69. Pittenger A, Doering A. Influence of motivational design on completion rates in online self - study pharmacy - content courses. Distance Educ 2010;31(3):275-293. [doi: 10.1080/01587919.2010.513953]
70. EL Machtani EL Idrissi W, Chemsi G, EL Kababi K, Radid M. The impact of serious game on the nursing students' learning, behavioral engagement, and motivation. Int J Emerg Technol Learn 2022;17(01):18-35. [doi: 10.3991/ijet.v17i01.26857]
71. Rondon-Melo S, Andrade CRFD. Educação mediada por tecnologia em Fonoaudiologia: impacto na motivação para aprendizagem sobre o Sistema Miofuncional Orofacial. Codas 2016;28(3):269-277 [FREE Full text] [doi: 10.1590/2317-1782/20162015143] [Medline: 27305632]
72. Su C. The effects of students' learning anxiety and motivation on the learning achievement in the activity theory based gamified learning environment. Eurasia J Math Sci Technol Educ 2016;13(5):1229-1258. [doi: 10.12973/eurasia.2017.00669a]
73. Hedman M, Schlickum M, Felländer-Tsai L. Surgical novices randomized to train in two video games become more motivated during training in MIST-VR and GI Mentor II than students with no video game training. Stud Health Technol Inform 2013;184:189-194. [Medline: 23400154]
74. Maag M. The effectiveness of an interactive multimedia learning tool on nursing students' math knowledge and self-efficacy. Comput Inform Nurs 2004;22(1):26-33. [doi: 10.1097/00024665-200401000-00007] [Medline: 15069846]
75. Pekrun R. The control-value theory of achievement emotions: assumptions, corollaries, and implications for educational research and practice. Educ Psychol Rev 2006;18(4):315-341. [doi: 10.1007/s10648-006-9029-9]
76. Wingo MT, Thomas KG, Thompson WG, Cook DA. Enhancing motivation with the "virtual" supervisory role: a randomized trial. BMC Med Educ 2015;15:76 [FREE Full text] [doi: 10.1186/s12909-015-0348-8] [Medline: 25889758]
77. Mohan D, Fischhoff B, Angus DC, Rosengart MR, Wallace DJ, Yealy DM, et al. Serious games may improve physician heuristics in trauma triage. Proc Natl Acad Sci U S A 2018;115(37):9204-9209 [FREE Full text] [doi: 10.1073/pnas.1805450115] [Medline: 30150397]
78. Mohan D, Farris C, Fischhoff B, Rosengart MR, Angus DC, Yealy DM, et al. Efficacy of educational video game versus traditional educational apps at improving physician decision making in trauma triage: randomized controlled trial. BMJ 2017;359:j5416 [FREE Full text] [doi: 10.1136/bmj.j5416] [Medline: 29233854]
79. Lee MK. Effects of mobile phone-based app learning compared to computer-based web learning on nursing students: pilot randomized controlled trial. Healthc Inform Res 2015;21(2):125-133 [FREE Full text] [doi: 10.4258/hir.2015.21.2.125] [Medline: 25995965]
80. Peterson AT, Roseth CJ. Effects of four CSCL strategies for enhancing online discussion forums: Social interdependence, summarizing, scripts, and synchronicity. Int J Educ Res 2016;76:147-161. [doi: 10.1016/j.ijer.2015.04.009]
81. Seibert D, Guthrie J, Adamo G. Improving learning outcomes: Integration of standardized patients & telemedicine technology. Nurs Educ Perspect 2004;25(5):232-237. [Medline: 15508562]
82. Buijs-Spanjers KR, Hegge HH, Jansen CJ, Hoogendoorn E, de Rooij SE. A web-based serious game on delirium as an educational intervention for medical students: randomized controlled trial. JMIR Serious Games 2018;6(4):e17 [FREE Full text] [doi: 10.2196/games.9886] [Medline: 30368436]
83. Allen EB, Walls RT, Reilly FD. Effects of interactive instructional techniques in a web-based peripheral nervous system component for human anatomy. Med Teach 2008;30(1):40-47. [doi: 10.1080/01421590701753518] [Medline: 18278650]
84. Berndt M, Thomas F, Bauer D, Härtl A, Hege I, Kääb S, et al. The influence of prompts on final year medical students' learning process and achievement in ECG interpretation. GMS J Med Educ 2020;37(1):Doc11 [FREE Full text] [doi: 10.3205/zma001304] [Medline: 32270025]
85. Blackmore C, Tantam D, Van DE. The role of the eTutor - evaluating tutor input in a virtual learning community for psychotherapists and psychologists across Europe. Int J Psychother 2006;10(2):35-46 [FREE Full text]
86. Bock A, Thomas C, Heitzer M, Winnand P, Peters F, Lemos M, et al. Transferring the sandwich principle to instructional videos: is it worth the effort? BMC Med Educ 2021;21(1):525 [FREE Full text] [doi: 10.1186/s12909-021-02967-3] [Medline: 34627213]
87. Booth R, Sinclair B, McMurray J, Strudwick G, Watson G, Ladak H, et al. Evaluating a serious gaming electronic medication administration record system among nursing students: protocol for a pragmatic randomized controlled trial. JMIR Res Protoc 2018;7(5):e138 [FREE Full text] [doi: 10.2196/resprot.9601] [Medline: 29807885]
88. Brull S, Finlayson S, Kostelec T, MacDonald R, Krenzischeck D. Using gamification to improve productivity and increase knowledge retention during orientation. J Nurs Adm 2017;47(9):448-453. [doi: 10.1097/NNA.0000000000000512] [Medline: 28834805]
89. Buijs-Spanjers KR, Hegge HHM, Cnossen F, Hoogendoorn E, Jaarsma DADC, de Rooij SE. Dark play of serious games: effectiveness and features (G4HE2018). Games Health J 2019;8(4):301-306. [doi: 10.1089/g4h.2018.0126] [Medline: 30964340]

90. Cao R, Sunaga M, Miyoshi T, Kinoshita A. Development and evaluation of a study level announcement system in e-learning. J Med Dent Sci 2018;65(3):113-122. [doi: 10.11480/jmds.650301]
91. Dankbaar MEW, Alsma J, Jansen EEH, van Merrienboer JJG, van Saase JLCM, Schuit SCE. An experimental study on the effects of a simulation game on students' clinical cognitive skills and motivation. Adv Health Sci Educ 2016;21(3):505-521 [FREE Full text] [doi: 10.1007/s10459-015-9641-x] [Medline: 26433730]
92. Dankbaar MEW, Richters O, Kalkman CJ, Prins G, Ten Cate OTJ, van Merrienboer JJG, et al. Comparative effectiveness of a serious game and an e-module to support patient safety knowledge and awareness. BMC Med Educ 2017;17(1):30. [doi: 10.1186/s12909-016-0836-5] [Medline: 28148296]
93. Frith KH, Kee CC. The effect of communication on nursing student outcomes in a web-based course. J Nurs Educ 2003;42(8):350-358. [doi: 10.3928/0148-4834-20030801-06] [Medline: 12938897]
94. Goldingay S, Land C. Emotion: the 'e' in engagement in online distance education in social work. JOFDL 2014;18(1):58-72. [doi: 10.61468/jofdl.v18i1.226]
95. Hwang GJ, Chang CY. Facilitating decision-making performances in nursing treatments: a contextual digital game-based flipped learning approach. Interact Learn Environ 2023;31(1):156-171. [doi: 10.1080/10494820.2020.1765391]
96. Inangil D, Dincer B, Kabuk A. Effectiveness of the use of animation and gamification in online distance education during pandemic. Comput Inform Nurs 2022;40(5):335-340 [FREE Full text] [doi: 10.1097/CIN.0000000000000902] [Medline: 35266898]
97. Jones EP, Wahlquist AE, Hortman M, Wisniewski CS. Motivating students to engage in preparation for flipped classrooms by using embedded quizzes in pre-class videos. Innov Pharm 2021;12(1):6 [FREE Full text] [doi: 10.24926/iip.v12i1.3353] [Medline: 34007679]
98. Karaksha A, Grant G, Anoopkumar-Dukie S, Nirthanan SN, Davey AK. Student engagement in pharmacology courses using online learning tools. Am J Pharm Educ 2013;77(6):125 [FREE Full text] [doi: 10.5688/ajpe776125] [Medline: 23966728]
99. Koop CFA, Marschollek M, Schmiedl A, Proskynitopoulos PJ, Behrends M. Does an audiovisual dissection manual improve medical students' learning in the gross anatomy dissection course? Anat Sci Educ 2021;14(5):615-628. [doi: 10.1002/ase.2012] [Medline: 33460300]
100. Metz CJ, Metz MJ. The benefits of incorporating active learning into online, asynchronous coursework in dental physiology. Adv Physiol Educ 2022;46(1):11-20 [FREE Full text] [doi: 10.1152/advan.00110.2021] [Medline: 34709946]
101. Pereira AC, Dias da Silva MA, Patel US, Tanday A, Hill KB, Walmsley AD. Using quizzes to provide an effective and more enjoyable dental education: a pilot study. Eur J Dent Educ 2022;26(2):404-408. [doi: 10.1111/eje.12716] [Medline: 34510674]
102. Rajan KK, Pandit AS. Comparing computer-assisted learning activities for learning clinical neuroscience: a randomized control trial. BMC Med Educ 2022;22(1):522 [FREE Full text] [doi: 10.1186/s12909-022-03578-2] [Medline: 35780115]
103. Scales CD, Moin T, Fink A, Berry SH, Afsar-Manesh N, Mangione CM, et al. A randomized, controlled trial of team-based competition to increase learner participation in quality-improvement education. Int J Qual Health Care 2016;28(2):227-232 [FREE Full text] [doi: 10.1093/intqhc/mzw008] [Medline: 26857941]
104. Sward KA, Richardson S, Kendrick J, Maloney C. Use of a web-based game to teach pediatric content to medical students. Ambul Pediatr 2008;8(6):354-359. [doi: 10.1016/j.ambp.2008.07.007] [Medline: 19084784]
105. Woelber JP, Hilbert TS, Ratka-Krüger P. Can easy-to-use software deliver effective e-learning in dental education? A randomised controlled study. Eur J Dent Educ 2012;16(3):187-192. [doi: 10.1111/j.1600-0579.2012.00741.x] [Medline: 22783845]
106. Hardré P, Ge X, Thomas M. Toward a model of development for instructional design expertise. Educ Technol 2005;45(1):53-57.
107. Ertmer PA, Stepich DA, York CS, Stickman A, Wu X, Zurek S, et al. How instructional design experts use knowledge and experience to solve ill-structured problems. Perform Improv Q 2008;21(1):17-42. [doi: 10.1002/piq.20013]
108. Kusurkar RA. Self-determination theory in health professions education research and practice. In: Ryan RM, editor. The Handbook of Self-Determination Theory. New York, NY: Oxford University Press; 2023:665-683.
109. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ 2023;9:e46885 [FREE Full text] [doi: 10.2196/46885] [Medline: 36863937]
110. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
111. Cook DA. The failure of e-learning research to inform educational practice, and what we can do about it. Med Teach 2009;31(2):158-162. [doi: 10.1080/01421590802691393] [Medline: 19330674]
112. Ellaway RH, Pusic M, Yavner S, Kalet AL. Context matters: emergent variability in an effectiveness trial of online teaching modules. Med Educ 2014;48(4):386-396. [doi: 10.1111/medu.12389] [Medline: 24606622]

113. Brisson BM, Hulleman CS, Häfner I, Gaspard H, Flunger B, Dicke A, et al. Who sticks to the instructions—and does it matter? Antecedents and effects of students' responsiveness to a classroom-based motivation intervention. Z Erziehungswissenschaft 2020;23(1):121-144. [doi: 10.1007/s11618-019-00922-z]
114. Rutledge C, Walsh CM, Swinger N, Auerbach M, Castro D, Dewan M, et al. Gamification in action: theoretical and practical considerations for medical educators. Acad Med 2018;93(7):1014-1020. [doi: 10.1097/ACM.0000000000002183] [Medline: 29465450]
115. Singhal S, Hough J, Cripps D. Twelve tips for incorporating gamification into medical education. MedEdPublish (2016) 2019;8:216 [FREE Full text] [doi: 10.15694/mep.2019.000216.1] [Medline: 38089323]
116. Davis K, Lo H, Lichliter R, Wallin K, Elegores G, Jacobson S, et al. Twelve tips for creating an escape room activity for medical education. Med Teach 2022;44(4):366-371. [doi: 10.1080/0142159X.2021.1909715] [Medline: 33872114]

Abbreviations

HPE: health professions education
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by B Lesselroth; submitted 10.07.24; peer-reviewed by S Papadakis, C Lo; comments to author 24.11.24; revised version received 28.02.25; accepted 06.03.25; published 11.04.25.

Please cite as:
Gavarkovs A, Miller E, Coleman J, Gunasegaran T, Kusurkar RA, Kulasegaram K, Anderson M, Brydges R
Motivation Theories and Constructs in Experimental Studies of Online Instruction: Systematic Review and Directed Content Analysis
JMIR Med Educ 2025;11:e64179
URL: https://mededu.jmir.org/2025/1/e64179
doi:10.2196/64179
PMID:

Review

Online-Based and Technology-Assisted Psychiatric Education for Trainees: Scoping Review

Mohd Amiruddin Mohd Kassim[1,2,3*], MD; Sidi Muhammad Yusoff Azli Shah[2*], MB Bch BAO; Jane Tze Yn Lim[1,2*], MD, DrPsych; Tuti Iryani Mohd Daud[1,2*], MBBS, DrPsych

[1]Department of Psychiatry, Faculty of Medicine, Universiti Kebangsaan Malaysia, Bandar Tun Razak, Kuala Lumpur, Malaysia

[2]Department of Psychiatry, Hospital Canselor Tuanku Muhriz, Bandar Tun Razak, Kuala Lumpur, Malaysia

[3]Department of Psychiatry and Psychological Health, Faculty of Medicine and Health Sciences, Universiti Malaysia Sabah, Kota Kinabalu, Sabah, Malaysia

[*]all authors contributed equally

Corresponding Author:
Tuti Iryani Mohd Daud, MBBS, DrPsych
Department of Psychiatry
Faculty of Medicine
Universiti Kebangsaan Malaysia
Jalan Yaacob Latif
Bandar Tun Razak, Kuala Lumpur, 56000
Malaysia
Phone: 60 3 9145 6143
Email: tutimd@hctm.ukm.edu.my

Abstract

Background: The concept of online learning in medical education has been gaining traction, but whether it can accommodate the complexity of higher-level psychiatric training remains uncertain.

Objective: This review aims to identify the various online-based and technology-assisted educational methods used in psychiatric training and to examine the outcomes in terms of trainees' knowledge, skills, and levels of confidence or preference in using such technologies.

Methods: A comprehensive search was conducted in PubMed, Cochrane, PsycINFO, Scopus, and ERIC to identify relevant literature from 1991 until 2024. Studies in English and those that had English translations were identified. Studies that incorporated or explored the use of online-based or technology-assisted learning as part of psychiatric training in trainees and had outcomes of interest related to changes in the level of knowledge or skills, changes in the level of preference or confidence in using online-based or technology-assisted learning, and feedback of participants were included. Studies were excluded if they were conducted on populations excluding psychiatric trainees or residents, were mainly descriptive of the concept of the intervention without any relevant study outcome, were not in English or did not have English translations, or were review articles.

Results: A total of 82 articles were included in the review. The articles were divided into 3 phases: prior to 2015, 2015 to 2019 (prepandemic), and 2020 onward (postpandemic). Articles mainly originated from Western countries, and there was a significant increase in relevant studies after the pandemic. There were 5 methods identified, namely videoconference, online modules/e-learning, virtual patients, software/applications, and social media. These were applied in various aspects of psychiatric education, such as theory knowledge, skills training, psychotherapy supervision, and information retrieval.

Conclusions: Videoconference-based learning was the most widely implemented approach, followed by online modules and virtual patients. Despite the outcome heterogeneity and small sample sizes in the included studies, the application of such approaches may have utility in terms of knowledge and skills attainment and could be beneficial for the training of future psychiatrists, especially those in underserved low- and middle-income countries.

(JMIR Med Educ 2025;11:e64773) doi:10.2196/64773

KEYWORDS

online learning; telepsychiatry; remote learning; virtual; training; education; psychiatry; trainees; residents

Introduction

The incorporation of online-based or technology-assisted methods in medical education is not new. Virtual grand rounds, web-based learning, online journal clubs, and virtual clinical cases and labs are among the many examples of their ubiquitous implementation [1]. The mass adoption of technology-based education is attributed to its numerous perceived advantages, including the ability to transcend geographical boundaries, the presence of learner-centered approaches, the development of students' self-directed learning skills, and the asynchronous interaction between teachers and students. As education is getting more globalized due to increasing connectivity, these benefits are being increasingly valued [2].

However, questions remain about whether the many advantages of online education are as intuitively apparent and relatable in the field of psychiatry. Traditionally considered as a face-to-face medical discipline, concerns arise regarding the unique interpersonal nature of psychiatry, with its emphasis on empathetic responsiveness toward patients. These concerns are particularly relevant when considering a virtual or simulated patient, and this represents one of the frontier aspects of online education [3]. It is an undeniable fact that appreciating cues from patients is an experiential aspect of knowledge, which is often deemed irreplicable in online sessions.

As technology-based education is increasingly recognized as being noninferior to physical education in undergraduate studies [4], it is imperative to investigate its application in training postgraduate students. The outcome is significant, as it pertains to the production of future psychiatric specialists. This inquiry is especially relevant today, given the radical and drastic transition to technology-based education at all levels due to the recent COVID-19 pandemic [5]. Most medical fraternities are able to integrate online-based and technology-assisted components in their syllabi to enhance the training of trainees or residents without much difficulty. However, acknowledging the distinct nature of psychiatry, which often can be rather ambiguous and subject to nuance, it is important to evaluate the suitability of such an approach to augment the training of future psychiatrists.

Given these considerations, the overarching goal of this study is to systematically map and summarize the existing literature on online-based and technology-assisted psychiatric education for trainees. Specifically, this review aims to identify the various online-based and technology-assisted educational methods used in psychiatric training and to examine the outcome of the aforementioned technologies in terms of trainees' knowledge, skills, or levels of confidence or preference in using such technologies. We hypothesized that online-based and technology-assisted education can be integrated into psychiatric training to improve trainees' knowledge, skills, and competency levels.

This review is aimed at psychiatric educators and training directors looking for ways to incorporate technology into their programs, psychiatric trainees who want to understand how online learning fits into their training, and policymakers or accreditation bodies shaping the future of psychiatric education. It is also relevant for researchers and academics interested in digital learning and medical education trends.

Methods

Search Strategy

The scoping review was conducted according to the recent methodological framework by Westphaln et al [6], which was derived from the earlier work of Arksey and O'Malley [7]. Five databases (PubMed, PsycINFO, Cochrane, Scopus, and ERIC) were searched from March until June 2024. The keywords applied in PubMed were as follows: ((“resident*”[Title/Abstract] OR “trainee*”[Title/Abstract] OR “postgrad*”[Title/Abstract] OR “graduate*”[Title/Abstract]) AND (“psychiatr*”[Title/Abstract] OR “psychologic* medicine”[Title/Abstract]) AND (“education”[Title/Abstract] OR “training”[Title/Abstract] OR “development”[Title/Abstract] OR “learning”[Title/Abstract] OR “teaching”[Title/Abstract] OR “internship”[Title/Abstract] OR “traineeship”[Title/Abstract] OR “residency”[Title/Abstract] OR “course”[Title/Abstract] OR “lesson”[Title/Abstract] OR “program”[Title/Abstract] OR “programme”[Title/Abstract] OR “class”[Title/Abstract] OR “workshop”[Title/Abstract] OR “module”[Title/Abstract] OR “mooc”[Title/Abstract] OR “academic”[Title/Abstract] OR “clerkship”[Title/Abstract] OR “curriculum”[Title/Abstract]) AND (“on-line”[Title/Abstract] OR “online”[Title/Abstract] OR “digital”[Title/Abstract] OR “virtual”[Title/Abstract] OR “internet-based”[Title/Abstract] OR “internet based”[Title/Abstract] OR “web-based”[Title/Abstract] OR “web based”[Title/Abstract] OR “telepsychiatry”[Title/Abstract] OR “tele-psychiatry”[Title/Abstract] OR “cyber”[Title/Abstract] OR “electronic”[Title/Abstract] OR “e-learning”[Title/Abstract] OR “tele-education”[Title/Abstract] OR “videoconferencing”[Title/Abstract] OR “elearning”[Title/Abstract] OR “distance”[Title/Abstract])).

Different search configurations were used for the databases, and the search strategies are presented in Multimedia Appendix 1.

To ensure completeness, the authors also conducted backward citation searches from key articles and performed searches in Google Scholar to look for grey literature, such as conference proceedings and theses, relevant to the topic. Google Scholar was adopted as it has extensive coverage of academic work and is one of the commonly used search engines for grey literature [8].

Inclusion and Exclusion Criteria

Acknowledging the emergence of the field and the relatively limited number of studies, the authors made a conscious decision to include a variety of publication types in this review, including original articles, empirical and brief reports, case reports, and short communications. Studies were included in the review if they met the following criteria: (1) incorporated or explored the use of online-based or technology-assisted learning as part of psychiatric training or education; (2) were conducted in populations that included psychiatric trainees or residents; (3) had an outcome of interest related to changes in the level of knowledge or skills or the level of preference or confidence in

using online-based or technology-assisted learning, or included feedback of participants on the aforementioned approach (regardless of qualitative or quantitative results); and (4) were written in English or had English translations.

Studies were excluded if they (1) were conducted on populations excluding psychiatric trainees or residents; (2) were mainly descriptive of the concept of the intervention without an evaluation or any relevant study outcome related to the application of online-based or technology-assisted psychiatric education; or (3) were review articles. In addition, studies that only used online questionnaires to conduct pre-post assessments for psychiatric education, which were otherwise not delivered through an online or technology-assisted platform, and studies that primarily assessed the learning needs in online-based or technology-assisted psychiatric education without an evaluation of the intervention itself were also excluded. Studies conducted in languages other than English and those without an English translation were omitted due to limitations in language proficiency and to prevent inaccuracies or misinterpretation of the study findings.

Data Screening and Extraction

The search results from the 5 databases were exported to Rayyan online reference manager. Duplicates of similar articles detected by Rayyan were screened manually by MAMK and SMYAS to minimize errors in excluding articles. Prior to the title and abstract screening process, both MAMK and SMYAS underwent screening training to promote standardization and to identify possible conflicts. Then, according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) approach, the titles and abstracts were screened by MAMK and SMYAS to assess suitability for further examination based on the following predetermined criteria:

- Did the study use online-based or technology-assisted instruments as part of the education approach?
- Was the study focused on psychiatric education or related necessary skills?
- Was the study conducted in samples that included psychiatric trainees or residents?

In this phase, articles were divided into the following categories: accept, maybe, and exclude. Subsequently, the full texts of eligible articles (accept and maybe categories) were retrieved. This was followed by a blinded screening phase during which MAMK and SMYAS independently examined the articles in accordance with the inclusion and exclusion criteria. Any disputes regarding the acceptability of the articles in the title and abstract screening phase and in the full-text screening phase were resolved by an impartial third referee (JTYL or TIMD). Data extraction from all included studies was then conducted, gathering parameters such as author names, study year and country, aims and objectives, interventions applied, and key outcomes of interest. The data were extracted by MAMK and SMYAS to Excel (Microsoft Corp) sheets with predefined data fields.

The quality of the studies was assessed according to the 10-item Medical Education Research Study Quality Instrument (MERSQI) for quantitative studies [9] and the Standards for Reporting Qualitative Research (SRQR) for qualitative studies [10]. Previously, a scoping review adopted the SRQR as a 21-score checklist to assess the quality of included qualitative studies within the review [11].

Ethical Considerations

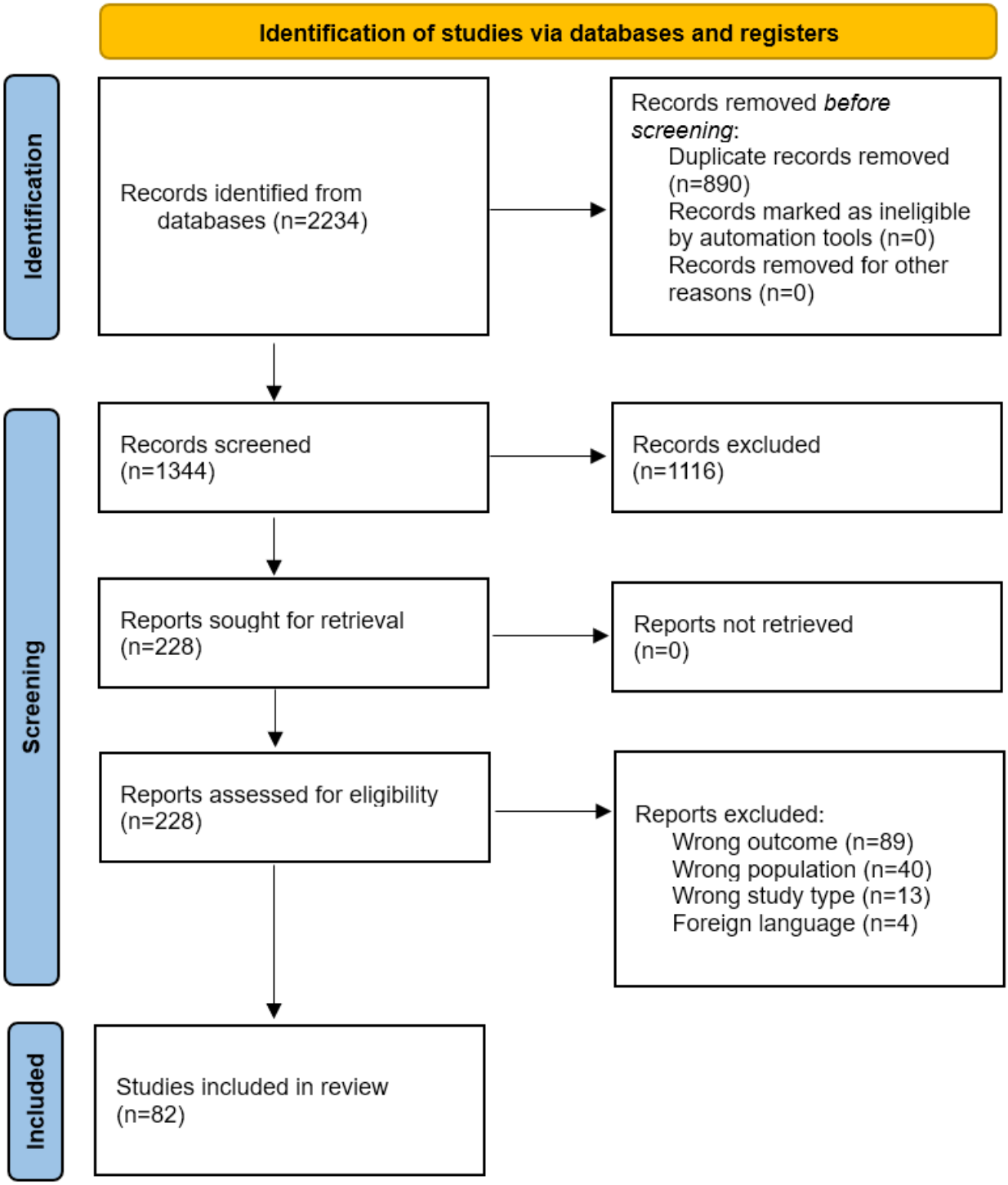
This review received ethical approval from the Research Ethics Committee of the National University of Malaysia (JEP-2023-789).

Results

Overview of the Included Studies

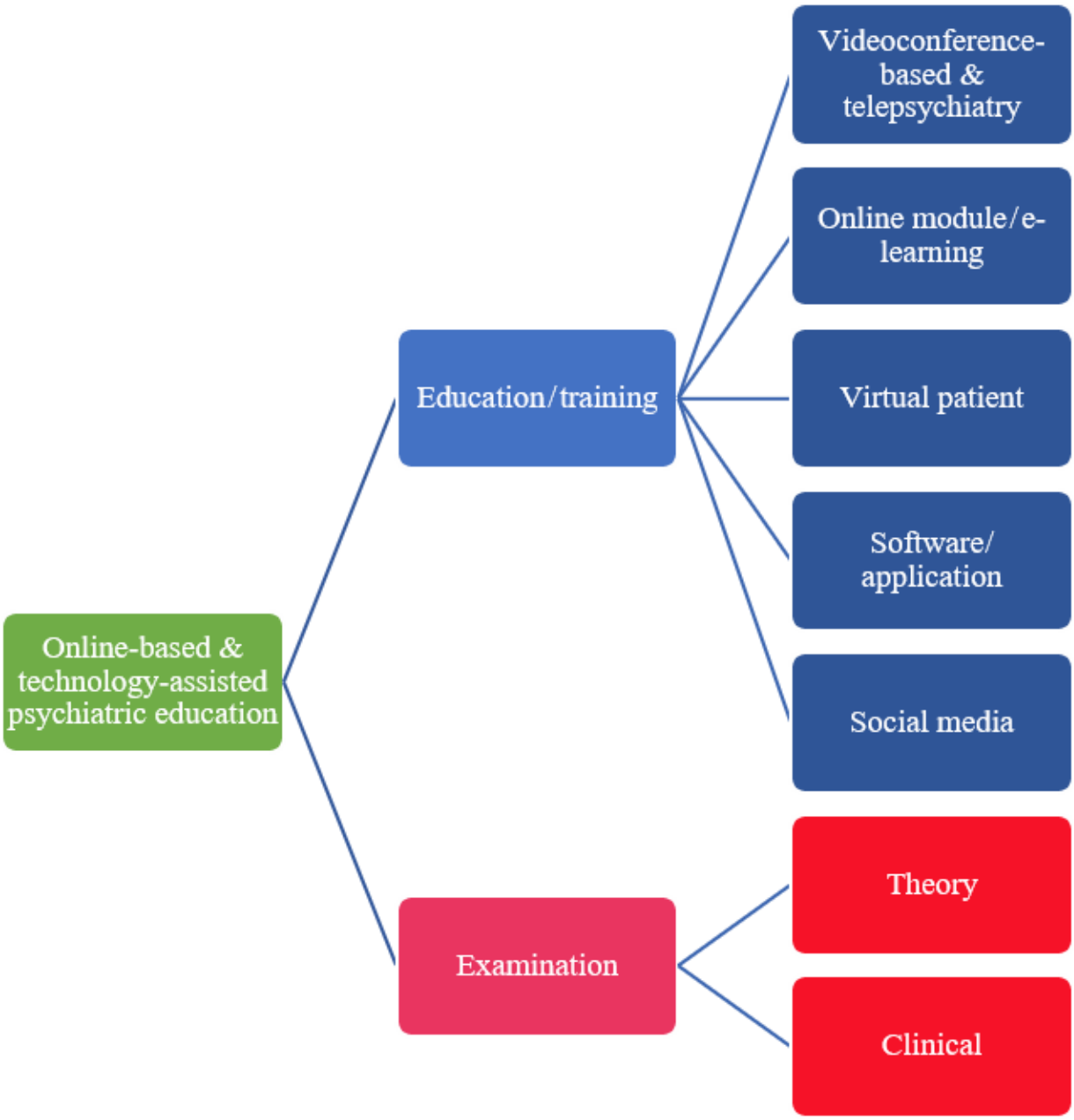
The initial search across the 5 databases yielded a total of 2234 articles, from which 890 duplicates were subsequently removed. After screening the titles and abstracts, 1344 articles were excluded and 228 articles proceeded to full-text screening. Of these, 89 articles were excluded for wrong outcomes (not evaluating the outcome of interest), 40 articles for wrong population (samples excluded psychiatric trainees or residents), 13 articles for wrong study type (review study design), and 4 articles for being in a foreign language. Thus, 82 articles were included in this review (Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.



Among the 82 articles included, 2 themes were identified: education and assessment (Figure 2). Under education, the trend could be divided into 5 subthemes, namely online software (or e-learning platform or massive open online course [MOOC]), videoconference (or telepsychiatry), virtual patient (or simulation), software/application, and social media. Under assessment, there were 2 subthemes, namely theory and clinical examination.

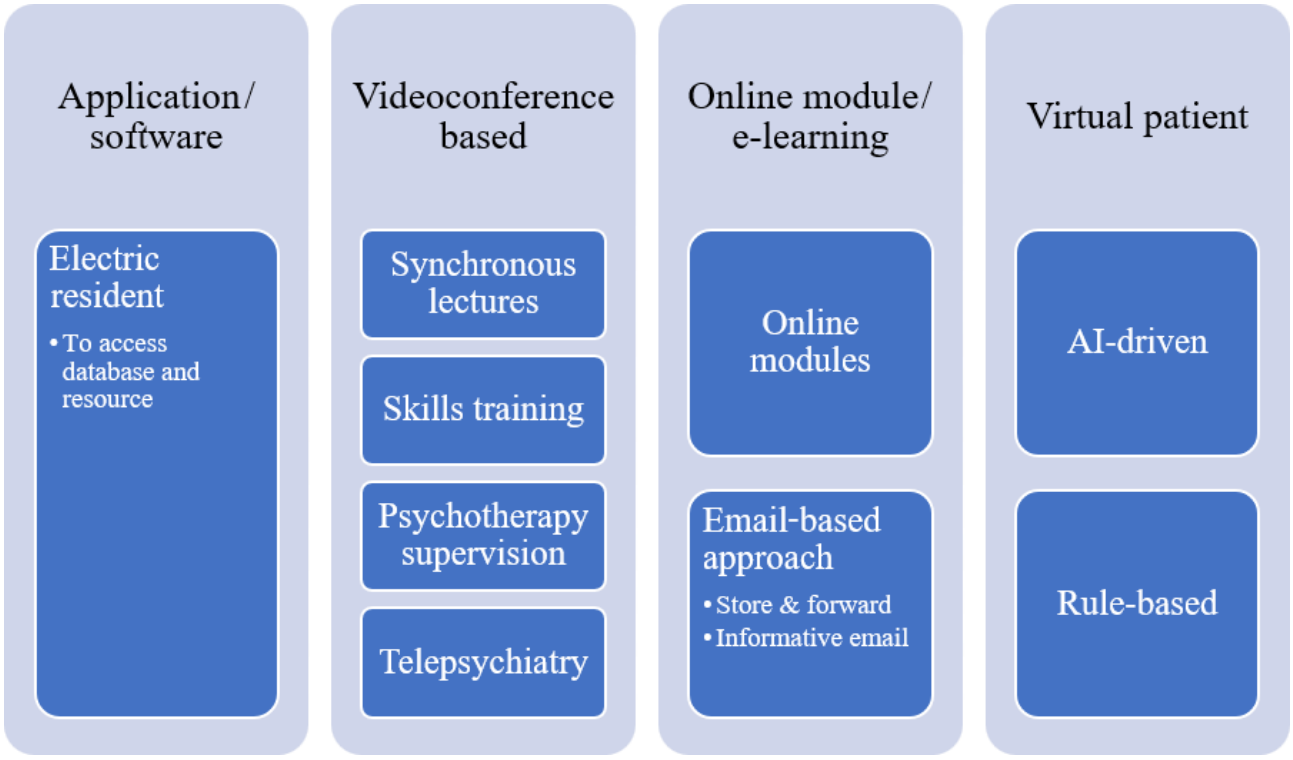
Figure 2. Trend of the implementation of online-based and technology-assisted psychiatric education.



Studies Prior to 2015

A total of 20 articles published before 2015 were included (Table S1 in Multimedia Appendix 2). Specifically, 2 articles were from the 1990s, 9 articles were from the 2000s, and 9 articles were from 2011 to 2014. These articles were mainly from Western countries, such as the United States and Canada. The articles primarily involved telepsychiatry and videoconferencing as tools for education, training, and supervision (n= 10/20, 50%), followed by web-based approaches or e-learning (n= 5/20, 25%), virtual patients (n= 3/20, 15%), and software (n=1/20, 5%). The taxonomy of the methods applied in psychiatric education during this period is presented in Figure 3. Studies mainly had a 1-group, posttest-only study design; 1-group, pretest-posttest study design; and randomized controlled trial design. Out of the 20 articles, 8 had objective measures, such as the change in the level of knowledge before and after the intervention, while the other 12 had subjective measures, such as the level of satisfaction and participant feedback.

Figure 3. Taxonomy of online-based and technology-assisted learning prior to 2015. AI: artificial intelligence.



Despite the limitation of the internet in earlier years, there have been some attempts to integrate IT as part of training. Within this period, depending on financial capabilities and regions, few types of internet connections were available, for example, dial-up connection (especially in the 1990s), Integrated Service Digital Network (ISDN), and broadband connection. The type of internet connection has influenced the strategy and the experience in the education process. With slower internet, an indirect instructional model or independent study was used, with trainees searching for relevant information to enhance their knowledge in particular topics or cases that they had been consulted on. One of the earliest studies highlighted the use of computers and the internet for accessing MEDLINE to assist residents with their consultations, checking drug interactions, and reviewing literature pertinent to cases that they were consulted on [12]. In a survey, printed materials were preferred when learning something new, but digital media or online resources were preferred when revising or searching for resources during patient care [13].

As seen in other fields of medicine, the advancement of relevant infrastructure enabled faster internet connectivity, which allowed widespread and accessible knowledge sharing through videoconference-based seminars, and this is also applicable in the field of psychiatry. This has allowed direct instructional models through approaches such as online lectures and seminars. However, these approaches were met with mixed feedback. One study in the United States in 2004 highlighted that satisfaction with videoconference-based lecturers was contingent on the internet speed, with trainees or residents in centers having a slower internet speed reporting less satisfaction with the lecture experience and the overall sound quality [14]. Another study in Australia in 2008 reported higher preference among participants to attend seminars from remote sites, and most participants felt that the videoconference-based seminars were beneficial for their practice [15]. Meanwhile, in a study in South Africa by Chipps et al [16], while videoconferencing was perceived as an excellent education tool by half of the psychiatric registrars, only 39% of them felt that it was as effective as face-to-face teaching. This led to decreased interest in further videoconference-based training. Additionally, another randomized controlled trial in Iran in 2014, which aimed to compare the effectiveness of face-to-face communication skills training sessions against distant learning in improving empathy, found that the level of empathy was significantly increased in the attending group but not in the distant learning group [17].

One study in Norway in 1998 explored the use of videoconferencing technology in terms of psychotherapy supervision [18]. The psychiatric trainees conducted face-to-face psychotherapy sessions with their patients and later had alternating face-to-face and online psychotherapy supervisions with their supervisors. Through semistructured interviews after the completion of the psychotherapy session, it was noted that while the reduced nonverbal cues were an issue, the limitations of the videoconferencing supervision paradoxically had some positive effects among the trainees in terms of the supervision process, such as verbalization and structure. The positive effects were also contributed by the ease of logistics and by having a neutral space separate from the supervisor's office.

As an extension to videoconference use, telepsychiatry serves as a valuable tool to expand the reach of psychiatric services. As such, it has been incorporated as part of training for psychiatric trainees or residents. In terms of supervision, most of the studies included direct, side-by-side supervision by attendings for assessing patients [19-22]. On the other hand, 1 study adopted a different approach, with attending psychiatrists sitting in with residents during their first session to help

familiarize them with conducting treatment via telepsychiatry, and in later sessions, the involvement of supervising attendings was on an "as needed basis" [23]. Most feedback by trainees on telepsychiatry programs indicated that telepsychiatry enhanced their skills and knowledge, with majority of trainees stating that it was interesting and enhanced their training. However, some trainees mentioned technical issues with this approach and the difficulty in assessing the influence on patients.

On the other hand, improved access to the internet has expanded the utility of asynchronous learning methods. In earlier years, a web-based email approach was applied to promote exposure or learning about stigma education [24] and child and psychiatry cases [25] through approaches such as the "store and forward" concept. However, in Western countries where technology was more advanced compared to the rest of the world at that point of time, e-learning materials were typically in the form of slides of didactic content with recorded audios and videos. This delivery method was used in learning evidence-based medicine [26] and to improve electrocardiogram reading skills [27]. A study by Garside et al [28] in 2009 managed to introduce direct and interactive instruction strategies to learn about how to fill Form 1 of the Mental Health Act. This was achieved by integrating slides of relevant materials regarding Form 1 and the laws related to it, together with interactive Flash animations and practice cases, using questions, and there was immediate expert feedback for each question. Throughout these studies, there were statistically significant improvements in the levels of knowledge and skills of the trainees, suggesting the potential of such an approach to augment the training of psychiatric trainees.

Interest in a virtual patient as an education tool for psychiatric training emerged in the 2000s and 2010s, and facilitated more immersive learning. Within this period, 2 types of virtual patients were studied: artificial intelligence (AI)-driven virtual patients and rule-based virtual patients. Kenny et al [29] and Pataki et al [30] described the use of an AI-driven virtual patient to simulate an adolescent patient with posttraumatic stress disorder (PTSD) ("Justina"). The virtual patient was developed according to the criteria of PTSD based on the Diagnostic and Statistical Manual of Mental Disorders (DSM) [31] and involved technologies such as voice recognition, response selection, behavior generation, and a visual graphics engine. "Justina" received good feedback from residents who mentioned that the experience they had from assessing the virtual patient closely matched their actual experience, but there were times when the virtual patient was not able to understand the questions from the residents. In another study, the rule-based virtual patient concept was applied to assess the doctor's competence in obtaining informed consent before prescribing antipsychotics in a simulated patient with psychosis [32]. A Flash-based video was shown to introduce the clinical scenario, followed by a series of menu options from which they could choose their next action. After completion of the scenario, the program provided a feedback screen regarding the appropriateness of their actions, as well as a link to relevant resources to improve their knowledge. There were statistically significant improvements in all items of the Confidence Scale (pretest to posttest).

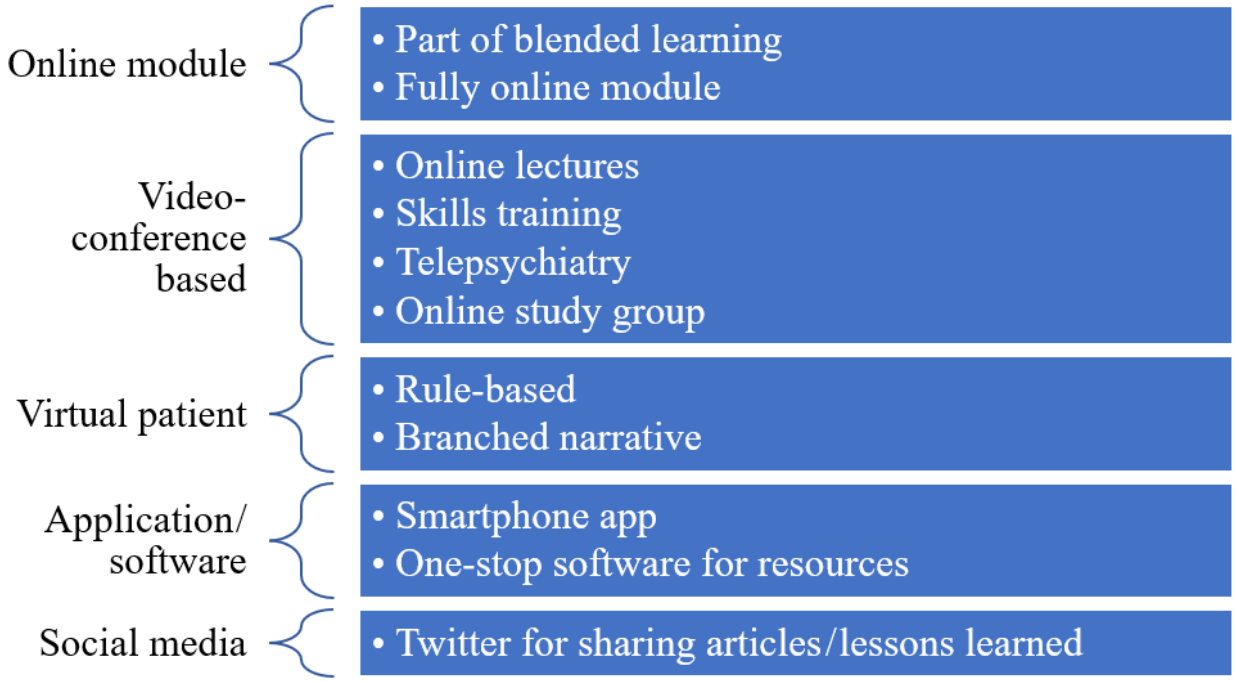
Studies From 2015 to 2019 (Before the COVID-19 Pandemic)

During the period from 2015 to 2019, there was an acceleration of internet-based or technology-assisted education in psychiatry, with 21 studies identified over 5 years (Table S2 in Multimedia Appendix 2). Western countries, particularly the United States and Canada, were at the forefront of these initiatives. Majority of the studies involved videoconference-based learning (n= 7/21, 33.3%), followed by use of online modules or e-learning platforms (n= 5/21, 23.8%), development of software or applications (n= 3/21, 14.3%), adoption of virtual patient approaches (n= 3/21, 14.3%), and use of social media as a learning tool (n= 1/21, 4.8%) (Figure 4). The most common study design was a 1-group, pretest-posttest design, followed by a 1-group, posttest-only design. There was only 1 randomized controlled trial and 1 crossover intervention study. Studies mainly applied subjective measures to assess the outcome of interest, and there was an increasing trend of objective measures in 9 studies, mainly to assess changes in knowledge.

There was more organized and comprehensive use of online resources, with UpToDate, PubMed, Wikipedia, and e-journals being popular among trainees [33]. However, some residents cited insufficient time, insufficient faculty guidance, and lack of resources specific to psychiatry as barriers to using online resources in their practice. Moreover, 86% of respondents felt that there is a need for more psychiatry-specific online resources, and 79% believed that online resources should be more visual and interactive. In another survey, some trainees preferred to read or take notes on paper for academic purposes [34]. However, they still preferred on-screen reading for checking medication dosing and information.

As accessibility to the internet improved with wider coverage and greater stability, there was an increased effort to conduct online modules or e-learning to complement the traditional methods of psychiatric learning, at least for delivering theoretical aspects. For example, Hickey et al [35] conducted a blended course of traditional lectures, online modules, and videotape reviews in psychotherapy education. The authors developed 2 modules on Davanloo Intensive Short-term Dynamic Psychotherapy (IS-DTP), which consisted of PowerPoint presentations, videos, and pre- and posttests. In a crossover intervention design, residents were divided into 2 groups, with each group receiving an online module and a face-to-face lecture but in a different order for 2 topics. It was found that there was a statistically significant improvement in knowledge acquisition in both the online module group and face-to-face lecture group, and there was no significant difference in comparison between both groups regardless of how the topics were delivered.

Figure 4. Taxonomy of online-based and technology-assisted learning in 2015-2019.



Three studies described blended learning but with a different model compared to the earlier study [36-38]. For example, a bookend blended learning model was adopted, in which there was self-paced virtual learning via online modules in the beginning, followed by synchronous in-person interactive discussion, and the approach ended with case supervision or reflection. This approach was used in learning PTSD [36] and in an integrative psychiatry curriculum [37], and residents appreciated the convenience and suitability to augment their training owing to 2 main factors. The first factor was that the self-paced online modules allowed for more time for personalized training and to process materials, and the second factor was the ability to access quality information linked directly to published sources. In another variation of blended learning, another study applied the flipped classroom model, in which the required reading was provided earlier, and then, participants engaged in interactive asynchronous discussion [38]. Interestingly, less than half of the participants indicated improvement in their knowledge, while the rest felt no difference, and a small number of participants felt that the approach had a negative impact on their knowledge, indicating the need to re-evaluate the suitability of the asynchronous flipped classroom approach for psychiatric training.

Fully online courses or modules have been continuously adapted to different aspects of psychiatric training. Brief online courses or modules were employed in rather specific topics under psychiatric-related theoretical knowledge, such as catatonia [39], tobacco use disorder [40], and substance use disorder [41]. Certain modules can be completed in 10 minutes, and brief modules often include slide presentations and relevant videos (recorded webinars, patient interviews, or demonstrations of symptoms and signs), which trainees can complete at their own pace. Despite the brevity of the approach, there were significant increases in knowledge [39,40] or improved levels of stigma [41] between pre- and posttests, with some of the improvements being retained at the 3-month or 6-month follow-up, suggesting the utility of such an approach.

Videoconference-based discussions or webinars remain useful tools in psychiatric education. Few studies used videoconference platforms to deliver online lectures and online case discussions tailored for aspects of psychiatric training, such as career development activities [42] and continuous education in conflict zones [43]. Videoconference platforms have also been used for skills training. Puspitasari et al [44] conducted a study on online behavioral activation training over 4 weeks with the aim to compare trainer-led online training and self-paced online training. There was a significant increase in behavioral activation skill assessment total scores in both groups, and there were significant between-group differences favoring the trainer-led online group at both the posttraining assessment and 3-month follow-up. Although direct instructional strategies were applied in both groups and both approaches were conducted online, the interactive component of the session in the trainer-led group caused a significant difference in the improvement of the skills of the trainees.

Building on the earlier work of Pignatiello et al [21] and Volpe et al [22], a study by Teshima et al [45] focused on feedback from over 300 residents in telelink telepsychiatry training in Canada. The trainees appreciated the opportunity to learn about different approaches to interviewing clients. Although the residents found the technology a “bit unnatural” at the beginning of the session and realized that it was “challenging to interview [patients] at a distance,” they were still able to obtain nonverbal cues, which were important to understand their patients. Almost all participants agreed that the telepsychiatry experience was interesting, while 97% agreed with the statement “the experience helped me understand more about providing psychiatric services to underserved areas.” As the telelink program was an established and mandatory program for psychiatric trainees in University of Toronto, this study had one of the largest sample

sizes in comparison to other studies, enhancing its reliability despite its single study site.

Davidson and Evans [46] illustrated how a videoconference platform was employed for an online study group as an augmentation tool in preparing for the Royal Australian and New Zealand College of Psychiatrist (RANZCP) objective structured clinical examination (OSCE). Four New Zealand trainees used Google Hangout (now known as Google Meet) for their online OSCE practice, with the exam questions based on online past papers on the RANZCP website [46]. The members rotated their role to be a candidate, examiner, or role-player, adding to the experiential learning. The 4 trainees passed their OSCE and acknowledged the benefit of a virtual study group in enhancing their preparation.

As demonstrated by earlier studies, the concept of a rule-based virtual patient was proven useful in learning PTSD and significantly improved the knowledge and confidence level of residents in managing PTSD [47,48]. In another variation, Wilkening et al [49] employed a branched-narrative virtual patient for advanced psychopharmacology sessions, in which residents were presented with a challenge and given choices, and the consequences would depend on the choice selected. The integration of the virtual patient concept led to statistically significant improvements in knowledge levels in advanced psychopharmacology, supporting the efficacy of the virtual patient concept as part of psychiatric education.

Abundance of reliable resources is a boon to evidence-based medicine; however, due to the hectic nature of clinicians, a quick decision must be made quite often. Thus, few software programs were developed to act as a 1-stop center for reference to help expedite and guide their practice. Adeponle et al [50] developed the Psychiatry Toolkit, which allowed direct, immediate, and full access (institutional login) to desired journals, articles, and relevant databases, including PubMed, PsycINFO, and UpToDate [50]. Another study by Dirlam et al [51] described Mental Health EMR Tools, which is a large database that allows residents to access prevetted, curated, and continuously updated information to help with their clinical practice. In addition, acknowledging the potential of a smartphone as an educational tool, Zhang et al [52] developed the Delirium University Health Network Application as a tool for delirium education. It was initially developed as an online application and was later piloted as a smartphone app via the Android Play Store. The app included many important contents related to delirium, such as the DSM-5 diagnostic criteria for delirium, common causes of delirium, pharmacological and nonpharmacological interventions for delirium, and objective assessment questionnaires (eg, Confusion Assessment Method [CAM]). Overall, Mental Health EMR Tools and the delirium smartphone app received good feedback from users, who appreciated the convenience of getting information from a consolidated source. On the other hand, while the Psychiatry Toolkit helped residents to look for answers to their clinical questions, the adoption rate of the toolkit among respondents was relatively low at 47% [50].

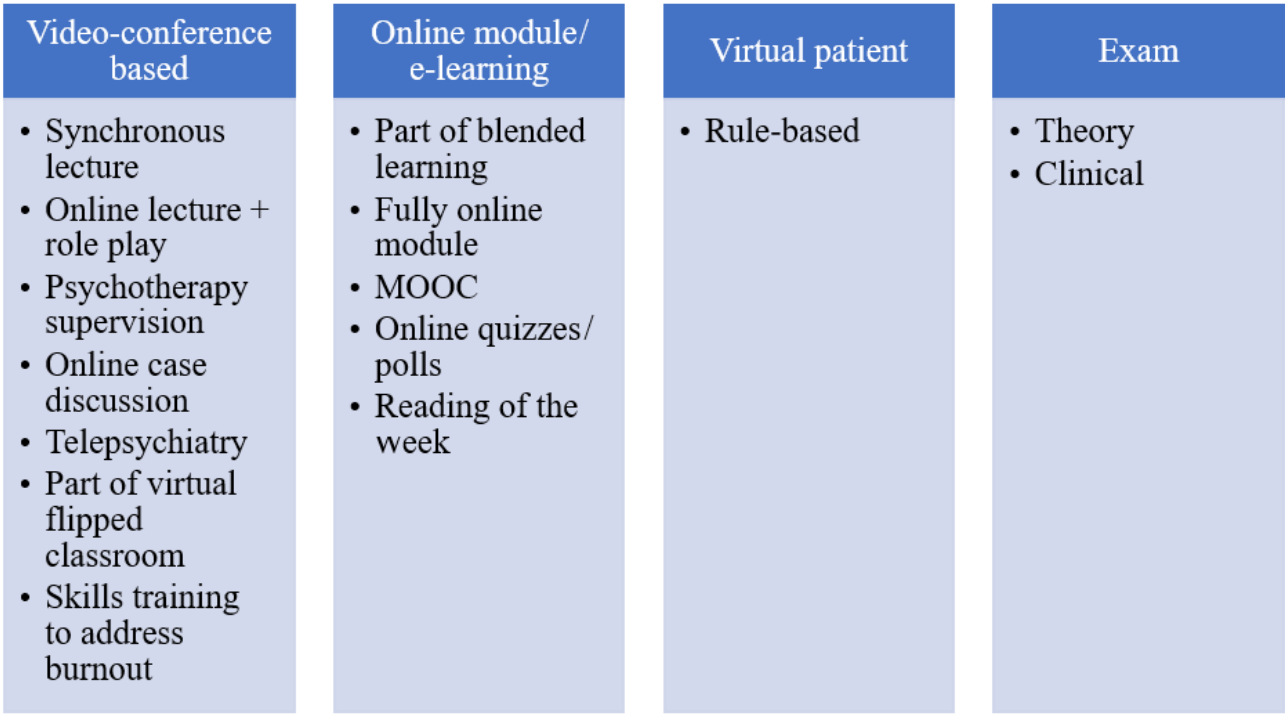
An interesting study described an innovative approach to adjunct psychiatric education using social media. Walsh et al [53] described the use of Twitter to disseminate education resources considered helpful in training. Under Twitter account @PhippsPsych, residents took turns to post tweets or retweet contents, such as take-home points from psychiatry grand rounds, links to journal articles, and references to psychiatry in current events. While the study had a rather small sample size with 49 residents, there was a significant increase in the proportion of participants using Twitter for medical education from 8.2% to 28.6%. However, residents' ratings regarding the usefulness of social media in medical education did not change from pre- to postsurvey, and corroborated by the fact that 60% of residents reported that the knowledge gained from following the account had no impact on their clinical practice, 37.2% reported a minimal or average impact and only 2.8% reported a great impact.

Studies From 2020 Onward (After the COVID-19 Pandemic)

The number of articles on online-based or technology-assisted learning in psychiatry education for psychiatric residents or trainees saw a significant spike during and after the COVID-19 pandemic. There were 41 relevant articles from 2020 until June 2024 (Table S3 in Multimedia Appendix 2). While the United States, Canada, and the United Kingdom had the highest number of studies, there was also notable involvement from Global South countries, such as Malaysia, Thailand, Pakistan, India, and Tunisia. Majority of the articles involved the application of videoconference-based learning or webinar concepts (n= 24/41, 58.5%) and online modules (n= 8/41, 19.5%) (Figure 5). Studies mainly had subjective measures as outcomes and had a 1-group posttest-only study design. There were 11 studies with a 1-group pretest-posttest design and 1 study with a nonequivalent group posttest-only design.

Transition to videoconference-based learning and webinars was necessary for the continuation of training and education during the COVID-19 pandemic, including for psychiatric residents. There were few variations in how videoconference-based learning was applied. In some articles, it was rather straightforward with synchronous online lectures through the videoconference platform, and some of the lectures were then followed by virtual group discussions or brainstorming sessions in the breakout rooms of the platform to make it more interactive. This approach was commonly used in the theoretical aspects of psychiatric training, such as in alcohol use disorder [54], fundamentals of remote psychotherapy [55], research in psychiatry [56], digital psychiatry [57], complex child and adolescent cases [58], biostatistics and methodology courses [59], and journal clubs [60].

Figure 5. Taxonomy of online-based and technology-assisted learning in 2020-2024. MOOC: massive open online course.



In another variation, the online lecture was paired with virtual role-play or simulation sessions. With the added experiential learning component, residents were able to apply their knowledge accordingly in case scenarios. In a study by Blamey et al [61], psychiatric trainees attended a 2-hour virtual lecture on the necessary skills for their on-call work, and then, the trainees participated in a series of 2-hour simulated on-call shifts once a week for 10 weeks, covering 10 common scenarios for psychiatric on-call work. Acknowledging the importance of understanding the complexities of health systems in delivering effective and safe patient care, Li et al [62] developed an online curriculum for core competencies in health systems science. The residents underwent 10-minute virtual didactics prior to the virtual simulation of case scenarios using the Zoom platform.

Looking at the outcomes, trainees or residents perceived a high level of satisfaction with the program and its online delivery, as well as an increase in confidence in skills and perceived learning gains [56,57,60]. Among the studies, 1 study had an objective outcome, in which significant improvements in the knowledge level regarding alcohol use were noted among residents after the videoconference-based lecture and recorded training video session, indicating the potential efficacy of such a program [54]. The virtual role-play concept was especially credited to be a useful technique to enhance the interactive learning of residents [61]; however, the virtual format can be awkward owing to the need for turn-taking, which in turn affects the interactivity, especially when there are few residents involved concurrently in a single case scenario [62].

As demonstrated by Gammon et al [18], a videoconference platform may be used for psychotherapy supervision. Due to the COVID-19 pandemic, this was of value, and there were 2 studies involving psychiatric trainees receiving psychotherapy supervision virtually. There were concerns regarding the loss of nonverbal cues or subtleties of communication during remote supervision, in addition to the "Zoom fatigue" phenomenon, all of which influenced residents to favor face-to-face supervision more than remote supervision [63]. However, the flexibility of remote supervision and the option to allow residents to attend the supervision even when they were busy with their ward work or were on-call were certainly advantageous, and the quality of psychotherapy skills attainment based on subjective assessments by supervisors was not significantly different between remote supervision and traditional face-to-face supervision [63,64]. In fact, as reported by Famina et al [65], supervising attendings noted that the quality of psychiatric care was not different between remote sessions and in-person sessions, and there was not much difference in terms of the ability to empathize and to interpret nonverbal cues.

With regard to telepsychiatry, the authors of an article mentioned their experience of a sudden unprecedented change to their service, which involved a transition to telepsychiatry due to the COVID-19 pandemic [66]. The service initially involved phone consultation, and residents and attendings subsequently switched to video consultation after obtaining approval to use a video platform [66]. While both trainees and attendings strongly agreed that the change to virtual care was necessary, the attendings felt that trainee supervision and training worsened during the pandemic. The trainees also felt less comfortable conducting virtual care and less confident in their assessments to the extent that they found video consultations "frustrating," especially when attempting to interview patients who had difficulty engaging in virtual interactions (eg, those with delirium, neurocognitive disorders, or mania) [66]. This

experience was echoed in a survey by Cruz et al [67], which showed that the top 5 concerns shared by residents and the faculty about telepsychiatry were the inability to perform a physical exam, poor internet connection, unknown liability risks related to telepsychiatry, certain cultures being less accepting, and nonverbal cues being missed.

Contrary to that experience, the authors of another study described their telepsychiatry experience in a rather positive note. Because of a rapid shift, telepsychiatry sessions were still following the prior model of in-person direct supervision involving attendings and residents, and both had to don a mask while conducting the telepsychiatry sessions, which affected the voice projection and the ability of the patient to hear the treatment plan [68]. Subsequently, with further understanding of the videoconference platform, they were able to continue direct supervision with slight modification as attendings joined the session from their private offices, allowing residents and attendings to remain mask-free and improving the audio for patients. Majority of the residents felt that telepsychiatry had positively impacted their clinical education experience, and it was significantly associated with comfort with practicing telepsychiatry in the future [68].

The flipped classroom model, which is a type of blended learning, reversed the settings, with direct instructions to be provided at home and learning activities involving higher order thinking to be done at school. COVID-19 restrictions necessitated the change to a fully virtual flipped classroom, and psychiatric training was not an exception. A study from Pakistan described a program using the flipped classroom model for an online trauma curriculum [69]. Under this program, reading materials and videos were provided to trainees earlier, and then, the trainees participated in virtual brainstorming, role-play, and case-based discussions. On a larger scale, the Metis didactic courses for psychiatric residents in Sweden (the pedagogical model has been in line with the flipped classroom concept from its inception in 2007) were switched to a digital format to ensure continuation of learning [70]. Each course consists of 3 phases: distance-based self-study, classroom-based meeting days for lectures and supervision, and distance-based examination. The second phase was subsequently transitioned to an online classroom for the same activities. While the fully online flipped classroom concept improved the level of knowledge and skills of psychiatric trainees, some residents preferred to return to face-to-face learning [69,70]. Interestingly, female participants and those aged younger than 50 years were more inclined to continue with online-based course meetings [70].

There was a growing issue of burnout among psychiatric trainees due to the pandemic and its consequences. As such, training programs included skills training as a necessary curriculum component to address burnout, and these were delivered virtually using videoconference platforms, such as the virtual Balint group [71,72], Mind-Body skills program [73], brief mindfulness-based cognitive behavioral therapy (CBT)-informed virtual well-being program [74], and virtual medical improvement program [75]. An interesting example was the virtual Balint group, an initiative that provides a cathartic space and helps to improve morale. With the idea to improve the understanding of patients' problems rather than finding solutions, residents were encouraged to participate with the camera on and the mic on mute when someone was presenting, and the presenter was free to express their experience of doctor-patient interactions, with the guarantee of nonjudgment and confidentiality [72]. During the discussion, the hands-up function of the Zoom platform was used when someone wanted to speak in order to control flow and avoid interruption. Participants were very positive of the virtual Balint group, with trainees feeling well supported by this initiative. However, many participants preferred face-to-face sessions but nevertheless would choose an online session over no session at all. In another study, the virtual Balint group was credited for promoting a sense of connectedness among peers and providing freedom to speak without needing to censor themselves. However, the virtual nature of the Balint group itself led some participants to feel an abrupt ending to the session, which does not occur in face-to-face sessions [71].

In terms of online modules, the most common design was problem-focused case vignettes, alongside interactive presentation and audiovisual content. Some of the modules also had tests at the end. This design was adopted in few of the studies to cover various aspects of psychiatry training, such as forensic psychiatry [76], catatonia [77], neuropsychiatry [78], tobacco use disorder [79], cultural sensitivity [80], and antiracism intervention [81]. A study by Owais et al [82] used the same online module concept in a blended learning approach for an electroconvulsive therapy curriculum, together with didactic seminars and hands-on clinical management. By combining indirect and direct instruction strategies, there was significant improvement in terms of knowledge attainment after the modules (smaller sample sizes) [76-78,82], and generally, the modules had high satisfaction levels reported by residents [80,81].

While most modules were confined to certain training institutions or regions, 1 article described a larger-scale MOOC to augment psychiatric training. Gargot et al [83] described the First European Psychiatric Association MOOC on CBT, which lasted for a month. With a focus on the theoretical aspects of CBT through recorded lectures, presentations, online forums, and online examinations, the self-paced MOOC had large participation, with 7116 participants enrolling from at least 49 countries. Although the eventual completion rate was 26%, a large number of participants (n=1828) completed the MOOC and the average score for the tests increased steadily from 21.4 out of 25 in the first week to 23.13 out of 25 in the final week, indicating the potential of the MOOC to fill the training gap.

In another study, a website was developed as an innovative, free psychiatry Continuing Professional Development (CPD) resource for Canadian psychiatrists and residents. Referred to as Reading of the Week, this website summarizes the latest psychiatric literature, provides expert commentaries, and promotes discussions on social media platforms [84]. The innovations in psychiatric education as described in these articles received good feedback from trainees. For the Reading of the Week initiative, in which the survey evaluation was based on the 6-level evaluation framework by Moore et al [85], positive feedback and satisfaction were reported by participants across

the 6 levels, including knowledge outcomes (level 3), behavior outcomes (levels 4 and 5), and practice outcomes (level 6) [84].

The pandemic also forced the examination process to be performed in a digital format. Generally, examinations in psychiatric training can be divided into theory examinations and clinical examinations (including OSCE). The Royal College of Psychiatrists conducted Member of the Royal College of Psychiatrists theory examinations via a digital platform using a combination of AI and in-person online proctoring [86]. Multiple choice questions were directly assessed, but for questions involving very short answers, smart algorithms were developed to recognize versions of correct answers, and answers that were nonexact matches were reviewed by a designated examiner. On the other hand, for the assessment of clinical psychiatry skills, there was mixed feedback from both examiners and trainees. Depending on the format of the clinical examination, examiners generally manned the stations or the breakout rooms. Integrating videoconference technology for the purpose of clinical examinations had inherent issues, such as connectivity problems, a sense of disconnect, lack of a framework to mentally reset, difficulty in building rapport, and an inadequate capacity to assess clinical skills [87]. However, some residents believed that online assessments were convenient for both participants and patients, reducing anxiety by being in a familiar environment and improving patient access [88]. Some of the candidates even stated that virtual communication was nearly as good as face-to-face communication and online examination was “better than expected” [89].

There is a paucity of studies comparing online training or learning with face-to-face or in-person training. In a quasiexperimental study in Germany to explore whether the satisfaction of online CBT training is noninferior to that of in-person CBT training, the 2 study groups had the same theoretical CBT content, a similar duration of training, comparable audiences, and an identical trainer [90]. The online training was conducted according to the inverted-classroom concept, with participants being required to watch recorded video lectures on the Moodle platform and then have a Zoom discussion at a fixed time for 6 to 7 sessions. It was found that evaluations of the online training group were noninferior to those of the in-person group in terms of information content, didactic presentation, assessment of the trainer as a suitable role-model, working atmosphere, own commitment, and practical relevance, suggesting that the delivery of CBT knowledge through an online platform may be sufficient.

In another study, Hewson et al [91] described a rather indirect comparison between face-to-face basic psychiatry skills simulation training and synchronous online training. The transition to online training via Zoom was due to evolving COVID-19 restrictions at that time and was not in the initial plan. In subgroup analyses, the face-to-face group showed statistically significant improvements in confidence across all psychiatry skills tested, whereas the online group showed significant improvements in confidence in all but 2 skills, namely psychiatric risk assessment and assessment of physical health problems in elderly patients with cognitive impairment. However, the face-to-face group included foundation doctors (junior doctors) and the online group included psychiatry and general practitioner trainees, suggesting that the lack of a significant improvement in confidence in those 2 skills could be related to a higher baseline self-confidence level prior to the simulation training.

A rule-based virtual patient appears to be a mainstay model of a virtual patient in psychiatric education. Rakofsky et al [92] developed a virtual patient-based assessment simulator as a tool to assess the proficiency of residents regarding psychopharmacological knowledge and practice. Combining virtual human avatars, AI, and an advanced pedagogical design, it allows for a realistic interaction, including live voice communication. According to the rule-based virtual patient concept, residents had choices of questions and answers to choose from, and they were given immediate feedback on all their choices alongside the rationale. Looking at the performance of the residents, the mean total score of the simulator by class correlated significantly with the mean score of the somatic therapies subscale of the Psychiatry Residency in Training Exam (PRITE), suggesting construct validity of the virtual patient simulator.

In a survey assessing residents’ perceptions of the pandemic’s impact on their didactic experience and training preferences, it was found that trainees appreciated several positive aspects of virtual didactics, such as being easy to attend and being engaging, and they were able to invite guest speakers from other institutions easily [93]. However, some negative experiences were also reported, including the “Zoom fatigue” phenomenon and frequent distractions, and some topics did not translate well to a virtual environment. Residents from Thailand, which was hit hard by the pandemic and had a significant shift in psychiatric training to online sessions, also reported mixed experiences. Although all residents had good results and passed their examinations, they felt that studying online and the uncertainty with virtual psychotherapy were major inconveniences in their training [94]. In another study, residents were ambivalent. They perceived face-to-face teaching to be superior, but majority of them did not think a complete return to in-person learning would be the most effective option when this becomes possible, implying a preference to continue with some online components in the training [95].

A summary of the key benefits and limitations of the 5 different online-based and technology-assisted educational methods (videoconference, online module/e-learning, virtual patient, software/applications, and social media) is provided in Table 1.

Table 1. Summary of the implementation of online-based and technology-assisted psychiatric education.

Method	Key benefits	Limitations
Videoconference	• Flexible and applicable for various objectives (lectures, skills training, psychotherapy supervision, etc) • Accommodates different instructional strategies (direct, indirect, interactive, and experiential) • Convenience of attending sessions regardless of location or schedule	• Relies on the internet speed • Struggles with nonverbal cues • Frequent distractions and Zoom fatigue
Online module	• Allows self-paced learning • Numerous relevant materials can be designed and included (animations, prerecorded videos, and quizzes) • Possibility of reaching a wide range of audiences via a MOOC[a]	• Lacks direct supervision • More suitable for theoretical aspects of training than clinical aspects • Often requires collaboration and resources to design the modules
Virtual patient	• Valuable for learning uncommon cases • Engaging learning experience	• Requires high levels of resources for development • Can be frustrating to interact in case of speech recognition issues
Software/applications	• Serves as a convenient point of reference • Integration with a smartphone • Highly favored in checking medication dosing and information	• Requires certain levels of resources for development • Software that primarily acts as a gateway for institutional logins to certain websites is not often used
Social media	• Promotes a continuous learning opportunity • Possibility of greater dissemination of knowledge to a larger audience • Provides information on current evidence-based studies	• Reported knowledge gain that translates into clinical practice is still less significant • Relies on the effort of the individual to follow the account and review the shared resource

[a]MOOC: massive open online course.

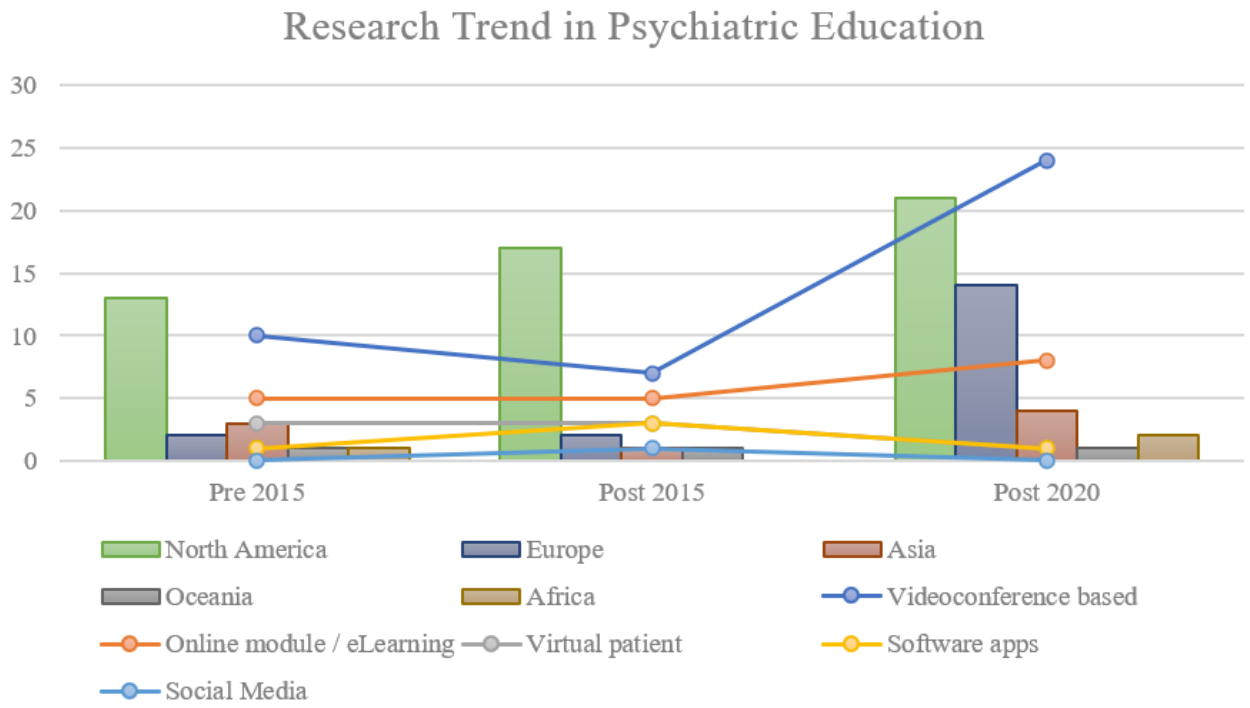
Discussion

Summary of the Results

The findings across the 3 phases (prior to 2015, 2015-2019 [prepandemic], and 2020 onward [postpandemic]) illustrated the creative integration of online-based and technology-assisted learning in psychiatric education for trainees (Figure 6). Five approaches were identified: videoconference, online module/e-learning, virtual patient, software/applications, and social media. These methods were used for various objectives, including but not limited to teaching theory knowledge, skills training, psychotherapy supervision, and information retrieval. North American countries were leading in research output, followed by European countries. There was consistent research presence from countries in Asia, Africa, and Oceania throughout the 3 phases, with a slight increase after 2020.

The trend of online-based and technology-assisted psychiatric education showed changes from one phase to another. Videoconference-based learning was consistently the preferred way of integrating technology into the learning or training of trainees. Meanwhile, there was a growing trend of online module/eLearning platform use, virtual patient use, and development of applications or software from the "prior to 2015" phase until the prepandemic phase (2015-2019). The trend however stagnated in the postpandemic phase. Videoconference-based learning was more dominantly used due to 3 factors. First, as seen in other fields, the sudden shift to online was not accompanied by the readiness of other technology modalities to assume the responsibility to continue the learning process [96]. Second, the improved accessibility and availability of videoconference platforms, such as Zoom, Google Meet, and Microsoft Teams, served as a plus point [97], and less resources were needed to shift learning to those platforms rather than developing online modules in an expedited manner. The third factor corresponds to the nature of psychiatric learning, which is preferably face-to-face, but among the choices that were presented, videoconference-based learning was a method that may offer some compromise in terms of allowing direct and synchronous instructional methods despite physical distance [98].

Figure 6. Trend of online-based and technology-assisted psychiatric education in the 3 study phases (prior to 2015, 2015-2019 [prepandemic], and 2020 onward [postpandemic]).



In terms of didactic teaching, evidence showed that online-based or technology-assisted learning is beneficial and well-tolerated by residents. Barring minor issues pertaining to connectivity, a recurring problem reported by several studies [59,65,78], theoretical learning of core knowledge in psychiatry through online platforms has been helpful in improving knowledge, and the impact cannot be understated. Some trainees or residents even expressed their preference for receiving academic or continuous professional development activities virtually, citing convenience and ease of access as important factors that encourage participation despite their busy schedules [72,93]. Nevertheless, psychiatry is not purely a theoretical field in medicine. Psychiatry knowledge must be paired with competency in necessary skills, such as communication skills and psychotherapy skills, to ensure robust and quality training. Learning these skills through an online platform is possible, and a study showed that learning CBT online was not inferior to learning it through in-person training [90]. Nonetheless, proper hands-on guidance remains necessary for mastering these skills, as it might not be possible to fully replicate or demonstrate those subtle, nuanced techniques through a screen. Hence, harmonizing virtual theoretical learning and practical hands-on learning to develop comprehensive blended learning may be a more interesting proposition [82]. Nevertheless, it must be highlighted that the approach of providing knowledge about certain skills and evaluating knowledge levels in trainees is vastly different from teaching skill competencies and then evaluating the levels of competencies of trainees. Because of this issue, some practitioners preferred in-person learning to ensure an optimum level of skill attainment, as compared to through online learning.

As demonstrated by many studies successfully integrating simulation in their medical training, the virtual patient concept (essentially simulating the experience of seeing a patient with the assistance of technology) has been quite helpful in psychiatry education. This concept has a unique strength: the ability to simulate cases that may be uncommon in clinical practice [99]. For example, a study used this concept to portray a refugee with PTSD symptoms, enriching the training of psychiatric trainees and increasing their confidence in managing PTSD cases [48]. Adjacent to the virtual patient concept is the virtual reality (VR) concept. VR allows for an immersive experience, frequently described as "being there," which involves more senses beyond just sight [100]. It has been used as part of the training curricula in medical fields, including orthopedics [101], surgery [102], and ophthalmology [103], with varying degrees of success. In psychiatric services, VR has been implemented as part of therapy or treatment, for example, exposure therapy for phobic disorders [104] and social skills training in patients with autism spectrum disorder [105]. Unfortunately, VR has not been extensively used in psychiatric education yet, perhaps due to the limitation of the current technology in grasping the complexity of psychiatric cases. As technology rapidly evolves, it remains an exciting avenue to explore in the future.

Competency is often assessed through an examination process. During the pandemic, the transition to online examinations became common worldwide, and psychiatry was no different. Online assessments were applied to various aspects of the psychiatry curriculum. While there were few issues with online theory examinations [86], the same cannot be said for other aspects. For example, an article described the challenges in sitting for the virtual Clinical Assessment of Skills and Competencies (CASC) under the Royal College of Psychiatrists, United Kingdom, where constant worry about internet issues, a sense of disconnect, and an inability to mentally reset between stations affected performance and the overall experience [87].

Another article described the experience of the online Basic Specialist Training examination under the College of Psychiatrists of Ireland [89]. Despite acknowledging the superiority of face-to-face examinations, the online examination was described as nearly as good and more favorable in view of respondents being in a familiar environment as well as saving cost and time to travel [89]. Although it is not possible to compare these examinations directly, the Basic Specialist Training examination highlighted that such examinations could be conducted virtually, but thorough preparation and strong technical support are warranted.

Another interesting aspect of how technology can be valuable in psychiatric training is through the development of cultural competence. Cultural competence refers to the capacity to respond to the unique needs of the population. In the context of psychiatry, it refers to the development of knowledge, skills, and attitude, which can enable the formulation of an intervention that considers the sociocultural backgrounds and sensitivities of psychiatric patients. In turn, this allows for a comprehensive and tailored treatment for patients, especially those from racial and minority ethnic groups. Previous measures to promote cultural competence included learning trips and student exchange programs. However, technology can also offer interesting and possibly cheaper options to achieve the same goals. Trinh et al [80] described an online module program to promote culture sensitivity in a psychiatry department. Three modules were developed, including presentation slides, case vignettes, and recorded videos, covering important topics, such as DSM-5 Outline for Cultural Formulation, Cultural Formulation Interview, and cultural identity as a multidimensional construct. This program was initially met with surprising feedback, with most clinicians indicating that they were not familiar with what questions to ask to elicit a cultural history; however, after completion, the respondents endorsed the module as useful and reported that they would change their practice, suggesting that a brief online module may have potential in this area.

Exploring the application of online-based or technology-assisted learning in psychiatry education for trainees holds significance in the training of future psychiatrists, especially in low- and middle-income countries (LMICs). To put this in context, most African countries have a massive shortage of psychiatrists, with an average of 0.1 per 100,000 people in the 47 countries across the World Health Organization African region [106]. Moreover, both Liberia [107] and Timor-Leste [108] had only 2 trained psychiatrists, indicating the pressing need to support the mental health systems in these countries. While there is no easy fix for this situation, efforts to increase the number of psychiatrists is of utmost importance. In this context, a concerted and collaborative effort among universities or training programs across regions or continents for the training of future psychiatrists in LMICs is needed to alleviate this issue, and the use of online platforms could be the key to bridging the gap.

Limitations

Our review has some limitations. First, we did not limit the types of publications, resulting in variations in the quality and rigor of the studies. Additionally, majority of the articles were from Western countries, with very few from LMICs, which might reduce generalizability. It is important to note that while a healthy number of studies were included in this scoping review, ultimately obvious heterogeneity was present in terms of the outcomes measured. Majority of the studies had a 1-group intervention study design and had a rather small sample size. Therefore, while some of the included studies might have shown positive responses or outcomes, the findings need to be interpreted carefully in the context of these factors, which might affect generalizability. Moreover, the lack of well-designed comparative intervention studies limits the understanding of the effectiveness of online learning in comparison to traditional face-to-face learning. Another key limitation is the profound lack of an objective assessment as part of the outcome measure within the included studies. Most of the studies assessed satisfaction and attitudes toward the interventions, rather than the actual impact of the interventions.

Moving forward, more well-designed studies in psychiatric education for trainees are needed, especially with objective assessments, to truly evaluate the suitability of online-based and technology-assisted learning. There is a clear paucity of studies evaluating the efficacy of psychotherapy skills training delivered virtually, which may be of significance to LMICs that need a higher number of competent mental health professionals. Lastly, the emergence of AI systems, such as ChatGPT and DeepSeek, can be a game changer for the psychiatric education of trainees, and further exploration is required on how to maximize the benefits of these systems while developing safe and competent psychiatrists in the future.

Conclusion

Videoconference-based learning was the most widely implemented approach, followed by online modules and virtual patients. Despite the outcome heterogeneity and small sample sizes in the included studies, the application of such approaches may have utility in terms of knowledge and skills attainment. With further fine-tuning, these approaches could become effective solutions to address the significant deficiency of psychiatrists, especially in LMICs.

Acknowledgments

The authors would like to thank Professor Dr Samy Azer for his input in preparing the manuscript. No funding was received to assist with the preparation of this manuscript.

Data Availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

MAMK and SMYAS: conceptualization, data collection, data analysis, and writing – original draft preparation. TIMD and JTYL: conceptualization, data analysis, writing – review and editing, and supervision. All authors contributed to the article and approved the submitted version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategies.
[DOCX File , 14 KB - mededu_v11i1e64773_app1.docx]

Multimedia Appendix 2
Data of the 3 study phases.
[DOCX File , 90 KB - mededu_v11i1e64773_app2.docx]

Multimedia Appendix 3
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 101 KB - mededu_v11i1e64773_app3.pdf]

References

1. Ahmady S, Kallestrup P, Sadoughi MM, Katibeh M, Kalantarion M, Amini M, et al. Distance learning strategies in medical education during COVID-19: A systematic review. J Educ Health Promot 2021;10:421 [FREE Full text] [doi: 10.4103/jehp.jehp_318_21] [Medline: 35071627]
2. Saiyad S, Virk A, Mahajan R, Singh T. Online Teaching in Medical Training: Establishing Good Online Teaching Practices from Cumulative Experience. Int J Appl Basic Med Res 2020;10(3):149-155 [FREE Full text] [doi: 10.4103/ijabmr.IJABMR_358_20] [Medline: 33088735]
3. Brenner AM. Uses and limitations of simulated patients in psychiatric education. Acad Psychiatry 2009 Apr 27;33(2):112-119. [doi: 10.1176/appi.ap.33.2.112] [Medline: 19398623]
4. Rauch C, Utz J, Rauch M, Kornhuber J, Spitzer P. E-learning is not inferior to on-site teaching in a psychiatric examination course. Front Psychiatry 2021 Apr 13;12:624005 [FREE Full text] [doi: 10.3389/fpsyt.2021.624005] [Medline: 33927651]
5. Beketov V, Menshikova I, Khudarova A. Fast track to full online education in the medical field: evaluating effectiveness and identifying problems from the COVID-19 experience. Int J Web-Based Learn Teach Technol 2022;17(1):1-24 [FREE Full text] [doi: 10.4018/IJWLTT.315824]
6. Westphaln KK, Regoeczi W, Masotya M, Vazquez-Westphaln B, Lounsbury K, McDavid L, et al. From Arksey and O'Malley and Beyond: Customizations to enhance a team-based, mixed approach to scoping review methodology. MethodsX 2021;8:101375 [FREE Full text] [doi: 10.1016/j.mex.2021.101375] [Medline: 34430271]
7. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol Theory Pract 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
8. Haddaway NR, Collins AM, Coughlin D, Kirk S. The role of google scholar in evidence reviews and its applicability to grey literature searching. PLoS One 2015 Sep 17;10(9):e0138237 [FREE Full text] [doi: 10.1371/journal.pone.0138237] [Medline: 26379270]
9. Reed DA, Cook DA, Beckman TJ, Levine RB, Kern DE, Wright SM. Association between funding and quality of published medical education research. JAMA 2007 Sep 05;298(9):1002-1009. [doi: 10.1001/jama.298.9.1002] [Medline: 17785645]
10. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251 [FREE Full text] [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
11. McCloskey K, Neuzil K, Basak R, Chan KH. Quality of reporting for qualitative studies in pediatric urology-A scoping review. J Pediatr Urol 2023 Oct;19(5):643-651. [doi: 10.1016/j.jpurol.2023.04.027] [Medline: 37481426]
12. Powsner S, Byck R. Implementing a computer system for psychiatric training : the electric resident. Acad Psychiatry 1991 Jun;15(2):100-105. [doi: 10.1007/BF03341304] [Medline: 24430518]
13. Briscoe GW, Fore Arcand LG, Lin T, Johnson J, Rai A, Kollins K. Students' and residents' perceptions regarding technology in medical training. Acad Psychiatry 2006 Dec 01;30(6):470-479. [doi: 10.1176/appi.ap.30.6.470] [Medline: 17139018]
14. Walter DA, Rosenquist PB, Bawtinhimer G. Distance learning technologies in the training of psychiatry residents: a critical assessment. Acad Psychiatry 2004 Mar 01;28(1):60-65. [doi: 10.1176/appi.ap.28.1.60] [Medline: 15140810]
15. Greenwood J, Williams R. Continuing professional development for Australian rural psychiatrists by videoconference. Australas Psychiatry 2008 Aug 01;16(4):273-276. [doi: 10.1080/10398560801982994] [Medline: 18608161]

16. Chipps J, Ramlall S, Mars M. Videoconference-based education for psychiatry registrars at the University of KwaZulu-Natal, South Africa. Afr J Psychiatry (Johannesbg) 2012 Jul 19;15(4):248-254. [doi: 10.4314/ajpsy.v15i4.32] [Medline: 22829227]
17. Nasr Esfahani M, Behzadipour M, Jalali Nadoushan A, Shariat SV. A pilot randomized controlled trial on the effectiveness of inclusion of a distant learning component into empathy training. Med J Islam Repub Iran 2014;28:65 [FREE Full text] [Medline: 25405130]
18. Gammon D, Sørlie T, Bergvik S, Sørensen Høifødt T. Psychotherapy supervision conducted via videoconferencing: A qualitative study of users' experiences. Nordic Journal of Psychiatry 2009 Jul 12;52(5):411-421. [doi: 10.1080/08039489850139445]
19. Szeftel R, Hakak R, Meyer S, Naqvi S, Sulman-Smith H, Delrahim K, et al. Training psychiatric residents and fellows in a telepsychiatry clinic: a supervision model. Acad Psychiatry 2008;32(5):393-399. [doi: 10.1176/appi.ap.32.5.393] [Medline: 18945978]
20. Dzara K, Sarver J, Bennett JI, Basnet P. Resident and medical student viewpoints on their participation in a telepsychiatry rotation. Acad Psychiatry 2013 May 01;37(3):214-216. [doi: 10.1176/appi.ap.12050101] [Medline: 23632939]
21. Pignatiello A, Teshima J, Boydell KM, Minden D, Volpe T, Braunberger PG. Child and youth telepsychiatry in rural and remote primary care. Child Adolesc Psychiatr Clin N Am 2011 Jan;20(1):13-28. [doi: 10.1016/j.chc.2010.08.008] [Medline: 21092909]
22. Volpe T, Boydell KM, Pignatiello A. Attracting Child Psychiatrists to a Televideo Consultation Service: The TeleLink Experience. Int J Telemed Appl 2013;2013:146858 [FREE Full text] [doi: 10.1155/2013/146858] [Medline: 23864854]
23. DeGaetano N, Greene C, Dearaujo N, Lindley S. A pilot program in telepsychiatry for residents: initial outcomes and program development. Acad Psychiatry 2015 Feb;39(1):114-118. [doi: 10.1007/s40596-014-0122-y] [Medline: 24777712]
24. Bayar MR, Poyraz BC, Aksoy-Poyraz C, Arikan MK. Reducing mental illness stigma in mental health professionals using a web-based approach. Isr J Psychiatry Relat Sci 2009;46(3):226-230 [FREE Full text] [Medline: 20039525]
25. Rahman A, Nizami A, Minhas A, Niazi R, Slatch M, Minhas F. E-Mental health in Pakistan: a pilot study of training and supervision in child psychiatry using the internet. Psychiatr Bull 2006;30(4):149-152. [doi: 10.1192/pb.30.4.149]
26. Kulier R, Hadley J, Weinbrenner S, Meyerrose B, Decsi T, Horvath AR, et al. Harmonising evidence-based medicine teaching: a study of the outcomes of e-learning in five European countries. BMC Med Educ 2008 Apr 29;8:27 [FREE Full text] [doi: 10.1186/1472-6920-8-27] [Medline: 18442424]
27. DeBonis K, Blair T, Payne S, Wigan K, Kim S. Viability of a web-based module for teaching electrocardiogram reading skills to psychiatry residents: learning outcomes and trainee interest. Acad Psychiatry 2015 Dec;39(6):645-648. [doi: 10.1007/s40596-014-0249-x] [Medline: 25391493]
28. Garside S, Levinson A, Kuziora S, Bay M, Norman G. Efficacy of teaching clinical clerks and residents how to fill out the Form 1 of the Mental Health Act using an e-learning module. Electron J e-Learning 2009;7(3):239-246 [FREE Full text]
29. Kenny P, Parsons T, Pataki C, Pato M, St George C, Sugar J, et al. Virtual Justina: A PTSD virtual patient for clinical classroom training. Annu Rev CyberTherapy Telemed 2008;6(1):113-118 [FREE Full text]
30. Pataki C, Pato MT, Sugar J, Rizzo AS, Parsons TD, St George C, et al. Virtual patients as novel teaching tools in psychiatry. Acad Psychiatry 2012 Sep 01;36(5):398-400. [doi: 10.1176/appi.ap.10080118] [Medline: 22983473]
31. Diagnostic and Statistical Manual of Mental Disorders, 4th Edition, Text Revision (DSM-IV-TR). Washington, DC: American Psychiatric Association; 2000.
32. Gorrindo T, Baer L, Sanders K, Birnbaum RJ, Fromson JA, Sutton-Skinner KM, et al. Web-based simulation in psychiatry residency training: a pilot study. Acad Psychiatry 2011;35(4):232-237. [doi: 10.1176/appi.ap.35.4.232] [Medline: 21804041]
33. Torous J, Franzan J, O'Connor R, Mathew I, Keshavan M, Kitts R, et al. Psychiatry residents' use of educational websites: a pilot survey study. Acad Psychiatry 2015 Dec;39(6):630-633. [doi: 10.1007/s40596-015-0335-8] [Medline: 26077007]
34. Reddy A, Duenas H, Luker A, Thootkur M, Kim JW. Preference of electronic versus paper reading resources among trainees in psychiatry. Acad Psychiatry 2018 Oct 3;42(5):743-745. [doi: 10.1007/s40596-018-0930-6] [Medline: 29725992]
35. Hickey C, McAleer S, Khalili D. E-learning: is it any better than traditional approaches in psychotherapy education? Arch Psychiatry Psychother 2015;17(4):48-52. [doi: 10.12740/APP/60644]
36. Kuhn E, Hugo E. Technology-based blended learning to facilitate psychiatry resident training in prolonged exposure (PE) therapy for PTSD. Acad Psychiatry 2017 Feb 5;41(1):121-124. [doi: 10.1007/s40596-016-0648-2] [Medline: 27921266]
37. Ranjbar N, Ricker M, Villagomez A. The integrative psychiatry curriculum: development of an innovative model. Glob Adv Health Med 2019;8:2164956119847118 [FREE Full text] [doi: 10.1177/2164956119847118] [Medline: 31080697]
38. Law M, Rapoport MJ, Seitz D, Davidson M, Madan R, Wiens A. Evaluation of a national online educational program in geriatric psychiatry. Acad Psychiatry 2016 Dec 25;40(6):923-927 [FREE Full text] [doi: 10.1007/s40596-015-0377-y] [Medline: 26108395]
39. Cooper J, Roig Llesuy J. Catatonia education: needs assessment and brief online intervention. Acad Psychiatry 2017 Jun;41(3):360-363. [doi: 10.1007/s40596-016-0632-x] [Medline: 27837452]
40. Williams JM, Poulsen R, Chaguturu V, Tobia A, Palmeri B. Evaluation of an online residency training in tobacco use disorder. Am J Addict 2019 Jul 16;28(4):277-284. [doi: 10.1111/ajad.12885] [Medline: 30993797]

41. Avery J, Knoepflmacher D, Mauer E, Kast KA, Greiner M, Avery J, et al. Improvement in residents' attitudes toward individuals with substance use disorders following an online training module on stigma. HSS J 2019 Feb 01;15(1):31-36 [FREE Full text] [doi: 10.1007/s11420-018-9643-3] [Medline: 30863230]
42. Kupfer D, Schatzberg A, Dunn L, Schneider A, Moore T, DeRosier M. Career development institute with enhanced mentoring: a revisit. Acad Psychiatry 2016 Jun;40(3):424-428 [FREE Full text] [doi: 10.1007/s40596-015-0362-5] [Medline: 26048460]
43. Hameed Y, Al Taiar H, O'Leary D, Kaynge L. Can Online Distance Learning improve access to learning in conflict zones? The Oxford Psychiatry in Iraq (OxPIQ) Experience. Br J Med Pract 2018;11(2):a1114 [FREE Full text]
44. Puspitasari AJ, Kanter JW, Busch AM, Leonard R, Dunsiger S, Cahill S, et al. A randomized controlled trial of an online, modular, active learning training program for behavioral activation for depression. J Consult Clin Psychol 2017 Aug;85(8):814-825. [doi: 10.1037/ccp0000223] [Medline: 28726481]
45. Teshima J, Hodgins M, Boydell KM, Pignatiello A. Resident evaluation of a required telepsychiatry clinical experience. Acad Psychiatry 2016 Apr;40(2):348-352. [doi: 10.1007/s40596-015-0373-2] [Medline: 26122350]
46. Davidson D, Evans L. Virtual study groups and online Observed Structured Clinical Examinations practices - enabling trainees to enable themselves. Australas Psychiatry 2018 Aug 03;26(4):429-431. [doi: 10.1177/1039856218765886] [Medline: 29609477]
47. Pantziaras I, Fors U, Ekblad S. Innovative training with virtual patients in transcultural psychiatry: the impact on resident psychiatrists' confidence. PLoS One 2015;10(3):e0119754 [FREE Full text] [doi: 10.1371/journal.pone.0119754] [Medline: 25794169]
48. Pantziaras I, Fors U, Ekblad S. Training with virtual patients in transcultural psychiatry: do the learners actually learn? J Med Internet Res 2015 Feb 16;17(2):e46 [FREE Full text] [doi: 10.2196/jmir.3497] [Medline: 25689716]
49. Wilkening GL, Gannon JM, Ross C, Brennan JL, Fabian TJ, Marcsisin MJ, et al. Evaluation of branched-narrative virtual patients for interprofessional education of psychiatry residents. Acad Psychiatry 2017 Feb 14;41(1):71-75. [doi: 10.1007/s40596-016-0531-1] [Medline: 26976401]
50. Adeponle A, Skakum K, Cooke C, Fleisher W. The University of Manitoba Psychiatry Toolkit: Development and evaluation. Acad Psychiatry 2016 Aug;40(4):608-611. [doi: 10.1007/s40596-015-0419-5] [Medline: 26443030]
51. Dirlam C, Vallera V, Nelson KJ, Bass D. A psychiatry resident's perspective of the "virtual preceptor" as an electronic medical record clinical education support tool. Acad Psychiatry 2018 Oct 10;42(5):741-742. [doi: 10.1007/s40596-017-0880-4] [Medline: 29322307]
52. Zhang MW, Ho RC, Sockalingam S. Methodology of development of a Delirium clinical application and initial feasibility results. Technol Health Care 2015;23(4):411-417. [doi: 10.3233/THC-150904] [Medline: 25735309]
53. Walsh A, Peters M, Saralkar R, Chisolm M. Psychiatry Residents Integrating Social Media (PRISM): Using Twitter in graduate medical education. Acad Psychiatry 2019 Jun;43(3):319-323. [doi: 10.1007/s40596-018-1017-0] [Medline: 30635806]
54. Kumar PR, Yee A, Francis B. Impact of alcohol withdrawal training program on knowledge, attitude, and perception among healthcare providers in a hospital setting. J Subst Use 2022;27(1):80-85. [doi: 10.1080/14659891.2021.1897696]
55. Shekunov J, Swintak C, Somers K, Kolla BB, Ruble A, Bhatt-Mackin S, et al. The Virtual Couch: a curriculum on the question of the fundamentals of remote psychotherapy-pilot study. Acad Psychiatry 2024 Feb;48(1):52-56 [FREE Full text] [doi: 10.1007/s40596-023-01805-6] [Medline: 37365485]
56. Tapoi C, de Filippis R, Di Lodovico L, Filip M, Fusar-Poli L, Salas DG, et al. 10th EPA Summer School on Research 2021: sharing experience of the first online edition. Inf Psychiatr 2022;98(6):469-474. [doi: 10.1684/ipe.2022.2443]
57. Noori S, Khasnavis S, DeCroce-Movson E, Blay-Tofey M, Vitiello E. A curriculum on digital psychiatry for a US-based psychiatry residency training program: pilot implementation study. JMIR Form Res 2024 May 13;8:e41573 [FREE Full text] [doi: 10.2196/41573] [Medline: 38739423]
58. Kiing J, Feldman H, Ladish C, Srinivasan R, Donnelly CL, Chong SC, et al. International Interprofessional Collaborative Office Rounds (iiCOR): Addressing children's developmental, behavioral, and emotional health using distance technology. Front Public Health 2021;9:657780 [FREE Full text] [doi: 10.3389/fpubh.2021.657780] [Medline: 34055722]
59. Ouanes S, Larnaout A, Jouini L. Use of modern technology in psychiatry training in a middle-income country. Asia Pac Psychiatry 2021 Dec 07;13(4):e12496. [doi: 10.1111/appy.12496] [Medline: 34873857]
60. de Cates AN, Mullin D, Stirland L, Pinto da Costa M, Tracy D. Breaking down barriers: promoting journals beyond the page with open access journal clubs. BJPsych Bull 2024 Apr 01;49(1):1-4 [FREE Full text] [doi: 10.1192/bjb.2024.3] [Medline: 38557559]
61. Blamey H, Harrison C, Roddick A, Malhotra T, Saunders K. Simulated virtual on-call training programme for improving non-specialised junior doctors' confidence in out-of-hours psychiatry: quantitative assessment. BJPsych Bull 2023 Oct;47(5):287-295 [FREE Full text] [doi: 10.1192/bjb.2022.40] [Medline: 36073524]
62. Li L, Ray JM, Bathgate M, Kulp W, Cron J, Huot SJ, et al. Implementation of simulation-based health systems science modules for resident physicians. BMC Med Educ 2022 Jul 30;22(1):584 [FREE Full text] [doi: 10.1186/s12909-022-03627-w] [Medline: 35906583]

63. Collin G, Turner O, Luthra V. Evaluation of doctors' experience of psychodynamic psychotherapy training at Leeds and York Partnership NHS Foundation Trust during the COVID-19 pandemic. Br J Psychother 2023 Jun 05;39(3):425-447. [doi: 10.1111/bjp.12833]
64. Shanley I, Jones C, Reddi N. Medical psychotherapy training and the COVID-19 pandemic. Br J Psychother 2022 May;38(2):338-352 [FREE Full text] [doi: 10.1111/bjp.12719] [Medline: 35601048]
65. Famina S, Farooqui AA, Caudill RL. Early use of telepsychotherapy in resident continuity clinics-our experience and a review of literature. Mhealth 2020 Jan;6:1-1 [FREE Full text] [doi: 10.21037/mhealth.2019.09.11] [Medline: 32190612]
66. Beran C, Sowa NA. Adaptation of an academic inpatient consultation-liaison psychiatry service during the SARS-CoV-2 pandemic: effects on clinical practice and trainee supervision. J Acad Consult Liaison Psychiatry 2021 Mar;62(2):186-192 [FREE Full text] [doi: 10.1016/j.psym.2020.11.002] [Medline: 33288272]
67. Cruz C, Orchard K, Shoemaker EZ, Hilty DM. A survey of residents/fellows, program directors, and faculty about telepsychiatry: clinical experience, interest, and views/concerns. J Technol Behav Sci 2021 Feb 09;6(2):327-337 [FREE Full text] [doi: 10.1007/s41347-020-00164-5] [Medline: 33585672]
68. McCutcheon S. Putting the cart before the horse: outcomes following rapid implementation of telepsychiatry in an outpatient resident clinic. Acad Psychiatry 2020 Dec;44(6):655-658 [FREE Full text] [doi: 10.1007/s40596-020-01316-8] [Medline: 32944873]
69. Nadeem T, Asad N, Hamid SN, Gul B, Aftab R. Experiences from implementing an Lessons from teaching psychiatry trainees at a tertiary care hospital in Karachi, Pakistan. Asian J Psychiatr 2021 Nov;65:102865. [doi: 10.1016/j.ajp.2021.102865] [Medline: 34560566]
70. Knez R, El Alaoui S, Ivarson J, Risö Bergerlind LL, Stasinakis S, Ahlgren AM, et al. Medical residents' and teachers' perceptions of the digital format of nation-wide didactic courses for psychiatry residents in Sweden: a survey-based observational study. BMC Med Educ 2023 Jan 05;23(1):9 [FREE Full text] [doi: 10.1186/s12909-022-03989-1] [Medline: 36604728]
71. Nalan P, Manning A. The juice is worth the squeeze: psychiatry residents' experience of Balint group. Int J Psychiatry Med 2022 Nov;57(6):508-520 [FREE Full text] [doi: 10.1177/00912174221127084] [Medline: 36112941]
72. Elzain M, Murthy S, Omer S, McCarthy G. Reflective practice in psychiatric training: Balint groups during COVID-19. Ir J Psychol Med 2023 Sep 15;40(3):326-329. [doi: 10.1017/ipm.2022.51] [Medline: 36519310]
73. Ranjbar N, Erb M, Tomkins J, Taneja K, Villagomez A. Implementing a mind-body skills group in psychiatric residency training. Acad Psychiatry 2022 Aug 02;46(4):460-465 [FREE Full text] [doi: 10.1007/s40596-021-01507-x] [Medline: 34341965]
74. Chacko E, Vara A, Cheung G, Naskar C, Ramalho R, Bell R. A mindfulness-based cognitive therapy informed virtual psychiatry trainee wellbeing programme: Development and preliminary feedback. Australas Psychiatry 2022 Oct;30(5):663-667. [doi: 10.1177/10398562221119090] [Medline: 35973679]
75. Westcott S, Simms K, van Kampen K, Jafine H, Chan T. Off-script, online: virtual medical improv pilot program for enhancing well-being and clinical skills among psychiatry residents. Acad Psychiatry 2023 Aug;47(4):374-379 [FREE Full text] [doi: 10.1007/s40596-023-01778-6] [Medline: 37101105]
76. Wasser T, Hu J, Danzig A, Yarnell-MacGrory S, Guzman J, Michaelsen K. Teaching forensic concepts to residents using interactive online modules. J Am Acad Psychiatry Law 2020 Mar;48(1):77-83. [doi: 10.29158/JAAPL.003890-20] [Medline: 31753964]
77. Wortzel JR, Maeng DD, Francis A, Oldham MA. Evaluating the effectiveness of an educational module for the Bush-Francis Catatonia Rating Scale. Acad Psychiatry 2022 Apr 07;46(2):185-193. [doi: 10.1007/s40596-021-01582-0] [Medline: 34997564]
78. Jacoby N, Gullick M, Sullivan N, Shalev D. Development and evaluation of an innovative neurology e-learning didactic curriculum for psychiatry residents. Acad Psychiatry 2023 Jun 14;47(3):237-244 [FREE Full text] [doi: 10.1007/s40596-023-01769-7] [Medline: 36918470]
79. Williams JM, Steinberg ML, Wang H, Chaguturu V, Poulsen R, Tobia A, et al. Practice change after training psychiatry residents in tobacco use disorder. Psychiatr Serv 2020 Feb 01;71(2):209-212 [FREE Full text] [doi: 10.1176/appi.ps.201900272] [Medline: 31690223]
80. Trinh N, O'Hair C, Agrawal S, Dean T, Emmerich A, Rubin D, et al. Lessons learned: developing an online training program for cultural sensitivity in an academic psychiatry department. Psychiatr Serv 2021 Oct 01;72(10):1233-1236. [doi: 10.1176/appi.ps.202000015] [Medline: 34106742]
81. Brown TR, Amir H, Hirsch D, Jansen MO. Designing a novel digitally delivered antiracism intervention for mental health clinicians: exploratory analysis of acceptability. JMIR Hum Factors 2024 Apr 03;11:e52561 [FREE Full text] [doi: 10.2196/52561] [Medline: 38568730]
82. Owais S, Saperson K, Levinson AJ, Payne S, Lamont R, Brown MV, et al. Evaluation of the online component of a blended learning electroconvulsive therapy curriculum for psychiatry residents to treat depression in older adults. Acad Psychiatry 2024 Feb 26;48(1):36-40. [doi: 10.1007/s40596-023-01825-2] [Medline: 37493958]

83. Gargot T, Arnaoutoglou NA, Costa T, Sidorova O, Liu-Thwaites N, Moorey S, et al. Can we really teach cognitive behavioral therapy with a massive open online course? Eur Psychiatry 2020 Mar 10;63(1):e38 [FREE Full text] [doi: 10.1192/j.eurpsy.2020.29] [Medline: 32151289]
84. Gratzer D, Islam F, Sockalingam S, Beckett R. Reading of the Week: a continuing professional development program for psychiatrists and residents that Osler would have liked. Can Med Educ J 2022 Mar;13(1):81-85 [FREE Full text] [doi: 10.36834/cmej.72089] [Medline: 35291453]
85. Moore D, Green J, Gallis HA. Achieving desired results and improved outcomes: integrating planning and assessment throughout learning activities. J Contin Educ Health Prof 2009;29(1):1-15. [doi: 10.1002/chp.20001] [Medline: 19288562]
86. Scheeres K, Agrawal N, Ewen S, Hall I. Transforming MRCPsych theory examinations: digitisation and very short answer questions (VSAQs). BJPsych Bull 2022 Feb;46(1):52-56 [FREE Full text] [doi: 10.1192/bjb.2021.23] [Medline: 33752773]
87. Chu K, Sathanandan S. The virtual Clinical Assessment of Skills and Competence: the impact and challenges of a digitised final examination. BJPsych Bull 2023 Apr;47(2):110-115 [FREE Full text] [doi: 10.1192/bjb.2021.112] [Medline: 34937596]
88. Gentry MT, Murray AP, Altchuler SI, McKean AJ, Joyce JB, Hilty DM. Development and implementation of a virtual clinical skills examination in general psychiatry. Acad Psychiatry 2023 Feb 02;47(1):48-52 [FREE Full text] [doi: 10.1007/s40596-022-01691-4] [Medline: 35918600]
89. Usman M, Adamis D, McCarthy G. Perspectives of psychiatric trainees and examiners on the assessment of communication skills during an online clinical examination: a qualitative study. Ir J Psychol Med 2023 Jun 15;41(4):1-7. [doi: 10.1017/ipm.2023.19] [Medline: 37318020]
90. Soll D, Fuchs R, Mehl S. Teaching cognitive behavior therapy to postgraduate health care professionals in times of COVID 19 - An asynchronous blended learning environment proved to be non-inferior to in-person training. Front Psychol 2021 Sep 27;12:657234 [FREE Full text] [doi: 10.3389/fpsyg.2021.657234] [Medline: 34646190]
91. Hewson T, Foster H, Sanderson R. Using socially distanced and online simulation training to improve the confidence of junior doctors in psychiatry. BJPsych Bull 2023 Aug 18;47(4):235-241 [FREE Full text] [doi: 10.1192/bjb.2022.18] [Medline: 35300744]
92. Rakofsky JJ, Talbot TB, Dunlop BW. A virtual standardized patient-based assessment tool to evaluate psychiatric residents' psychopharmacology proficiency. Acad Psychiatry 2020 Dec;44(6):693-700 [FREE Full text] [doi: 10.1007/s40596-020-01286-x] [Medline: 32681418]
93. Ben Ammer A, Bryan J, Asghar-Ali A. The impact of COVID-19 in reshaping graduate medical education: harnessing hybrid learning and virtual training. Cureus 2024 Mar;16(3):e56790 [FREE Full text] [doi: 10.7759/cureus.56790] [Medline: 38650783]
94. Kalayasiri R, Wainipitapong S. Training of psychiatry and mental health in a low- and middle-income country: Experience from Thailand before and after COVID-19 outbreak. Asia Pac Psychiatry 2021 Dec 07;13(4):e12493 [FREE Full text] [doi: 10.1111/appy.12493] [Medline: 34873871]
95. Heldt JP, Agrawal A, Loeb R, Richards MC, Castillo EG, DeBonis K. We're not sure we like it but we still want more: trainee and faculty perceptions of remote learning during the COVID-19 pandemic. Acad Psychiatry 2021 Oct 17;45(5):598-602 [FREE Full text] [doi: 10.1007/s40596-021-01403-4] [Medline: 33594628]
96. Hodges C, Moore S, Lockee B, Trust T, Bond M. The Difference Between Emergency Remote Teaching and Online Learning. EDUCAUSE. 2020 Mar 27. URL: https://er.educause.edu/articles/2020/3/the-difference-between-emergency-remote-teaching-and-online-learning [accessed 2025-01-03]
97. Adipat S. Why web-conferencing matters: rescuing education in the time of COVID-19 pandemic crisis. Front Educ 2021 Sep 22;6:752522. [doi: 10.3389/feduc.2021.752522]
98. Sterling J. 10 Ways Video Conferencing Changed the Education Industry for Good. Usherwood Office Technology. 2024 Mar 18. URL: https://www.usherwood.com/blog/10-ways-video-conferencing-changed-the-education-industry-for-good [accessed 2025-01-07]
99. Ayaz O, Ismail F. Healthcare Simulation: A Key to the Future of Medical Education - A Review. Adv Med Educ Pract 2022;13:301-308 [FREE Full text] [doi: 10.2147/AMEP.S353777] [Medline: 35411198]
100. Slater M, Banakou D, Beacco A, Gallego J, Macia-Varela F, Oliva R. A separate reality: an update on place illusion and plausibility in virtual reality. Front Virtual Real 2022 Jun 27;3:914392. [doi: 10.3389/frvir.2022.914392]
101. Combalia A, Sanchez-Vives MV, Donegan T. Immersive virtual reality in orthopaedics-a narrative review. Int Orthop 2024 Jan 11;48(1):21-30 [FREE Full text] [doi: 10.1007/s00264-023-05911-w] [Medline: 37566225]
102. Desselle MR, Brown RA, James AR, Midwinter MJ, Powell SK, Woodruff MA. Augmented and virtual reality in surgery. Comput Sci Eng 2020 Feb 11;22(3):18-26. [doi: 10.1109/mcse.2020.2972822]
103. Ma M, Saha C, Poon S, Yiu R, Shih K, Chan Y. Virtual reality and augmented reality- emerging screening and diagnostic techniques in ophthalmology: A systematic review. Surv Ophthalmol 2022;67(5):1516-1530. [doi: 10.1016/j.survophthal.2022.02.001] [Medline: 35181279]
104. van Loenen I, Scholten W, Muntingh A, Smit J, Batelaan N. The effectiveness of virtual reality exposure-based cognitive behavioral therapy for severe anxiety disorders, obsessive-compulsive disorder, and posttraumatic stress disorder: meta-analysis. J Med Internet Res 2022 Feb 10;24(2):e26736. [doi: 10.2196/26736] [Medline: 35142632]

105. Ke F, Moon J, Sokolikj Z. Virtual reality–based social skills training for children with autism spectrum disorder. J Spec Educ Technol 2020 Sep 15;37(1):49-62. [doi: 10.1177/0162643420945603]
106. Mental health a human right, but only 1 psychiatrist per 1,000,000 people in sub-Saharan Africa – UNICEF/WHO. UNICEF. 2023 Oct 10. URL: https://www.unicef.org/esa/press-releases/mental-health-a-human-right [accessed 2024-07-01]
107. Ghebrehiwet S, Ogundare T, Owusu M, Harris BL, Ojediran B, Touma M, et al. Building a postgraduate psychiatry training program in Liberia through cross-country collaborations: initiation stages, challenges, and opportunities. Front Public Health 2023;11:1020723 [FREE Full text] [doi: 10.3389/fpubh.2023.1020723] [Medline: 37727607]
108. Mathur A. Let's work to recognize mental health as a universal human right. World Health Organization. 2023 Oct 10. URL: https://www.who.int/timorleste/news/detail/10-10-2023-let-s-work-to-recognize-mental-health-as-a-universal-human-right [accessed 2024-07-01]

Abbreviations

AI: artificial intelligence
CBT: cognitive behavioral therapy
DSM: Diagnostic and Statistical Manual of Mental Disorders
LMIC: low- and middle-income country
MOOC: massive open online course
OSCE: objective structured clinical examination
PTSD: posttraumatic stress disorder
RANZCP: Royal Australian and New Zealand College of Psychiatrist
SRQR: Standards for Reporting Qualitative Research
VR: virtual reality

Edited by B Lesselroth; submitted 26.07.24; peer-reviewed by D Hilty, K Drude; comments to author 30.09.24; revised version received 31.01.25; accepted 25.02.25; published 15.04.25.

Please cite as:
Mohd Kassim MA, Azli Shah SMY, Lim JTY, Mohd Daud TI
Online-Based and Technology-Assisted Psychiatric Education for Trainees: Scoping Review
JMIR Med Educ 2025;11:e64773
URL: https://mededu.jmir.org/2025/1/e64773
doi:10.2196/64773
PMID:

Review

Gender Equality Training for Students in Higher Education: Scoping Review

Claire Condron[1], PhD, MBA, PG Dip Med Ed; Mide Power[2], MPhil; Midhun Mathew[1], MB, BCh, BAO, MCh; Siobhan Lucey[3], PhD, H Dip Ed; Patrick Henn[3]; Tanya Dean[4], BA (Hons); Michelle Kirrane Scott[1], MPharm, MSc Clin Pharm; Walter Eppich[5], MD, PhD; Siobhan M Lucey[3], MClinDent, PG Cert Ed

[1]RCSI SIM, Royal College of Surgeons in Ireland University of Medicine and Health Sciences, Dublin, Ireland

[2]Department of Sociology, School of Social Sciences and Philosophy, Trinity College Dublin, Dublin, Ireland

[3]School of Medicine, University College Cork, Cork, Ireland

[4]Conservatoire, Technology University of Dublin, Dublin, Ireland

[5]Faculty of Medicine, Dentistry and Health Sciences, Department of Medical Education & Collaborative Practice Centre, The University of Melbourne, Melbourne, Australia

Corresponding Author:
Claire Condron, PhD, MBA, PG Dip Med Ed
RCSI SIM
Royal College of Surgeons in Ireland University of Medicine and Health Sciences
26 York St
Dublin, D02 P796
Ireland
Phone: 353 0863004343
Email: ccondron@rcsi.com

Abstract

Background: Despite recent improvements, gender inequality persists within the higher education sector, as evidenced by the proportionally greater number of student and academic leadership positions occupied by male students and staff. Gender equality education and training for students may help to develop awareness, knowledge, and skills among individual students, building capacity to address biases and accelerate culture change in higher education institutions.

Objective: We aimed to identify and explore the existing literature on gender equality training interventions for students in tertiary education, with a particular emphasis on training content, methodology, and outcome evaluation.

Methods: The 6-stage framework developed by Arskey and O'Malley was used to map the literature related to current best practice in gender equality training for students in higher education. Systematic database searches of peer-reviewed literature were carried out and 3142 titles, 333 abstracts, and 52 full-text articles were screened for eligibility with 14 (27%) articles selected for inclusion in this review.

Results: The selected studies detailed a range of pedagogical approaches, including didactic lectures, participatory and co-design workshops, reflective writing, and service-learning, with durations ranging from a single interaction to 1 year. Most articles reviewed did not explicitly state their study aims or research question, and the theoretical underpinnings were generally vaguely described. The longer-term impact of most interventions was unclear, as evaluation metrics seldom go beyond the level of adoption.

Conclusions: This scoping review shows that the literature base for gender equality training for tertiary students lacks coherence, highlighting the need for further work to evaluate its impact. This work provides a foundation for developing training design recommendations.

(JMIR Med Educ 2025;11:e60061) doi:10.2196/60061

KEYWORDS

gender equality; tertiary education; students; training; higher education institutions

Introduction

Background

Higher education institutions (HEIs) can be effective allies in the fight for diversity, inclusion, and gender equality in the education context and in society as a whole. The leadership, academic and administrative staff, and students of HEIs are increasingly mobilized by the United Nations 2030 Agenda for Sustainable Development Goals [1]. The Higher Education Sustainability Initiative [2] recommends that all formal education curricula should feature education for sustainable development. The principles of gender equality are integral to the goals, targets, and indicators of all sustainable development goals and goal 5, "Achieve gender equality and empower all women and girls," is of particular importance. Regrettably, HEIs continue to be organizations that are both gendered and gendering [3]. Gender equality at HEIs is persistently hindered by structural, institutional, and cultural barriers [4].

Recent statistical reports on gender equality data in the European Union show that female students make up most of the undergraduate population in HEIs, yet several published reports highlight ongoing cultural, institutional, and structural barriers inhibiting gender equality from true realization in this sector [5]. Male students remain a majority among postgraduates [5]. In addition, research highlights the gender inequality that exists in student leadership, demonstrating that although women made up 55% of the student body, they represented only 33% of the leadership in student organizations [6].

In Ireland, the undergraduate student population is comprised of approximately equal numbers of women and men. However, in 2016, the Higher Education Authority [7] reported that most of the country's student unions' officers tended to be young, White, and male, an inequality also evident internationally. This gender imbalance has been replicated in senior leadership in tertiary institutions [7]). While much has been achieved since 2016, the subsequent report maintains the recommendation to embed gender equality, and equality more broadly, into teaching, learning, research, and quality assurance processes [8].

The assertion that education and training foster durable change is supported by a substantial body of research emphasizing the transformative potential in both individual and societal contexts. The concept of critical pedagogy by Freire [9] underscores education as a means for empowering individuals to challenge existing inequalities and contribute to social transformation. The United Nations Educational, Scientific and Cultural Organization [1] identifies education as a cornerstone for achieving sustainable development, emphasizing its role in fostering long-term changes in health, gender equality, and environmental sustainability. The European Institute for Gender Equality (EIGE) argues that gender equality training should empower participants to define gender equality principles, identify inequalities, incorporate gender into planning, monitor progress, and assess work from a gender perspective [10]. EIGE defines gender equality training as "A tool and strategy to develop awareness, knowledge, and skills among individuals and to influence organizational processes to promote gender equality and tackle gender-based discrimination" [10].

Coe et al [11] advocate for the integration of equity, diversity, and inclusion training into medical curricula to embed diversity and inclusion as foundational institutional and cultural values. Connolly et al [12] highlight the necessity of training to raise awareness and develop competencies in student leaders, enabling them to proactively tackle gender inequality. These authors assert that such training will have a broader impact on the entire HEI community. Acai et al [13] argue that change must start early in a student's academic career to address and combat gender inequality in the higher education setting.

Preliminary literature reviews [14-16] suggest that gender equality-based training is being conducted in secondary and tertiary education settings, which includes didactic teaching, face-to-face collaboration projects, site visits, case studies, and coaching. To date, a comprehensive review collating and synthesizing the available evidence on gender equality training for tertiary students has not yet been carried out.

Scoping reviews are designed to examine the scope and limits of knowledge within an emerging field. The scoping review approach allows identification and synthesis of all relevant literature regardless of study design and is useful in clarifying key concepts and definitions in the literature. As detailed in our published protocol [17] we adopt the theory of change (ToC) as the theoretical framework to guide the analysis and interpretation of our findings [18]. The ToC provides a deliberate process for analyzing and outlining how an intervention is likely to be effective, who will benefit, in what ways, and the conditions necessary for its success.

The United Nation women's approach to gender equality training is grounded in the ToC recognizing that achieving gender equality requires a long-term institutional commitment, in addition to the development of key competencies for individuals [19]. ToC offers a conceptual foundation for how gender equality can be achieved through training, by emphasizing the interconnected roles of knowledge, desire, and ability as the necessary components for lasting change. It also aligns with the broader goal of embedding gender equality into institutional and cultural norms.

Objectives

Our work aimed to support gender equality in HEIs by working with the leaders of the future (men, women, and nonbinary) to address biases and accelerate culture change. This review is a component of a detailed needs assessment following the Kern's 6-step framework for curriculum development and implementation [20]. Understanding the nature of interventions and approaches currently being used to improve the knowledge, skills, and attitudes of students is an integral step in the design of an effective training program. The findings of this scoping review will inform education researchers, faculty, and academic administrators on the application of gender equality training, pinpoint gaps in the literature, and help identify opportunities for instructional designers and subject matter experts to improve course content.

The primary objectives of this scoping review were to produce a descriptive overview of gender equality training and interventions for students in higher education or postsecondary

education, which will inform curriculum development for skills training in the domain of gender equity. The secondary objectives were (1) to determine the methodology and content of gender equality training delivered to students; (2) to establish the skills and competences that are required by students to promote gender equality; (3) to ascertain how gender equality training is evaluated; (4) to review the extent to which the concept of leadership is included in gender equality interventions; and (5) to establish how gender equality leadership skills are fostered among students.

Methods

Overview

The review was structured using the 6-stage framework developed by Arksey and O'Malley [21], as follows: (1) identifying the research question; (2) identifying relevant studies; (3) study selection; (4) charting the data; (5) collating, summarizing, and reporting the results; and (6) expert consultation. This approach was chosen due to its well-established rigor and effectiveness [22]. Our review protocol has been peer reviewed to ensure the appropriateness and effectiveness of our methods [17].

This review involved the analysis of publicly available empirical research and the production of secondary data; therefore, ethics approval was not required.

Identifying the Research Question

Using the population, concept, context framework [23], the primary and secondary research questions were developed (Textbox 1) in alignment with the previously stated objectives. The population included students in higher education; the concept, gender equality training; and the context encompassed all HEIs.

Textbox 1. Primary and secondary research questions that guide this scoping review.

Primary

- What is the current nature and scope of gender equality training and interventions delivered to students in higher education?

Secondary

- How is gender equality training delivered to students, and what are the key topics included in the intervention?
- What specific skills and competences are taught to students to enable them to promote gender equality?
- How is gender equality training evaluated? To what extent is the concept of leadership included in gender equality interventions?
- How are gender equality leadership skills fostered among students?

Identifying Relevant Studies

Following a preliminary search to identify key terms, the search strategy was designed in consultation with an experienced research librarian. The search terms comprised 3 thematic combinations, including: (1) gender equality training, (2) HEIs, and (3) students, each separated with the Boolean operator AND. Within each thematic combination, search terms were separated using the Boolean operator OR. Wildcards were used to ensure the inclusion of plurals and variation in spelling across the search terms. The search was limited to studies and other sources published between January 2011 and November 2021.

An example of the search strategy used in 2 of the key databases is shown in Textbox 2.

Textbox 2. Search strategy showing the search string for the APA, Psycinfo, and CINAHL databases.

1. TI (gender N2 training OR bias N2 training OR discrimination N2 training OR diversity N2 training OR equality N2 training OR inclusion N2 training OR sexuality N2 training) OR AB (gender N2 training OR bias N2 training OR discrimination N2 training OR diversity N2 training OR equality N2 training OR inclusion N2 training OR sexuality N2 training)
2. TI (gender N2 course$ OR bias N2 course$ OR discrimination N2 course$ OR diversity N2 course$ OR equality N2 course$ OR inclusion N2 course$ OR sexuality N2 course$) OR AB (gender N2 course$ OR bias N2 course$ OR discrimination N2 course$ OR diversity N2 course$ OR equality N2 course$ OR inclusion N2 course$ OR sexuality N2 course$)
3. TI (gender N1 program* or bias N1 program* or discrimination N1 program* OR diversity N1 program* or equality N1 program* OR inclusion N1 program* OR sexuality N1 program*) OR AB (gender N1 program* OR bias N1 program* OR discrimination N1 program* OR diversity N1 program* OR equality N1 program* OR inclusion N1 program* OR sexuality N1 program*)
4. TI (gender N1 awareness* or gender N1 bias or gender N1 equality OR gender N1 inclusion OR gender N1 equity OR sex N1 bias) OR AB (gender N1 awareness* or gender N1 bias or gender N1 equality OR gender N1 inclusion OR gender N1 equity OR sex N1 bias)
5. 1 OR 2 OR 3 OR 4
6. TI ("higher education" OR "third level education" OR "tertiary education" OR university* OR college$) OR AB ("higher education" OR "third level education" OR "tertiary education" OR university* OR college$)
7. TI (Undergraduate$ OR postgraduate$ OR student$) OR AB (Undergraduate$ OR postgraduate$ OR student$)
8. 1 AND 2 AND 3
9. LIMIT 8 to 2010-2021

Systematic searches were carried out by the librarian in 6 databases of peer-reviewed research, including APA PsycInfo, CINAHL, Embase, MEDLINE (Ovid), Scopus, and Web of Science, and 3 additional databases, MedEdPortal, MedEdPublish, and Open Grey, to identify any gray literature that could further inform the review. The search was limited to titles, abstracts, and key words, to optimize the return of articles and sources of evidence with an appropriate focus on the topic in hand. All identified citations were collated in EndNote 20.2.1 (Clarivate Analytics) and duplicates were removed. This EndNote library was exported to Rayyan (Rayyan Systems, Inc) [24] to facilitate collaborative evidence screening.

Study Selection

In line with the recommendations from Peters et al [23], evidence selection was based exclusively on agreed eligibility criteria. These inclusion and exclusion criteria were developed in accordance with the previously stated research questions, building on the elements of the population, concept, context framework (Textbox 3). A reflexive approach was used throughout the selection process to adapt and develop the eligibility criteria in an iterative fashion through discussion within the research team throughout the course of the screening process.

Using Rayyan, CC and MP screened the titles of all sources independently and disagreements were resolved through discussion. The included abstracts were then screened by CC and MP in a similar fashion. The full-text studies were retrieved and CC, MP, SL, PH and TD collectively reviewed the first 10 texts to pilot the eligibility criteria framework. Group discussion was used to further clarify and develop the criteria. The remaining texts were then reviewed by 2 reviewers independently, with any disagreements between these reviewers resolved by discussion. CC and MP acted as third reviewers if a consensus could not be reached.

The reference lists of the included texts were back searched by MM and CC for further relevant studies and sources. Title and abstract screening by MM and CC of articles from all issues of 2 key journals, *Gender, Work & Organization*, and the *Journal of Gender Studies*, published between January 2011 and November 2021, did not yield any additional studies which met our inclusion criteria.

Textbox 3. Eligibility criteria for evidence selection.

Inclusion criteria

- Population: undergraduate or postgraduate students in higher education
- Concept:
  - interventions (eg, campaign or workshop) promoting gender equality awareness or gender equality competences
  - intervention that includes a specific goal or objective to educate or raise awareness of gender equality or to foster competence in gender equality
  - interventions inside or outside of the academic curriculum, for example, project within a module or exercise within a module, or course or program external to course of study
  - articles or sources focusing on the experience of participation in an intervention
  - interventions can also focus on other aspects of equity, diversity, and inclusion training, provided gender equality is included
  - interventions involving gender equality in terms of gender diversity (transgender, nonbinary, and gender diverse students)
  - training in gender equality for student teachers
- Context: higher education institutions
- Publication dates: between January 2011 and November 2021
- Sources:
  - peer-reviewed literature and gray literature (dissertations, websites, conference papers, corporate documents, government reports, preprints, proceedings, research reports, and periodicals)
  - secondary research, that is, literature, systematic or scoping reviews
- Language: studies in English

Exclusion criteria

- Population: primary or secondary students, higher education staff, and general public
- Concept:
  - measuring gender balance in higher education courses
  - strategies to promote equality among applicants to courses
  - experiences of gender or gender inequality among students
  - impact of gender on subject-specific competences, for example, programming or spatial awareness, among others
  - recruitment or retention of female students in academic courses
  - gender equality in health care
  - mentorship programs
  - diversity or equity, diversity, and inclusion programs that do not include a gender aspect
  - gender studies modules in which the goal is to inform on a variety of gender-related theories
  - articles focusing on a whole-of-campus institutional change
  - research on students' attitudes to gender equality unless these form part of a specified intervention
  - exploring the perceptions or experiences of people delivering the intervention
- Context: primary or secondary schools, youth services, and community services
- Publication dates: before January 2011
- Sources: books and book and film reviews
- Language: studies for which no English translation is available

Charting the Data

A standardized data charting tool (Textbox 4) was developed to extract data. This tool was adapted by the research team from the Joanna Briggs Institute template data extraction instrument [23] to align with the review objectives and questions.

Textbox 4. Data charting form to collect extra data and chart data for each paper.

Heading

- Study details (authors, year, title, and citation)
- Location
- Context
- Type of source
- Study aims or intervention aims
- Participants details and sample size
- Research questions addressed
- Study design
- Intervention style and duration
- Intervention content
- Intervention methods
- Outcome evaluation measures
- Key findings

Miscellaneous

- Interesting observations
- Gender equality theories employed
- Incorporation of gender equality and leadership
- Incorporation of intersectionality
- Incorporation of transgender and gender diverse inclusivity

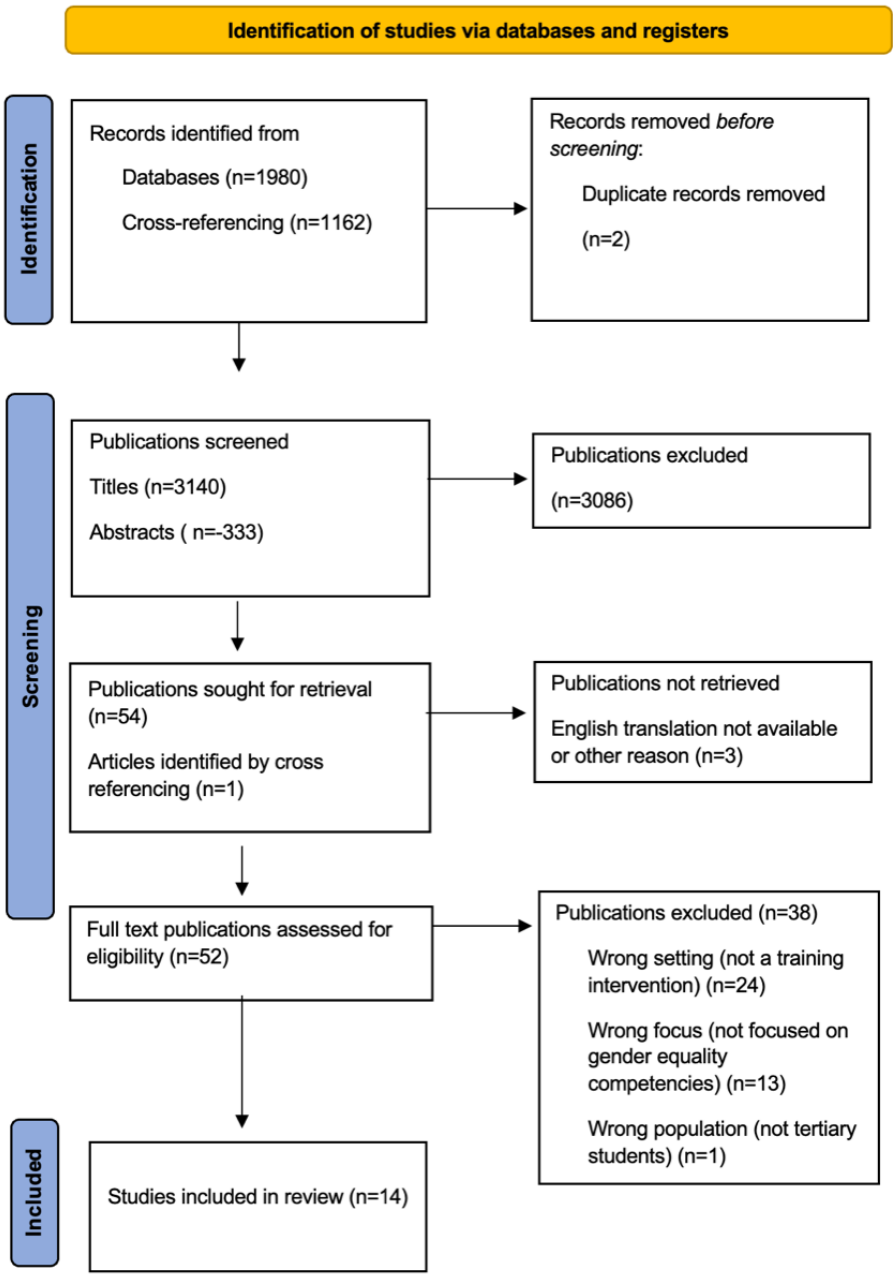
The charting form was expanded and refined in an iterative fashion by all authors during the full-text screening process. Data extraction was carried out by 2 reviewers working independently, with a single finalized form for each study agreed through collaborative discussion and communication.

Results

Collating, Summarizing, and Reporting the Results

The study selection process is presented in a flow diagram (Figure 1) as per the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) in Multimedia Appendix 1 [25]. On the basis of the database searches, and the subsequent hand search of other sources as described in the Identifying Relevant Studies section, 3142 titles were screened. This yielded 333 abstracts for further screening. In total, 14 full-text articles were selected for inclusion in the review.

Figure 1. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) diagram detailing the process for identification and selection of articles included in this study.



Study Characteristics

Basic descriptive analysis is the most suitable approach to data analysis given the exploratory nature of a scoping review [23]. In total, 13 of the publications included in the final sample were peer-reviewed academic articles and 1 was a conference paper. The details of each article, including the country in which the study took place, the participants, aims, methodology, evaluation, and outcomes, are summarized in Table 1. Much (5/14, 36%) of the work in this field has been conducted in the United States and in Europe, with Spain (3/14, 21%) and Sweden (1/14, 7%) contributing. A further 3 (21%) studies were conducted in Turkey, with a single study located in each of Taiwan (1/14, 7%) and South Africa (1/14, 7%). In keeping with the agreed search strategy for this review, the included studies were published from 2011 to 2021. The published studies demonstrate a growing interest in the topic, as 57% (8/14) of the studies were published in the latter 3 years of the selected period (2019 to 2021).

Table 1. Summary of included studies.

Study	Country	Population	Aims	Methodology	Evaluation	Outcomes
Shields et al [26], 2011	United States	N=118; female: n=62, 52.5%; male: n=53, 44.9%; nonresponders: n=3, 2.5%; mean age: 19 y	To evaluate game-like simulation in teaching the nature and consequences of unconscious biases and stereotyping, which underlie gender inequity.	Experimental group divided into 2 teams, who participated in WAGES[a]–Academic, a game that simulates academic career progression with and without advantages. Separate control group played Chutes and Ladders; 3 phase study; pretest evaluations, immediate and delayed posttest evaluations; Kolb's experiential learning model used [27].	Neosexism scale at pretest; KGE[b] scale at all 3 stages; single open-ended question in final evaluation, "Since you played the game, have you thought about issues or made observations that you might not have before? If so, what are they?"	No difference in sexism scores among WAGES–Academic participants; no difference in KGE scores at baseline. Significant increase in KGE scores for intervention (both teams) versus control. KGE scores remained higher for intervention group at third evaluation; 24 participants indicated that they had thought about or made observations related to issues raised by WAGES–Academic; 15 noted gender bias, 5 wanted to learn more, 4 were unsure.
Kennedy et al [28], 2011	Turkey	N=546 students taking 2 introductory courses in Sociology and Psychology; gender: not reported; age: not reported	To assess the merit of showing movies as critical pedagogy; to develop student interest in serious social issues, including gender equality, and encourage critical agency in society.	In total, 3 cohorts of students watched 3 different movies over 3 successive academic terms; no comparator or control group; post-movie discussion undertaken	Postevent analysis of 3 student essays completed after watching the movies; 14-item survey completed by 112 students	Student essays indicated a development of consciousness in 3 critical subject matters, the death penalty, gender inequality and prejudice. Students rated the utility of watching movies favorably concluded that "movies help the students build critical perspectives on the 'social' through interest development on sociological topics;" 105 students reported development of a high level of critical perspective or a critical perspective; 7 students indicated that they had not developed a critical perspective.
Falk et al [29], 2012	Sweden	N=9 students participated in a gender module; gender: not reported; age: not reported	To evaluate Euro-Education: Employability for all, a multi-center project in which 4 course modules relating to different aspects of employability were developed: gender, age, disability, and ethnicity.	The module consisted of tutorial groups, lectures, tasks, and seminars. Students used a variety of electronic learning applications. Activities included a critical review of the labor market, an "exchange" in a university Occupational Therapy Department, and development of a "course" to address gender impact as the final assignment.	Continuous evaluation by students and teachers as part of feedback process throughout; postcourse oral feedback; questionnaire; overall assessment by EU[c] Commission (method not stated)	Students' evaluation was positive about increasing awareness and knowledge of gender theory, its application to employability, and development of new skills; technical aspect was challenging at times; teachers recommended precourse training on the electronic learning applications; the project received a global score of 8/10 from the EU Commission.
Case et al [30], 2014	United States	Study 1: n=177, 80% female; study 2: n=131, 71% female	To examine the effectiveness of 2 interventions (privilege list handout and testimonial video), coupled with reflective writing, in raising awareness of heterosexual (study 1) and male (study 2) privilege.	Study 1: participants assigned to control and 2 comparator groups, who received either a privilege list or watched a video); study 2: same method with different content	Study 1: pre- and postsurvey with 3 scales (heterosexual privilege awareness, internal and external motivation to respond without prejudice) and a reflective summary; study 2: pre- and postsurvey with 5 scales (male privilege awareness, modern sexism, hostile and benevolent sexism, and the motivation scales as above) and reflective summary	Study 1: Neither intervention had a more significant impact than the other. The authors indicated that both techniques were helpful in enhancing learning about heterosexual privilege; study 2: The video intervention increased male privilege awareness, but the handout did not. The authors concluded that further research is required to target aspects of sexism which neither intervention addressed.

Study	Country	Population	Aims	Methodology	Evaluation	Outcomes
Abrams et al [31], 2016	United States	N=13; male undergraduate (n=10, 77%) and graduate (n=3, 23%) engineering students invited to participate. One withdrew, one deemed not suitable, one went on placement for second semester; 10 students participated in full program	To evaluate the AWE[d] program, which aims to improve women's retention in engineering fields by promoting allyship among student population. Content focused on gender equity, implicit bias, micro aggressions, and sociocultural conversations.	Participants completed the 1-y AWE program consisting of initial training (workshops, seminars, and speakers) over 4 d; up to 5 hr. per wk. in semester 1 spent on reflection, further training sessions and planning outreach activities, which were then delivered to 447 students and staff on campus; Broido's Model of Allyship was used as framework for training design [32].	Authors intended to assess using a mixed methods approach, including qualitative and quantitative techniques; pre- and postsurveys, focus groups and interviews; this paper presents the results of the presurvey and 2 free-text response questions	After the initial 4 d training, most participants agreed that they could identify and explain microaggressions, knew how to address bias and discrimination, and were comfortable to do so; participants gained insight, awareness, confidence, and new perspectives from attending the training; participants felt they would change their behaviors, have increased awareness, have more confidence to speak up, become more inclusive daily, and become a role model for their peers.
Freedman et al [33], 2018	United States	N=143 college students; female: n=65, 45.4%; male: 77, 53.8%; nonresponder: 1, 0.7%; mean age: 20 y (a second study involved high-school students, which is not reported here)	To examine the impact of experiencing an "aha" moment about assumptions about women in science on subsequent attitudes toward women in science.	Randomized controlled trial in which participants play a logic-puzzle game that is won by making a realization about a character. In the control game, the character is a professor; in the intervention game, the character (a scientist) is a woman. Numbers of participants in each group not reported.	Questionnaire included a monetary allocation task in which participants had to allocate US $500 among 14 college organizations, 2 of which supported women in STEM[e]; shortened version of the Attitudes toward Women in Science Scale; the ambivalent sexism inventory	Most students who used gendered pronouns assumed that nongendered scientist characters were men; contrary to hypotheses, the intervention condition did not increase positive attitudes toward women in science, decrease sexism, or increase donations to women in STEM organizations.
Altınova et al [34], 2016	Turkey	N=65 social-work students; female: n=39, 65%; 26 male: n=26, 40%); age: not reported	To examine the impact of the HREP[f] on improving the gender perceptions. The HREP aims to raise social consciousness concerning human rights violations that women encounter.	Pretest posttest design with experimental group (n=32) participating in HREP and control group (n=33) who did not participate; participants engaged in narration, role play, case study, and problem-based learning during 12 two-h training sessions.	25-item Gender Perception Scale	GP[g] levels of the experimental group were increased; no change in GP levels for control group; concluded that the HREP is effective
Segovia-Pérez et al [16], 2019	Spain	N=50 female students from a variety of academic backgrounds over 2 consecutive years with 25 students in each cohort; mean age: 22 y	To evaluate a women's leadership program for university students which sought to ensure acquisition of skills, competencies, and tools for leadership by the participants, in addition to increased self-confidence.	The program involved 24 h of classes, a case study, visits to companies, and the European Parliament, and a coaching system. Subjects included personal branding, communication, networking, public speaking, negotiation, leadership techniques, and business management; the second cohort also completed a leadership test.	A self-administered survey with 9 questions using a 10-point Likert scale related to the program, prior knowledge, potential future impact, and a global evaluation; 3 focus groups with 8 participants in each group; 4 individual informal interviews	The participants rated the course highly, with scores of 8 out of 10 in all areas of the survey, other than prior knowledge; the qualitative data indicated an improvement in specific tools and skills, and a positive change in attitudes and self-confidence of the participants.

Study	Country	Population	Aims	Methodology	Evaluation	Outcomes
Toraman and Özen [14], 2019	Turkey	N=433 students in the Faculty of Education; female: n=319, 73.7%; male: n=144, 3.2%; age: not reported	To determine the opinions of the participants regarding gender equality, and to compare their opinions before and after taking a compulsory gender equality course; later, described third goal to explore association with participant variables.	Students participated in a training course over one semester, which included the following topics; the concept of gender, sociology of gender, gender and family, gender and religion, gender and language, gender and media, gender and body images, gender, work life and labor, feminist movements, and social change. Limited description of learning and teaching activities.	Descriptive study with pre- and postintervention administration of the Gender Equality Scale, with validation of the scale for university students included in the methods; data were also collected on background participant variables; logistic regression used to assess the extent to which background variables affected participants' opinions	Initially, most participants did not believe that men were superior to women. Unexpectedly, after the course, participants were more likely to develop the opinion that men are superior to women, whereas there was no impact on opinions relating to women being dependent on men; men were more likely to believe that they are superior, and that women are dependent on men; significant variables in the regression analysis included academic background, father's education status, newspaper reading by family members, and location of family home.
Gorrotxategi et al [35], 2020	Spain	N=64 social education students; female: n=52, 81%; male: 11, 17%; and 1 did not identify themselves as male or female; mean age: 20 y	To measure the students' knowledge about transgender people, and the attitudes of students toward gender and transgender people, before and after an education program.	Training was based on the "Creative Factory" intervention model and consisted of a weekly training session on gender and transgender learning over a 4-mo period. The goal of the Creative Factory model is to "enable students...to analyze social realities to generate discussion and innovate ideas to design successful practices."	Pre- and postintervention questionnaires; 12-item short version of the Gender and Transphobia Scale; transgender knowledge assessed with single item scale	Level of transgender knowledge was increased by the intervention (P<.05); improvements were noted in gender bashing and transphobia dimensions but these were not statistically significant; the data suggests that men had higher levels of gender bashing attitudes (P<.05) and transphobia, but this latter difference was not statistically significant.
Liao and Wang [36], 2020	Taiwan	N=82 medical students; gender: not reported; age: not reported	To investigate whether the integration of the gender perspective into literature studies would create any difference among students in gender awareness and critical thinking.	Intervention entailed twice weekly literature study sessions over 15 wk, self-study, electronic discussion between both groups. Experimental group received gender perspective training and were introduced to gender-related terms to facilitate discussion. Through literature, this group were encouraged to consciously reflect on traditional and socially constructed gender norms.	Quasi-experimental study with a control group (n=41) and experimental group (n=41); both groups completed the CTDA[h] and the Chinese version of the Gender Awareness Scale pre- and postintervention	With respect to gender awareness, the findings show that following the integration of the gender perspective into literature studies, medical university students had significantly higher posttest scores for "public gender consciousness" and "private gender consciousness;" regarding critical thinking, they also had significantly better posttest scores in "systematicity and analyticity," "maturity and skepticism," and "inquisitiveness and conversance."

Study	Country	Population	Aims	Methodology	Evaluation	Outcomes
Locke et al [37], 2021	United States	Male undergraduate and graduate STEM students; this paper includes the reflections of a male engineering student ally, and a research adviser	Update of Abrams et al [31]; the success of the AWE program led to the development of a leadership course that has been offered in the College of Engineering every semester since autumn 2016.	A total of 14-week course incorporating videos, workshops, case studies, and group discussions; participants engaged in: information gathering to develop awareness of gender equity challenges in engineering; meaning making to examine personal biases; and, contextual application of strategies that promote inclusive engineering climates.	Pre- and postcourse survey with Likert-type scales	Increase in self-reported efficacy of participants; this course helped the male ally to "develop the skills and mindset needed to make...difficult conversations productive." It gave him "the tools needed to identify the various kinds of situations that contribute to a hostile environment and how to better diffuse them;" created "an enhanced supportive environment" in a research laboratory, as reported by a female research adviser.
Bosch et al [38], 2021	Spain	N=19 final year students taking a sociology of gender course female: 15, 79%; male: 4, 21%; age: not reported	To evaluate a SL[i] project, in which students deliver workshops in high schools on gender and technology. The optional project within the sociology of gender course was intended to enhance student understanding of topics covered in the course.	For each workshop, students created a presentation with interactive activities, and showed a video. Their assessment consisted of an oral presentation to the rest of their university class and a self-reflection; 13 workshops were carried out over 4 y.	Qualitative analysis of 16 university student self-assessment reports; quantitative survey data from high-school students (n=284) and teachers (n=13) on a range of items, such as usefulness, alignment with student need, evaluation of pedagogical tools, opportunities for participation and a global evaluation. This quantitative data are not relevant for this review but the abovementioned description is included for completeness.	All students were satisfied with the experience and the course grades for these students were higher than the class average. The authors felt that these students went "far beyond the curriculum" and "reached a deep understanding of the multiple dimensions on the topics of gender and technology." Other themes which emerged included social transformation through SL, acquisition of competences, such as improved communication, putting empathy into practice, empowerment and "professionalization" of the participants.
de Villiers et al [39], 2021	South Africa	N=15 "student leaders" from various faculties; all male; age range: 20-25 y	To evaluate the OMC[j] intervention for university settings. OMC aims to encourage development of equitable relationships between men and women, thereby preventing gender-based violence and HIV transmission.	A case study design describing the adaptation and implementation of OMC, in which 5 participatory workshops were conducted. Content was related to personal values, belief systems, societal gender-based norms, rape and consensual sex, courage and bystander intervention, and healthy relationships.	Qualitative data collection methods, including pre- and postintervention focus groups, discussion content and researchers' field notes documented during the intervention workshops, participant reflections were collected after each workshop via open-ended questions and 5 semistructured interviews were conducted 6 mo after the intervention; thematic data analysis was used.	The data demonstrated "increasing accountability" on the part of the male student leaders to prevent sexual violence. The authors note that critical engagement and dialogue on sexual violence is shown to shift key norms on gender equality, on being a man and reflection on their role in preventing sexual violence; this work resulted in the development of the Men with Conscience intervention; a 6-h adapted intervention based on OMC.

[a]WAGES: Workshop Activity for Gender Equity Simulation.

[b]KGE: knowledge of gender equity.

[c]EU: European Union.

[d]AWE: allies for women engineers.

[e]STEM: science, technology, engineering, and mathematics.

[f]HREP: Human Rights Education Program for Women.

[g]GP: gender perception.

[h]CTDA: Critical Thinking Disposition Assessment.

[i]SL: service-learning.

[j]OMC: One Man Can.

Settings

All these educational programs (N=14) took place at individual tertiary academic institutions. One involved collaboration with secondary schools within the same country requiring participants to present to school students, 2 programs incorporated workplace visits, and 1 used an online discussion forum. Most studies did not provide information on how the program was financially supported. Some of the training interventions were undertaken as official curricula to which the participants were enrolled as part of their required studies. Most studies recruited volunteers, one via a competitive recruitment process and another accepted nomination from teachers. One study reported that learners were incentivized to participate in the training experience with a US $300 book token per term [31].

Participants

All studies provided some demographic information on participants and the number of participants in the studies ranged from 9 to 546, with an average of 134 participants. With respect to gender of the participants, some of the interventions specifically targeted either women [16] or men [31,37,39]. In the remainder of the studies which reported participant gender (7/14, 50%), the majority were female. When reported, the age range was typical for higher education as most participants were in their early twenties. Most studies involved participants from single faculties. There was a slight predilection for interventions involving students in education and the social science programs [14,28,34,35,38]. There was also representation from health sciences, including medicine [36] nursing [39] psychology [26,30] and occupational therapy [29], in addition to engineering faculties [31,37]. In 2 (14%) studies, the participants were from a variety of academic programs: Business, Social Sciences, Law, Engineering, and Architecture [16,37]. Freedman et al [33] recruited participants from student accommodation. One study did not provide information on participant specialty.

Content

The works reviewed were from countries with considerable variation in political-cultural contexts, civil liberties, and personal freedoms and thus the content of the courses was widely varied reflecting local needs.

The language used to describe content themes was also disparate. The main content areas covered in the training programs are listed in Table 2. In total, 5 (36%) of the studies from this heterogeneous group explored the impact of courses, modules, or programs, which directly aimed to increase awareness of gender and gender-based issues or enhance gender equality knowledge [14,29,34,35,37]. Another group of studies sought to enhance knowledge and skills indirectly through another medium, such as movies [28], testimonial videos [30], games [26,33], and literature studies [36]. In addition, 2 (14%) studies reported on the same initiative, an allyship program in a college of engineering that encompassed gender equity issues [31]. A subsequent paper [37] described the extension of the initial project. A single study was identified which related specifically to a leadership program for women [16]. The final study related to a service-learning project on gender and technology [38].

Table 2. Content themes identified and the frequencies of inclusion in the individual curricula from studies examined in this study (N=14).

Theme	Frequency, n (%)
Allyship	2 (14)
Gender awareness	2 (14)
Gender barriers	1 (7)
Gender equality	5 (36)
Implicit bias	5 (36)
Intersectionality	1 (7)
LGBT[a] marginalization	2 (14)
Male privilege	1 (7)
Microaggressions	1 (7)
Stereotypes	1 (7)

[a]LGBT: lesbian, gay, bisexual, and transgender.

Gender Equality Leadership Skills

The concept of leadership featured explicitly in 2 (14%) of the included articles [16,37]. Both publications detailed how gender equality leadership skills were fostered among the participants. The concepts of social identity, privilege, microaggressions, and implicit bias were explored by Locke et al [37] in a leadership program for male student allies. In contrast, Segovia-Pérez et al [16] ran a leadership program for female students which covered topics such as personal branding, communication, networking, public speaking, negotiation, leadership techniques, and business management.

Study Design: Inputs and Activities

The most common (4/14, 29%) study type was an interventional mixed methods pre-post design, wherein a single group of participants completed self-assessments of confidence, ability, and knowledge before the intervention and then participated in a module or workshops. The learners completed the same assessments after the intervention in addition to participating in focus groups or interviews [14,26,31,35]. Interventional

studies using pre- and postintervention questionnaires were the next most common (3/14, 21%) study design [16,30,37]. Moreover, 2 (14%) studies used a quasi-experimental design [36,39], and 1 (7%) used a randomized controlled trial design [33]. The final pool of articles comprised 5 (36%) qualitative studies using after event analysis only ranging from educators' reflections (2/14, 14%) to analysis of student diaries or essays (2/14, 14%) [28,38] and an after event questionnaire with oral feedback (1/14, 7%) [29].

Delivery Methods

Methods of delivery varied widely. Most studies (11/14, 79%) were described as longitudinal events that occurred over multiple sessions, ranging from several months to a year. One study consisted of 15 European Credit Transfer and Accumulation System (ECTS) where 1 ECTS is equal to between 25 and 30 hours. In addition, 2 (14%) studies described single event game base interventions. The format of teaching was also wide-ranging and included lectures and workshops, guest speakers, small group discussions, workplace visits, role play, videos, movies and literature analysis, student presentations, and online discussion.

Assessment and Evaluation: Outcomes

Most studies reported no assessments of learner competency after training. Most studies (8/14, 57%) measured learner outcomes via survey instruments examining perceptions, sympathies, or confidence. In total, 2 (14%) studies assessed outcomes with participant focus groups. Moreover, 2 (14%) studies measured learning by thematically analyzing reflection writing or diaries. No study reported on transfer of learning.

Significant increase in knowledge of gender equity scores [26], gender perception scores [34], and higher posttest scores for "public gender consciousness" and "private gender consciousness" [36] were demonstrated for intervention participants versus control.

Students self-reported the development of critical perspective [28], increased male privilege awareness [30], increased awareness and knowledge of gender theory [29], and increased level of transgender knowledge [35]. Students agreed that they could identify and explain microaggressions and knew how to address bias and discrimination [31].

In qualitative studies, data from focus groups with participants following training indicated "increasing accountability" on the part of the male student leaders to prevent sexual violence [39], and a positive change in attitudes and self-confidence [16].

Unexpectedly, participants from a course in the faculty of education in a Turkish university were more likely to hold the opinion that men are superior to women postintervention and there was no impact demonstrated on opinions relating to women being dependent on men. Regression analysis included academic background, father's education status, newspaper reading by family members, and location of family home [14]. In addition, contrary to the study hypotheses, a logic-puzzle game intervention from the United States did not increase positive attitudes toward women in science, decrease sexism, or increase in-game donations to women in science, technology, engineering, and mathematics organization [33].

Theoretical and Conceptual Foundations

The array of theoretical and conceptual foundations cited by the authors is displayed in Table 3. Most studies (12/14, 86%) introduced and discussed theoretical underpinnings as a reason to provide training and less so for instructional design and program content. Furthermore, there was little or no reference to educational learning theories or pedagogical approaches when designing learning objectives, delivery methods, or evaluation and assessment of the programs. Exceptions include the use of Broido's model of allyship [32] by Abrams et al [31] and Kolb's experiential learning model [27] by Shields et al [26].

Table 3. Underpinning theories.

Study	Gender equality theories used
Shields et al [26], 2011	Work-family balance, salary, mentoring, workplace climate, and token status
Kennedy et al [28], 2011	Critical teaching
Falk et al [29], 2012	Gender theory and gender as a social construction
Case et al [30], 2014	Privilege, modern sexism, hostile and benevolent sexism, and prejudice
Abrams et al [31], 2016	Gender equity, implicit bias, micro aggression, and sociocultural conversations
Altınova et al [34], 2019	Gender equality from a human rights perspective
Freedman, et al [33], 2020	Gender stereotypes; implicit and explicit gender bias; social psychology—challenging the assumption of invulnerability to bias
Segovia-Pérez et al [16], 2019	Stereotypes and social role theory. Inclusive and equitable quality education, gender equality, and empowerment
Toraman and Özen [14], 2019	Social cognitive theory; gender and its inequalities originate from learned or taught behaviors
Gorrotxategi et al [35], 2020	Gender bashing, transphobia, and genderism
Liao and Wang [36], 2020	Socially constructed gender norm, patriarchal system, gender politics, and ideology
de Villiers et al [39], 2021	No explicit gender equality theories

Discussion

Principal Findings

This scoping review explored and mapped the evidence related to gender equality training and interventions for university students using a ToC lens. We described inputs and activities in terms of content areas, delivery modes, participants, and settings. We described outcomes and impact in terms of evaluation and assessment.

It was not possible to measure the long-term systemic impacts of the interventions due to limited postintervention follow-up and the absence of transfer of learning assessments. The studies collectively indicate potential for improved gender equity knowledge and leadership skills in diverse educational contexts and increased capacity to address microaggressions, implicit bias, and gender-based discrimination. However, it is important to note that self-reported data are typically considered low-quality evidence for evaluating the effectiveness of teaching interventions [40], especially if the questions focus primarily on learner satisfaction rather than assessing actual knowledge acquisition or behavioral change. Improvement or impact on skills to support gender equality is less frequently observed or demonstrated as an outcome in the literature we reviewed, as is any change at the organization level. Critical assumptions underpinning these studies included the willingness of participants to engage actively with the material and the alignment of local cultural contexts with course content.

The study by Toraman and Özen [14] highlights the unintended reinforcement of gender stereotypes, emphasizing the impact of cultural and political contexts. Similarly, the adverse effects of the Freedman et al [33] training intervention on explicit attitudes toward women in science serve as a cautionary tale, highlighting the importance of educational research in designing interventions that reduce biases without provoking defensiveness.

The usefulness of the studies we identified to inform curriculum design and provide actionable insights for educators is constrained by a lack of theoretical grounding, limited long-term follow-up, and reliance on self-reported data. This aligns with the findings of Guthridge et al [41] who, in their systematic review of interventions, aimed at enhancing gender equality and questioned whether the interventions they examined effectively led to substantive change. The EIGE understands developing competence in gender equality as being able to identify and change gender stereotypes and gendered roles [10]. Thus, gender equality training should enable and empower participants to (1) define and understand gender equality principles, (2) identify gender inequalities in their field, (3) incorporate gender in their planning and policy implementation, (4) monitor progress, and (5) review and assess their work from a gender perspective [10]. We have incorporated these concepts with developments from the wider gender equality training literature in subsequent sections and used a ToC to provide a lens through which we can consider and synthesize our findings. From this synthesis, combined with our own experience as educators and consultation with experts in gender equality, we have created recommendations for both practice and future research to help close the gap between the current practice and desired outcomes.

Instructional Design

Pedagogical frameworks should guide the design and implementation of educational activities. These frameworks are based on established principles of teaching and learning and provide a structured approach to designing effective and engaging learning experiences. Without a solid theoretical foundation, it can be challenging to determine the best instructional strategies to use, how to sequence and structure the content, and how to assess whether learners have achieved the desired learning outcomes. A lack of underpinning theory in the instructional design of training can lead to an inconsistent approach and potentially ineffective training. By grounding an approach in theory, educators aiming to upskill university students in gender equality skills can design training that is more likely to be effective in achieving the EIGE 5 categories of learning outcomes. Interventions to combat biases are particularly effective when they involve active participation rather than passive learning [26,33].

Diversity of Content

Ensuring diversity of content to cover knowledge, skills, and attitudes is critical for designing effective learning experiences. Not only must learners develop a deeper understanding of the subject matter and its various dimensions, but they must also develop the skills and attitudes to bring about change in behaviors [19]. The terminology used to describe content themes varies widely in the reviewed papers. The key content areas addressed in the training programs include knowledge of gender and gender-based issues, allyship, leadership, and bias. Communication skills, teamwork, and leadership skills training can support learners to become more adaptable and better equipped to promote a gender equality agenda.

Leadership

Within the body of evidence included in this scoping review, the concept of leadership appears underrepresented in most gender equality training initiatives. One notable exception is the women's leadership program for female university students described by Segovia-Pérez et al [16]. The authors demonstrated that leadership training enhanced learner self-confidence and their view of their own capacities, providing tools and guidelines for professional communication and personal branding. Enhancing leadership skills were also considered by Locke et al [37] who described a course that focused on gender equity and the practice of inclusive leadership for male allies in the STEM fields. The authors stated that this course would continue to be offered, with plans to explore changes in student behaviors in addition to investigating potential trends or differences in students in varying engineering majors and academic careers stages. Furthermore, a similar workshop for new engineering staff at a prominent state company has also been offered, thereby demonstrating the potential of gender equality training initiatives to promote linkages between higher education and industry, and indeed society more widely.

Intersectionality

Case et al [30] recommends an intersectional approach for future work: "Future research considering several forms of privilege within the same intervention would provide information about the learning process." They conclude that "teaching and learning about LGBT psychology through an intersectional lens allows students from a variety of backgrounds to connect seemingly irrelevant systems of oppression and privilege to their own social identities and social locations." These authors suggest that training which addresses intersections of identity to allow further understanding of sexism, heterosexual privilege, and male privilege will further raise awareness of not only privilege but also the complexities of identity and the matrix of oppression [30]. Gorrotxategi et al [35] demonstrates that interactive training in gender education using the Creative Factory methodology creates a context of reflection and knowledge generation and promotes a significant improvement in knowledge about transgender and slight improvements in transphobia. Bosch et al [38] report on a service-learning intervention, using community engagement pedagogies that facilitate learners to volunteer with an agency and engage in reflection activities to deepen understanding. After the experience in schools with a great diversity of social backgrounds, learners were motivated to reflect on intersectionalities, such as origin or class.

Delivery, Implementation, and Evaluation

The nature of gender equality training interventions identified for this review are heterogeneous. This can be seen in terms of the format of interventions, and whether gender equality knowledge and skills were directly seen in terms of the population, concepts, and contexts. Delivery, implementation, and evaluation are critical components of any effective learning program, and standardization of these components can help ensure that training is effective, efficient, and consistent and will allow evaluation of its impact. Impact is best identified through a range of evidence that provides robust verification for enhanced knowledge, behaviors, and practices, Furthermore, researchers are advised to consider both quantitative and qualitative forms of evidence. A longitudinal approach to evaluation is recommended in preference to singular measurements to ensure that the full value of an intervention can be captured over time and skills retention is measured. Program designers may also wish to consider evaluation at both participant and institutional level.

Faculty Development

Verge et al [42] stated that it is crucial to enhance the teaching staff's required qualifications, account for institutional opposition to gender-related change, and implement monitoring and evaluation systems. These steps are important to ensure that training outcomes are regularly assessed and improved when integrating gender equality training into higher education programs.

Compulsory Training

The United Nations Educational, Scientific and Cultural Organization [1] recommends that gender equality should be integrated into the whole program of faculties of education, rather than being included as a course. Toraman and Özen [14] found that the gender equality training offered as a compulsory-elective course in the faculty of education did not produce the expected results for their students. De Villers et al [39] noted that student leaders who volunteer may have leadership value systems and may not reflect the student body. Abrams et al [31] indicated plans to embed their program in the engineering curriculum and participants would be recruited as future teaching assistants to enhance sustainability. It has been demonstrated that elective courses in general receive more favorable responses than the required courses using both scaled response evaluation formats and open‐ended response evaluation forms [43] and these observations have relevance for gender equality training. When participation is voluntary, those who opt in are likely to have a preexisting interest, motivation, or alignment with the course's objectives, which can positively influence their engagement and satisfaction. Self-selection bias may skew evaluations and limit the generalizability of findings regarding the course's effectiveness. In addition, voluntary participants are typically intrinsically motivated, which enhances their willingness to engage deeply with course material and reflect on their learning. In contrast, students who are required to take part in such courses may approach the material with skepticism, resistance, or apathy, potentially lowering their engagement and perceptions of the course [44].

Mandating participation can broaden the reach of training; however, course evaluations should account for differences in participant motivations. Care should be taken to ensure that the course's status as elective or required is considered for evaluation purposes [43]. Combining qualitative and quantitative assessment methods can provide a more nuanced understanding of the course's impact, particularly on participants who might initially be less engaged. Furthermore, longitudinal studies tracking behavior change or attitude shifts after training can help determine the broader efficacy of these programs beyond immediate satisfaction ratings [45].

Lau et al [46] proposed that a key reason for the lack of progress in gender parity in organizations lies in the predominance of empirical research focusing on the causes and manifestations of gender inequality, while insufficient attention has been given to exploring solutions. To advance gender equality, they argue that a paradigm shift from problems to solutions is critical and urgent.

Significant research is required to bring gender mainstreaming to higher education. The Swedish Secretariat for Gender Research [47] distinguishes between gender-mainstreamed teaching as a pedagogical practice and the integration of gender and gender equality knowledge into the subject content. They call for resources for research-based pedagogical development for the implementation of teaching activities.

ToC Framework

Overview

Applying the ToC framework can enhance the practical value of our findings, offering a structured approach for educators to better understand the pathways through which gender equality

training can lead to desired outcomes. Viewed through the ToC lens, the findings from this review demonstrate the importance of aligning gender equality training programs with a clear, structured pathway from inputs and activities to outcomes and impact. To achieve transformative change in attitudes, behaviors, and institutional practices, the following elements emerge as critical:

Inputs and Activities

Effective programs must begin with carefully designed interventions rooted in robust pedagogical frameworks. These interventions should actively engage participants, address intersectional perspectives, and incorporate diverse methodologies, such as service-learning or active, immersive simulation-based activities, and leadership-focused training.

Outputs

Outputs should be aimed at improving participants' awareness, knowledge, and attitudes toward gender equity. Training activities must also equip participants with practical skills, such as leadership and communication skills, and encourage and motivate the application of these skills.

Outcomes

The outcomes from training programs should focus on fostering shifts in individual behaviors and attitudes, including a heightened ability to identify and address implicit biases, privileges, and microaggressions. Critical success outcomes would include embedding these skills and behaviors within institutional practices to promote systemic change.

Impact

To move toward the desired systemic impact of gender equality within higher education and beyond, interventions need to demonstrate sustainable behavior changes at the organizational and societal levels. Embedding gender equity principles into institutional structures—such as curriculum design, faculty development, and organizational policies—will help to create a ripple effect, influencing broader cultural norms and reducing structural barriers to equality.

Assumptions and Contextual Factors

The success of training interventions depends on several assumptions, such as participant willingness to engage, institutional buy-in, and adequate resources for program implementation and evaluation. The evidence suggests that voluntary participation may enhance learner engagement, but care must be taken to mitigate self-selection biases by complementing elective offerings with strategically designed mandatory components.

To bridge the gap between current practice and desired outcomes, a ToC-driven approach should prioritize longitudinal, multilevel evaluations that measure not only immediate knowledge gains but also the sustainability and transferability of learning outcomes. This includes tracking participants' ability to influence institutional and societal change over time. By connecting theory, evidence, and practice, this framework provides a cohesive road map for designing and implementing impactful gender equality training programs in higher education.

Strengths and Limitations

The search strategy was comprehensively developed by all authors with the support of an experienced librarian to facilitate a thorough and extensive database search. The searches carried out were limited to the 9 databases available to the authors. However, following consultation with the librarian, the process of hand-searching the reference lists of the included articles to identify important studies and gray literature may be considered a strength of this review. It is also important to note the lack of relevant literature available, highlighting a significant gap in the evidence base for this area and this should encourage further research. Considering the accelerating interest in equality, diversity, and inclusion issues in most recent years, relevant work may have been missed in the months since the main literature search was conducted. Excluding studies which focused on mentorship programs, gender equality in health care delivery, and gender studies modules may have excluded potentially useful information. However, following the preliminary database searching and screening, we concluded that these studies did not fully align with the research objectives for this review.

Implication for Practice

This review aimed to map the depth and breadth of gender equality training for student leaders in HEI to support curriculum development for training. Our findings suggest room for improvement in the conduct and reporting of research on training interventions with particular attention to theoretically informed decisions about the development of learning activities, the choice of instructional methods, and tools and resources to implement the interventions. In addition, a more effective approach to evaluation that goes beyond the immediate reaction of learners and assesses behavior change is required to allow continuous improvements in this field. Educational programs for gender equality can play a significant role in fostering awareness, knowledge, and skills necessary to address gender disparities in the higher education sector. However, they cannot operate in isolation. There is a real risk that stand-alone educational initiatives will fail to create lasting, systemic change if they are not integrated into a broader, multipronged strategy that addresses the structural, cultural, and institutional barriers perpetuating gender inequality [40].

Conclusions

Initiatives such as the Athena Swan Charter, a framework to support gender equality within higher education and research, seek to advance equal opportunities for all genders in HEIs across the globe. However, significant gender inequality remains and encouraging positive leadership among students may help to build capacity to support equality. Appropriate sustainable and effective skills training is needed to increase awareness and nurture competencies for male, female, and nonbinary student leaders to actively address gender inequality. Investments in gender equity training demonstrate organizational commitment to inclusivity. Long-term dedicated financial support is essential for the sustainability of training interventions aimed at promoting gender equity. Adequate funding enables the development of high-quality training materials, the engagement of skilled facilitators, and the integration of innovative

methodologies. Funding also facilitates the evaluation and monitoring of programs, ensuring that interventions remain effective and adaptable to changing needs and contexts. Without sustainable funding, training programs risk becoming fragmented or short-lived, limiting their impact on fostering institutional change.

Acknowledgments

This review forms part of the study 'LIBRA: Future-proofing Gender Equality in Irish Higher Education Institutions'. This research is funded by the Irish Higher Education Authority.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist.
[PDF File (Adobe PDF File), 84 KB - mededu_v11i1e60061_app1.pdf]

References

1. A Guide for gender equality in teacher education policy and practices. United Nations Educational, Scientific and Cultural Organization. 2017. URL: https://unesdoc.unesco.org/ark:/48223/pf0000231646 [accessed 2023-04-02]
2. Assessments of higher education's progress towards the UN sustainable development goals (volume 2). Higher Education Sustainability Initiative. URL: https://sdgs.un.org/sites/default/files/2021-09/HEI%20assessment%20for%20the%20SDGs%20-%20Volume%202%20HEIs1.pdf [accessed 2023-04-03]
3. Rosa R, Drew E, Canavan S. An overview of gender inequality in EU universities. In: Drew E, Canavan S, editors. The Gender-Sensitive University: A Contradiction in Terms?. New York, NY: Routledge; 2020:5.
4. Gender equality in academia and research - GEAR tool. The European Institute for Gender Equality. 2016. URL: https://eige.europa.eu/sites/default/files/documents/mh0716096enn_1.pdf [accessed 2023-01-07]
5. Equality in higher education: statistical report 2021. Advance HE. URL: https://www.advance-he.ac.uk/knowledge-hub/equality-higher-education-statistical-report-2021 [accessed 2024-04-29]
6. Handayani T, Widodo W. Gender gaps in student leadership at a University in Portugal. In: Proceedings of the 6th International Conference on Community Development. 2019 Presented at: ICCD '19; July 24-25, 2019; Bandar Seri Begawan, Indonesia p. 578-581 URL: https://www.academia.edu/104647287/Gender_Gaps_in_Students_Leadership_at_a_University_in_Portugal?uc-sb-sw=15368025
7. HEA national review of gender equality in Irish higher education institutions. Higher Education Authority. 2016. URL: https://hea.ie/assets/uploads/2017/06/HEA-National-Review-of-Gender-Equality-in-Irish-Higher-Education-Institutions.pdf [accessed 2021-06-20]
8. Higher education institutional staff profiles by gender. Higher Education Authority. URL: https://hea.ie/assets/uploads/2019/07/Higher-Education-Institutional-Staff-Profiles-by-Gender-2020.pdf [accessed 2021-06-20]
9. Freire P. Pedagogy of the Oppressed. New York, NY: Continuum International Publishing; 1970.
10. Gender equality training. European Institute for Gender Equality. URL: https://eige.europa.eu/sites/default/files/documents/mh0716093enn.pdf [accessed 2024-04-29]
11. Coe IR, Wiley R, Bekker L. Organisational best practices towards gender equality in science and medicine. Lancet 2019 Feb;393(10171):587-593. [doi: 10.1016/s0140-6736(18)33188-x]
12. Connolly FF, Goossen M, Hjerm M. Does gender equality cause gender differences in values? Reassessing the gender-equality-personality paradox. Sex Roles 2019 Dec 04;83(1-2):101-113 [FREE Full text] [doi: 10.1007/s11199-019-01097-x]
13. Acai A, Mercer-Mapstone L, Guitman R. Mind the (gender) gap: engaging students as partners to promote gender equity in higher education. Teach High Educ 2019 Apr 08;24(6):1-21. [doi: 10.1080/13562517.2019.1696296]
14. Toraman C, Özen F. An Investigation of the Effectiveness of the Gender Equality Course with a Specific Focus on Faculties of Education. Educ Policy Anal Strateg Res 2019 Jun 27;14(2):6-28. [doi: 10.29329/epasr.2019.201.1]
15. Dür M, Keller L. Education for sustainable development through international collaboration. A case study on concepts and conceptual change of school - students from India and Austria on gender equality and sustainable growth. Educ Sci 2018 Oct 27;8(4):187. [doi: 10.3390/educsci8040187]
16. Segovia-Pérez M, Laguna-Sánchez P, de la Fuente-Cabrero C. Education for sustainable leadership: fostering women's empowerment at the university level. Sustainability 2019 Oct 09;11(20):5555. [doi: 10.3390/SU11205555]
17. Condron C, Power M, Mathew M, Lucey SM. Gender equality training for students in higher education: protocol for a scoping review. JMIR Res Protoc 2023 Sep 20;12:e44584 [FREE Full text] [doi: 10.2196/44584] [Medline: 37728987]

18. Weiss CH, Connell J, Kubisch A, Schorr L, Weiss C. Nothing as practical as good theory: exploring theory-based evaluation for comprehensive community initiatives for children and families. In: Connell A, Kubisch L, Schorr L, Weiss C, editors. New Approaches to Evaluating Community Initiatives. Washington, DC: Aspen Institute; 1995:62-92.
19. Ferguson L. A theory of change for training for gender equality. UN Women Training Centre. 2019. URL: https://trainingcentre.unwomen.org/RESOURCES_LIBRARY/Resources_Centre/01%20Theory%20of%20Change.pdf [accessed 2024-06-13]
20. Thomas PA, Kern DE, Hughes MT, Chen BY. Curriculum Development for Medical Education – A Six–Step Approach. 3rd edition. Baltimore, MD: Johns Hopkins University Press; 2016.
21. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
22. Peterson J, Pearce PF, Ferguson LA, Langford CA. Understanding scoping reviews: definition, purpose, and process. J Am Assoc Nurse Pract 2017 Jan;29(1):12-16. [doi: 10.1002/2327-6924.12380] [Medline: 27245885]
23. Peters MD, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth 2020 Oct;18(10):2119-2126. [doi: 10.11124/JBIES-20-00167] [Medline: 33038124]
24. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 05;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
25. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
26. Shields SA, Zawadzki MJ, Johnson RN. The impact of the workshop activity for gender equity simulation in the academy (WAGES–Academic) in demonstrating cumulative effects of gender bias. J Divers High Educ 2011;4(2):120-129. [doi: 10.1037/A0022953]
27. Kolb DA. Experiential Learning: Experience as the Source of Learning and Development. New York, NY: Prentice Hall; 1984.
28. Kennedy NF, Şenses N, Ayan P. Grasping the social through movies. Teach High Educ 2011 Feb;16(1):1-14. [doi: 10.1080/13562517.2010.507305]
29. Falk AL, Sandqvist JL, Liedberg GM. Euro-education: employability for all (EEE4all). Design and implementation of an international course for future health-care professionals. Work 2012;41(4):433-438. [doi: 10.3233/WOR-2012-1311] [Medline: 22495414]
30. Case KA, Hensley R, Anderson A. Reflecting on heterosexual and male privilege: interventions to raise awareness. J Soc Issues 2014 Dec 09;70(4):722-740. [doi: 10.1111/josi.12088]
31. Abrams L, Shoger SG, Corrigan L, Nozaki SY, Narui M. Empowering male students as allies for gender equity within an engineering college. In: Proceedings of the 2016 ASEE Annual Conference & Exposition. 2016 Presented at: ASEE '16; June 26, 2016; New Orleans, LA URL: https://monolith.asee.org/public/conferences/64/papers/15952/view [doi: 10.18260/p.26945]
32. Broido EM, Reason RD. The development of social justice attitudes and actions: an overview of current understandings. New Drctns for Student Srvcs 2005 Jun 09;2005(110):17-28. [doi: 10.1002/ss.162]
33. Freedman G, Seidman M, Flanagan M, Kaufman G, Green MC. The impact of an “aha” moment on gender biases: limited evidence for the efficacy of a game intervention that challenges gender assumptions. J Exp Soc Psychol 2018 Sep;78:162-167. [doi: 10.1016/j.jesp.2018.03.014]
34. Altınova HH, Duyan V, Megahead HA. The impact of the human rights education program for women on gender perceptions of social work students. Res Soc Work Pract 2016 Dec 08;29(1):113-121. [doi: 10.1177/1049731516679889]
35. Gorrotxategi MP, Ozamiz-Etxebarria N, Jiménez-Etxebarria E, Cornelius-White JH. Improvement in gender and transgender knowledge in university students through the creative factory methodology. Front Psychol 2020 Mar 13;11:367 [FREE Full text] [doi: 10.3389/fpsyg.2020.00367] [Medline: 32231615]
36. Liao HC, Wang YH. Integrating the gender perspective into literature studies to enhance medical university students' gender awareness and critical thinking. Int J Environ Res Public Health 2020 Dec 10;17(24):9245 [FREE Full text] [doi: 10.3390/ijerph17249245] [Medline: 33321913]
37. Locke JW, Pope J, Abrams JM. Reflections from a male engineering student ally, his professor, and his advisor. JOM 2021 Aug 03;73(9):2588-2590. [doi: 10.1007/S11837-021-04830-8]
38. Bosch VN, Freude L, Camps Calvet CC. Service learning with a gender perspective: reconnecting service learning with feminist research and pedagogy in sociology. Teach Sociol 2021 Apr 01;49(2):136-149. [doi: 10.1177/0092055x21993465]
39. de Villiers T, Duma S, Abrahams N. "As young men we have a role to play in preventing sexual violence": development and relevance of the men with conscience intervention to prevent sexual violence. PLoS One 2021;16(1):e0244550 [FREE Full text] [doi: 10.1371/journal.pone.0244550] [Medline: 33411823]
40. Rosenman R, Tennekoon V, Hill LG. Measuring bias in self-reported data. Int J Behav Healthc Res 2011 Oct;2(4):320-332 [FREE Full text] [doi: 10.1504/IJBHR.2011.043414] [Medline: 25383095]
41. Guthridge M, Kirkman M, Penovic T, Giummarra M. Promoting gender equality: a systematic review of interventions. Soc Just Res 2022 Sep 01;35(3):318-343 [FREE Full text] [doi: 10.1007/s11211-022-00398-z]

42. Verge T, Ferrer-Fons M, González MJ. Resistance to mainstreaming gender into the higher education curriculum. Eur J Women's Stud 2017 Jan 09;25(1):86-101 [FREE Full text] [doi: 10.1177/1350506816688237]
43. Darby JA. The effects of the elective or required status of courses on student evaluations. J Vocat Educ Train 2006 Mar;58(1):19-29. [doi: 10.1080/13636820500507708]
44. Dobbin F, Kalev A. Why diversity programs fail and what works better. The Harvard Business Review. URL: https://hbr.org/2016/07/why-diversity-programs-fail [accessed 2024-04-29]
45. Kirkpatrick DL, Kirkpatrick JD. Evaluating Training Programs: The Four Levels. 3rd edition. New York, NY: Berrett-Koehler Publishers; 2006.
46. Lau VW, Scott VL, Warren MA, Bligh MC. Moving from problems to solutions: a review of gender equality interventions at work using an ecological systems approach. J Organ Behavior 2022 Jul 08;44(2):399-419 [FREE Full text] [doi: 10.1002/job.2654]
47. Guidelines for gender mainstreaming academia. Swedish Secretariat for Gender Research. URL: https://eige.europa.eu/sites/default/files/sscr_guidelines-for-gender-mainstreaming-academia.pdf [accessed 2023-03-15]

Abbreviations

EIGE: European Institute for Gender Equality
HEI: higher education institution
ToC: theory of change

Edited by B Lesselroth; submitted 30.04.24; peer-reviewed by J Mattsson, SSC Herrick, D Verran; comments to author 09.11.24; revised version received 03.01.25; accepted 06.03.25; published 11.07.25.

Please cite as:
Condron C, Power M, Mathew M, Lucey S, Henn P, Dean T, Kirrane Scott M, Eppich W, Lucey SM
Gender Equality Training for Students in Higher Education: Scoping Review
JMIR Med Educ 2025;11:e60061
URL: https://mededu.jmir.org/2025/1/e60061
doi:10.2196/60061
PMID:40132179

Review

Applications of Artificial Intelligence in Psychiatry and Psychology Education: Scoping Review

Julien Prégent[1], BAA; Van-Han-Alex Chung[2], BSc; Inès El Adib[1]; Marie Désilets[3], MSc; Alexandre Hudon[1,3,4,5], BEng, MD, PhD

[1]Department of Psychiatry and Addictology, Faculty of Medicine, Université de Montréal, Montreal, QC, Canada

[2]Faculty of Medicine, McGill University, Montreal, QC, Canada

[3]Department of Psychiatry, Institut universitaire en santé mentale de Montréal, Montréal, QC, Canada

[4]Department of Psychiatry, Institut national de psychiatrie légale Philippe-Pinel, Montreal, QC, Canada

[5]Centre de recherche de l'Institut universitaire en santé mentale de Montréal, Montreal, QC, Canada

Corresponding Author:
Alexandre Hudon, BEng, MD, PhD
Department of Psychiatry
Institut universitaire en santé mentale de Montréal
7401 Boulevard Hochelaga
Montréal, QC, H1N 3M5
Canada
Phone: 1 5142514000
Email: alexandre.hudon.1@umontreal.ca

Abstract

Background: Artificial intelligence (AI) is increasingly integrated into health care, including psychiatry and psychology. In educational contexts, AI offers new possibilities for enhancing clinical reasoning, personalizing content delivery, and supporting professional development. Despite this emerging interest, a comprehensive understanding of how AI is currently used in mental health education, and the challenges associated with its adoption, remains limited.

Objective: This scoping review aimed to identify and characterize current applications of AI in the teaching and learning of psychiatry and psychology. It also sought to document reported facilitators of and barriers to the integration of AI within educational contexts.

Methods: A systematic search was conducted across 6 electronic databases (MEDLINE, PubMed, Embase, PsycINFO, EBM Reviews, and Google Scholar) from inception to October 2024. The review followed Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) guidelines. Studies were included if they focused on psychiatry or psychology, described the use of an AI tool, and discussed at least 1 facilitator of or barrier to its use in education. Data were extracted on study characteristics, population, AI application, educational outcomes, facilitators, and barriers. Study quality was appraised using several design-appropriate tools.

Results: From 6219 records, 10 (0.2%) studies met the inclusion criteria. Eight categories of AI applications were identified: clinical decision support, educational content creation, therapeutic tools and mental health monitoring, administrative and research assistance, natural language processing (NLP), program/policy development, students' study aid, and professional development. Key facilitators included the availability of AI tools, positive learner attitudes, digital infrastructure, and time-saving features. Barriers included limited AI training, ethical concerns, lack of digital literacy, algorithmic opacity, and insufficient curricular integration. The overall methodological quality of included studies was moderate to high.

Conclusions: AI is being used across a range of educational functions in psychiatry and psychology, from clinical training to assessment and administrative support. Although the potential for enhancing learning outcomes is clear, its successful integration requires addressing ethical, technical, and pedagogical barriers. Future efforts should focus on AI literacy, faculty development, and institutional policies to guide responsible and effective use. This review underscores the importance of interdisciplinary collaboration to ensure the safe, equitable, and meaningful adoption of AI in mental health education.

(JMIR Med Educ 2025;11:e75238) doi:10.2196/75238

KEYWORDS

artificial intelligence; medical education; psychiatry education; psychology education; machine learning; digital health; chatbot; clinical decision support; e-learning; mental health training

Introduction

Cloud computing, artificial intelligence (AI), machine learning (ML), telehealth, digitally assisted diagnosis and treatment, and consumer-focused mobile health applications have reshaped the landscape of health care delivery. These technologies are now widely used in clinical care, scientific research, and self-management [1]. These developments offer the potential to improve treatment outcomes, promote greater patient engagement, and enable earlier diagnosis and intervention [1]. In addition to enhancing conventional clinical procedures, such as teleconsultation and patient record management, these advancements have made way for fresh, data-driven methods of diagnosing and treating patients with a wide range of illnesses [2]. Particularly in the fields of psychology and psychiatry, new research highlights the expanding impact of AI online learning environments, and e-therapies, all of which could potentially reshape clinical practice, educational routes, and health policy [3,4]. As they provide new avenues for psychiatric condition screening, diagnosis, and monitoring, AI and ML have, in fact, attracted a lot of attention in the field of mental health care [3]. Self-guided mental health apps and web-based cognitive behavioral therapy (CBT) modules are examples of digital interventions that can help address systemic gaps in mental health care by improving access in underserved areas, reducing stigma, and lowering infrastructure costs. Still, limitations, such as weakened therapeutic alliance, limited support for complex conditions, and digital literacy barriers persist, highlighting the need for complementary innovations, such as AI [5]. The goal of ML, a branch of AI, is to empower computers to learn from data and make predictions or judgments by using statistical models and algorithms [6]. Potential applications of ML include predicting the effectiveness of antidepressant medicine, defining depression, estimating the risk of suicide, and predicting psychotic episodes in people with schizophrenia [7-14]. Beyond individual diagnosis and prognosis, ML has also been explored for system-level uses, such as optimizing service triage, forecasting population mental health trends, and informing resource allocation strategies [3]. Although promising, these uses raise critical concerns about the limits of algorithmic decision-making in capturing complex human experiences, therapeutic nuance, and clinical judgment.

In any case, any increase in accuracy or efficiency must be balanced against possible drawbacks, including algorithmic bias, concerns about data privacy, and the requirement for open systems to maintain patient confidence [15]. There are concerns over chatbots and other automated therapy tools' ability to give compassionate care and uphold therapeutic boundaries, despite being promoted as affordable ways to provide mental health help [16]. Therefore, the use of AI-based technologies (eg, chatbots, virtual assistants, or automated therapy platforms) necessitates careful evaluation of any potential negative effects, such as algorithmic bias, transparency issues, patient confidentiality, and wider ethical issues [16,17]. These tools are already transforming the patient-clinician relationship and raise complex questions about how therapeutic alliances can be established and maintained in digital environments [18]. The traditional roles and responsibilities of health care providers may change because of the advent of automated evaluations and AI-driven therapy recommendations. As a matter of fact, clinicians will soon have to balance the results of algorithms with their professional judgment [2].

These factors highlight the need for careful consideration for psychologists, psychiatrists, and other medical practitioners to gain a thorough understanding of the ethical and therapeutic aspects of developing technology [19]. Numerous experts stress that substantial training for aspiring professionals is necessary for a meaningful integration of AI into mental health treatment, especially when it comes to e-therapy techniques or semiautomated diagnostics [19,20]. Beyond the purview of one's initial licensing, lifelong learning in order to keep up with technological advancements to guarantee that mental health practitioners stay knowledgeable about the rapidly changing field of e-therapies and digital health solutions could be seen as nonnegotiable [21]. This need aligns with constructivist and experiential learning theories, which emphasize that learners build knowledge most effectively through active engagement and real-world application, both of which are necessary for the safe and ethical use of AI in clinical practice. The requirements for certification and continuing education programs, which may include specific workshops, learning modules centered on digital proficiency, or periodic re-examinations, reflect these needs [22]. Education researchers warn that lectures or brief seminars alone will not be sufficient to close the knowledge gap; clinical and research experiences that enable practitioners to assess, adjust, and use AI technologies with confidence in a safe and ethical way are also necessary [23].

The purpose of this scoping review was to determine the types of AI that are currently used in academic programs and educational curricula in psychology and psychiatry. Furthermore, it attempted to identify the barriers to and facilitators of such uses. By evaluating various apps, this review aimed to synthesize existing apps and highlight areas where further development or inquiry may be warranted. These included not only content delivery and assessment but also more challenging domains, such as teaching clinical judgment, fostering empathy, assessing nuanced interpersonal interactions, and supporting therapeutic decision-making—areas that are often context dependent and difficult to replicate algorithmically. Finally, suggestions were offered for future research directions to support the responsible and effective incorporation of AI into mental health education.

Methods

Search Strategies

A broad scoping search was systematically performed to retrieve the recent literature from multiple electronic databases, including

Medline, PubMed, Embase, PsycNet (PsycINFO), EBM Reviews – Cochrane Database of Systematic Reviews, and Google Scholar, covering records from database inception through October 2024. The review adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) guidelines [24]. The search strategy integrated both free-text keywords and controlled vocabulary (Medical Subject Headings [MeSH] terms), centering on themes related to AI (eg, "artificial intelligence," "AI") and medical education (eg, "education," "medical," "students"), in accordance with the study's aims. Full search strategies are detailed in Multimedia Appendix 1. The search strategies were designed by an experienced librarian in the field of mental health (author MD). They were also cross-validated by another librarian using the Peer Review of Electronic Search Strategies (PRESS) approach. The methodology was designed by the corresponding author; searches were conducted by author AH and independently verified by author JP. No filters were applied concerning geographical location or institutional setting. The completed PRISMA-ScR checklist is available in Multimedia Appendix 2.

Study Eligibility

Studies were eligible for inclusion if they met the following criteria: (1) study focused on a topic within the fields of psychiatry or psychology education; (2) involved the use of an AI tool, model, or approach; (3) included a discussion of facilitators of or barriers to the use of AI; and (4) were available in either English or French. Papers that did not pertain to psychiatry or psychology or that referenced AI technologies without a clearly defined implementation or application were excluded. In addition, studies were excluded if they featured AI tools outside the domain of relevance, such as rule-based expert systems or search algorithms unrelated to data-driven models. Unpublished studies and gray literature were not considered for inclusion.

Data Extraction

A standardized data extraction form was developed to systematically collect and organize relevant information from each included study. Data extraction was performed independently by at least 2 of the authors, with discrepancies resolved through discussion and consensus or by consulting a third author, when necessary. The process was guided by the research objectives for examining the integration of AI in health profession education.

The following elements were extracted from each study:

- Authors: citation details, including the first author and year of publication.
- Population: description of the study participants, including the sample size, educational level (eg, undergraduate, postgraduate, continuing professional development), and discipline (eg, psychiatry, psychology, medical education).
- Use of AI: specific application or role of AI within the educational context, such as adaptive learning platforms, natural language processing (NLP) tools, intelligent tutoring systems, virtual patient simulations, or predictive analytics.
- Main outcomes: primary findings related to educational effectiveness, learner satisfaction, knowledge acquisition, clinical reasoning, engagement, or other measured outcomes. Where applicable, outcomes were categorized by study design (qualitative, quantitative, or mixed methods).
- Facilitators: factors that supported the successful implementation or perceived value of AI-enhanced educational interventions, such as user-friendliness, institutional support, personalization of learning, or alignment with curricular goals.
- Barriers: reported challenges or limitations in the adoption of AI tools, including technical constraints, ethical concerns, paucity of digital literacy among users, limited evidence of effectiveness, or resistance to change within educational environments.

Extracted data were tabulated in Microsoft Excel (version 17.0) to allow comparison across studies and to identify patterns, gaps, and emerging themes related to the integration of AI in teaching and learning within the fields of psychiatry and psychology.

Quality Assessment

To examine the included studies' methodology, clarity, and transparency, we carried out a systematic evaluation. Considering the variety of research designs seen in the literature on AI in psychiatry and addiction education, this phase attempted to improve the findings' interpretability. We used evaluation instruments that were specific to each research type's design to guarantee methodological adequacy.

The Joanna Briggs Institute (JBI) Checklist for Analytical Cross-Sectional Studies was used to assess both quantitative and observational research [25]. The methodological soundness of studies looking at correlations between exposures and outcomes at a specific moment in time can be evaluated with this tool. Items on the checklist evaluate the appropriateness of statistical analyses, the identification and control of confounding factors, the validity and reliability of exposure and outcome measurements, and the clarity of the inclusion criteria.

The JBI Checklist for Qualitative Research was also used [25]. This tool assesses how well research methodology and research topics align, how data are collected, how participant voices are represented and interpreted, and how much the researcher has influenced the study. To guarantee that qualitative insights are obtained through a thorough and reliable process, it also evaluates ethical issues and the openness of data analysis techniques.

Mixed methods studies were appraised using the Mixed Methods Appraisal Tool (MMAT), 2018 version [26]. The MMAT is a validated tool specifically developed for the assessment of studies that combine qualitative and quantitative approaches. It allows for concurrent evaluation of the components of each methodology within a single study and includes criteria for the integration of qualitative and quantitative data, the appropriateness of the design to the research questions, and the coherence of interpretations drawn from the combined methods.

For nonempirical contributions, such as viewpoints, editorials, and conceptual papers, we used the authority, accuracy, coverage, objectivity, date, and significance (AACODS) checklist [27]. This framework assesses 6 key domains: authority (credibility of the author or source), accuracy (evidence supporting claims), coverage (comprehensiveness of content), objectivity (balance and absence of bias), date (currency and relevance), and significance (contribution to the field). It is particularly useful for evaluating gray literature and opinion-based texts where conventional empirical criteria may not apply.

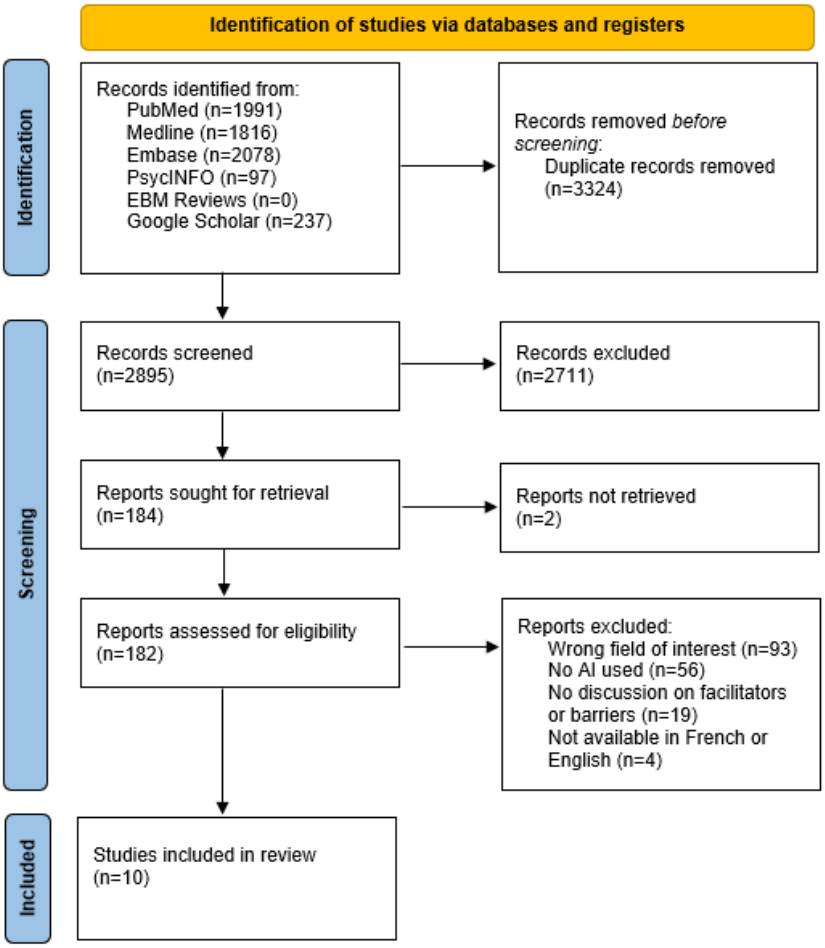
All studies were reviewed independently by 2 researchers (JP and AH). Each appraisal tool includes a set of key domains that were rated as "yes," "partial," or "no" based on the extent to which the methodological criteria were clearly addressed and appropriately implemented in each study.

Results

Description of Studies

The scoping review explored the use of AI in teaching and learning within psychiatry and psychology. The initial search across 6 electronic databases yielded 6219 records. After removing 3324 (53.4%) duplicates, 2895 (46.6%) records were screened by title and abstract. Of these, 2711 (93.6%) papers were excluded for not meeting the inclusion criteria. A total of 184 (6.4%) full-text papers were sought for retrieval, with 2 (1.1%) reports not successfully retrieved. The remaining 182 (98.9%) papers were assessed for eligibility in full. Following detailed evaluation, 172 (94.5%) papers were excluded due to being outside the field of interest (n=93, 54.1%), lacking the use of AI (n=56, 32.6%), not addressing facilitators or barriers (n=19, 11%), or not being available in English or French (n=4, 2.3%). Ultimately, 10 (5.5%) studies met all inclusion criteria and were included in the final analysis. A flowchart summarizing the selection process is presented in Figure 1, and details of the included studies can be found in Multimedia Appendix 3.

Figure 1. PRISMA-ScR flowchart for the inclusion of studies.



Main Uses of Artificial Intelligence

Across the 10 studies [28-37] included in this scoping review, AI was applied to a variety of educational functions in psychiatry and psychology. The most frequently observed category was *clinical decision support* (n=5, 50%), where AI tools were used to train learners in diagnosis, prognosis, risk assessment, and early intervention strategies. This was followed by 5 categories that each appeared in 3 (30%) studies: *educational content creation and enhancement*, *therapeutic tools and mental health monitoring*, *administrative and research*

assistance, *NLP applications*, and *program/policy development*. These categories reflect AI's growing role in the development of educational materials, therapeutic simulations, and institutional planning. Less frequently, studies addressed AI's applications in *professional development and assessment* (n=1, 10%) and *student/applicant support* (n=1, 10%), highlighting emerging but less explored domains. This distribution suggests a strong current emphasis on clinical reasoning, content automation, and digital service integration in educational contexts.

Clinical Decision Support

AI tools are increasingly integrated into psychiatry and psychology education to train learners in diagnosis, prognosis, and risk assessment. Through exposure to AI-driven systems, such as ML models and NLP tools, trainees can learn how to identify suicide risk, detect patterns in substance use disorders, or evaluate the progression of neurodegenerative conditions [28]. Banerjee et al [29] emphasized the importance of training programs that illustrate how AI can triage patients or generate diagnostic suggestions, helping doctors understand both the power and limitations of these technologies. Furthermore, students are being introduced to AI's role in treatment personalization, such as adapting therapy plans or medication regimens using predictive modeling [30]. These systems also facilitate early intervention, teaching clinicians how AI can flag high-risk patients in real time, thus embedding preventive care principles into training [31].

Educational Content Creation and Enhancement

Generative AI is transforming the design and delivery of educational content in psychiatry and psychology. One example is the use of ChatGPT to develop script concordance tests (SCTs), which promote clinical reasoning in undergraduate medical education. Hudon et al [32] demonstrated that AI-generated SCTs are nearly indistinguishable from those written by human experts, highlighting AI's potential to rapidly generate quality training materials aligned with psychiatric diagnostic frameworks. In addition, AI can support exam preparation, providing structured explanations, study plans, and interactive feedback for licensing exams and clinical cases [28]. Spallek et al [31] noted that although AI tools require human oversight, they offer accessible and customizable resources that align with the best practices in mental health communication and literacy, ultimately empowering educators and enhancing learner engagement.

Student/Applicant Support

AI can now be used to assist learners with the residency and fellowship application process, offering new educational opportunities in self-presentation and professional writing. Mangold and Ream [33] explored how students and faculty use AI to draft and refine personal statements and letters of recommendation, improving clarity and grammar and reducing biased language. From a pedagogical perspective, this practice teaches students to reflect critically on AI-generated outputs and engage in iterative editing processes. Moreover, AI is being proposed for application screening, potentially reducing human bias and increasing efficiency in admissions, which prompts institutions to educate both applicants and reviewers on ethical and equitable AI use [33]. These developments signal a shift in how learners engage with professional identity formation and how educators must adapt guidance accordingly.

Therapeutic Tools and Mental Health Monitoring

A body of literature highlighted AI's use in therapeutic education, particularly around digital interventions, such as CBT- or acceptance and commitment therapy (ACT)-based apps and chatbots. Blease et al [30] and Gratzer and Goldbloom [34] emphasized the need to train future psychiatrists to evaluate and potentially integrate these tools into treatment plans. E-therapy technologies, including AI-powered chatbots, provide psychoeducation on demand and demonstrate how therapy can be delivered asynchronously [34]. Trainees also learn about real-time symptom tracking, which is increasingly used in digital platforms to monitor patient well-being and guide interventions. These digital tools expose students to new care modalities, challenging them to assess efficacy, ethics, and clinical utility in both individual and population-based care models [34].

Administrative and Research Assistance

AI is also reshaping how clinicians and students interact with administrative and scholarly tasks, including documentation, summarization, and literature reviews. Banerjee et al [38] observed that AI is viewed as helpful in reducing the time spent on clinical documentation, allowing more focus on direct learning and patient care. Additionally, AI tools are being used in academic psychiatry to assist with automated literature searches, summarization, and even initial drafting of manuscripts, offering educational insights into how information is synthesized and presented in research [28]. These uses favorize critical thinking and teach students to evaluate AI-generated content, thereby strengthening their roles as both users and producers of scientific knowledge.

Professional Development and Assessment

In postgraduate and continuing education contexts, AI supports competency-based assessment and professional development. Anzia [35] discussed how longitudinal assessment platforms, potentially enhanced by AI, are replacing traditional high-stakes exams in psychiatry, promoting lifelong learning and more reflective practice. These tools can track learning progress, provide feedback, and recommend tailored educational pathways. Moreover, AI may be integrated into clinical skills evaluations, simulating patient scenarios or providing automated assessments of diagnostic reasoning. This shift calls for educators to incorporate digital literacy and AI fluency into curricula, ensuring that learners are prepared to navigate evolving certification and assessment landscapes [35].

Natural Language Processing Applications

NLP tools powered by AI are being introduced in educational settings to illustrate how clinical language can be analyzed, interpreted, and used for real-time support. Banerjee et al [29] noted NLP's utility in automating clinical documentation, reducing the clerical burden and highlighting the importance of clear, structured input. In psychiatry training, NLP is also applied to suicide risk detection, teaching learners how digital footprints and language cues from online interactions can signal

crisis [28]. Additionally, NLP technologies are integrated into telehealth platforms, enhancing communication between patients and providers, and offering opportunities for students to learn how AI can support culturally sensitive, evidence-based interactions [31].

Program/Policy Development

Finally, AI's influence on institutional teaching structures is growing, particularly through the design of AI-compatible teaching modules and the creation of policies to guide AI use. Manjunatha et al [36] illustrated how telepsychiatry programs in India are incorporating AI into remote learning for primary care doctors, serving as a scalable model for underserved settings. These modules reflect a shift toward digital curricula that embed clinical translation and evidence-based AI use. Similarly, Mangold and Ream [33] emphasized the need for training programs to define guidelines for AI use in admissions and evaluation, prompting educators to prepare learners for ethical dilemmas and policy engagement in the evolving digital landscape.

Facilitators

A number of facilitators support the integration of AI into teaching and learning in psychiatry and psychology.

Technological Readiness and Tool Availability

A recurring theme across the 10 studies was the availability and accessibility of AI tools, particularly large language models, such as ChatGPT, which simplify the creation of educational content and customization of learning materials for diverse learner needs [31,32]. These tools are perceived as valuable due to their ability to generate structured and accessible outputs, supporting educators in preparing mental health scenarios or assessments rapidly and effectively. Moreover, other technologies, such as telepsychiatry and mobile-based learning platforms, are supported by the widespread use of smartphones and digital infrastructure, increasing scalability and access to remote or underserved areas [36].

Educational and Efficiency Enhancement

Several studies have emphasized that structured prompting or prompt engineering significantly enhances output quality, improving both relevance and accuracy for educational use [33]. In clinical training contexts, AI is seen as a time-saving facilitator, capable of reducing administrative burdens, such as documentation, and allowing learners to focus on clinical reasoning and decision-making [29]. In addition, generative AI technologies and related models can facilitate the production of highly realistic synthetic data and the seamless integration of unstructured content across diverse formats [37]. These innovations have the potential to transform core practices, such as risk assessment, diagnostic decision-making, and treatment planning, while simultaneously creating new opportunities in educational and training environments.

Learner Engagement and Openness

Finally, positive attitudes from students and trainees, particularly their willingness to explore new tools, and their recognition of AI's role in increasing efficiency, accuracy, and engagement, are essential to AI adoption in educational environments [30]. These facilitators reflect a confluence of technological readiness, user engagement, and curricular flexibility.

Barriers

Despite growing enthusiasm, several barriers hinder the seamless integration of AI into psychiatry and psychology education.

Digital and Educational Gaps

A dominant concern is the absence of formal training and digital literacy among students and educators, which limits their ability to interpret and critically evaluate AI-generated outputs [29,30]. Many studies have noted a limited presence of AI content in medical curricula, resulting in missed opportunities to prepare learners for evolving clinical environments where AI plays a central role [30].

Ethical and Legal Issues

There are also ethical and legal concerns, particularly around data privacy, algorithmic bias, and informed consent, which raise questions about the responsible use of AI tools, such as chatbots or diagnostic aids [31,34]. The scarcity of high-quality data and the opacity of AI algorithms, often referred to as the "black box" problem, are also cited as major obstacles to trust and widespread adoption [28]. Additionally, concerns persist about the accuracy and reliability of AI-generated educational materials, particularly when outputs are not subject to expert review, posing risks of misinformation or oversimplification [31,32].

Pedagogical Limitations

Finally, several papers highlighted that AI outputs can lack nuance, empathy, or personalization, making them less suitable for teaching relational and humanistic aspects of psychiatric care [33].

These barriers highlight the need for comprehensive strategies that include AI literacy, ethical guidance, and faculty support to ensure safe and effective integration into educational practice.

Quality Assessment of the Identified Studies

Overall, the methodological quality of the included literature was moderate to high. Mixed methods and empirical studies demonstrated clear objectives, coherent data collection and analysis strategies, and ethical transparency. However, several studies lacked detailed descriptions of sampling procedures or integration of data types. Nonempirical papers were generally strong in authority and relevance but limited by the absence of primary data or systematic methodology. Despite variability in design and depth, most studies provided valuable insights into the educational applications of AI in psychiatry and psychology. The full quality appraisal is presented in Multimedia Appendix 3.

Discussion

Principal Findings

This scoping review aimed to identify the different ways AI is currently used in the teaching and learning of psychiatry and psychology. A total of 10 studies were fully analyzed, and 8 categories of AI applications were identified: clinical decision

support, educational content creation and enhancement, therapeutic tools and mental health monitoring, administrative and research assistance, NLP applications, program and policy development, student/applicant support, and professional development and assessment. These categories reflect the diverse roles AI plays in shaping educational strategies, curricular design, and learner engagement in mental health training. The studies included were overall of moderate-to-high quality. The most notable facilitators to AI integration in teaching and learning in psychiatry and psychology are technological readiness and tool availability, educational and efficiency enhancement, and learner engagement and openness. The barriers that hinder the integration of AI into psychiatry and psychology education are digital and educational gaps, ethical and legal issues, and pedagogical limitations.

Comparison With Prior Work

The findings of this scoping review confirm and expand on other studies that demonstrate the growing integration of AI into psychological and psychiatric education, especially through clinical decision support technologies. Previous research has shown, for example, that ML models can help forecast the risk of depression, schizophrenia, and suicide. Our study also noted that similar predictive technologies are already being used in educational contexts [38,39]. This is consistent with the findings of Rajkomar et al [40], who observed that clinical AI tool exposure aids in the development of important data literacy in health care trainees. Our work showcases the potential of AI in supporting students as they develop their critical reasoning. However, it is important to keep in mind that these critical reasoning skills are still mostly shaped over time through clinical exposure and case discussion, both of which cannot be replaced by algorithms or AI. The multidisciplinary team also holds a significant place in clinical management and risk sharing. AI can give theoretical advice to students regarding how to lead a team, but it cannot teach a student how to lead a team in real time. Although previous research has often emphasized clinical outcomes, we focused here on the learning process itself, with the understanding that AI is only one of many educational tools.

The literature on AI also supports the increasing use of generative AI in the production of instructional materials. Recent research has switched to looking at how tools such as ChatGPT, GPT-4, and Claude might scaffold learning in medical education, whereas the majority of earlier applications were on patient-facing educational interventions (eg, mental health apps) [41]. These studies support our findings by demonstrating that AI is capable of creating excellent test questions, simulating clinical situations, and even co-creating course curricula. However, critical gaps still exist, notably the possibility that students will passively accept AI outputs without engaging sufficiently. In fact, some of these critical knowledge gaps might be due to the current state of research, which remains limited when it comes to assessing how heavy reliance on dialogue AI may affect decision-making, critical thinking, and analytical reasoning in both educational and research contexts [42]. Meskó [43] shared this concern and called for clear instruction in AI prompt engineering, appropriate medical professional tutorials, and verification skills. The findings of this scoping review highlight that such competencies are increasingly essential in psychiatry and psychology, where nuance and context matter deeply.

According to certain studies, an important area of AI integration in training is mental health apps and therapeutic chatbots, which is consistent with previous research [44,45]. According to these studies, chatbot-based psychotherapy and psychoeducational tools are beneficial teaching tools, in addition to being successful for patients. When included in clinical simulations, they give students an opportunity to analyze intervention results, gauge the therapeutic tone, and practice making moral decisions. Our results highlight the need for directed education to guarantee that students acquire abilities in digital empathy, data protection, and cultural adaptation, even when these tools show promise. Training programs need to change in a way that assists clinicians in interpreting AI-driven outputs, while upholding person-centered treatment and therapeutic alliances, as D'Alfonso et al [46] contended.

Lastly, the findings confirm that faculty development, institutional preparedness, and ethical advice are drivers of AI adoption, which is in line with research on digital transformation in health professions education [47]. Although students frequently use AI tools for convenience, educators are nevertheless worried about the loss of critical thinking, professional identity, and interpersonal skills. These difficulties point to the necessity of organized curricula, such as one that incorporates AI literacy into psychology and psychiatry undergraduate and graduate education programs [48].

Directions for Further Research

This scoping review revealed that slowly but surely, the integration of AI, although remaining nascent in psychology and psychiatry education, is nonetheless there to stay. Future research should prioritize rigorous, outcome-based studies that evaluate further the impact of the AI-enhanced educational tools that this paper described, such as diagnostic stimulation, e-therapies, AI-assisted clinical decision-making, and the impact on real-life learning performance. Another important aspect that would be crucial to investigate is the management of algorithmic bias and transparency and understanding the extent of protecting one's private data. Research that captures the perspectives of students and educators could shed light on readiness, perceived barriers, and opportunities for meaningful adoption of AI.

Furthermore, given the fact that the majority of studies analyzed are from Euro-American and high-income contexts, there is a clear need for research that centers on diverse populations, including Indigenous, racialized, and culturally distinct groups. Together, these directions can help guide the development of inclusive and ethically grounded approaches to AI integration into mental health education.

Recommendations for Institutions

To ensure that psychology and psychiatry programs prepare trainees for a rapidly evolving clinical landscape, educational institutions should take proactive steps toward integrating AI literacy into core curricula. For example, digital literacy could be integrated early into medical education, with foundational topics such as ML and data ethics. This effort would benefit from interdisciplinary collaboration among health sciences,

computer science, and bioethics departments to ensure a well-rounded approach.

The extent to which a person believes that a technology can enhance their performance at work (or enhance their learning experience, for that matter) is commonly referred to as "perceived usefulness," and it plays a pivotal role in determining whether individuals are likely to adopt new technologies [49]. A meta-analysis by Scherer and Teo [50] identified perceived usefulness as a strong predictor of a teacher's readiness to engage with digital tools. Conversely, barriers included anxiety and a lack of AI-specific training [51]. Thus, to effectively deal with the growing threats posed by AI, it seems crucial to promote the various uses of AI among the professors. Teachers who understand AI better are more equipped to use it in ways that meet the varied needs of their students [52]. In addition, when educators develop a solid understanding of AI, they are better equipped to tackle its ethical issues, such as algorithmic bias, data privacy, and the risk of becoming overly dependent on AI [53].

Hence, educators require adequate training, resources, and institutional support to effectively teach and oversee the responsible use of emerging digital tools in clinical and academic settings.

How will universities effectively manage the growing threats that AI presents to medical education? Those threats, as previously mentioned, include academic dishonesty in assessments (eg, plagiarism), the spread of misinformation, and the difficulty AI-using students face in discerning some of the nuances in fields where human interaction is key, such as psychiatry and psychology. Because these threats are often multifactorial in nature, their management requires an inclusive approach involving all key actors (students, educators, institutions) [54]. In practice, at the university level, an interesting avenue would be the constitution of a task force (or a committee) comprising students, faculty members, AI scholars, and IT personnel. This task force would be aimed at determining permissible uses of AI versus prohibited uses of AI and how to detect the latter. The detection of AI-generated content can be facilitated by applications such as GPTZero and QuillBot, for example. Higher education institutions (HEI) should not only establish such thorough guidelines but also review them periodically to ensure their continued relevance [54]. Bozkurt [55] suggested that HEI also require transparent disclosure of AI usage by their personnel and students (eg, in a given assessment, the students would be expected to explicitly declare the sections drafted by ChatGPT with human oversight). HEI could also offer some new job positions focused on AI or hire professionals specifically trained to identify AI-related academic misconduct and sanction such ethical issues. These professionals would ideally also have some degree of experience with the efficient implementation of AI in an academic context.

Strengths and Limitations

The authors brought diverse disciplinary perspectives to this review, including training in psychology, psychiatry, medical education, and digital health. These backgrounds informed the framing of the review and the interpretation of its findings.

This scoping review also has a few limitations. First, although a comprehensive search strategy was used across multiple databases, it is possible that relevant studies were missing, particularly those published in nonindexed journals or categorized under broader terms not captured by the search keywords. Second, the authors acknowledge the potential for disciplinary bias rooted in Western academic systems. An effort was made to include literature from a range of geographic regions and institutional contexts; however, due to the limited published research available on the subject, the review was limited to studies published in English or French, potentially excluding valuable insights from non-English/French literature; in addition, the majority of published research available in English originated from Euro-American or high-income settings, which may limit the generalizability of the review's conclusions. Third, due to the heterogeneity of study designs and the inclusion of both empirical and conceptual papers, no formal meta-analysis was conducted, and the synthesis remained primarily narrative. In addition, although efforts were made to assess study quality using validated tools, some included papers, particularly perspectives and editorials, lacked sufficient methodological detail, which may limit the generalizability of their conclusions. Finally, because the field of AI in mental health education is rapidly evolving, newer studies and technologies may have emerged since the time of data collection, warranting future updates to this review.

Conclusion

To conclude, this scoping review provided an overview of how AI is being integrated into the teaching and learning of psychiatry and psychology. By analyzing 10 studies, 8 distinct categories of AI uses were identified, ranging from clinical decision support and educational content generation to digital therapeutic tools and policy development. These findings highlight the emerging role of AI not only as a clinical adjunct but also as a transformative educational tool that can support adaptive learning, promote efficiency, and broaden access to mental health training. Although the overall quality of the studies included was moderate to high, important challenges remain, particularly related to ethical considerations, digital literacy, and institutional readiness. As AI technologies continue to evolve, future research and curriculum development efforts should focus on promoting safe, equitable, and pedagogically sound integration of AI in mental health education. Equipping educators and students will require the integration of AI literacy into core curricula, embedding foundational topics such as ML and data ethics and fostering interdisciplinary collaboration between various departments. Faculty personnel development, institutional preparedness, and student involvement are essential drivers for successful AI adoption. Future research should prioritize rigorous, outcome-based studies and capture diverse perspectives in order to guide the development of inclusive and effective AI integration strategies in mental health education. This review underscores the need for ongoing interdisciplinary collaboration between educators, clinicians, technologists, and policymakers to ensure that future practitioners are well equipped to engage with AI in meaningful and responsible ways.

Acknowledgments

This study was funded indirectly by the La Fondation de l'Institut universitaire en santé mentale de Montréal and operating funds from IVADO for AH.

Authors' Contributions

AH and JP were involved in conceptualization. Data curation, writing—original draft, and writing—review and editing were performed by all the authors; formal analysis by AH and JP; funding acquisition, investigation, project administration, resources, and supervision by AH; methodology by AH and MD; and validation by AH and JP.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Electronic search strategy for the scoping review conducted.
[DOCX File , 17 KB - mededu_v11i1e75238_app1.docx]

Multimedia Appendix 2
PRISMA-ScR checklist.
[DOCX File , 85 KB - mededu_v11i1e75238_app2.docx]

Multimedia Appendix 3
Scoping review study selection detailed results and quality assessment.
[DOCX File , 35 KB - mededu_v11i1e75238_app3.docx]

References

1. Abernethy A, Adams L, Barrett M, Bechtel C, Brennan P, Butte A, et al. The promise of digital health: then, now, and the future. NAM Perspect 2022:1-24 [FREE Full text] [doi: 10.31478/202206e] [Medline: 36177208]
2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56 [FREE Full text] [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
3. Lee EE, Torous J, De Choudhury M, Depp CA, Graham SA, Kim H, et al. Artificial intelligence for mental health care: clinical applications, barriers, facilitators, and artificial wisdom. Biol Psychiatry Cogn Neurosci Neuroimaging 2021 Sep;6(9):856-864 [FREE Full text] [doi: 10.1016/j.bpsc.2021.02.001] [Medline: 33571718]
4. Thakkar A, Gupta A, De Sousa A. Artificial intelligence in positive mental health: a narrative review. Front Digit Health 2024 Mar 18;6:1280235 [FREE Full text] [doi: 10.3389/fdgth.2024.1280235] [Medline: 38562663]
5. Gan DZQ, McGillivray L, Han J, Christensen H, Torok M. Effect of engagement with digital interventions on mental health outcomes: a systematic review and meta-analysis. Front Digit Health 2021;3:764079 [FREE Full text] [doi: 10.3389/fdgth.2021.764079] [Medline: 34806079]
6. Hudon A, Beaudoin M, Phraxayavong K, Potvin S, Dumais A. Exploring the intersection of Schizophrenia, machine learning, and genomics: scoping review. JMIR Bioinform Biotechnol 2024 Nov 15;5:e62752 [FREE Full text] [doi: 10.2196/62752] [Medline: 39546776]
7. Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, Trivedi MH, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. Lancet Psychiatry 2016 Mar;3(3):243-250. [doi: 10.1016/S2215-0366(15)00471-X] [Medline: 26803397]
8. Schnyer DM, Clasen PC, Gonzalez C, Beevers CG. Evaluating the diagnostic utility of applying a machine learning algorithm to diffusion tensor MRI measures in individuals with major depressive disorder. Psychiatry Res Neuroimaging 2017 Jun 30;264:1-9 [FREE Full text] [doi: 10.1016/j.pscychresns.2017.03.003] [Medline: 28388468]
9. Reece AG, Danforth CM. Instagram photos reveal predictive markers of depression. EPJ Data Sci 2017 Aug 8;6(1):1-12 [FREE Full text] [doi: 10.1140/epjds/s13688-017-0110-z]
10. Wager TD, Woo CW. Imaging biomarkers and biotypes for depression. Nat Med 2017 Jan 06;23(1):16-17. [doi: 10.1038/nm.4264] [Medline: 28060802]
11. Walsh CG, Ribeiro JD, Franklin JC. Predicting risk of suicide attempts over time through machine learning. Clin Psychol Sci 2017 Apr 11;5(3):457-469 [FREE Full text] [doi: 10.1177/2167702617691560]
12. Franklin JC, Ribeiro JD, Fox KR, Bentley KH, Kleiman EM, Huang X, et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. Psychol Bull 2017 Feb;143(2):187-232. [doi: 10.1037/bul0000084] [Medline: 27841450]

13. Just MA, Pan L, Cherkassky VL, McMakin DL, Cha C, Nock MK, et al. Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth. Nat Hum Behav 2017 Oct 30;1(12):911-919 [FREE Full text] [doi: 10.1038/s41562-017-0234-y] [Medline: 29367952]
14. Chung Y, Addington J, Bearden CE, Cadenhead K, Cornblatt B, Mathalon DH, North American Prodrome Longitudinal Study (NAPLS) Consortium and the Pediatric Imaging, Neurocognition, and Genetics (PING) Study Consortium. Use of machine learning to determine deviance in neuroanatomical maturity associated with future psychosis in youths at clinically high risk. JAMA Psychiatry 2018 Sep 01;75(9):960-968 [FREE Full text] [doi: 10.1001/jamapsychiatry.2018.1543] [Medline: 29971330]
15. Fiske A, Henningsen P, Buyx A. Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. J Med Internet Res 2019 May 09;21(5):e13216 [FREE Full text] [doi: 10.2196/13216] [Medline: 31094356]
16. Chin H, Song H, Baek G, Shin M, Jung C, Cha M, et al. The potential of chatbots for emotional support and promoting mental well-being in different cultures: mixed methods study. J Med Internet Res 2023 Oct 20;25:e51712 [FREE Full text] [doi: 10.2196/51712] [Medline: 37862063]
17. Mennella C, Maniscalco U, De Pietro G, Esposito M. Ethical and regulatory challenges of AI technologies in healthcare: a narrative review. Heliyon 2024 Feb 29;10(4):e26297 [FREE Full text] [doi: 10.1016/j.heliyon.2024.e26297] [Medline: 38384518]
18. Malouin-Lachance A, Capolupo J, Laplante C, Hudon A. Does the digital therapeutic alliance exist? Integrative review. JMIR Ment Health 2025 Feb 07;12:e69294 [FREE Full text] [doi: 10.2196/69294] [Medline: 39924298]
19. Oudin A, Maatoug R, Bourla A, Ferreri F, Bonnot O, Millet B, et al. Digital phenotyping: data-driven psychiatry to redefine mental health. J Med Internet Res 2023 Oct 04;25:e44502 [FREE Full text] [doi: 10.2196/44502] [Medline: 37792430]
20. Gutierrez G, Stephenson C, Eadie J, Asadpour K, Alavi N. Examining the role of AI technology in online mental healthcare: opportunities, challenges, and implications, a mixed-methods review. Front Psychiatry 2024 May 7;15:1356773 [FREE Full text] [doi: 10.3389/fpsyt.2024.1356773] [Medline: 38774435]
21. Yeo G, Reich SM, Liaw NA, Chia EYM. The effect of digital mental health literacy interventions on mental health: systematic review and meta-analysis. J Med Internet Res 2024 Feb 29;26:e51268 [FREE Full text] [doi: 10.2196/51268] [Medline: 38421687]
22. Orsolini L, Jatchavala C, Noor IM, Ransing R, Satake Y, Shoib S, et al. Training and education in digital psychiatry: a perspective from Asia-Pacific region. Asia Pac Psychiatry 2021 Dec 07;13(4):e12501 [FREE Full text] [doi: 10.1111/appy.12501] [Medline: 34873845]
23. Bekbolatova M, Mayer J, Ong CW, Toma M. Transformative potential of AI in healthcare: definitions, applications, and navigating the ethical landscape and public perspectives. Healthcare (Basel) 2024 Jan 05;12(2):125 [FREE Full text] [doi: 10.3390/healthcare12020125] [Medline: 38255014]
24. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
25. Joanna Briggs Institute. Checklist for systematic reviews and research syntheses. JBI Global. 2017. URL: https://jbi.global/critical-appraisal-tools [accessed 2025-02-12]
26. Hong QN, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Educ Inf 2018 Dec 18;34(4):285-291 [FREE Full text] [doi: 10.3233/efi-180221]
27. Tyndall J. AACODS checklist. Flinders University, Adelaide. 2010. URL: https://dspace.flinders.edu.au/jspui/bitstream/2328/3326/4/AACODS_Checklist.pdf [accessed 2025-02-12]
28. López-Ojeda W, Hurley RA. Medical metaverse, part 2: artificial intelligence algorithms and large language models in psychiatry and clinical neurosciences. J Neuropsychiatry Clin Neurosci 2023 Oct;35(4):316-320. [doi: 10.1176/appi.neuropsych.20230117] [Medline: 37840258]
29. Banerjee M, Chiew D, Patel KT, Johns I, Chappell D, Linton N, et al. The impact of artificial intelligence on clinical education: perceptions of postgraduate trainee doctors in London (UK) and recommendations for trainers. BMC Med Educ 2021 Aug 14;21(1):429 [FREE Full text] [doi: 10.1186/s12909-021-02870-x] [Medline: 34391424]
30. Blease C, Kharko A, Annoni M, Gaab J, Locher C. Machine learning in clinical psychology and psychotherapy education: a mixed methods pilot survey of postgraduate students at a Swiss University. Front Public Health 2021;9:623088 [FREE Full text] [doi: 10.3389/fpubh.2021.623088] [Medline: 33898374]
31. Spallek S, Birrell L, Kershaw S, Devine EK, Thornton L. Can we use ChatGPT for mental health and substance use education? Examining its quality and potential harms. JMIR Med Educ 2023 Nov 30;9:e51243 [FREE Full text] [doi: 10.2196/51243] [Medline: 38032714]
32. Hudon A, Kiepura B, Pelletier M, Phan V. Using ChatGPT in psychiatry to design script concordance tests in undergraduate medical education: mixed methods study. JMIR Med Educ 2024 Apr 04;10:e54067 [FREE Full text] [doi: 10.2196/54067] [Medline: 38596832]

33. Mangold S, Ream M. Artificial intelligence in graduate medical education applications. J Grad Med Educ 2024;16(2):115-118. [doi: 10.4300/jgme-d-23-00510.1]
34. Gratzer D, Goldbloom D. Therapy and e-therapy—preparing future psychiatrists in the era of apps and chatbots. Acad Psychiatry 2020 Apr 02;44(2):231-234. [doi: 10.1007/s40596-019-01170-3] [Medline: 31898301]
35. Anzia JM. Lifelong learning in psychiatry and the role of certification. Psychiatr Clin North Am 2021 Jun;44(2):309-316. [doi: 10.1016/j.psc.2021.03.001] [Medline: 34049651]
36. Manjunatha N, Kumar C, Math S, Thirthalli J. Designing and implementing an innovative digitally driven primary care psychiatry program in India. Indian J Psychiatry 2018;60(2):236-244. [doi: 10.4103/psychiatry.indianjpsychiatry_214_18]
37. Tortora L. Beyond discrimination: generative AI applications and ethical challenges in forensic psychiatry. Front Psychiatry 2024 Mar 8;15:1346059 [FREE Full text] [doi: 10.3389/fpsyt.2024.1346059] [Medline: 38525252]
38. Dwyer DB, Falkai P, Koutsouleris N. Machine learning approaches for clinical psychology and psychiatry. Annu Rev Clin Psychol 2018 May 07;14(1):91-118. [doi: 10.1146/annurev-clinpsy-032816-045037] [Medline: 29401044]
39. Shatte ABR, Hutchinson DM, Teague SJ. Machine learning in mental health: a scoping review of methods and applications. Psychol Med 2019 Jul;49(9):1426-1448. [doi: 10.1017/S0033291719000151] [Medline: 30744717]
40. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019 Apr 04;380(14):1347-1358. [doi: 10.1056/nejmra1814259]
41. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb 9;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
42. Zhai C, Wibowo S, Li L. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. Smart Learn Environ 2024 Jun 18;11(1):28 [FREE Full text] [doi: 10.1186/s40561-024-00316-7]
43. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res 2023 Oct 04;25:e50638 [FREE Full text] [doi: 10.2196/50638] [Medline: 37792434]
44. Miner AS, Milstein A, Hancock JT. Talking to machines about personal mental health problems. JAMA 2017 Oct 03;318(13):1217-1218. [doi: 10.1001/jama.2017.14151] [Medline: 28973225]
45. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. Can J Psychiatry 2019 Jul;64(7):456-464 [FREE Full text] [doi: 10.1177/0706743719828977] [Medline: 30897957]
46. D'Alfonso S, Lederman R, Bucci S, Berry K. The digital therapeutic alliance and human-computer interaction. JMIR Ment Health 2020 Dec 29;7(12):e21895 [FREE Full text] [doi: 10.2196/21895] [Medline: 33372897]
47. Cureton D, Jones J, Hughes J. The postdigital university: do we still need just a little of that human touch? Postdigit Sci Educ 2021 Dec 21;3(1):223-241 [FREE Full text] [doi: 10.1007/s42438-020-00204-6] [Medline: 40477229]
48. Tolentino R, Baradaran A, Gore G, Pluye P, Abbasgholizadeh-Rahimi S. Curriculum frameworks and educational programs in AI for medical students, residents, and practicing physicians: scoping review. JMIR Med Educ 2024 Jul 18;10:e54793 [FREE Full text] [doi: 10.2196/54793] [Medline: 39023999]
49. Sanusi I, Ayanwale M, Chiu T. Investigating the moderating effects of social good and confidence on teachers' intention to prepare school students for artificial intelligence education. Educ Inf Technol 2023 Nov 06;29(1):273-295 [FREE Full text] [doi: 10.1007/s10639-023-12250-1]
50. Scherer R, Teo T. Unpacking teachers' intentions to integrate technology: a meta-analysis. Educ Res Rev 2019 Jun;27:90-109 [FREE Full text] [doi: 10.1016/j.edurev.2019.03.001]
51. Granström M, Oppi P. Assessing teachers' readiness and perceived usefulness of AI in education: an Estonian perspective. Front Educ 2025 Jun 19;10:1622240. [doi: 10.3389/feduc.2025.1622240]
52. Pörn R, Braskén M, Wingren M, Andersson S. Attitudes towards and expectations on the role of artificial intelligence in the classroom among digitally skilled Finnish K-12 mathematics teachers. LUMAT: Int J Math Sci Technol Educ 2024;12(3):53-77 [FREE Full text] [doi: 10.31129/lumat.12.3.2102]
53. Molefi R, Ayanwale M, Kurata L, Chere-Masopha J. Do in-service teachers accept artificial intelligence-driven technology? The mediating role of school support and resources. Comput Educ Open 2024 Jun;6:100191 [FREE Full text] [doi: 10.1016/j.caeo.2024.100191]
54. Rasul T, Nair S, Kalendra D, Balaji M, Santini FDO, Ladeira WJ, et al. Enhancing academic integrity among students in GenAI Era: a holistic framework. Int J Manag Educ 2024 Nov;22(3):101041. [doi: 10.1016/j.ijme.2024.101041]
55. Bozkurt A. GenAI et al.: cocreation, authorship, ownership, academic ethics and integrity in a time of generative AI. Open Praxis 2024:1-10 [FREE Full text] [doi: 10.55982/openpraxis.16.1.654]

Abbreviations

AI: artificial intelligence
CBT: cognitive behavioral therapy
HEI: higher education institutions
JBI: Joanna Briggs Institute

ML: machine learning
MMAT: Mixed Methods Appraisal Tool
NLP: natural language processing
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
SCT: script concordance test

Edited by AH Sapci; submitted 30.03.25; peer-reviewed by L Ng, JJ Beunza; comments to author 01.05.25; revised version received 24.06.25; accepted 15.07.25; published 28.07.25.

Please cite as:
Prégent J, Chung VHA, El Adib I, Désilets M, Hudon A
Applications of Artificial Intelligence in Psychiatry and Psychology Education: Scoping Review
JMIR Med Educ 2025;11:e75238
URL: https://mededu.jmir.org/2025/1/e75238
doi:10.2196/75238
PMID:

Review

Effects of the Hidden Curriculum in Medical Education: Scoping Review

Sebastian Parra Larrotta[1*], MD; Erwin Hernando Hernández Rincón[2*], MSc, MD, PhD; Daniela Niño Correa[1*], MD; Claudia Liliana Jaimes Peñuela[2], MHPE, MD; Alvaro Enrique Romero Tapia[3], MSc, MD

[1]School of Medicine, Universidad de La Sabana, Chia, Colombia

[2]Department of Family Medicine and Public Health, School of Medicine, Universidad de la Sabana, Chia, Colombia

[3]Department of Mental Health and Psychiatry, School of Medicine, Universidad de La Sabana, Chía, Colombia

[*]these authors contributed equally

Corresponding Author:
Erwin Hernando Hernández Rincón, MSc, MD, PhD
Department of Family Medicine and Public Health
School of Medicine
Universidad de la Sabana
Campus Universitario Puente del Común, Km 7 de la Autopista Norte
Chia, 250001
Colombia
Phone: 57 1 8615555
Email: erwinhr@unisabana.edu.co

Abstract

Background: Medical education now focuses on developing skilled and dependable professionals, with particular attention to the hidden curriculum and its influence on professionalism and humanism.

Objective: This scoping review aimed to analyze the available evidence on the benefits and adverse effects of the hidden curriculum in medical education.

Methods: A scoping review of the literature available in the indexed databases PubMed, Scopus, ScienceDirect, and Latin American and Caribbean Health Sciences Literature (LILACS) with MeSH (Medical Subject Headings) descriptors was conducted on the effects of the hidden curriculum in medical education between January 2000 and April 2024. A total of 29 papers were selected for the review.

Results: Our review included studies from 10 countries, most of which were descriptive and cross-sectional, revealing both positive and negative impacts of the hidden curriculum in medical education. These include the transmission of implicit values and the influence on forming skills and professional identity. It was found that some elements contributed to the integral development of students, and others generated challenges that affected the quality of medical education. Likewise, the need for further research to design implementation strategies in different medical schools was described.

Conclusions: The hidden curriculum proves to have both a positive and negative impact on the attitudes and values of medical students. The findings highlight the need to generate greater awareness and proactive strategies in educational institutions to improve the quality of training and promote the holistic development of future health professionals.

(JMIR Med Educ 2025;11:e68481) doi:10.2196/68481

KEYWORDS

hidden curriculum; medical education; professionalism; concept formation; professional ethics; humanism

Introduction

Background

Over the years, medical education has undergone essential transformations to train skilled and reliable professionals who promote health in all people without forgetting the humanistic attitude that the profession demands [1]. In response, medical schools have committed themselves to building a curriculum that allows them to provide society with physicians who respond adequately to the population's health needs and who are efficient in the practice of their profession [2].

Thus, the curriculum has been the subject of research, where it has been explored in such a way that more than one type of curriculum has been found within the educational process [3]. This has led to the recognition that medical education is a cultural process influenced by external forces, which is how, in 1968, the term 'hidden curriculum' gained importance. Jackson [4] described it as "the tacit ways in which knowledge and behavior are constructed, outside the formally programmed courses and subjects" [5]. A study by Hafferty and O'Donnell [6] first documented the hidden curriculum phenomenon in medical education by observing how students develop their professional identity through hidden curriculum instead of formal learning experiences. According to Hafferty and O'Donnell [6], "Hidden curriculum are the customs, rituals, and taken-for-granted aspects of education in the health professions, particularly those that learners experience during interactions with faculty and clinicians in practice settings." In contrast to formal and null curricula with clear rules and expectations, the hidden curriculum is rarely planned or stated. Through the use of hidden curriculum, instructors were able to shape students' ideas about the significance and applicability of patients' knowledge. During this process, learning is generated with no apparent relation to what has been previously established, which is why it can be seen as not very significant; however, evidence shows that it plays an essential role in the fulfillment of educational goals and that it should be included in academic training programs [6,7].

The term hidden curriculum is frequently, though inaccurately, used interchangeably with informal or implicit curriculum; however, these terms have distinct conceptual meanings in medical education literature. Lawrence et al [8] describe the hidden curriculum as the implicit influences arising specifically from institutional culture, organizational structure, and social interactions, significantly shaping professional development and medical identity. In contrast, the informal curriculum refers to spontaneous, unstructured learning that occurs without formal planning, either inside or outside formal educational environments. Meanwhile, the implicit curriculum includes educational content integrated implicitly into institutional practices and expectations, though never explicitly documented [8]. Clearly distinguishing these terms is essential, as the unique nature of the hidden curriculum presents particular methodological challenges for objective measurement and evaluation of its impacts. To facilitate understanding and conceptual differentiation among hidden, informal, and implicit curricula, Table 1 summarizes their key characteristics and provides illustrative examples within medical education.

Table 1. Differentiating hidden, informal, and implicit curricula in medical education[a].

Curriculum type	Definition and key concept	Examples
Hidden	Implicit influences arising from institutional culture, organizational structure, and social interactions	Implicit hierarchy among medical specialties and gender biases in clinical practice
Informal	Spontaneous, unstructured learning experiences occurring without formal planning	Informal discussions among students and faculty and spontaneous clinical discussions in corridors or cafeterias
Implicit	Educational content implicitly integrated into institutional practices, not formally documented	Tacit expectations of professionalism and implicitly assumed ethical behaviors in clinical practice

[a]Adapted from the study by Lawrence et al [8].

Considering the points mentioned previously in the text, it is known that some skills, behaviors, and values inherent to the medical profession are learned through the hidden curriculum [9]. Empathy, respect, confidentiality, and care are among the values learned through experiences in clinical settings. In the hospital setting, the medical community faces challenges ranging from supervised and respectful knowledge construction to witnessing hierarchy, mistreatment, and exploitation by teaching and care staff during their academic process [10].

Objectives

The primary objective of this review was to conduct a comprehensive analysis of the available evidence regarding the impact of the hidden curriculum in medical education, examining both its benefits and adverse effects. This includes assessing how informal learning experiences, unstructured interactions, and observed behaviors within clinical and educational environments can positively influence the professional and personal development of medical students. In addition, the review aims to investigate potential negative outcomes, such as the perpetuation of undesirable values or stereotypes, and their impact on educational quality and student well-being.

Despite the evident impact of the hidden curriculum, much of the existing research focuses primarily on formal teaching practices, neglecting the hidden dimensions of medical education. As a result, this literature remains scattered and fails to provide a comprehensive analysis of the impact of the hidden curriculum in shaping medical education.

This study aims to address that gap by focusing on specific research questions:

- What are the main effects of the hidden curriculum on medical training, both positive and negative?
- What strategies have been proposed or implemented to mitigate the negative effects of the hidden curriculum in medical education?

By showcasing the varied importance of the hidden curriculum across different medical schools, the review seeks to provide evidence that will encourage other institutions to recognize and incorporate the hidden curriculum into their learning objectives and goals. Ultimately, the review aims to offer practical recommendations for integrating the hidden curriculum in a way that enhances overall medical education and better prepares future health care professionals.

Methods

Ethical Considerations

Written informed consent and ethics approval were not required due to the nature of the study.

Study Design

The PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) statement was used to perform this scoping review.

Eligibility Criteria

This review primarily included qualitative studies with an interpretative perspective, featuring articles that used thematic analysis and grounded theory methodology, as well as opinion pieces, systematic reviews, scoping reviews, and literature reviews. The study excluded quantitative research, focusing instead on gaining deeper insights and understanding the nuanced impact of the hidden curriculum rather than measuring it statistically. Editorial comments and letters to the editor were excluded. Eligible participants were undergraduate and postgraduate medical students. Studies involving other health professions, such as nursing, pharmacy, nutrition, dentistry, and psychology, were not included. Articles included those published between 2000 and April 2024, with full-text availability, whose titles or abstracts incorporated research on the effect of the hidden curriculum in medical education, including values such as professionalism, medical ethics, and humanism. The 2000 to 2024 timeframe was chosen to capture the evolution of research on the hidden curriculum in medical education. Other curricula different than the hidden curriculum were excluded. We included postgraduate students in residency programs such as anesthesia, surgery, family medicine, emergency medicine, and radiology because of their relevance to study objectives and the impact of the hidden curriculum within these fields.

Information Sources

The PubMed, Scopus, ScienceDirect, and Latin American and Caribbean Health Sciences Literature (LILACS) databases were systematically searched for literature published between January 2000 and April 2024.

Search Strategy

The search strategy followed systematic search principles and was guided by the population, intervention, comparison, and outcome (PICO) framework. The search terms included the keywords “Education,” “Medical,” “Curriculum,” and “Professionalism” using the Boolean operator AND to obtain results for the search query “Medical Education” [MeSH] AND “Hidden Curriculum” [MeSH] AND “Professionalism” [Mesh]. Experts were not formally consulted during the development of the search strategy. Only qualitative studies published between 2000 and 2024 were eligible for inclusion in this review. The search was limited to studies published in English, Spanish, or Portuguese. The review did not use a standardized quality assessment tool, as its primary aim was to map existing research rather than conduct a detailed critical assessment of individual studies.

Selection of Sources of Evidence

Two authors (SPL and DNC) screened the titles and abstracts of identified studies based on the inclusion and exclusion criteria. The full texts of the shortlisted studies were analyzed and evaluated independently for eligibility by the same 2 authors (SPL and DNC). In instances of uncertainty, the other 3 authors (EHHR, CLJP, and ART) were consulted, and decisions were made by consensus.

Data Charting Process (Data Extraction)

Two reviewers (SPL and DN) independently gathered relevant data from the articles included in the review. These data covered several aspects: title, article characteristics (eg, the publication year, country, and field of education), participant details (eg, educational degree and medical residency), and specifics related to the intervention (eg, actual impact and future outcomes in professional skills). Any disagreements between the reviewers were resolved through consultation with the other 3 authors (EHHR, CLJP, and ART), and decisions were reached by consensus.

Data Items

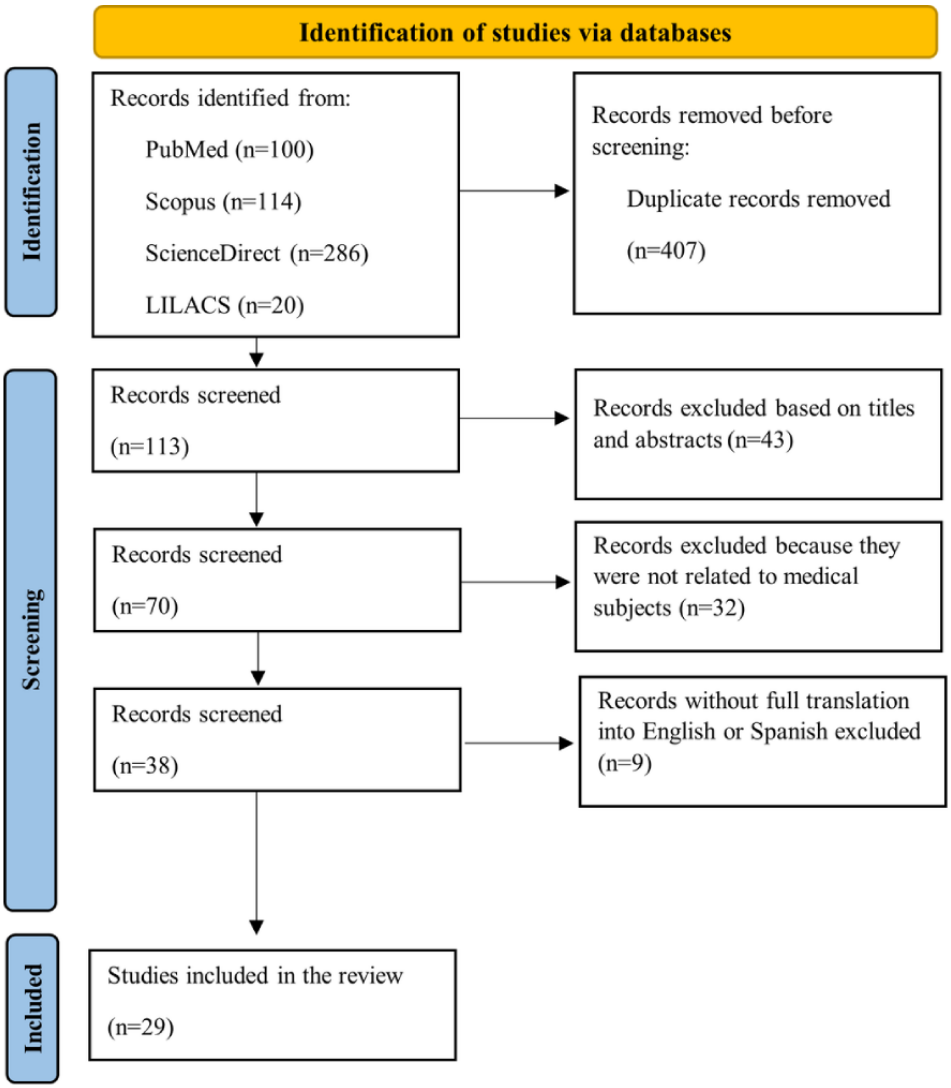
Articles were included if they featured any independent variable related to the following: presence and characteristics of the hidden curriculum (ie, implicit values, behaviors, and attitudes within medical education settings), educational contexts (ie, academic institutions and hospital settings), and educational stages (ie, undergraduate and postgraduate).

Results

Study Selection

A total of 520 documents were identified in the databases through an electronic literature search, including 100 from PubMed, 114 from Scopus, 286 from ScienceDirect, and 20 from LILACS. After duplicates were removed, 113 articles were screened. The initial title and abstract screening excluded 43 articles, leaving 70. Of these, 32 focused on health areas other than medicine, and 9 lacked full translation into English or Spanish. Excluding these 41 articles left 29 for inclusion in the qualitative synthesis of the scoping review (Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram for new systematic review that include searches of databases and registers only.



Study Characteristics

The characteristics of all included studies are summarized in Multimedia Appendix 1 [11-39]. Publication dates ranged from 2007 to 2024. Most studies were conducted in the United States (14/29,48%) [11,16-20,24,25,28,31,33,34,38,39], followed by the United Kingdom (4/29, 14%) [12,13,36,37], Iran (3/29, 10%) [14,15,35], Brazil (2/29, 7%) [29,32], Canada (1/29, 3%) [22], Sweden (1/29, 3%) [21], Ireland (1/29, 3%) [23], the Netherlands (1/29, 3%) [27], Israel (1/29, 3%) [26], and Argentina (1/29, 3%) [30]. In terms of methodology, qualitative studies were most common (15/29, 52%), using different designs such as thematic analysis and grounded theory (6/29, 21%) [12,13,15,24,26,29], evaluation design (3/29, 10%) [17,19,33], exploratory and observational approaches (3/29, 10%) [14,23,30], intervention studies (2/29, 7%) [16,33], and discourse analysis (1/29, 3%) [34]; other study types included literature reviews (3/29, 10%) [20,27,33], mixed methods (2/29, 7%) [18,28], systematic and scoping reviews (3/29, 10%) [21,25,35], a position paper (1/29, 3%) [11], integrative reviews (2/29, 7%) [32,36], an innovation report (1/29, 3%) [38], a retrospective study (1/29, 3%) [31], and a perspective article (1/29, 3%) [39].

Synthesis of Results

In terms of content, the articles included in this scoping review revealed different benefits [13-25] and adverse effects [26-31] of the hidden curriculum present in medical education during different moments of academic training, such as undergraduate [8,10,11,13-19,24-31] and postgraduate education [20-23], as well as tools and strategies for its implementation within the learning objectives and goals of different medical schools for the development of professional skills in medical personnel in training [11,33-37]. Similarly, 3 categories were considered for the analysis of the available literature: “benefits of the hidden curriculum in medical education,” “negative effects of the hidden curriculum in medical education,” and “implementation strategies and limitations within the medical education process.”

Categorization and Frequency of Effects of the Hidden Curriculum

The effects of the hidden curriculum were categorized and quantified into 2 primary domains: positive (beneficial) and negative (adverse). Among the 29 studies analyzed, (27/29, 93%) reported at least one positive outcome, including the development of professional identity (12/29, 41%), ethical and

moral reasoning (7/29, 24%), empathy and humanistic values (5/29, 17%), and reflective thinking (4/29, 14%). Meanwhile, (17/29, 59%) described adverse effects, such as reinforcement of medical hierarchies (6/29, 21%), emotional detachment or stress (5/29, 17%), and value incongruence between formal and informal teachings (5/29, 17%). A detailed summary of these categorizations is presented in Table 2. In this process, multiple studies reported more than one effect.

Table 2. Categorization and frequency of effects of the hidden curriculum (N=29).

Effect category and subcategory	Studies, n (%)	References
Positive[a]		
Professional identity formation	12 (41)	[11-13,16,18,20,26,29,30,32,34,37]
Ethical and moral development	7 (24)	[11,12,15,20,25,32,37]
Empathy and humanistic values	5 (17)	[13,24,25,31,36]
Reflective thinking and critical judgment	4 (14)	[12,17,18,23]
Negative[b]		
Reinforcement of hierarchy or authoritarianism	6 (21)	[10,19,20,22,28,33]
Emotional stress and detachment	5 (17)	[13,16,17,23,31]
Incongruence between taught and modeled values	5 (17)	[11,21,26,33,37]

[a]Categories are not mutually exclusive; some studies contributed to more than one subtheme.

[b]Seventeen of the 29 studies reported at least one adverse effect. The subcategories listed capture most of these, although some studies contributed negative aspects not easily classified under the three main themes.

Discussion

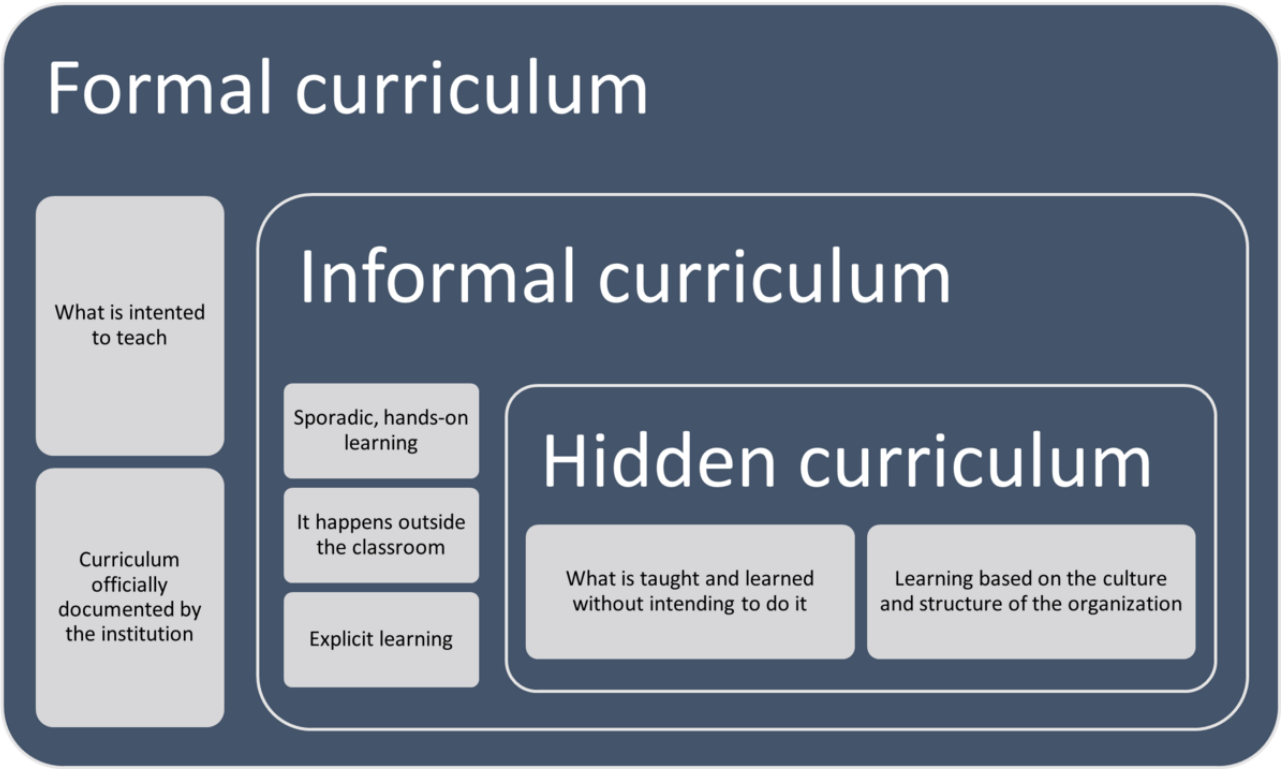
Principal Findings

This scoping review investigated the role of the hidden curriculum in medical education and its positive and negative aspects. The results highlight critical insights: one of the key findings of this study is that the hidden curriculum is an unavoidable, powerful force that shapes how medical trainees develop their professional identities, medical professionalism, and humanism. The findings of this study also show that a hidden curriculum can effectively strengthen the physician-patient relationship. The hidden curriculum proved to be the dominant factor that determines professional adaptability for medical practitioners who face ongoing changes in clinical practice. Medical students learn essential professional attitudes and behaviors through the pervasive influences of the hidden curriculum. The hidden education system surpasses traditional classroom learning because it shapes students' moral values and ethical conduct as well as their approach to making important decisions. This review also suggests that medical institutions should create proactive approaches and refine procedures for an implicit hidden curriculum that affects medical professionals.

Comparison With Prior Literature

The review aligns with the existing research showing how the hidden curriculum's power remains unrecognized in its ability to mold medical student experiences. The first description of the hidden curriculum was made in 1970 by Jackson [5]; however, it was not until 1994 that its relationship with medical education was defined by Hafferty [10], who stated that the training and learning process of the medical profession was influenced by cultural aspects and by the context in which the students developed and attributed to the hidden curriculum the reason for most of their mistakes [7]. Culture significantly shapes the hidden curriculum by influencing medical students to incorporate professional standards and core values through educational settings that differ across societies. Some cultural influences encourage patient-centered care and professionalism, but other cultural elements maintain hierarchical structures while promoting discrimination and emotional suppression. Likewise, learning has been approached through 3 dimensions of the curriculum, which are represented in Figure 2 [6,10]: the formal curriculum, what is documented to offer; the informal curriculum, how the curriculum is carried out through the interactions of the community; and the hidden curriculum, all that learning that occurs outside the expected and stipulated [11], which is where skills that cannot be taught through the formal curriculum are developed.

Figure 2. Types of curricula described in the literature (adapted from the studies by Hafferty and O'Donnell [6] and Hafferty [10]).



Given the previously mentioned positive and negative effects, several authors in the analyzed literature explore the hidden curriculum. They highlight both the benefits and drawbacks of the hidden curriculum in medical training, where values such as empathy, respect, and professionalism are cultivated. Nevertheless, even the physician-patient relationship can be affected by negative interactions and experiences in clinical practice settings [12].

However, this review expands on previous research by highlighting the benefits of the hidden curriculum. First, a benefit is described in the search, construction, and strengthening of professional identity [13]. When students face negative experiences and behaviors in the hospital environment, it is necessary to confront their own identity with the one they are expected to acquire, which generates a reflection about the professionals they wish to become in the future. The hidden curriculum gives students the skills they need to adjust to different professional situations in addition to helping them establish their identities. The development of critical soft skills, including empathy, teamwork, and communication, is positively impacted by the hidden curriculum [13]. Medical students who master the interpretation and understanding of the hidden curriculum become better prepared for dynamic health care environments, demonstrating greater resilience, innovative problem-solving, and a proactive commitment to lifelong learning [18].

This review finding aligns with the findings by Azmand et al [14] that a hidden curriculum can effectively strengthen physician-patient relationships, where the role models of teachers motivate them to dedicate adequate time to patients, recognizing their concerns beyond their medical conditions and needs. Physician-patient relationships embedded within the hidden curriculum substantially enhance patient outcomes. The hidden curriculum enhances formal medical education by incorporating patient-centered approaches into medical students' everyday lives, which eventually helps both practitioners and the patients they treat. These relationships were characterized by a sense of ethics, responsibility toward others, and the medical environment [15,16].

By contrast, the influence of the hidden curriculum on the construction of professionalism and the trust that patients and their families place in training professionals has been explored [17]. Teachers and students emphasized that the formal curriculum does not teach them the necessary skills to interact in a dynamic environment such as a hospital and that lived experiences allow them to develop skills and attitudes that are inherent to the medical profession, among which respect, responsibility, compassion, and communication, among others, stand out [18]. The hidden curriculum promotes the value of moral decision-making, transparency, and confidentiality, all of which help to foster trust with patients as well as colleagues. The hidden curriculum is evident at all stages of training [19], from undergraduate to postgraduate programs such as general surgery [20], family medicine [21], radiology [22], and anesthesia [23].

Finally, regarding its benefits, the hidden curriculum has been shown to influence dimensions such as religion, spirituality [24], and humanism [25]. Through these experiences, students developed coping strategies for emotional stress when treating patients facing serious illness, family difficulties, or death. They also acquired values such as empathy, communication, respect, and emotional stability—skills learned through real-world practice rather than the formal curriculum.

Through a scoping review of the literature, it was determined that the available evidence, to a great extent, revealed the adverse effects of the hidden curriculum in medical education. Mistreatment [26] and hierarchical dynamics toward students and residents were among the most frequently reported issues, often linked to decreased motivation and commitment to the training process [27]. The students mentioned that 2 different teachers in the same service could improve or affect the learning environment and their performance. Thus, mistreatment by colleagues and teachers during the selection process of medical-surgical specialties has even been described [28], which negatively impacts their decisions. Studies link mistreatment exposure to higher stress levels and burnout symptoms, along with diminished empathy, which harms both the students' mental state and their competency in providing patient care [28].

In addition, another component of the hidden curriculum that generates adverse effects is the burden and pressure on undergraduate and graduate students [29]. They explained that they were required to respond in an accelerated manner to their obligations to their patients, which affected their care in terms of time and quality. This attitude of immediacy in practice scenarios is often generated in teachers by a conflict between the teaching role and the caring role [30], in which, by neglecting one or appropriating the other, attention and respect for the patient or their teaching activity with students are affected. The implicit demands of the hidden curriculum may cause stress, which could have a negative impact on students' general well-being and long-term ability to adjust professionally. Addressing this role conflict is essential, as medical training should prioritize proper practice, integrating clear instruction with comprehensive patient care.

The dual impact of the hidden curriculum on medical education, its capacity to promote humanistic values and simultaneously induce stress, has been recently explored. Auckley et al [31], through their analysis of the Humanism in Medicine Initiative, demonstrate that while extracurricular humanism-focused activities significantly enhance student well-being and professional identity, moderate engagement paradoxically correlates with increased stress. Such findings underline the nuanced role of the hidden curriculum in shaping medical students' experiences and highlight the need for balanced engagement.

Regarding humanism, although some studies highlight the benefit of the hidden curriculum in integrating these qualities with medical professionalism, others suggest a disconnect between clinical skills and the humanistic aspects of medical education [32]. Skills such as communication, decision-making, respect, and autonomy are often acquired through implicit learning, which tends to receive less emphasis. This limited focus may reduce the humanistic approach to patient care when these students later practice as physicians [32]. Table 3 compares the positive and negative effects of the hidden curriculum.

Table 3. Comparison of the positive and negative effects of the hidden medical education curriculum.

Feature	Positive effects	Negative effects
Professional identity	Consolidation of professional identity through confrontation between personal and expected roles [13]	Imposed identity leading to dissonance and emotional distress [13]
Medical professionalism	Reflection spaces to analyze emotions and promote respect and empathy toward teachers, colleagues, and patients [11,17]	Mistreatment, humiliation, and hierarchy normalized as acceptable cultural patterns [11]
Physician-patient relationship	Recognition of patient concerns and needs beyond disease, fostering trust-based relationships [14-16]	Loss of patient interaction opportunities due to efficiency culture and high workload [29,30]
Wellness and mental health	Promotion of self-care, spirituality, and mental well-being [24]	Clinical distress, burnout, low well-being, and depression affecting performance [11]
Humanism	Adoption of communication, empathy, respect, and autonomy through the hidden curriculum [25]	Dehumanization in medical education and patient care perpetuated across generations [11]

Implementation Strategies and Challenges in Addressing the Hidden Curriculum

To date, the benefits and adverse effects of the hidden curriculum in medical education have been analyzed. This leads to identifying strategies to enhance the virtues and mitigate the unfavorable effects it may have on medical students and the limitations of its implementation. First, studies highlight the importance of constantly evaluating the learning environments in educational institutions and clinical practice scenarios [33,34] so that they become a safe and adequate environment for the acquisition of professional and personal skills for patient care, recognizing that the educational process is based not only on the knowledge to be acquired but also on multicultural education that strengthens professional identity and ethics [34].

Likewise, the importance of strengthening institutional culture in practice scenarios has been described based on humanistic attitudes, such as compassion, curiosity, respect, and empathy, enabling students to identify with these values and carry them into their professional practice [11]. This process should involve the student in situations where they must make ethically complex decisions and face emotions that impact them, supported by teacher guidance and opportunities for reflection [11].

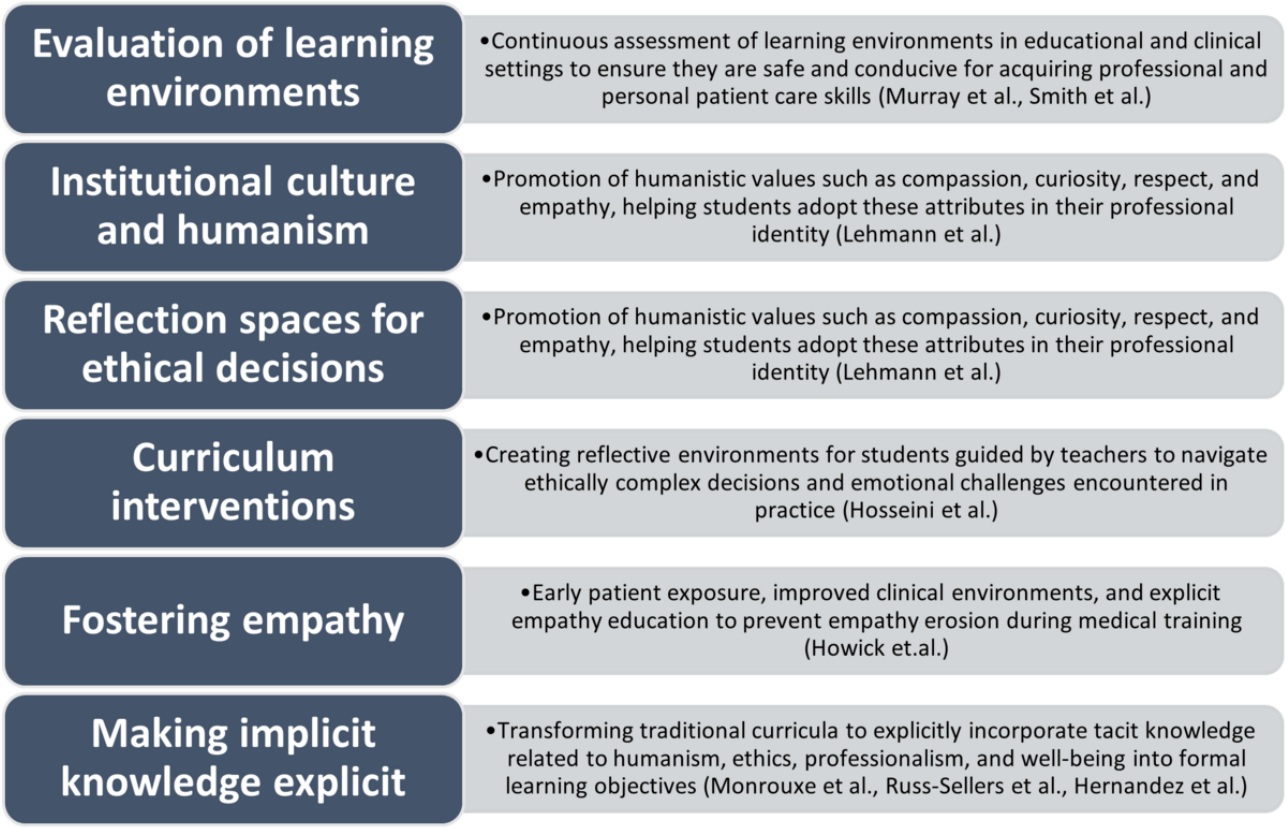
Moreover, recent literature highlights specific interventions aimed at mitigating negative impacts and optimizing the hidden curriculum. Hosseini et al [35] identify key strategies, such as the implementation of new curricula, team-based clinical clerkships, and longitudinal faculty development workshops. These strategies are essential to manage the implicit values and unintended messages conveyed through informal learning

experiences and interactions, reducing the adverse outcomes associated with the hidden curriculum.

In addition, Howick et al [36] propose a transformative approach to leverage the hidden curriculum as a tool for fostering empathy among medical students. They suggest evidence-based interventions, such as early patient exposure, improving clinical environments, and explicit empathy education as critical components of what they term an "empathic hidden curriculum." Such approaches are crucial in countering empathy erosion frequently observed during medical training.

Finally, one of the most significant challenges, with several limitations, is to include some of the knowledge acquired through the hidden curriculum within an explicit curricular objective [37]. Educational environments are increasingly aware of the dissociation that exists between what they wish to teach and what their students learn; strategies of early immersion of students with the communities they are going to serve, including an early approach to hospital environments [38] and accompaniment based on humanism, professionalism, and ethics [39], make it possible to put into practice some of the tacit learning and make it visible. The main limitation during this process and a source of future research is the transformation of traditional curriculum and strategies to include aspects of humanism, well-being, and medical professionalism within the learning objectives of future professionals [39]. Some of these strategies are illustrated in Figure 3.

Figure 3. Hidden curriculum implementation strategies and limitations in medical education.



Implications

The findings of this study have significant implications. Aspects of the hidden curriculum should be incorporated into official educational frameworks by educational policymakers. Universities ought to spend money on extracurricular activities that promote the development of professional skills.

To guide future research, several key recommendations are proposed. First, comparative studies evaluating the effectiveness of different educational approaches, such as integrating the hidden curriculum versus more explicit methods or practical versus theoretical strategies, would be valuable. In addition, incorporating digital interventions could provide insights into how technology can influence medical training, given recent advancements in this area. It is also crucial to consider implementation costs and discuss resource accessibility to enable financial and logistical analyses of proposed interventions. Long-term follow-up studies of students would be beneficial to assess how they apply learned lessons in their clinical practice. This will help determine if medical education achieves its goals of producing competent and ethical professionals and explore whether the hidden curriculum practices result in tangible benefits for patients and improvements in educational quality.

Strengths

The study's strength lies in its use of qualitative research methods, which capture detailed perspectives on the hidden curriculum in medical education. The research investigates medical education from 2000 to 2024 to provide a comprehensive analysis of historical and recent developments. The extended timeframe enables researchers to detect both evolving patterns and enduring difficulties that come from the hidden curriculum. The comprehensive range of themes studied

across different publications emerges as a key strength in this evaluation. This assessment brings together research from different medical disciplines to demonstrate how the hidden curriculum expresses itself distinctly between specialties, thus advancing our collective understanding of its effects.

Limitations

This scoping review focused primarily on qualitative studies with interpretative perspectives and excluded quantitative research, which may have limited the breadth of evidence regarding the hidden curriculum's impact. Certain studies, such as those involving other health professions (eg, nursing and pharmacy), were excluded, potentially omitting relevant insights that could provide a broader understanding of the hidden curriculum's impact across different fields. The exclusion of Web of Science, Embase, and Google Scholar may have led to the omission of relevant studies, potentially limiting the comprehensiveness of the findings. The review was limited to studies published in English, Spanish, and Portuguese, which might exclude relevant research published in other languages. In addition, the inclusion of only published studies could introduce publication bias, as studies with negative findings may be less likely to be published. One methodological limitation was not using specific Boolean operators, such as NOT, to explicitly differentiate between hidden, informal, and implicit curricula during the literature search, potentially affecting the breadth and specificity of identified studies. While the review discussed strategies for integrating the hidden curriculum into formal curriculum, it did not fully address the practical challenges and barriers that institutions might face when implementing these strategies.

Conclusions

This scoping review has described various aspects of the hidden curriculum in medical education and its potential consequences. A synthesis of the available literature shows that the hidden curriculum conveys explicit knowledge and has many subtle but significant effects on medical students. From forming attitudes and values to influencing professional identity, the hidden curriculum has emerged as a crucial factor in shaping the educational experience.

As understanding of these effects progresses, the need for greater awareness and reflection within different medical schools is highlighted. This review underscores the importance of proactively addressing the hidden curriculum, recognizing it as an essential component for the holistic development of future health professionals. Implementing strategies to maximize positive aspects and minimize negative aspects could significantly improve the quality of medical education and contribute to future physicians' personal and professional growth.

Acknowledgments

This research was derived from project MED-341-2023 at Universidad de La Sabana, Colombia.

Authors' Contributions

SPL and DNC conceptualized and designed the study. SPL conducted the data extraction and drafted the initial version of the manuscript. DNC contributed to the data analysis. CLJP critically reviewed and revised the manuscript for important intellectual content. All authors read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Summary of the documents included in the review.
[DOCX File , 13 KB - mededu_v11i1e68481_app1.docx]

Multimedia Appendix 2
PRISMA-ScR checklist.
[PDF File (Adobe PDF File), 101 KB - mededu_v11i1e68481_app2.pdf]

References

1. Pinzón C. The great paradigms of medical education in Latin America. Acta Med Colomb 2008;33(1):33-41.
2. Vílchez NG. A review and update of the concept of curriculum. Telos 2004;6(2):194-208.
3. Reyes Y, Burgos M, de HL, Cascioli N. A look at strategic curriculum planning. Telos 2010;12(2):202-216. [doi: 10.4018/978-1-5225-9242-6.ch002]
4. Jackson PW. Life in Classrooms. New York, NY: New York, Holt, Rinehart and Winston; 1968.
5. Wear D. On white coats and professional development: the formal and the hidden curricula. Ann Intern Med 1998 Nov 01;129(9):734-737 [FREE Full text] [doi: 10.7326/0003-4819-129-9-199811010-00010] [Medline: 9841607]
6. Hafferty FW, O'Donnell JF. The Hidden Curriculum in Health Professional Education. Hanover, MD: Dartmouth College Press; 2015.

7. Centeno A, De la Paz Grebe M. The hidden curriculum and its influence on teaching in the health sciences. Inv Ed Med 2021 Apr 07;10(38):89-95 [FREE Full text] [doi: 10.22201/fm.20075057e.2021.38.21350]
8. Lawrence C, Mhlaba T, Stewart K, Moletsane R, Gaede B, Moshabela M. The hidden curricula of medical education: a scoping review. Acad Med 2018 Apr;93(4):648-656 [FREE Full text] [doi: 10.1097/ACM.0000000000002004] [Medline: 29116981]
9. Fortoul van der Goes TI, Núñez-Fortoul A. What we say and what we do, incongruence in the teaching of good habits: the hidden curriculum. Inv Ed Med 2013 Jul 01;2(7):119-121. [doi: 10.1016/S2007-5057(13)72699-1]
10. Hafferty F. Beyond curriculum reform: confronting medicine's hidden curriculum. Acad Med 1998 Apr;73(4):403-407 [FREE Full text] [doi: 10.1097/00001888-199804000-00013] [Medline: 9580717]
11. Lehmann L, Sulmasy L, Desai S. Hidden curricula, ethics, and professionalism: optimizing clinical learning environments in becoming and being a physician: a position paper of the American College of Physicians. Ann Intern Med 2018 Apr 03;168(7):506-508 [FREE Full text] [doi: 10.7326/m17-2058]
12. Mossop L, Dennick R, Hammond R, Robbé I. Analysing the hidden curriculum: use of a cultural web. Med Educ 2013 Feb;47(2):134-143 [FREE Full text] [doi: 10.1111/medu.12072] [Medline: 23323652]
13. Brown M, Coker O, Heybourne A, Finn G. Exploring the hidden curriculum's impact on medical students: professionalism, identity formation and the need for transparency. Med Sci Educ 2020 Sep;30(3):1107-1121 [FREE Full text] [doi: 10.1007/s40670-020-01021-z] [Medline: 34457773]
14. Azmand S, Ebrahimi S, Iman M, Asemani O. Learning professionalism through hidden curriculum: Iranian medical students' perspective. J Med Ethics Hist Med 2018;11:10 [FREE Full text] [Medline: 31346387]
15. Safari Y, Khatony A, Tohidnia M. The hidden curriculum challenges in learning professional ethics among Iranian medical students: a qualitative study. Adv Med Educ Pract 2020;11:673-681 [FREE Full text] [doi: 10.2147/AMEP.S258723] [Medline: 33061738]
16. Shorey JM. Signal versus noise on the wards: what "messages" from the hidden curriculum do medical students perceive to be importantly meaningful? Trans Am Clin Climatol Assoc 2013;124:36-45. [Medline: 23874008]
17. Mackin R, Baptiste S, Niec A, Kam A. The hidden curriculum: a good thing? Cureus 2019 Dec 06;11(12):e6305 [FREE Full text] [doi: 10.7759/cureus.6305] [Medline: 31938597]
18. Karnieli-Miller O, Vu T, Frankel R, Holtman M, Clyman S, Hui S, et al. Which experiences in the hidden curriculum teach students about professionalism? Acad Med 2011 Mar;86(3):369-377. [doi: 10.1097/ACM.0b013e3182087d15] [Medline: 21248599]
19. Nittur N, Kibble J. Current practices in assessing professionalism in United States and Canadian allopathic medical students and residents. Cureus 2017 May 22;9(5):e1267 [FREE Full text] [doi: 10.7759/cureus.1267] [Medline: 28652951]
20. Rogers D, Boehler M, Roberts N, Johnson V. Using the hidden curriculum to teach professionalism during the surgery clerkship. J Surg Educ 2012;69(3):423-427 [FREE Full text] [doi: 10.1016/j.jsurg.2011.09.008] [Medline: 22483148]
21. Rothlind E, Fors U, Salminen H, Wändell P, Ekblad S. The informal curriculum of family medicine – what does it entail and how is it taught to residents? a systematic review. BMC Fam Pract 2020 Mar 11;21(1):1-1 [FREE Full text] [doi: 10.1186/S12875-020-01120-1]
22. Van Deven T, Hibbert K, Faden L, Chhem R. The hidden curriculum in radiology residency programs: a path to isolation or integration? Eur J Radiol 2013 May;82(5):883-887 [FREE Full text] [doi: 10.1016/j.ejrad.2012.12.001] [Medline: 23305755]
23. Moran PJ, Bates JJ. The hidden curriculum in anaesthesia. Ir Med J 2019 Feb 14;112(2):872. [Medline: 30892005]
24. Balboni M, Bandini J, Mitchell C, Epstein-Peterson Z, Amobi A, Cahill J, et al. Religion, spirituality, and the hidden curriculum: medical student and faculty reflections. J Pain Symptom Manage 2015 Oct;50(4):507-515 [FREE Full text] [doi: 10.1016/j.jpainsymman.2015.04.020]
25. Martimianakis M, Michalec B, Lam J, Cartmill C, Taylor J, Hafferty F. Humanism, the hidden curriculum, and educational reform: a scoping review and thematic analysis. Acad Med 2015 Nov;90(11 Suppl):S5-13. [doi: 10.1097/ACM.0000000000000894] [Medline: 26505101]
26. Karnieli-Miller O, Vu T, Holtman M, Clyman S, Inui T. Medical students' professionalism narratives: a window on the informal and hidden curriculum. Acad Med 2010 Jan;85(1):124-133. [doi: 10.1097/ACM.0b013e3181c42896] [Medline: 20042838]
27. Boer C, Daelmans H. Team up with the hidden curriculum in medical teaching. Br J Anaesth 2020 Mar;124(3):e52-e54 [FREE Full text] [doi: 10.1016/j.bja.2019.12.031] [Medline: 31973827]
28. Oser T, Haidet P, Lewis P, Mauger D, Gingrich D, Leong S. Frequency and negative impact of medical student mistreatment based on specialty choice: a longitudinal study. Acad Med 2014 May;89(5):755-761. [doi: 10.1097/ACM.0000000000000207] [Medline: 24667501]
29. Silveira G, Campos L, Schweller M, Turato E, Helmich E, de Carvalho-Filho M. “Speed up”! the influences of the hidden curriculum on the professional identity development of medical students. Health Prof Educ 2019 Sep;5(3):198-209 [FREE Full text] [doi: 10.1016/j.hpe.2018.07.003]
30. Montesinos M. Hidden curriculum in surgery: what else do they learn when we teach? Rev Argent Cir 2012 Sep;103(1):9-15 [FREE Full text]

31. Auckley E, Barbee J, Verbeck N, McCambridge T, Stone L, Garvin J. Extracurricular humanism in medicine initiative and medical student wellness: retrospective study. JMIR Form Res 2022 Sep 16;6(9):e37252 [FREE Full text] [doi: 10.2196/37252]
32. Martins e Silva J. Educação Médica e Profissionalismo. Acta Med Port 2013 Aug 30;26(4):420-427. [doi: 10.20344/amp.1284]
33. Murray-García JL, García JA. The institutional context of multicultural education: what is your institutional curriculum? Acad Med 2008 Jul;83(7):646-652. [doi: 10.1097/ACM.0b013e3181782ed6] [Medline: 18580080]
34. Smith K, Saavedra R, Raeke J, O'Donell AA. The journey to creating a campus-wide culture of professionalism. Acad Med 2007 Nov;82(11):1015-1021. [doi: 10.1097/ACM.0b013e318157633e] [Medline: 17971683]
35. Hosseini A, Ghasemi E, Nasrabadi AN, Sayadi L. Strategies to improve hidden curriculum in nursing and medical education: a scoping review. BMC Med Educ 2023 Sep 11;23(1):658 [FREE Full text] [doi: 10.1186/s12909-023-04652-z]
36. Howick J, Slavin D, Carr S, Miall F, Ohri C, Ennion S, et al. Towards an empathic hidden curriculum in medical school: a roadmap. J Eval Clin Pract 2024 Jun;30(4):525-532 [FREE Full text] [doi: 10.1111/jep.13966] [Medline: 38332641]
37. Monrouxe L, Rees C, Hu W. Differences in medical students' explicit discourses of professionalism: acting, representing, becoming. Med Educ 2011 Jun;45(6):585-602. [doi: 10.1111/j.1365-2923.2010.03878.x] [Medline: 21564198]
38. Russ-Sellers R, Blackwell T. Emergency medical technician training during medical school: benefits for the hidden curriculum. Acad Med 2017 Jul;92(7):958-960. [doi: 10.1097/ACM.0000000000001579] [Medline: 28145946]
39. Hernandez S, Nnamani Silva ON, Conroy P, Weiser L, Thompson A, Mohamedaly S, et al. Bursting the hidden curriculum bubble: a surgical near-peer mentorship pilot program for URM medical students. J Surg Educ 2022;79(1):11-16 [FREE Full text] [doi: 10.1016/j.jsurg.2021.07.003] [Medline: 34315681]

Abbreviations

LILACS: Latin American and Caribbean Health Sciences Literature
PICO: population, intervention, comparison, and outcome
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

Edited by B Lesselroth; submitted 08.11.24; peer-reviewed by S Peng, KK Daugherty, A Elsenousi; comments to author 08.03.25; revised version received 08.04.25; accepted 16.07.25; published 15.09.25.

Please cite as:
Parra Larrotta S, Hernández Rincón EH, Niño Correa D, Jaimes Peñuela CL, Romero Tapia AE
Effects of the Hidden Curriculum in Medical Education: Scoping Review
JMIR Med Educ 2025;11:e68481
URL: https://mededu.jmir.org/2025/1/e68481
doi:10.2196/68481
PMID:

Review

Evaluating the Potential and Accuracy of ChatGPT-3.5 and 4.0 in Medical Licensing and In-Training Examinations: Systematic Review and Meta-Analysis

Anila Jaleel[1*], PhD; Umair Aziz[1*], MBBS; Ghulam Farid[1*], MPhil; Muhammad Zahid Bashir[1], MBBS; Tehmasp Rehman Mirza[1]; Syed Mohammad Khizar Abbas[1]; Shiraz Aslam[1*]; Rana Muhammad Hassaan Sikander[1*]

Shalamar Medical and Dental College, Lahore, Pakistan

[*]these authors contributed equally

Corresponding Author:
Anila Jaleel, PhD
Shalamar Medical and Dental College
Shalimar Link Road, Mughalpura
Lahore, 54000
Pakistan
Phone: 92 3009205779
Email: anilajaleel@gmail.com

Abstract

Background: Artificial intelligence (AI) has significantly impacted health care, medicine, and radiology, offering personalized treatment plans, simplified workflows, and informed clinical decisions. ChatGPT (OpenAI), a conversational AI model, has revolutionized health care and medical education by simulating clinical scenarios and improving communication skills. However, inconsistent performance across medical licensing examinations and variability between countries and specialties highlight the need for further research on contextual factors influencing AI accuracy and exploring its potential to enhance technical proficiency and soft skills, making AI a reliable tool in patient care and medical education.

Objective: This systematic review aims to evaluate and compare the accuracy and potential of ChatGPT-3.5 and 4.0 in medical licensing and in-training residency examinations across various countries and specialties.

Methods: A systematic review and meta-analysis were conducted, adhering to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. Data were collected from multiple reputable databases (Scopus, PubMed, JMIR Publications, Elsevier, BMJ, and Wiley Online Library), focusing on studies published from January 2023 to July 2024. Analysis specifically targeted research assessing ChatGPT's efficacy in medical licensing exams, excluding studies not related to this focus or published in languages other than English. Ultimately, 53 studies were included, providing a robust dataset for comparing the accuracy rates of ChatGPT-3.5 and 4.0.

Results: ChatGPT-4 outperformed ChatGPT-3.5 in medical licensing exams, achieving a pooled accuracy of 81.8%, compared to ChatGPT-3.5's 60.8%. In in-training residency exams, ChatGPT-4 achieved an accuracy rate of 72.2%, compared to 57.7% for ChatGPT-3.5. The forest plot presented a risk ratio of 1.36 (95% CI 1.30-1.43), demonstrating that ChatGPT-4 was 36% more likely to provide correct answers than ChatGPT-3.5 across both medical licensing and residency exams. These results indicate that ChatGPT-4 significantly outperforms ChatGPT-3.5, but the performance advantage varies depending on the exam type. This highlights the importance of targeted improvements and further research to optimize ChatGPT-4's performance in specific educational and clinical settings.

Conclusions: ChatGPT-4.0 and 3.5 show promising results in enhancing medical education and supporting clinical decision-making, but they cannot replace the comprehensive skill set required for effective medical practice. Future research should focus on improving AI's capabilities in interpreting complex clinical data and enhancing its reliability as an educational resource.

(JMIR Med Educ 2025;11:e68070) doi:10.2196/68070

KEYWORDS

ChatGPT; medical education; accuracy of ChatGPT; ChatGPT-3.5 performance; ChatGPT-4.0 performance; artificial intelligence; AI in health care; medical licensing examinations; clinical decision-making

Introduction

Artificial intelligence (AI) has penetrated virtually every field, ranging from telecommunications to medicine, as scientific literature has explored its potential impact, ramifications, limitations, and potential uses. In the health care industry, AI has made significant inroads: it has personalized treatment plans and simplified workflows [1]. Radiology has undergone a remarkable revolution: the US Food and Drug Administration approved AI-equipped devices in 2023 [2]. Through advanced algorithms and data processing techniques, it can aid in making informed clinical decisions [3]. Nevertheless, the complex procedure of clinical decision-making involves much more than pattern recognition and predictive analytics. However, some studies have highlighted the efficacy of artificial cognitive empathy, which highlights the growth of AI in the "soft skills" department, that is, empathy, ethics, and judgment, of clinical practice [3,4].

Perhaps AI's biggest breakthrough was ChatGPT, launched by OpenAI in November 2022 [5]. ChatGPT is a large language model (LLM) trained on a vast dataset designed to mimic human responses. A freely available conversational AI model, ChatGPT, has revolutionized health care and medical education. Through the simulation of clinical scenarios, students were able to gain therapeutic insights. Academic evaluation is another widely explored use. Some work has shown its potential in improving communication skills. The latest version of OpenAI's LLM has also allowed for the customization of teaching and learning plans by tailoring to individual preferences [6].

Despite the significant advancements in medical education and patient care, the performance of ChatGPT in special medical licensing examinations (MLEs) has been inconsistent, varying between countries and among specialties. One of the studies demonstrated that GPT showed the highest accuracy in Italian examinations (73% correct answers) and the lowest in French examinations (22% correct answers) [1]. Another study found that the LLM could not pass the National Specialty Examination in the Polish Education System. On the contrary, GPT-3 demonstrated tremendous performance in the United States Medical Licensing Examination (USMLE). These highlight the variance and disparity of ChatGPT's accuracy across different exams, raising questions about its reliability in particular contexts and the general structure of the examinations [7].

There is also much interest regarding the accuracy gap between ChatGPT-3.5 and ChatGPT-4. The latter version was the improved one due to its training on a larger dataset. In a study involving the USMLE, ChatGPT-4 outperformed ChatGPT, correctly answering 90% compared to ChatGPT-3.5's 2.5% [8,9].

In light of these advancements, it is evident that AI, particularly in the form of LLMs such as ChatGPT, has significantly impacted health care, medical education, and clinical decision-making. The disparity between ChatGPT-3.5 and ChatGPT-4 performances, as well as the variability between countries and specialties, underscores the importance of refining these technologies and adapting them to specific medical systems and environments. Although systematic reviews have been done, but to our knowledge, no systematic review and meta-analysis are available so far on comparative analysis and accuracy of ChatGPT versions [10-12].

Due to the inconsistent performance of AI models across various MLEs, there is a need for further research into the contextual factors that influence AI accuracy. Moreover, as AI continues to evolve, there is a pressing need to explore its potential in enhancing not only technical proficiency but also "soft skills" such as empathy and ethical judgment, ensuring that AI becomes a reliable, holistic tool in patient care and medical education. Despite the rapid advances in AI, especially LLMs such as ChatGPT, its performance on medical licensing and in-training examinations remains inconsistent across countries and specialties. This study set out to systematically evaluate and compare the accuracy of ChatGPT versions 3.5 and 4.0 when tested on real medical licensing and residency exams from around the world. Our research specifically asked: Does GPT-4 offer a meaningful improvement in answering medical exam questions compared to GPT-3.5? To answer this, we conducted a systematic review and meta-analysis of studies published between January 2023 and July 2024, following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. We expected, based on prior evidence, that GPT-4 would outperform GPT-3.5, given its expanded training and improved reasoning capabilities. The findings are intended to guide medical educators, exam boards, and AI developers in understanding the strengths and limitations of these tools, and how they might fit into medical training and assessment in the years ahead.

Methods

The review followed the PRISMA guidelines for conducting systematic reviews and meta-analyses.

Information Sources

Data were sourced from reputable databases such as Scopus, PubMed, Wiley Online Library, JMIR Publications, Wolters Kluwer OVID-SP, Hindawi, Taylor & Francis, Science Direct, ProQuest, Sage Publications, BMJ, and Google Scholar in July 2024.

Search String

We systematically reviewed the related literature using different sets of keywords, such as "ChatGPT" OR "medical license exam" OR "ChatGPT" OR "medical education" OR "ChatGPT" OR "USMLE" OR "AI in medical education" OR "artificial intelligence" OR "ChatGPT" OR "LLM" OR "exam performance" OR "machine learning" OR "ChatGPT accuracy" OR "Generative Pre-trained Transformer" OR "accuracy in medical exam" OR "license exam" OR "healthcare exam" OR

"clinical exam" OR "ChatGPT-3.5 and 4 in residency exam" OR "accuracy of ChatGPT" OR "ChatGPT and authenticity" AND "Generative Model" (Survey* OR qualitative* OR editorial* OR questionnaire* OR letter to editor* OR empirical study).

Inclusion Criteria

This review includes studies focused on the application of ChatGPT in medical education and its involvement in medical licensing exams, such as the USMLE and in-training residency programs, published between January 2023 and July 2024. Only English-language studies were considered to ensure clarity and accessibility. The studies included range from peer-reviewed journal articles, editorials, and case reports to letters to the editor, conference papers, meeting papers, and dissertations. By incorporating a variety of study types, the review aims to provide a comprehensive understanding of the role of ChatGPT-3.5 and 4.0 in medical education, capturing detailed research findings, expert perspectives, real-world case examples, emerging trends, and original research.

Exclusion Criteria

Studies that did not meet the specific focus of this review were excluded, including book chapters, as they tend to provide broad overviews rather than original research focused on ChatGPT's application in medical education. Additionally, any studies on ChatGPT's use in fields outside of medical education, such as higher education disciplines (eg, engineering and humanities) or unrelated health care domains, were excluded to maintain a sharp focus on the intersection of AI and medical education.

Study Selection and Data Extraction

The PRISMA diagram outlines the process of selecting 53 studies for inclusion. Each study was analyzed using a material extraction framework, which included author and publication year, title, country, methodology, key findings of ChatGPT-3.5 and 4.0, advantages and disadvantages, conclusions, and type of exam.

To ensure the reliability of the screening and eliminate duplicates, 3 independent reviewers (TRM, SA, and SMKA) reviewed abstracts. Any discrepancies between the 3 were reconciled by senior reviewers (UA and GF). Abstracts were downloaded and screened using the Rayyan.ai tool, with csv file formats. The selected manuscripts were also screened independently for full text by the authors (AJ and RMHS), and disagreements were resolved by discussion with the lead reviewer (MZB).

Authors (SA, TRM, UA, and GF) independently extracted and synthesized the comparative accuracy data from the included studies in the Rayyan.ai tool. These were done manually; no automation tools were used. Any discrepancies in extracted data were discussed and resolved by consensus with reviewers (AJ and MZB). Data extraction included sensitivity, accuracy, precision, and CIs. A meta-analysis of aggregated data was conducted with a random-effects inverse-variance model, with RevMan 5.4.

The PRISMA chart outlines the systematic study selection process for evaluating ChatGPT in MLEs. Initially, 1215 records were identified from multiple databases, including PubMed, Scopus, and Google Scholar. After removing 732 duplicate records, 496 studies underwent title and abstract screening, resulting in 79 relevant records. Of these, 75 full-text articles were assessed for eligibility, with 4 excluded due to unavailability. Nine additional articles were excluded for irrelevance, leaving 66 studies for the final review. The primary reasons for exclusion during screening included non-English language and studies unrelated to MLEs (eg, knowledge tests or library use). This rigorous process ensured a focused and comprehensive analysis of ChatGPT's performance in medical exams.

Quality Assessment

The Cochrane Collaboration tool was used to assess Risk of Bias in Nonrandomized Interventional Studies (ROBINS-I) by using the following domains (Figure 1): (1) bias due to confounding, (2) bias due to selection of participants, (3) bias in classification of interventions, (4) bias due to deviations from intended interventions, (5) bias due to missing data, (6) bias in measurement of outcome, and (7) bias in selection of the reported results. According to predefined criteria 2, the domains were rated as "low risk," "unclear risk," and "high risk." The risk in all domains was low across studies included in the comparative meta-analysis, including Flores-Cohaila et al [13], Takagi et al [14], and Lewandowski et al [15], and indicates good methodological quality and bias. A frequent finding in studies by Brin et al [16] and Meyer et al [17] is a moderate risk of confounding bias (D1) denoted by yellow symbols in this domain across multiple studies. This means that confounding factors were not entirely isolated, which possibly might affect the study outcomes [16,17].

Figure 1. Risk of bias assessment [13-17,19,20,28,31,32,35,40-52].

Studies presenting a serious risk of bias in one or more domains, such as Wang et al [46], were noted. In particular, any studies in which substantial bias was identified in either outcome measurement or in the selection of reported results (ie, domains D6 and D7) were marked with a red indicator. Examples include Angel et al [19], Moshirfar et al [43], and Toyama et al [45]. The serious risks substantially undermine the internal validity of the study findings and need to be duly considered in the context of the conclusions. Though many of the included studies were rated as having a low to moderate risk of bias, few general methodological flaws and a few critical concerns regarding confounding control and outcome reporting were identified. The strong accent on the importance of bias minimization, rigorous study design, and transparent reporting highlights the need for meticulous attention to detail to enhance the validity of evidence collated in systematic reviews and meta-analyses.

Results

Comparison of the Accuracy of ChatGPT Versions in Medical Licensing and Training Residency Examinations

Figure 2 presents the PRISMA flowchart illustrating the study selection process for ChatGPT in medical licensing and residency examinations.

Table 1 shows 53 studies carried out by several researchers showing the capability of ChatGPT-3.5 and ChatGPT-4 in medical licensing and in-training residency exams conducted in various countries around the globe. It also shows the comparative evaluation of the success rate of both versions of ChatGPT in licensing examinations. Advantages and disadvantages are also described in this table.

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of study selection process for ChatGPT in medical licensing examinations.

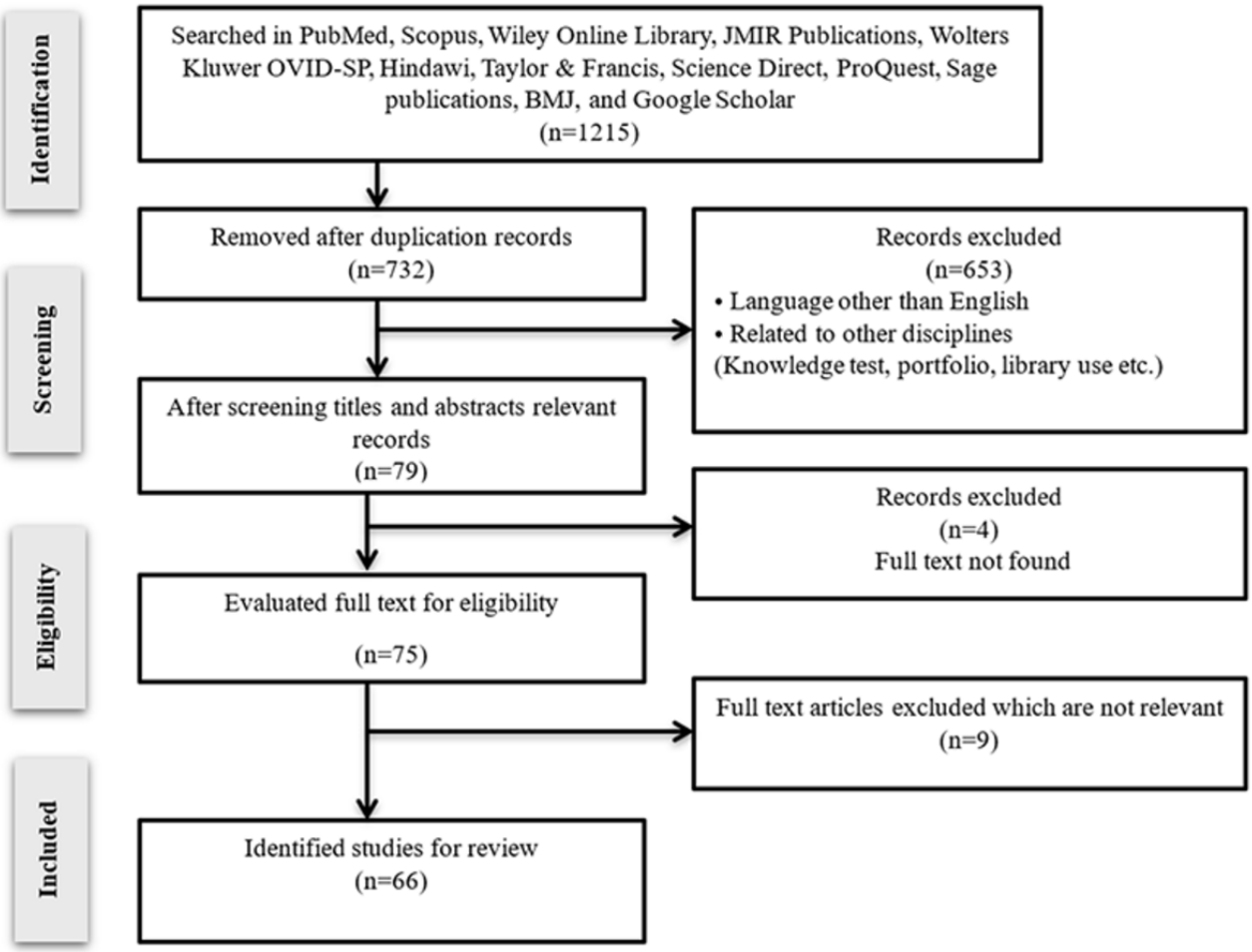


Table 1. Overview of studies comparing the accuracy of ChatGPT-3.5 and 4.0 in medical licensing and in-training residency exams.

Study ID	Exam type	Dataset and methods	ChatGPT-4.0 accuracy	ChatGPT-3.5 accuracy	Advantages	Disadvantages
Aljindan et al (2023) [18]	Saudi Medical Licensure Exam	A dataset of 220 questions across 4 medical disciplines.	88.6%	None	Useful supplementary educational tool.	A high margin of error (4-16%) emphasizes that it should be used with traditional study methodology.
Angel et al (2023) [19]	American Board of Anesthesiology (ABA) Exam	The basic exam used the full set of 60 MCQs[a] on the ABA website. The advanced exam was formulated using the book *Anesthesia Review: 1000 Questions and Answers to Blast the BASICS and Ace the ADVANCED* by randomly choosing 5/70 questions.	GPT-4 scored 47/60 (78.33%) in the basic exam. In the advanced exam, the model scored 80%.	GPT-3 scored 35/60 (58.33%).	The performance of the AI[b] in this process was impressive enough that it leads one to question whether we should not already, as highly trained clinicians, be using these AI systems to help us avoid some of the more common cognitive errors that may occur during critical events.	The models struggle to fully comprehend questions involving numerical calculations, leading to imprecise responses. Additionally, numerical calculations demand greater precision and exactness, which may not always be within the model's capabilities.
Antaki et al (2023) [20]	Ophthalmology	Two 260-question simulated exams from the Basic and Clinical Science Course (BCSC) Self-Assessment Program and the Ophthalmology Questions online question bank were solved by ChatGPT-3.5.	None	ChatGPT Plus's accuracy increased to 59.4. Accuracy improved with easier questions when controlling for the examination section and cognitive level.	Generally, we found that ChatGPT Plus provided highly consistent and repeatable results, but some variations occurred.	ChatGPT does not have the capability to process images. This is a significant limitation because ophthalmology is a field that heavily relies on visual examination and imaging to diagnose, treat, and monitor patients.
Bartoli et al (2024) [21]	Neurosurgical residents' written exam	51 questions (open-ended and MCQ) were included in the written exam, 46 questions were generated by humans (senior staff members), and 4 were generated by ChatGPT.	None	ChatGPT scored among the lowest ranks (9/11) among all the participants. ChatGPT answered correctly to all its self-generated questions.	Different ways to exploit AI for medical education and exams are imaginable; that is, a question may be posed by a human, an answer is generated by AI, and a critical appraisal is then elaborated by the human, based on their own logical thinking, knowledge, and personal experience and creativity.	At this stage, it seemed like AI did not "intend" or "think" to probe someone's specific knowledge on a topic when generating a question, like humans do.
Beaulieu-Jones et al (2024) [22]	Surgical knowledge assessments	Two surgical knowledge assessments, the Surgical Council on Resident Education (SCORE; 167 questions) and Data-B (112 questions), were solved by ChatGPT-4.0.	ChatGPT correctly answered 71% and 68% of multiple-choice SCORE and Data-B questions, respectively.	None	The study highlights the accuracy of ChatGPT within a highly specific and sophisticated field without specific training or fine-tuning in the domain.	The findings also underscore some of the current limitations of AI, including variable performance on the same task and unpredictable gaps in the model's capabilities.

Study ID	Exam type	Dataset and methods	ChatGPT-4.0 accuracy	ChatGPT-3.5 accuracy	Advantages	Disadvantages
Benirschke et al (2024) [23]	Questions from My Pathologist	61 general knowledge pathology questions were solved by ChatGPT-4.0.	98% of responses were "completely/mostly accurate," and 82% of responses had "all relevant information."	None	Answers to general pathology inquiries were accurate and complete, and LLMs[c] can potentially be used to save time for pathology professionals.	Differences were seen in the completeness of the answer, with clinical pathology answers judged as more complete than anatomic pathology answers.
Brin et al (2023) [16]	USMLE[d]	80 USMLE-style questions involving soft skills, from the AMBOSS question bank. ChatGPT-3.5 and 4.0 took the examination, and the results were compared.	90%	62.5%	LLMs demonstrate impressive results in questions that test soft skills required from physicians. GPT-4 surpassed the human performance benchmark.	Models operate based on calculated probabilities for an output, rather than human-like confidence.
Chen et al (2023) [24]	Neurology written board examination questions	509 eligible questions out of the 560 from BoardVitals QBank (without images) were solved by ChatGPT-4.0.	335/509 questions (65.8%) on the first attempt and 383/509 (75.3%) in the subsequent iterations.	None	ChatGPT was sensitive to questions regarding depression and suicide and referred us to a suicide hotline.	The base model lacks visual input and thus is unable to process image-based questions. There were differences in the accuracy of different subject fields.
Cheung et al (2023) [25]	A multinational prospective study	50 MCQs were generated by ChatGPT-4.0 with two standard undergraduate medical textbooks. Another 50 MCQs were drafted by 2 professors using the same.	None	The total time required for ChatGPT to create the 50 questions was 20 minutes and 25 seconds.	As demonstrated in our study, a reasonable MCQ can be written by ChatGPT using simple commands with a text reference provided.	Including ChatGPT, most AI models are trained by the vast content available on the internet, and their reliability and credibility are questionable. Moreover, many AI models were found to have significant bias due to their training data.
Ebrahimian et al (2023) [26]	Iranian Medical Licensing Examination	200 MCQs were translated into English. The accuracy of performance by ChatGPT-4.0 was assessed.	68.5% (surpassed the passing criteria of 45%)	None	ChatGPT is capable of answering MCQs, surpassing the pass mark by a significant margin.	The expertise and nuanced understanding of medical cases that human physicians possess are not yet matched by AI models.
Fang et al (2023) [27]	Chinese National Medical Licensing Examination	Out of 600 questions, 340 were common questions, and 260 case analysis questions were solved by ChatGPT-4.0.	442/600 (73.67%), surpassing the passing criteria, that is, 360.	None	ChatGPT exhibits a high level of answer-explanation concordance in the Chinese language.	ChatGPT's performance declines when handling encoded questions, declining by 40.5%, 9.7%, 32.4%, and 41.8% for Units 1, 2, 3, and 4, respectively.

Study ID	Exam type	Dataset and methods	ChatGPT-4.0 accuracy	ChatGPT-3.5 accuracy	Advantages	Disadvantages
Flores-Cohaila et al (2023) [13]	Peruvian National Licensing Medical Examination (ENAM)	180 multiple-choice questions were performed by ChatGPT-3.5 and 4.0, and the results were compared.	86% (155/180). GPT-4 surpassed almost 90% of examinees.	77% (139/180), with prompting. GPT-3.5 surpassed 80% of examinees.	Role-play and context-setting in prompts improved performance, reducing incorrect answers. Outperformed human performance.	LLM expertise is limited to passing a licensing exam. To be a practitioner requires communication skills, empathy, and so on.
Giannos, (2023) [28]	UK Neurology Specialty Certificate Examination	69 questions from the Pool—Specialty Certificate Examination (SCE) Neurology Web Questions bank were solved by both versions of ChatGPT.	ChatGPT-4.0 achieved the highest accuracy of 64%, surpassing the passing threshold and outperforming its predecessors across disciplines and subtopics.	ChatGPT-3.5 displayed an overall accuracy of 57%.	ChatGPT-4 has shown promise in attaining specialty-level medical knowledge. This sets a new benchmark for AI models in specialized medical education and practice.	ChatGPT3's performance in the specialized field of neurology and neuroscience is lower than that in general medical examinations. Specialty examinations require a deeper understanding of specific medical domains.
Gilson et al (2023) [29]	USMLE	4 datasets with 389 questions were solved by ChatGPT-3 and compared with medical students' performance in steps 1 and 2.	None	AMBOSS Step 1: 44% (44/100); AMBOSS Step 2: 42% (42/100); NBME[e] Step 1: 64.4% (56/87); NBME Step 2: 57.8% (59/102)	The model may facilitate the creation of an on-demand, interactive learning environment for students. Logical explanations for answers are always provided, even if incorrect.	Incorrect answers are related to logical and information-based errors mostly.
Gobira et al (2023) [30]	Brazilian National Examination for Medical Degree Revalidation	81 nonnullified and 14 nullified questions. ChatGPT-4.0 solved questions in various specialties.	For nonnullified questions, 71/81 (87.7%); for nullified questions, 71.4% (10/14).	None	There was no statistically significant difference in the performance across different specialties.	Encountered challenges when tackling questions that involved concepts related to the Brazilian public health care system and ethical decisions.
Guillen-Grima et al (2023) [31]	Spanish MIR (Medical Resident Intern)	182 questions were solved by both versions of ChatGPT and compared.	Spanish: 86.81% (81.13-90.98); English: 87.91% (82.38-91.88).	Spanish: 63.18% (55.98-69.85); English: 66.48% (59.35-72.94)	Improved consistency demonstrated by GPT-4 across multiple attempts presents the refinements in training and better underlying model architecture.	The error rate of 13.2% is of concern. 25 questions linked to an image and 3 challenge questions were excluded.
Haddad et al (2024) [32]	Ophthalmology Examinations	Questions from the USMLE Step 1 (n=44), Step 2 (n=60), and Step 3 3 (n=28) were extracted from AMBOSS, and 248 questions were extracted from the book *Ophthalmology Q&A Board Review* and solved by both versions, and results were compared for accuracy.	GPT-4.0 achieved a total of 70% of correct answers. GPT-4.0 answered 70.45% of questions correctly in step 1, 90.32% in step 2, 96.43% in step 3, and 62.90% in the remaining 248 questions.	GPT-3.5 achieved a total of 55% of correct answers. GPT-3.5 answered 75% of questions correctly in step 1, 73.33% in step 2, 60.71% in step 3, and 46.77% in the OB-WQE[f].	ChatGPT can be a great add-on to mainstream resources to study for board examinations. There have been reports of using it to generate clinical vignettes and board examination–like questions.	Many questions were excluded due to them containing images, which is a considerable limitation considering the visual nature of ophthalmology.

Study ID	Exam type	Dataset and methods	ChatGPT-4.0 accuracy	ChatGPT-3.5 accuracy	Advantages	Disadvantages
Hoch et al (2023) [33]	otolaryngology subspecialties	A dataset covering 15 otolaryngology subspecialties was collected from an online learning platform funded by the German Society of Oto-Rhino-Laryngology, Head and Neck Surgery, and was given to ChatGPT-3.5 for solving.	None	The dataset included 2576 questions (479 multiple-choice and 2097 single-choice), of which 57% (n=1475) were answered correctly by ChatGPT.	As an educational resource, the performance of ChatGPT indicated potential efficacy in offering educational assistance in specific subspecialties and question formats.	ChatGPT delivered a considerable number of incorrect responses within specific otolaryngology subdomains, rendering it unreliable as the sole resource for residents preparing for the otolaryngology board examination.
Huang et al (2024) [34]	SPTEMD[g] Stage 1	600 MCQs extracted from 3 separate tests conducted in February 2022, July 2022, and February 2023 were solved by ChatGPT-4.0.	525/600 (87.5%)	None	Potential to facilitate not only the preparation for exams but also improve the accessibility of medical education and support continuous education for medical professionals.	The model performed inconsistently across different subjects, exhibiting comparatively lower performance in anatomy, parasitology, and embryology.
Huang et al (2023) [35]	University of Toronto Family Medicine Residency Progress Test	The 108 questions were stratified into 11 areas of family medicine knowledge and solved by both versions of ChatGPT.	82.4% (89/108). It scored much higher than an average resident, that is, 56.9%.	57.4% (62/108). Its performance was comparable to the average resident, that is, 56.9%.	GPT-4 demonstrates a broad knowledge base and strong reasoning abilities in FM[h], as evidenced by its high level of accuracy and logical justification.	Hallucinations raise concerns about a model's integrity and overall accuracy, and may mislead learners into believing an incorrect response to be correct.
Jain et al (2023) [36]	Orthopedic In-Training Examination (OITE)	Of 635 questions, 360 were usable as inputs (56.7%) by ChatGPT-3.5.	None	ChatGPT-3.5 scored 55.8%, 47.7%, and 54% for the years 2020, 2021, and 2022, respectively.	The potential of machine learning in medicine is that it can automate tasks, assist in providing thought processes, and improve management.	Tendency to fabricate references or have incorrect reasoning when solving problems that require logic beyond this date.
Jang et al (2023) [37]	K-NLEKMD[i]	340 (114 questions for recall, 99 for diagnosis, and 127 for intervention) were solved by ChatGPT-4.0 and assessed.	66.18% (after model optimization)	None	We found that high consistency of response is associated with increased accuracy for questions.	Models had weak performance on questions that require understanding the Korean language and TKM[j]/Korea-adapted health care. LLMs are susceptible to hallucinations.
Knoedler et al (2023) [38]	USMLE	2069 USMLE Step 3 practice questions. 1840 entered into GPT-3.5, while a subset of 229 entered into GPT-4, and performance was compared.	84.7% (194/229)	56.9% (1047/1840)	ChatGPT-4 showcases its superiority as a newer model, with better accuracy.	ChatGPT-4 encountered difficulties in questions related to cardiology and neurology.

Study ID	Exam type	Dataset and methods	ChatGPT-4.0 accuracy	ChatGPT-3.5 accuracy	Advantages	Disadvantages
Kufel et al (2023) [39]	Polish specialty exam in radiology	This study used a PES[k] exam consisting of 120 questions, provided by the Medical Examinations Center in Lodz. The performance of ChatGPT-4.0 was assessed.	None	ChatGPT did not reach the pass rate threshold of the PES exam (52%).	We identified that questions for which ChatGPT provided a correct answer had a significantly higher confidence index. Therefore, the confidence index can be considered a parameter indicating a higher likelihood of ChatGPT providing a correct answer.	The performance of the ChatGPT model in passing the specialist in radiology and imaging diagnostics examination in Poland remains uncertain.
Kung et al (2023) [41]	Orthopedic In-Training Examination (OITE)	OITE 2020, 2021, and 2022 questions without images were input into ChatGPT version 3.5 and version 4.0 with zero prompting.	GPT-4 answered 265 of 360 questions correctly, corresponding to the average performance of a PGY-5[l].	ChatGPT answered 196 of 360 answers correctly (54.3%), corresponding to a PGY-1 level.	Multiple studies have suggested that deep learning models for automated image analysis can be synergistic with clinicians, resulting in superior predictions compared with those of clinicians alone.	Finally, in medicine specifically, there can be multiple potentially correct answers to a given question with only one best answer, which may cause difficulty for the AI when there is correct information supporting each answer.
Lewandowski et al (2023) [15]	Specialty Certificate Examination in Dermatology	Three Specialty Certificate Examination in Dermatology tests, in English and Polish, consisting of 120 single-best-answer, multiple-choice questions each, were solved by both versions of AI and compared.	The percentages of correct answers to questions in the Polish vs English versions obtained by ChatGPT-4.0 were 77.3% vs 84%, 75.8% vs 85% and 71.4% vs 80.7%, respectively.	The percentages of correct answers to questions in the Polish vs English versions obtained by ChatGPT-3.5 were 61.3% vs 68.9%, 60.8% vs 70%, and 54.6% vs 60.5%, respectively.	It is estimated that ChatGPT is capable of understanding and communicating in more than 100 languages at various levels. ChatGPT-4.0 proved to be highly effective in answering clinical picture-type questions.	An important limitation of ChatGPT and the other language models is a rare phenomenon of artificial hallucination, defined as self-conscious, seemingly realistic responses by an AI. We did not observe it in this study, but it is widely described in the literature.
Lin et al (2024) [53]	Taiwan Medical Licensing Examination	80 single-choice questions were assessed by ChatGPT-4.0.	Accuracy in medical exams ranged from 63.75% to 93.75%. The highest accuracy was in the February 2022 exam.	None	The implementation of the “Chain of Thought” prompt strategy proved effective, enabling the model to correct its initial incorrect responses and achieve an accuracy rate exceeding 90%.	Falters in questions related to surgical precautions and decision-making. It had a 100% failure rate in such questions. Overall, it is not ready for complex decision-making.
Long et al (2024) [54]	Otolaryngology Head and Neck Surgery Certification Examinations (OHNS)	21 open-ended questions were adopted from the Royal College of Physicians and Surgeons of Canada’s sample examination. The accuracy of ChatGPT-4.0 was assessed on them.	Average of 75% across 3 trials in the attempts and demonstrated higher accuracy with prompts.	None	Information provided by the AI was clinically valid; it could be used to provide equitable access to resources in low-resource settings where access to such information may not be readily available.	Hallucinations may present benign or harmful misinformation, with significant implications in the field of medicine.

Study ID	Exam type	Dataset and methods	ChatGPT-4.0 accuracy	ChatGPT-3.5 accuracy	Advantages	Disadvantages
Lum (2023) [55]	Orthopedic In-Training Examination (OITE)	Random selection of 400/3840 publicly available questions based on the Orthopedic In-Training Examination, and compared the mean score with that of residents.	None	ChatGPT selected the correct answer 47%	AI can handle large amounts of data that can be quickly accessed. The LLM performed better at recognition and recall-type questions, as well as comprehension and interpretation, than at problem-solving and application of knowledge.	The model may have limitations in terms of its ability to integrate, synthesize, generalize, and apply factual knowledge in more nuanced ways.
Mackey, et al (2024) [56]	USMLE	Questions from AMBOSS were used. 900 questions from 9 specialties, that is, 100 questions from each specialty, were randomly selected. Performance by ChatGPT-4.0 was assessed for accuracy.	89% of the total questions were answered accurately.	None	ChatGPT-4.0 had an impressive performance in areas like psychiatry, neurology, obstetrics, and gynecology.	LLMs have specialty-specific strengths and weaknesses. Its accuracy was notably lower in pediatrics, emergency medicine, and family medicine.
Mahajan et al (2023) [57]	Otolaryngology Residency In-Service Exam	1088 questions from the BoardVitals otolaryngology bank were solved by ChatGPT-3.5 and accuracy was assessed.	None	572/1088 (53%) yielded a correct answer, and 586/1088 (54%) yielded a correct explanation.	LLM can accurately answer complex multiple - choice patient care questions to an extent.	The accuracy rate is far below an acceptable level for it to be useful in a clinical or educational setting to aid in decision-making.
Maitland et al (2024) [58]	MRCP^{m} (UK)	Practice questions for MRCP 1 and 2 produced by MRCPUK were solved by ChatGPT-4.0 (images excluded).	For part 1, ChatGPT provided 170 accurate responses for 196 questions. In part 2, it provided responses for 127/128 questions correctly.	None	ChatGPT is able to answer MRCP written examination questions, without additional prompts, to a level that would equate with a comfortable pass for a human candidate.	LLMs are known to “hallucinate” plausible sounding, but false, facts, information, and even references.
Massey et al (2023) [42]	Orthopedic Assessment Examination	180 questions and answer choices from 9 different orthopedic subspecialties (excluding images). The accuracy of both versions was compared.	GPT-4 scored 47.2% with a 95% CI of 40%-54.5%. The average response time was 31.8 seconds.	GPT-3.5 scored 29.4% with a 95% CI of 23.2%-36.4%. The average response time for ChatGPT-3.5 was 12.1 seconds.	The implications and potential of ChatGPT are still quite promising. GPT-4 showed particular improvement when answering questions requiring no image interpretation.	Chatbots cannot decipher radiographic images in conjunction with clinical vignettes.
Meyer et al (2024) [17]	German Medical Licensing Examination	937 original MCQs from German medical licensing examinations were solved by ChatGPT-3.5 and 4.0, and the accuracy of both was compared.	796/937 (85%)	548/937 (58%)	There is potential for using GPT-3.5 and GPT-4.0 in a medical education tutoring environment.	The model's inconsistencies across different specialties restrain recommendations for its use by the general population for medical purposes.

Study ID	Exam type	Dataset and methods	ChatGPT-4.0 accuracy	ChatGPT-3.5 accuracy	Advantages	Disadvantages
Morjaria et al (2023) [59]	Undergraduate Medical Program	40 problems used in prior assessments were used: 30 submitted to ChatGPT-3.5. For the remaining 10 problems, we retrieved past minimally passing student responses.	None	ChatGPT-generated responses received a mean score of 3.29 out of 5, compared to 2.38 for a group of students meeting minimum passing marks, presenting higher performance.	This represents success on a different form of assessment compared to the multiple-choice format. ChatGPT's performance was shown to be comparable to students who are performing at the minimum standard.	ChatGPT's performance was shown to be comparable to students who are performing at the minimum standard on these assessments, but slightly lower when compared to a historical student cohort.
Moshirfar et al (2023) [43]	StatPearls ophthalmology questions	467 questions from the StatPearls Question Bank were solved by both versions of ChatGPT and compared.	73.2%	55.5%	GPT-4.0 outperformed humans who had an accuracy of 58.3%.	The “lens and cataract” category presented a unique challenge for the model.
Nakao et al (2024) [60]	Japanese National Medical Licensing Examination	108 questions that had 1 or more images as part of a question were solved by ChatGPT-4.0.	68% (73/108) when presented with images and 72% (78/108) when presented without images.	None	For the clinical questions, for which sufficient information was available in the text form, GPT-4V was able to choose the correct answers solely from the textual information (76/98, 78%).	GPT-4V cannot effectively interpret images related to medicine.
Oh et al (2023) [44]	Korean general surgery board exams	The dataset comprised 280 questions from the Korean general surgery board exams conducted between 2020 and 2022 and solved by both versions of ChatGPT.	GPT-4 demonstrated a significant improvement with an overall accuracy of 76.4%.	GPT-3.5 achieved an overall accuracy of 46.8%.	Active surgeons who completed their training over a decade ago may find LLMs helpful for continuous medical education, especially as a supplementary resource.	Instead of providing strictly accurate information, they generate responses based on the probability of the most appropriate words given the data they have been trained on.
Panthier et al (2023) [61]	European Board of Ophthalmology examination	GPT-4 was exposed to a series of EBO^{n} examination questions in French, covering various aspects of ophthalmology, and performance was compared to experts.	ChatGPT correctly answered 6188 out of 6785 questions, demonstrating a high level of competency in ophthalmology.	None	ChatGPT’s proficiency in clinical management and decision-making suggests that it could be a valuable resource for practicing ophthalmologists and other medical professionals seeking information and guidance on complex cases.	ChatGPT was unable to interpret figures, graphs, or tables. In addition, some epidemiological data were unavailable online, and some standards of medical care had recently changed; thus, for certain questions, ChatGPT could not give the correct answer.

Study ID	Exam type	Dataset and methods	ChatGPT-4.0 accuracy	ChatGPT-3.5 accuracy	Advantages	Disadvantages
Riedel et al (2023) [62]	German OB/GYN[o] exams	The dataset of questions from the OB/GYN course included 160 questions, and the dataset of the medical state exam included 104 questions that were solved by an AI machine.	None	ChatGPT provided correct answers 85.6% of the time. In the second round, ChatGPT achieved similarly good results for the dataset, with 88.7% of answers being correct from the OB/GYN course, and 70.4% correct in the Medical State Exam.	ChatGPT can perform intricate tasks related to handling complex medical and clinical information in the field of OB/GYN. ChatGPT provided consistent answers and explanations to medical problems and did not require help in finding the correct solutions.	Finding raises doubts about ChatGPT's adaptability to varying levels of difficulty and emphasizes the need both for further fine-tuning and for incorporating data. Another potential limitation is the inability of ChatGPT to process images.
Saad et al (2023) [63]	FRCS[p] Orthopedic Exam	Questions were sourced from a bank of FRCS (Orth) Part A mock questions compiled from various resources. We divided the questions into two mock examinations.	ChatGPT-4.0 achieved an overall score of 67.5% on Part A. For component one of the Part A exams, ChatGPT-4.0 scored 60/120. For the second component of Part A, ChatGPT-4.0 scored 102/120 (85%).	None	Its relatively stronger performance in anatomy-based and shorter questions may be attributed to its ability to recall and provide information from its training data.	The inability of ChatGPT-4.0 to effectively handle complex clinical scenarios and image-based questions suggests limitations in its understanding and interpretation of intricate medical information.
Sarangi et al (2024) [64]	Radiology Case Vignettes (FRCR2A examination)	120 MCQs were solved by both ChatGPT-3.5 and residents, and the results were compared.	None	45%	In some of the cases, AI could point out the answers, but residents could not.	AI models currently lack the accuracy levels of human professionals, as the residents outscored them (63.33% and 57.5%). AI gave an inaccurate explanation in 50% of the cases.
Scaioli et al (2023) [65]	Italian Medical Residency Exam	136 questions classified into clinical cases and notional questions. These were solved by ChatGPT-3.0, and the accuracy was determined.	None	90.44%, with higher performance on clinical cases (92.45%) than on notional questions (89.15%).	ChatGPT's performance was higher than 99.6% of the participants. Potential for being a tool for learning.	Overreliance on these tools may develop, leading to a decline in the physician's adaptive judgment.
Skalidis et al (2023) [66]	European Exam in Core Cardiology (EECC)	After filtering, 362 MCQ items (ESC[q] sample: 68; BHDRA[r]: 150; and StudyPRN[s]: 144) were included to be solved by ChatGPT-3.5.	None	ChatGPT answered 340 questions out of 362, with 22 indeterminate answers in total, and an overall accuracy of 58.8%.	ChatGPT correctly answered the majority of questions and showed consistency across all different MCQ sources, exceeding 60% in most analyses. It exceeds the passing mark.	ChatGPT is designed for natural language processing tasks and thus currently only accepts text-based inputs, resulting in the exclusion of all questions with image content.

Study ID	Exam type	Dataset and methods	ChatGPT-4.0 accuracy	ChatGPT-3.5 accuracy	Advantages	Disadvantages
Surapaneni (2023) [67]	Medical biochemistry	The performance of ChatGPT was evaluated in medical biochemistry using 10 randomly selected clinical case vignettes.	None	ChatGPT generated correct answers for 4 questions on the first attempt.	Large language models such as ChatGPT may enhance student engagement and learning by assisting in web-based learning by generating pertinent and comprehensive content.	According to the findings of our study, ChatGPT may not be considered an accurate information provider for application in medical education to improve learning and assessment.
Takagi et al (2023) [14]	Japanese Medical Licensing Examination (JMLE)	254 MCQs on essential knowledge and clinical skills were performed by ChatGPT-3.5 and 4.0 and compared (questions containing tables, images, and underlining were excluded).	79.9% overall, with 87.2% in essential knowledge, 73.3% in general clinical, and 81.7% in specific disease.	50.8% overall, with 55.1% in essential knowledge, 43.8% in general clinical, and 56.3% in specific diseases.	GPT-4 satisfied the JMLE passing criteria unlike GPT-3, showing better proficiency in Japanese.	Concerns related to hallucinations.
Toyama et al (2024) [45]	JRBE[t]	103 questions from JRBE 2022 were used. These questions were categorized by pattern, required level of thinking, and topic.	GPT-4 scored 65% (GPT-4 correctly answered 93.3% of the questions in nuclear medicine (n=15). In questions on radiological general knowledge, it scored 90%.	ChatGPT scored 40.8% (42/103). In the questions in nuclear medicine (n=15), ChatGPT scored 40% (*P*=.01). In questions on radiological general knowledge, it scored 30%.	GPT-4 passed the JRBE with an overall score of 65%.	They behaved confidently, although their responses differed from their previous choices for the same question. Such unfavorable responses, entirely grounded in incorrect evidence or factual inaccuracies, are commonly labeled as "hallucinations."
Wang et al (2023) [46]	Chinese National Medical Licensing Examination (NMLE) 2021 and 2022	4 units with 150 questions per unit were solved by ChatGPT-3.5. The performance of ChatGPT-3.5 was compared with medical students.	None	275/600 (45.8%) in the 2021 NMLE and 219/600 (36.5%) in the 2022 NMLE.	ChatGPT holds the possibility of serving as a virtual medical mentor.	ChatGPT's proficiency in questions pertaining to the Chinese NMLE is not yet at par with Chinese medical students.
Wang et al (2024) [47]	Pathology Domain-Specific Knowledge	Google Forms were sent out to 15 participants, who each asked 1 short-answer question from both versions of ChatGPT.	Met or exceeded expectations in 12/15 of the questions.	Met or exceeded expectations in 9/15 of the questions.	Can provide quick and easily accessible information about various pathology topics, with newer LLMs also having the capability to provide references for the data used in the response.	Answers included unnecessarily lengthy responses containing irrelevant material, awkwardness of language, and provision of incorrect information.
Watari et al (2023) [68]	GM-ITE[u] examination	The GM-ITE examination had 137 questions for the years 2020, 2021, and 2022 to conduct a comparative analysis with only single-choice answers, excluding audio and visual cues. These were solved by ChatGPT-4.0.	GPT-4 scores were significantly higher than the mean scores of residents.	None	GPT-4 demonstrated remarkable proficiency in the detailed disease knowledge section, which requires an in-depth understanding of diseases, as well as in more challenging questions and domains.	GPT-4 seemed to struggle with questions in the "medical interview and professionalism" and "psychiatry" categories, which are typically easier for residents.

Study ID	Exam type	Dataset and methods	ChatGPT-4.0 accuracy	ChatGPT-3.5 accuracy	Advantages	Disadvantages
Wójcik et al (2024) [69]	Polish medical specialization licensing exam (PES)	120 questions were solved by ChatGPT-4.0.	80/120 (67.1%)	None	ChatGPT may be used as an assistance tool in medical education.	Mastery over exam content based mostly on rote learning does not translate to the practice of medicine, which is fundamentally based on human interactions.
Yu et al (2024) [70]	Chinese Postgraduate Examination for Clinical Medicine	A dataset of 165 medical questions that were divided into three categories—(1) common questions, (2) case analysis questions, and (3) multiple-choice questions—was solved by ChatGPT-3.5.	None	ChatGPT scored 153.5 out of 300 for original questions in Chinese. However, ChatGPT had low accuracy in answering open-ended medical questions, with only 31.5% total accuracy.	ChatGPT demonstrated a high level of internal concordance, which suggests that the explanations provided by ChatGPT support and affirm the given answers. Moreover, ChatGPT generated multiple insights toward the same questions.	Poorer accuracy was linked to lower concordance and a lack of insight. Thus, it was hypothesized that inaccurate responses were primarily driven by missing information, which could result in reduced insight and indecision in the AI.
Zong et al (2024) [71]	Chinese NMLE[v] (2017-2021)	3000 questions across 5 exams were solved by ChatGPT-3.5.	None	Average accuracy of 53.05%.	ChatGPT excelled in questions related to clinical epidemiology, human parasitology, and dermatology. No significant difference in performance on case-based and non-case-based questions.	ChatGPT has limited training in languages like Chinese. It failed to meet the minimal benchmark of 60%.
Sahin et al (2024) [72]	Turkish Neurosurgical Society Proficiency Board Exams (TNSPBE)	100 questions from the last 6 TNSPBE were used, 77 with visual elements were excluded, leaving 523 for analysis.	79% (412 out of 523 questions answered correctly).	None	ChatGPT can quickly provide accurate explanations and supplementary learning material, making it a valuable tool for exam preparation.	ChatGPT may generate incorrect or misleading answers with convincing explanations, which can misguide candidates if not verified.
Yanagita et al (2023) [40]	National Medical Licensing Examination in Japan	All 400 questions from the 2022 NMLE in Japan were included; 292 targeted questions (after exclusion) were analyzed.	81.5% (237 out of 292 questions answered correctly).	42.8% (125 out of 292 questions answered correctly).	GPT-4 reached passing standard; performance could improve as the model relearns.	GPT is limited to written questions.
Guerra et al (2023) [73]	Congress of Neurological Surgeons Self-Assessment Neurosurgery Exam (SANS)	591 out of 643 board-style questions were included after exclusion of questions with no text.	76.6% accuracy for all questions, but 79% for text-only questions.	None	GPT-4's accuracy suggests applications in educational settings and clinical decision-making.	GPT-4 may have potential limitations without manual input for answer suggestions.
Isleem et al (2023) [74]	Orthopaedic In-Training Examination (OITE) developed by the American Academy of Orthopaedic Surgeons (AAOS)	301 self-assessment examination questions from AAOS were included.	None	60.8% (183 out of 301 questions answered correctly).	ChatGPT has the potential to provide orthopedic educators and trainees with accurate clinical conclusions for board exam questions.	Since GPT's reasoning needs to be carefully analyzed for clinical accuracy and validity, its usefulness in clinical educational contexts is limited.

Study ID	Exam type	Dataset and methods	ChatGPT-4.0 accuracy	ChatGPT-3.5 accuracy	Advantages	Disadvantages
Gupta et al (2023) [48]	Plastic Surgery Inservice Training Examination (PSITE)	250 sample assessment questions from the 2022 PSITE were obtained from the American Council of Academic Plastic Surgeons website; 242 were answered by GPT.	77.3% (187 out of 242 questions answered correctly).	None	GPT-4 has shown to be more accurate and reliable for plastic surgery resident education when compared to GPT-3.5 and should be used to enhance educational curriculum.	GPT-4 may rely on inaccurate sources and may misunderstand prompts.
Wang et al (2013) [75]	Chinese (CNMLE) and English (ENMLE) datasets of the China National Medical Licensing Examination	A total of 220 questions were extracted from the $SMLE^{W}$ test bank across 4 medical disciplines.	GPT-4 open-domain hallucinations accounted for 56% (9/16), 43% (6/14), and 83% (15/18), while close-domain hallucinations accounted for 44% (7/16), 57% (8/14), and 17% (3/18), respectively.	GPT-3.5 scored 56%, 76%, and 62% on CNMLE, ENMLE, and NEEPM, compared to GPT-4's 84%, 86%, and 82%. Verbal fluency exceeded 95% across responses.	GPT-4 is a highly valuable AI-assisted tool in medical education.	GPT-4 had difficulty in answering difficult and complex questions.
Ali et al (2023) [49]	Self-Assessment Neurosurgery Examination Indications Examination	149 questions from the Self-Assessment Neurosurgery Examination Indications Examination were used to query LLM accuracy.	Overall accuracy of 82.6%.	Overall accuracy of 62.4%.	Potential value and applications of LLMs such as GPT-4 in neurosurgical education and in clinical decision-making.	The use of multiple-choice questions to quantify LLM knowledge for higher-order neurosurgical topics incompletely captures the open-ended nature of the true neurosurgery oral board examination.
Gravina et al (2024) [76]	Italian National Residency Admission Exam (SSM23)	Multiple-choice gastroenterology-focused questions were chosen from the 140 questions in the 2023 Italian medical specialization exam.	None	Overall accuracy of 94.11%.	ChatGPT-3.5 exhibits promise in addressing gastroenterological queries, emphasizing potential educational roles.	ChatGPT-3.5 shows variable performance, mandating cautious use alongside other educational tools.
D'Souza et al (2023) [77]	Clinical Vignettes in Psychiatry	ChatGPT-3.5's responses to 100 clinical vignettes representing 100 different psychiatric illnesses were assessed by expert psychiatrists.	None	Grade A: 61/100 cases; Grade B: 31/100 cases; Grade C: 8/100 cases.	ChatGPT-3.5 has appreciable knowledge and interpretation skills in psychiatry.	Depending upon the query and information provided, ChatGPT-3.5 can provide varying responses.

Study ID	Exam type	Dataset and methods	ChatGPT-4.0 accuracy	ChatGPT-3.5 accuracy	Advantages	Disadvantages
Khan et al (2024) [50]	Anesthesiology Board-Style Examination Questions for the American Board of Anesthesiology (ABA)	A total of 884 multiple-choice questions were used from *Anesthesia: A Comprehensive Review* (6th edition), a question bank largely regarded as one of the premium study sources for ABA certification and recertification examinations.	Overall accuracy of 69.4%.	Overall accuracy of 47.9%.	GPT-4 significantly outperformed GPT-3.5.	Although GPT-4 shows promise, current LLMs are not sufficiently advanced to answer anesthesiology board examination questions with passing success.
Haze et al (2023) [51]	Japanese National Medical Examination	Questions available from three editions of the Japanese National Medical Examination administered in 2020, 2021, and 2022 were used; questions with images or those officially identified as inappropriate were excluded.	Accuracy rate of 81% and a consistency rate of 88.8%.	Accuracy rate of 56.4% and a consistency rate of 56.5%.	GPT-4 showed significantly higher percentages of correct and consistent responses than GPT-3.5.	Neither LLM reached the level required in real clinical practice.
Rizzo et al (2024) [52]	Orthopaedic In-Service Training Exams (OITEs)	All questions from the 2020–2022 Orthopaedic In-Service Training Exams (OITEs) were given to OpenAI's GPT-3.5 Turbo and GPT-4 LLMs.	2022: 67.63%; 2021: 58.69%; 2020: 59.53%.	2022: 50.24%; 2021: 47.42%; 2020: 46.51%.	GPT-4's performance is comparable to a second- to third-year resident, and GPT-3.5 Turbo's performance is comparable to a first-year resident.	The application of current LLMs can neither pass the OITE nor substitute orthopaedic training.

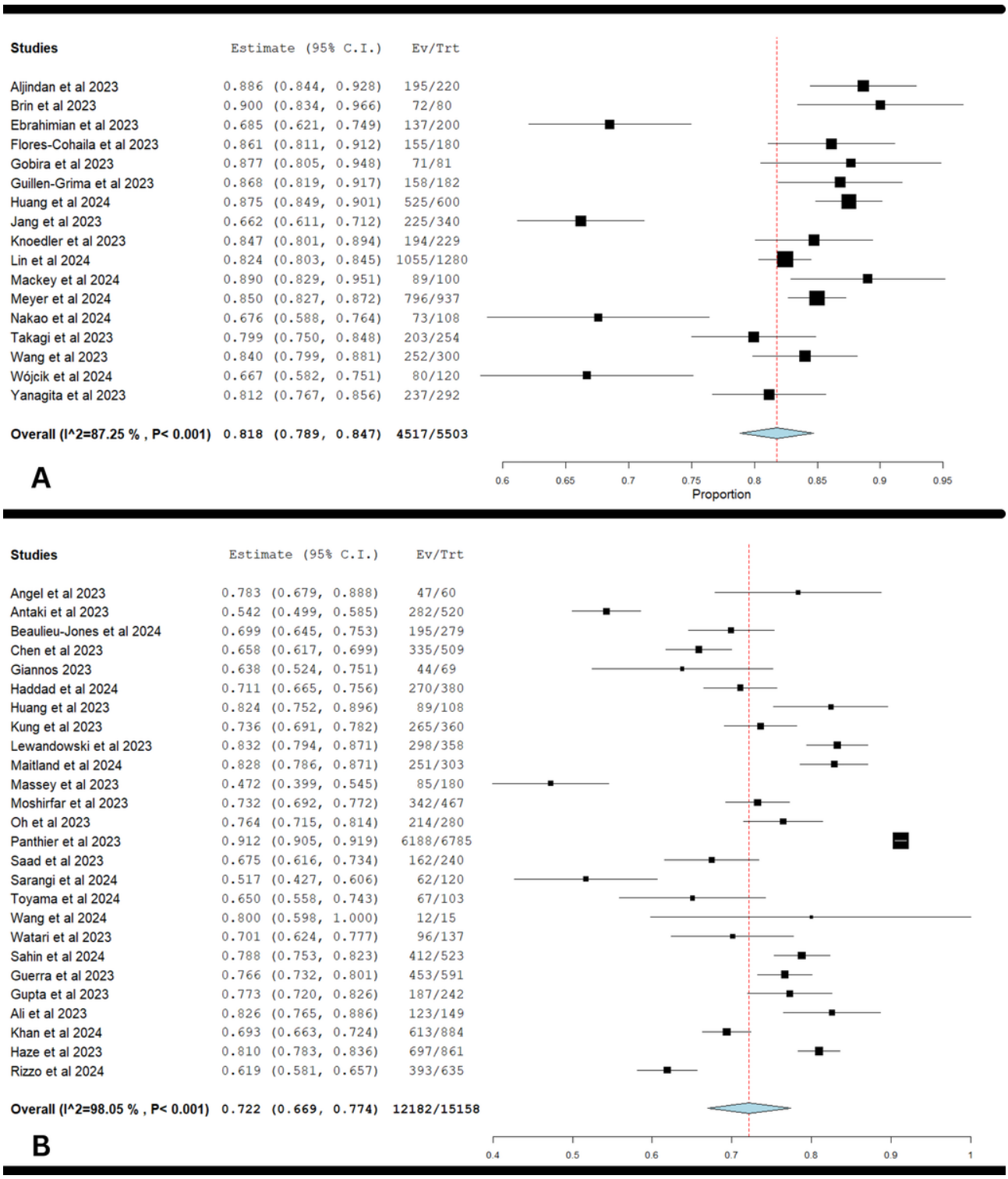
Study ID	Exam type	Dataset and methods	ChatGPT-4.0 accuracy	ChatGPT-3.5 accuracy	Advantages	Disadvantages
Mannam et al (2023) [78]	Congress of Neurological Surgeons (CNS) Self-Assessment Neurosurgery (SANS) Exam Board Review Prep Questions	Questions were obtained from the Congress of Neurological Surgeons (CNS) Self-Assessment Neurosurgery (SANS) Exam Board Review Prep; the ChatGPT Output Precision Ladde was developed to evaluate the quality and accuracy of the ChatGPT output.	None	ChatGPT achieved spot-on accuracy for 66.7% of prompted questions, 59.4% of unprompted questions, and 63.9% of unprompted questions with a leading phrase. In comparison to SANS explanations, ChatGPT output was considered better in 19.1% of questions, equal in 51.6%, and worse in 29.3%.	The authors envision a future where LLMs like ChatGPT are integrated into medical education, providing explanations based on the RIME (Reporter, Interpreter, Manager, and Educator) framework.	The language and phrasing of the pain questions may not have been optimally suited for the data ChatGPT was trained on.

[a]MCQs: multiple-choice questions.
[b]AI: artificial intelligence.
[c]LLMs: large language models.
[d]USMLE: United States Medical Licensing Examination.
[e]NBME: National Board of Medical Examiners.
[f]OB-WQE: Ophthalmology Board Written Qualifying Exam.
[g]SPTEMD: Senior Professional and Technical Examinations for Medical Doctors.
[h]FM: Family Medicine.
[i]K-NLEKMD: Korean National Licensing Examination for Korean Medicine Doctors.
[j]TKM: Traditional Korean Medicine.
[k]PES: Państwowy Egzamin Specjalizacyjny (English: Polish Medical Specialization Licensing Exam).
[l]PGY-5: postgraduate year 5.
[m]MRCP: Membership of the Royal Colleges of Physicians.
[n]EBO: European Board of Ophthalmology.
[o]OB/GYN: obstetrics and gynecology.
[p]FRCS: Fellowship of the Royal Colleges of Surgeons.
[q]ESC: European Society of Cardiology.
[r]BHDRA: Braunwald's Heart Disease Review and Assessment.
[s]StudyPRN: Study Professional Resource Network.
[t]JRBE: Japan Radiology Board Examination.
[u]GM-ITE: General Medicine In-Training Examination.
[v]NMLE: National Medical Licensing Examination in Japan.
[w]SMLE: Saudi Medical Licensing Exam.

ChatGPT-4 Performance in Medical Licensing Exams

The forest plot shown in Figure 3A indicated that the subject had a pooled accuracy proportion of 0.818 (95% CI 0.789-0.847) in medical licensing exams. This means that, on average, the subject correctly answered about 81.8% of the questions. However, there was significant heterogeneity among the studies, with I^2=87.25% and χ^2_{16}=125.54 (P<.001). This high variability suggests that differences in study design, population, and exam content significantly influenced the accuracy estimates. Despite this variability, the subject's high accuracy indicates its potential utility.

Figure 3. Performance of ChaGPT-4.0 in (A) medical licensing and (B) in-training residency exams [13-20,22,24,26,28,30-32,35,37,38,40-42,44-53,56,58,60,61,63,64,68,69,72,73].



ChatGPT-4 Performance in In-Training Residency Exams

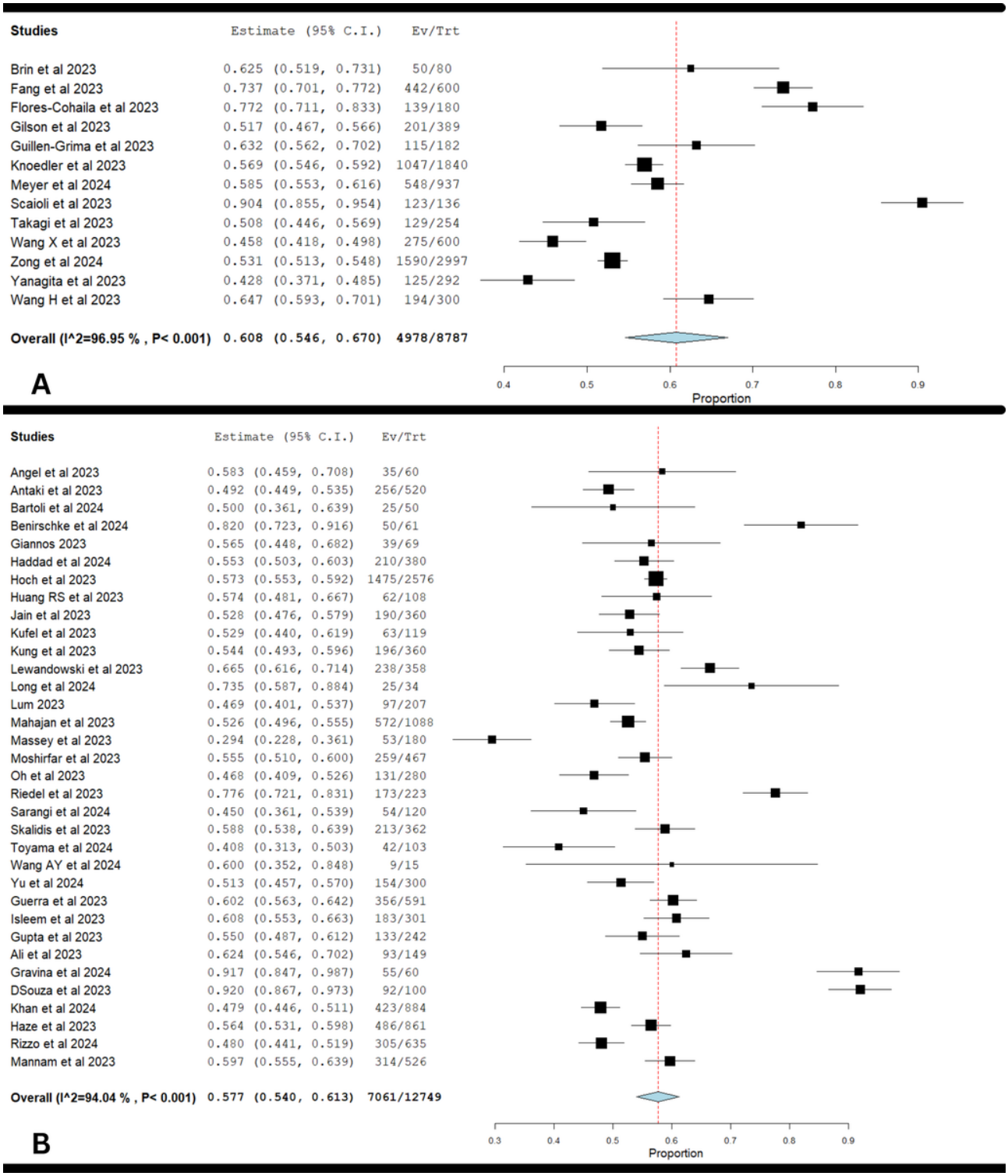
The forest plot shown in Figure 3B indicated that the subject had a pooled accuracy proportion of 0.722 (95% CI 0.669-0.774) in in-training residency exams. This means that, on average, the subject correctly answered about 72.2% of the questions. However, there was significant heterogeneity among the studies, with I^2=98.05% and χ^2_{25}=1280.36 (P<.001). This high variability suggests that differences in study design, population, and exam content significantly influenced the accuracy estimates. Despite this variability, the subject's high accuracy indicates its potential utility in medical education.

ChatGPT-3.5 Performance in Medical Licensing Exams

The forest plot showed a pooled accuracy proportion of 0.577 (95% CI 0.540-0.613), indicating a lower accuracy compared to ChatGPT-4 (Figure 4A). The heterogeneity was also high,

with I^2=94.04% and χ^2_{33}=554.02 (P<.001). This suggests considerable variability in how ChatGPT-3.5 performed across different studies. The lower accuracy and high variability underscore the need for improvements in ChatGPT-3.5's capabilities, particularly in standardizing the conditions under which it is tested to better understand its limitations and strengths.

Figure 4. Overall performance and evaluation of ChatGPT-3.5 in (A) medical licensing exams and (B) in-training residency exams [13-17,19-21,23,27,29,31-33,35,36,38-52,54,55,57,62,64-66,70,71,73,74,77,78].



ChatGPT-3.5 Performance in In-Training Residency Exams

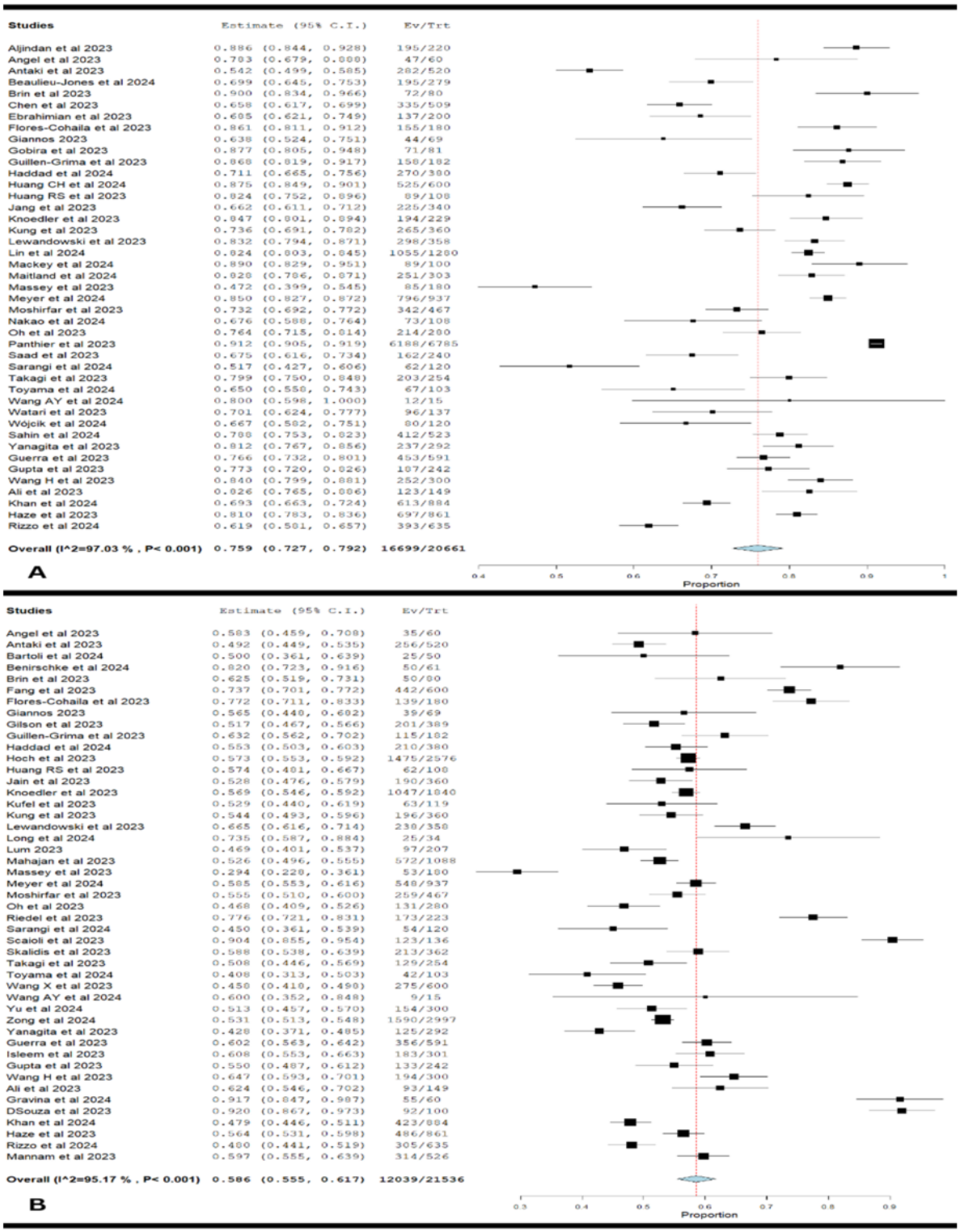
The forest plot showed a pooled accuracy proportion of 0.608 (95% CI 0.546-0.670), indicating a lower accuracy compared to ChatGPT-4 (Figure 4B). The heterogeneity was also high, with I^2=96.95% and χ^2_{12}=393.17 (P<.001). This suggests considerable variability in how ChatGPT-3.5 performed across different studies. The lower accuracy and high variability underscore the need for improvements in GPT-3.5's capabilities,

particularly in standardizing the conditions under which it is tested to better understand its limitations and strengths.

Overall Performance of ChatGPT-4

The forest plot shows that ChatGPT-4 had a pooled accuracy proportion of 0.759 (95% CI 0.727-0.792) in in-training residency exams (Figure 5A). This means that, on average, GPT-4 correctly answered about 75.9% of the questions. However, there was significant heterogeneity among the studies, with τ=0.01, χ^2_{42}=1414.01 (P<.001), and I^2=97.03%. This high variability suggests that differences in study design, population, and exam content significantly influenced the accuracy estimates. Despite this variability, ChatGPT-4's high accuracy indicates its potential utility in medical education, though further research is necessary to understand the sources of heterogeneity and enhance its performance.

Figure 5. Overall performance of (A) ChatGPT-4 and (B) ChatGPT-3.5 in medical licensing exams and in-training residency exams [13-24,26-58,61-66,70,71,73,74,77,78].



Overall Performance of ChatGPT-3.5

The forest plot showed a pooled accuracy proportion of 0.586 (95% CI 0.555-0.617), indicating a lower accuracy compared to ChatGPT-4 (Figure 5B). The heterogeneity was also high, with τ=0.05, χ^2_{46}=951.72 (P<.001), and I^2=95.17%. This suggests considerable variability in how ChatGPT-3.5 performed across different studies. The lower accuracy and high variability underscore the need for improvements in ChatGPT-3.5's capabilities, particularly in standardizing the conditions under which it is tested to better understand its limitations and strengths.

Comparative Analysis: ChatGPT-4 vs ChatGPT-3.5 in Medical Licensing and Residency Exams

The comparative forest plot presented a risk ratio (RR) of 1.36 (95% CI 1.30-1.43), demonstrating that ChatGPT-4 was 36% more likely to provide correct answers than ChatGPT-3.5 across both medical licensing and residency exams, as shown in Figure 6. For medical licensing exams specifically, the RR was 1.42 (95% CI 1.30-1.56), with high heterogeneity (I^2=85%). In residency examinations, the RR was 1.31 (95% CI 1.27-1.39), with moderate heterogeneity (I^2=49%). These results indicate that ChatGPT-4 significantly outperforms ChatGPT-3.5, but the performance advantage varies depending on the exam type. The findings suggest that while ChatGPT-4 is generally more reliable, the context of its application (eg, type of exam) plays a critical role in its accuracy. This highlights the importance of targeted improvements and further research to optimize ChatGPT-4's performance in specific educational and clinical settings.

Further subgroup analyses revealed distinct patterns in ChatGPT's performance based on exam origin or medical specialty. For surgical specialties, the pooled RR was 1.37 (95% CI 1.28-1.47), with moderate heterogeneity (I^2=49.8%) as shown in Figure 7. Orthopedic surgery showed similarly strong performance (RR 1.33, 95% CI 1.24-1.44), with notably low heterogeneity (I^2=10.8%), indicating consistent outcomes across studies. In contrast, the ophthalmology subgroup had a lower pooled RR of 1.24 (95% CI 1.11-1.38) but higher heterogeneity (I^2=65.9%), suggesting variability in ChatGPT performance possibly due to the visual or nuanced nature of the content. Performance in the Japanese Medical Licensing Exam subgroup was the most heterogeneous (RR 1.61, 95% CI 1.37-1.89; I^2=83.7%), likely reflecting the linguistic and cultural specificity of the test.

The only variable studied was the examination performance of the intervention group (ChatGPT-4.0) against the comparison group (ChatGPT-3.5). For dichotomous outcomes, RRs with 95% CIs were calculated. A random-effects model was used to pool data from both the intervention group and the comparison group. The meta-analysis was performed with the Review Manager 5.4. Subgroup analyses were performed using R software (version 4.4.1; R Foundation for Statistical Computing) based on examination level, country of origin, and specialties. OpenMeta[Analyst] (Brown University) was used to check the effect sizes of each arm, including noncomparator studies, giving a more holistic review of the existing literature.

Heterogeneity was assessed using the chi-square test and quantified with the I^2 statistic, with I^2>50% indicating substantial heterogeneity. To explore potential sources of heterogeneity, subgroup analyses and leave-one-out sensitivity analyses were conducted. A P value <.05 was considered statistically significant for all analyses.

Figure 6. Comparative analysis: ChatGPT-4 vs ChatGPT-3.5 in medical licensing and in-training residency exams [13-17,19,20,28,31,32,35,40-45,47-52,64,73].

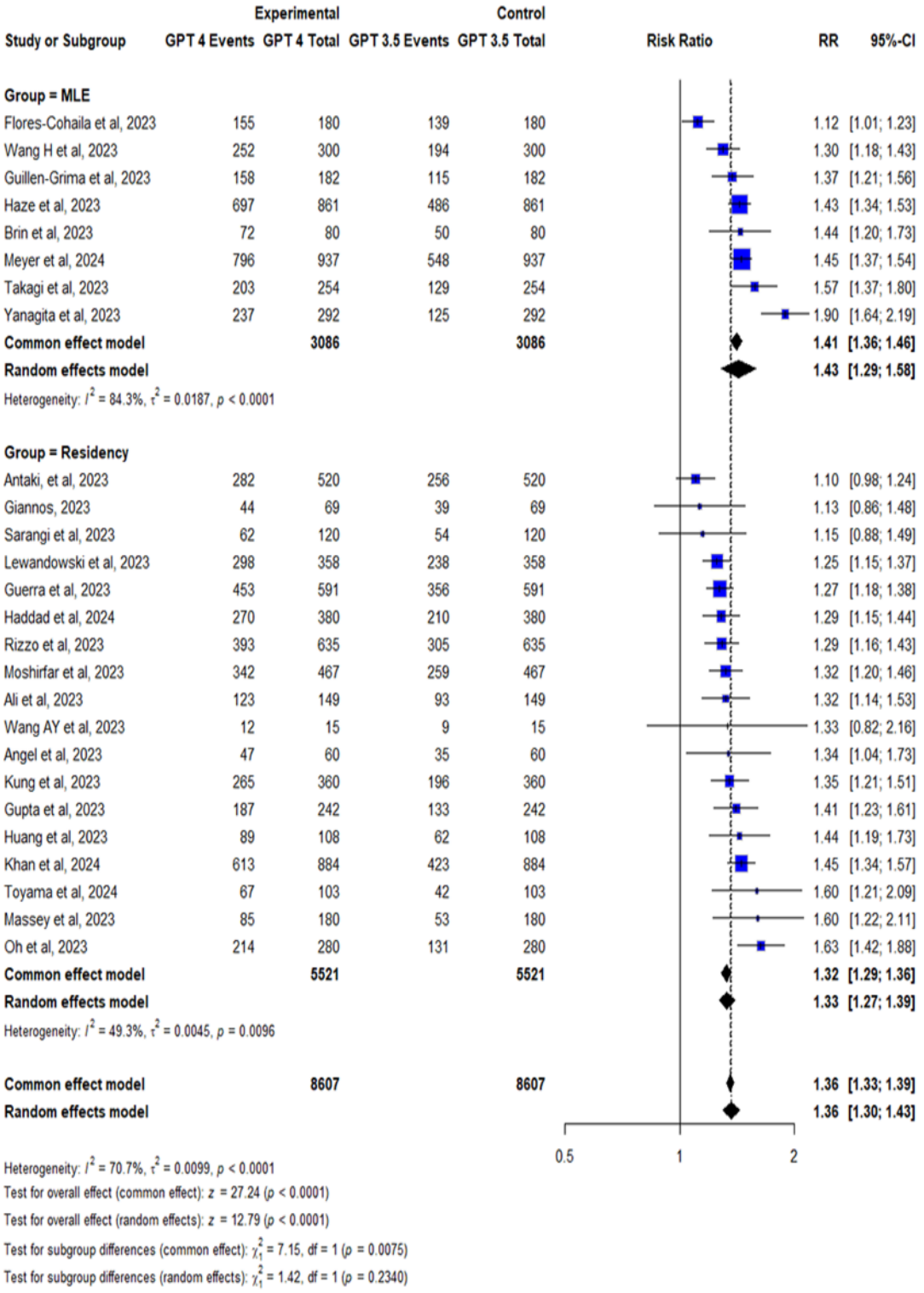
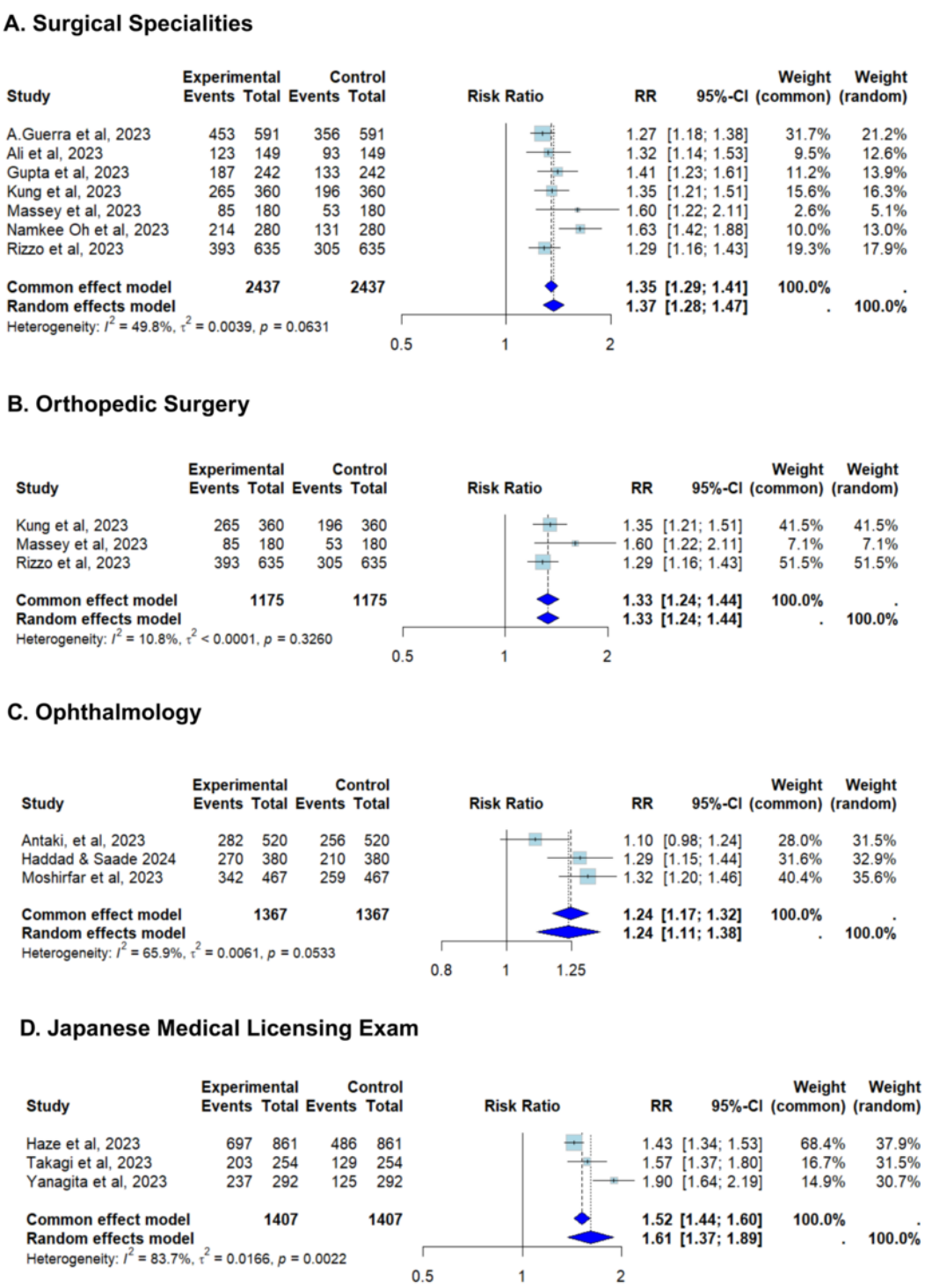


Figure 7. Further subgroup analyses: ChatGPT-4 vs ChatGPT-3.5 in medical licensing exams and residency exams [20,32,41-44,48,49,52,73].



Sensitivity Analyses and Publication Bias

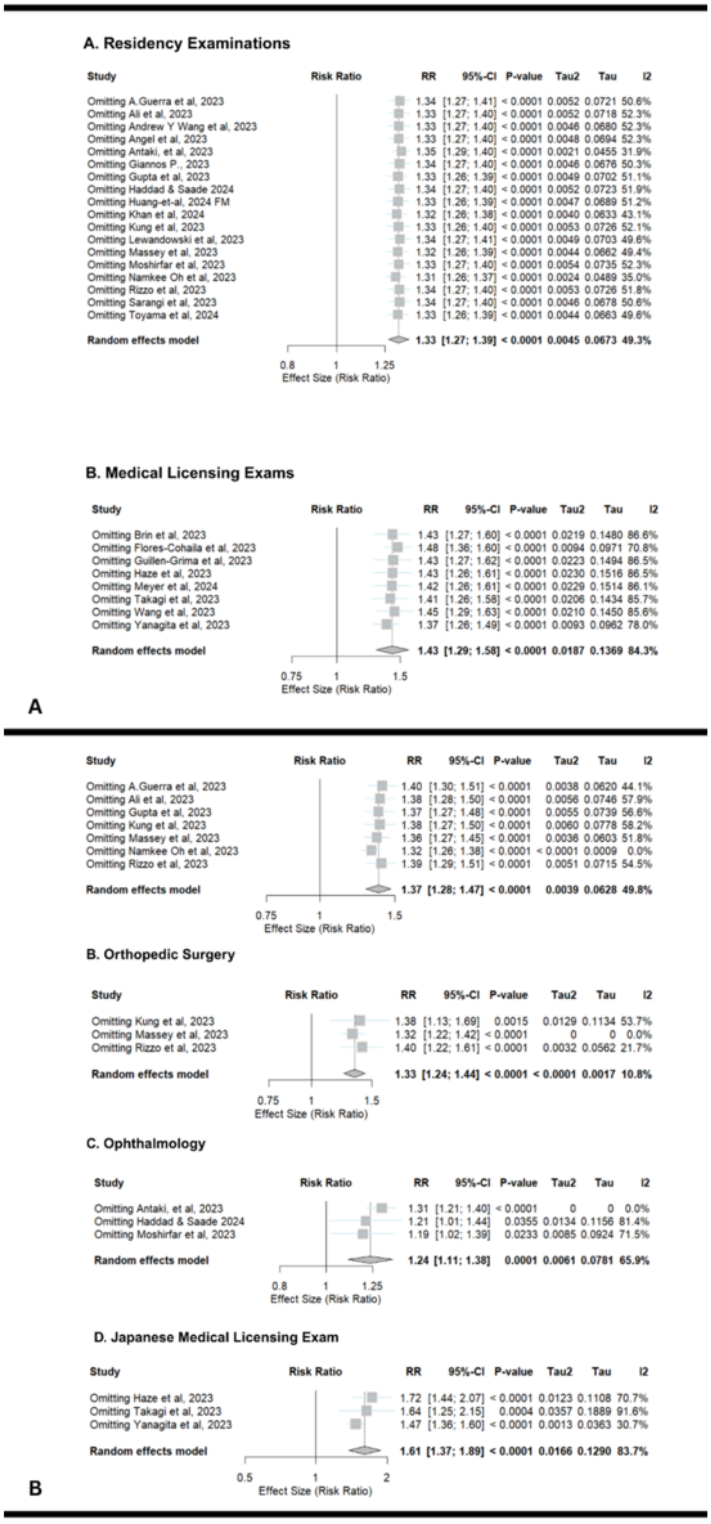
Sensitivity analyses were conducted across key subgroups to assess the robustness of the pooled RRs and the influence of individual studies on heterogeneity (Figure 8). Among broader categories, residency exams (RR 1.33, 95% CI 1.27-1.39; I^2=49.3%) exhibited consistent results across all exclusions, with negligible shifts in heterogeneity. In contrast, the medical licensing exams subgroup demonstrated persistent high heterogeneity (I^2=84.3%), even after omitting studies, indicating intrinsic variability across these assessments (Figure 8). Overall, the pooled estimates proved stable across all sensitivity models. The sources of heterogeneity appeared to be distributed across multiple studies rather than driven by a single outlier, reinforcing the internal consistency and robustness of the main findings.

For surgical specialties, no single study unduly influenced the overall estimate (RR 1.37, 95% CI 1.28-1.47), with

heterogeneity remaining moderate (I^2=49.8%) across exclusions. Notably, omitting [44] reduced heterogeneity to 0%, while other exclusions led to only marginal reductions in I^2 values. In orthopedic surgery, all leave-one-out models preserved significance (RR 1.33, 95% CI 1.24-1.44). The exclusion of [42] notably eliminated heterogeneity (I^2=0%), suggesting it was a moderate contributor to between-study variance. Ophthalmology showed more sensitivity to individual studies, with RR values ranging from 1.19 to 1.31 and I^2 fluctuating between 0% and 81.4%. Omitting [20] fully resolved heterogeneity (I^2=0%), while the exclusion of [43] led to substantial reductions in I^2, indicating a significant source of inconsistency. For the Japanese Medical Licensing Exam subgroup, heterogeneity remained high (I^2=83.7%) across all models. However, the exclusion of [14] resulted in notable reductions in I^2, confirming that all 3 studies contributed to the observed variability (Figure 8B).

Publication bias was evaluated using the Egger linear regression test for funnel plot (Figure 8) asymmetry. The test yielded a nonsignificant result (t_{24}=0.15, 2-tailed; P=.88), indicating no evidence of small-study effects or significant publication bias in the included studies. These findings suggest that the funnel plot asymmetry observed is unlikely due to publication bias and may instead reflect genuine heterogeneity across studies (Multimedia Appendix 1).

Figure 8. (A) Sensitivity analyses and (B) further subgroup sensitivity analyses of ChatGPT-3.5 and 4.0 in medical licensing exams and in-training residency exams [13-24,26-58,61-66,70,71,73,74,77,78].



Discussion

Performance Comparison of ChatGPT-3.5 and ChatGPT-4 in Medical Examinations

The study highlights the promising potential of AI-driven language models such as ChatGPT-3.5 and 4.0 in revolutionizing the medical field by assessing their performance across a range of medical licensing and in-training exams worldwide. The findings revealed that ChatGPT-4 consistently outperformed its predecessor, ChatGPT-3.5, in accuracy and proficiency [69,70], thanks to the more advanced training data and significant algorithmic improvements.

In this study, ChatGPT-4 demonstrated a pooled accuracy of approximately 75.9% in in-training residency exams, showcasing its strong overall performance. In comparison to ChatGPT-3.5, ChatGPT-4 was 36% more likely to provide

correct answers across both medical licensing and residency exams, with an RR of 1.36 (95% CI 1.30-1.43). This suggests that ChatGPT-4's advancements have translated into tangible improvements in accuracy, particularly in multiple-choice question formats. For medical licensing exams specifically, ChatGPT-4 had an even higher RR of 1.42 (95% CI 1.30-1.56), indicating a significant performance edge over ChatGPT-3.5 [71].

However, there was notable heterogeneity in the performance of both models, with high I^2 values (85% for licensing exams and 49% for residency exams), reflecting variations in study design, population, and exam content that impacted the accuracy estimates. These findings are similar to Zong et al [71]. These findings underscore the importance of contextual factors in applying AI models like ChatGPT-4. They also highlight the need for targeted improvements and further research to enhance its performance, ensuring that it becomes a more reliable and effective tool for specific educational and clinical settings [79,80,81].

Further subgroup and sensitivity analyses provide critical insights into the generalizability and robustness of ChatGPT's performance across the medical licensing exam. The lower heterogeneity among surgical and orthopedic specialties indicates that the ChatGPT models succeed in domains with structured procedural knowledge. These results suggest that ChatGPT may potentially serve as an adjunct learning tool in surgical education [35,45,59].

The above findings were in contrast with high heterogeneity in ophthalmology and Japanese licensing exams, which correspond to the fields with less predictable ChatGPT performance. Some potential explanations include the visual nature of ophthalmologic evaluations, discrepancies in linguistic constructs, and culturally specific clinical reasoning models. These findings highlight the critical requirement for more localized training datasets and a standardized mechanism to ensure generalizability [80]. The stability of effect estimates across sensitivity analyses reinforces the robustness of the meta-analytic findings. Small changes in pooled RRs after the removal of any individual study alleviate concerns about publication bias or undue influence from outliers [71,81].

These observations align closely with the findings of several systematic reviews and meta-analyses. Liu et al [10] conducted one of the most comprehensive meta-analyses to date, demonstrating that ChatGPT-4 generally outperforms its predecessors across a variety of standardized medical exams, although performance fluctuates based on question structure, language, and specialty content. The author highlighted that while LLMs are strong in knowledge-based tasks, their clinical reasoning capabilities are notably weaker, particularly in real-world scenarios or open-ended problem-solving formats [80].

The authors [80,81] presented important insights into the role of AI as an educational tool. Both studies emphasized the importance of viewing AI not as a replacement for human reasoning, but as a potential companion to strengthen formative assessment and feedback in medical education, provided its limitations are clearly understood and accounted for. LLMs' medical exam accuracy at 0.61, USMLE accuracy at 0.51, and ChatGPT's higher accuracy of 0.64 support our observation of ChatGPT-4.0's superior performance. The study also emphasized LLMs' potential to address health care challenges but stressed the need for rigorous evaluation frameworks. Their rubric framework provides a structured approach to ensure safe and ethical integration of LLMs into clinical and educational settings, aligning with our call for standardized evaluation metrics and greater transparency in future research [80].

Beyond technical accuracy, the issue of AI's ability to simulate human soft skills, such as empathy and ethical reasoning, has also been discussed extensively in the literature. The literature addressed the concept of artificial cognitive empathy, concluding that despite improvements in language generation, AI still lacks the depth of emotional understanding and situational awareness that human physicians develop through clinical experience [79,82]. Similarly, certain studies emphasized the need for careful integration of AI in both educational and clinical environments, cautioning that overconfidence in AI-generated answers could introduce safety risks if used without appropriate human oversight [83,84]. This is particularly relevant in high-stakes scenarios like licensing examinations, where even minor inaccuracies can have significant consequences. In addition, Artsi et al [11] highlighted the risk of "overreliance" on AI models during the learning process, which could unintentionally affect the development of independent critical thinking among medical students, an observation that complements the variability we observed in AI performance across exam types and specialties [11].

In our study, the performance of AI models in medical licensing exams varies widely across different countries. For example, while ChatGPT-3.5 performed relatively well in the Italian medical exam, scoring 73%, it scored significantly lower in the French exam, with only 22%. The marked difference in ChatGPT's performance between the Italian and French medical licensing exams likely reflects a combination of factors. The French exam's multiple-answer question format posed a particular challenge, as language models like ChatGPT typically perform better on single-answer questions, such as those used in the Italian exam. Additionally, longer question lengths, especially in the French set, were linked to lower accuracy. Language-specific factors, including tokenization and training data exposure, also influenced results, while differences in national exam design philosophies likely contributed to the observed performance gap [1,80]. To summarize, this variability could be attributed to differences in exam formats, language subtleties, and subject matter, highlighting the need for localization and customization of AI models for different regions and medical systems.

The analysis of various in-training exams across multiple medical specialties demonstrates both the advantages and limitations of LLMs in the medical field. These show potential in medical training exams by showing their efficiency in processing large amounts of data, coupled with improved performance in certain domains, showing potential for enhancing clinical workflows by augmenting human decision-making. However, their performance varies across specialties, with

notable limitations in problem-solving, nuanced reasoning, and explaining their logic. While ChatGPT-4 outperforms average human resident students in some cases, its accuracy is inconsistent, and it sometimes generates incorrect or misleading information [33]. These models can assist in medical education but still require human oversight, especially in complex decision-making areas.

A notable limitation identified in these models is their inability to handle questions involving figures and tables. This was demonstrated by studies on Japan's National Medical Licensing Examination, where ChatGPT-4.0 scored 81.5%, passing the exam, while ChatGPT-3.5 only scored 42.8%. Both versions struggled with visual data, which is crucial for medical exams [40,81]. Additionally, ChatGPT-4 showed limited capability in interpreting medical images, achieving only 68% accuracy on image-based questions in another Japanese exam study [9]. While textual comprehension has improved, significant gaps remain in processing and interpreting visual information, which is essential for comprehensive medical practice. The study by Yang et al [85] evaluated the performance of ChatGPT-4V, ChatGPT-4, and GPT-3.5 Turbo on medical licensing exams involving images, showing GPT-4V's high accuracy but significant limitations in image interpretation and explanation quality. While GPT-4V outperformed its predecessors in multiple-choice questions, its incorrect answers often featured poor image understanding and reasoning. The study highlights the need for further evaluation of GPT-4V's capabilities before clinical integration [86].

The performance variability of ChatGPT-3.5 and GPT-4 across different countries and medical licensing exams can be attributed to several factors. Language nuances, such as regional variations in medical terminology, phrasing, and dialect, may affect how the AI interprets and responds to exam questions [80,87]. Exam formats also play a role, with different countries prioritizing multiple-choice questions, clinical scenarios, or essay-based responses, each requiring distinct approaches from the AI. Cultural differences in health care practices, ethical considerations, and medical education systems can influence how questions are framed and what knowledge is emphasized, affecting the AI's ability to provide contextually appropriate answers. These factors combined can create significant variability in AI performance, highlighting the need for localized training and adaptation [87,85].

Despite significant advancements, ChatGPT's performance still lags behind that of trained medical students. ChatGPT performed below the student average, with a score of 80.5/100 compared to the student average of 86.21/100 in the medical microbiology exam [79]. However, in some cases, such as the Peruvian National Licensing Medical Examination, GPT-4 outperformed human students, scoring 86% and surpassing 90% of human examinees [29]. These discrepancies suggest that while ChatGPT can achieve expert-level performance on standardized medical exams, it remains limited in handling complex, domain-specific tasks requiring deep understanding and critical thinking [82].

ChatGPT-4's performance across different medical subjects further highlights its variability. In a study by Jang et al [16] on the Korean National Licensing Examination for Korean Medicine Doctors, ChatGPT-4 passed in 7 out of 12 subjects, excelling in herbology and neuropsychiatry but performing poorly in public health and acupuncture. It scored 85.9% on diagnosis-based questions, 63.2% on recall-based questions, and 53.5% on intervention-based questions, consistently outperforming ChatGPT-3.5. These results suggest the potential for AI models to specialize in certain medical fields, though challenges remain, especially with language nuances and culturally specific medical concepts [11,86,83,84].

While AI models provided mostly clinically valid information, they are not without flaws, particularly the risk of hallucinations or misinformation. Long et al [54] evaluated ChatGPT's performance on otolaryngology exams, finding that it achieved a passing mark and showed higher accuracy with specific prompts. However, caution is advised due to the risk of AI-generated hallucinations or misinformation [33]. Similarly, Takagi et al [14] found ChatGPT-4 outperformed GPT-3.5 on the Japanese Medical Licensing Exam, but both faced challenges with hallucinations and inaccurate information [14]. Maitland et al [58] identified a key reason for these misinformation issues: language models are trained "to predict the next token in a sequence rather than verify factual accuracy."

Concerns about AI misuse, such as cheating and misdiagnosis, remain significant in educational and clinical settings. A recent study highlighted the challenges posed by AI models like ChatGPT in providing relevant, readable, and accurate responses in medical exams [19,88,89]. The presence of multiple potentially correct answers poses a challenge for AI systems, as demonstrated in Kung et al's [41] study, which compared ChatGPT's performance to orthopedic surgery residents at various postgraduate levels. While AI models offer quick and accessible information, the risk of misuse and the need for further refinement cannot be overlooked. To address this issue, there should be guidelines and policies with clear ethical standards for AI usage. Transparency about AI decision-making should be ensured with accountability for misuse and error. Moreover, oversight committees must be formulated in each organization to monitor and review AI outputs and decision-making processes. There should be limited access to data with anonymization to protect privacy. Regular audits of AI models should be conducted for bias and inaccuracies, with diverse datasets being used. Training of staff on the ethical use of AI is necessary, with policies for responsible use. Validation of AI outputs with human experts' assessments should be done with continuous feedback for the improvement of AI models [87,85].

The role of AI in generating medical exam questions and educational content also presents both opportunities and challenges. A multinational prospective study conducted by Cheung et al [25] highlighted how ChatGPT demonstrated the ability to generate medical graduate exam multiple-choice questions efficiently [25]. However, the AI's questions lacked the depth, relevance, and specificity seen in human-generated content. This points to the nuanced complexities and limitations inherent in AI-driven content creation, including concerns about reliability, credibility, bias, and factual accuracy [33,90]. Thus, while ChatGPT shows potential, careful oversight is required

to ensure its effective integration into medical education. Furthermore, the findings underline the growing potential that LLMs can play in medical education and licensing exams, but highlight the need for contextualization and adaptation, particularly in any assessments that involve visual information, native language perception, or abstracted clinical reasoning.

Limitations

The study on ChatGPT-3.5 and 4.0 in medical education reveals several limitations. A primary concern is the models' inability to interpret visual data, such as medical images, which are crucial for effective practice and examination performance. Additionally, the findings may lack generalizability across different medical disciplines. Potential biases in AI-generated content raise questions about reliability and equity in diverse educational contexts.

Future Research and Gap

Further empirical data is essential to support the accuracy and utility of ChatGPT-3.5 and 4.0 in medical licensing and in-training examinations. While these versions show promising potential in processing vast amounts of medical knowledge, their application in critical, high-stakes environments requires rigorous validation. Current studies suggest variable performance, with both versions demonstrating strengths in general medical knowledge but limitations in reasoning and clinical decision-making. To solidify their role in medical education and assessment, more research is needed, especially focusing on advanced future iterations. These future versions must be scrutinized for improved understanding of complex medical scenarios, ethical considerations, and practical applications in real-world settings.

Implications and Contributions

This study highlights the implications and contributions of ChatGPT versions by analyzing their accuracy in medical examinations across the globe. By synthesizing data from diverse studies conducted in various countries, it offers a broader perspective on the performance of ChatGPT-3.5 and 4.0 in medical licensing and in-training exams. The cross-country analysis showcases the potential of these AI models in adapting to different medical curricula and standards, contributing valuable insights into their applicability in international medical education. This global evaluation established a more comprehensive understanding of the strengths and limitations of ChatGPT, providing a foundation for further research and development in medical AI tools.

Acknowledgments

We would like to acknowledge the management of Shalamar Institute of Health Sciences, Lahore, for their motivation and support. We would like to acknowledge the use of ChatGPT (OpenAI) for its assistance in drafting certain sections of this manuscript and providing support in the data processing stage. The contributions of ChatGPT were invaluable in streamlining the writing process and enhancing the clarity of the manuscript.

Authors' Contributions

Conceptualization: AJ, GF
Data curation: RMHS, SMKA
Formal analysis: SA, SMKA, TRM
Method: GF, TRM, SA
Funding acquisition: MZB
Investigation: AJ, MZB
Methodology: GF, TRM, RMHS
Project administration: AJ, MZB
Resources: UA, GF
Software: SA, SMKA, TRM, UA
Supervision: GF, AJ
Validation: AJ, MZB, UA
Visualization: TRM, GF
Writing – original draft: UA, TRM
Writing – review & editing: MZB

Conflicts of Interest

None declared.

Multimedia Appendix 1
Funnel plot for publication bias: ChatGPT-4 vs ChatGPT-3.5 in medical licensing exams and residency exams.
[PNG File , 15 KB - mededu_v11i1e68070_app1.png]

Multimedia Appendix 2

PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 177 KB - mededu_v11i1e68070_app2.pdf]

References

1. Alfertshofer M, Hoch CC, Funk PF, Hollmann K, Wollenberg B, Knoedler S, et al. Sailing the seven seas: a multinational comparison of ChatGPT's performance on medical licensing examinations. Ann Biomed Eng 2024;52(6):1542-1545 (forthcoming) [FREE Full text] [doi: 10.1007/s10439-023-03338-3] [Medline: 37553555]
2. Hirani R, Noruzi K, Khuram H, Hussaini AS, Aifuwa EI, Ely KE, et al. Artificial intelligence and healthcare: a journey through history, present innovations, and future possibilities. Life 2024 Apr 26;14(5):557 [FREE Full text] [doi: 10.3390/life14050557] [Medline: 38792579]
3. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. Research Square Preprint posted online on February 28, 2023 [FREE Full text] [doi: 10.21203/rs.3.rs-2566942/v1] [Medline: 36909565]
4. Lahat A, Sharif K, Zoabi N, Shneor Patt Y, Sharif Y, Fisher L, et al. Assessing generative pretrained transformers (GPT) in clinical decision-making: comparative analysis of GPT-3.5 and GPT-4. J Med Internet Res 2024;26:e54571 [FREE Full text] [doi: 10.2196/54571] [Medline: 38935937]
5. Whalen J, Mouza C. ChatGPT: challenges, opportunities, and implications for teacher education. Contemp Issues Technol Teacher Educ 2023;23(1):1-23 [FREE Full text]
6. Roumeliotis KI, Tselikas ND. ChatGPT and Open-AI models: a preliminary review. Future Internet 2023;15(6):192. [doi: 10.3390/fi15060192]
7. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023;9:e48291 [FREE Full text] [doi: 10.2196/48291] [Medline: 37261894]
8. Sumbal A, Sumbal R, Amir A. Can ChatGPT-3.5 pass a medical exam? A systematic review of ChatGPT's performance in academic testing. J Med Educ Curric Dev 2024;11:23821205241238641 [FREE Full text] [doi: 10.1177/23821205241238641] [Medline: 38487300]
9. Morrow E, Zidaru T, Ross F, Mason C, Patel KD, Ream M, et al. Artificial intelligence technologies and compassion in healthcare: a systematic scoping review. Front Psychol 2022;13:971044 [FREE Full text] [doi: 10.3389/fpsyg.2022.971044] [Medline: 36733854]
10. Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. J Med Internet Res 2024;26:e60807 [FREE Full text] [doi: 10.2196/60807] [Medline: 39052324]
11. Artsi Y, Sorin V, Konen E, Glicksberg BS, Nadkarni G, Klang E. Large language models for generating medical examinations: systematic review. BMC Med Educ 2024;24(1):354 [FREE Full text] [doi: 10.1186/s12909-024-05239-y] [Medline: 38553693]
12. Waldock WJ, Zhang J, Guni A, Nabeel A, Darzi A, Ashrafian H. The accuracy and capability of artificial intelligence solutions in health care examinations and certificates: systematic review and meta-analysis. J Med Internet Res 2024;26:e56532 [FREE Full text] [doi: 10.2196/56532] [Medline: 39499913]
13. Flores-Cohaila JA, García-Vicente A, Vizcarra-Jiménez SF, De la Cruz-Galán JP, Gutiérrez-Arratia JD, Quiroga Torres BG, et al. Performance of ChatGPT on the Peruvian national licensing medical examination: cross-sectional study. JMIR Med Educ 2023;9:e48039 [FREE Full text] [doi: 10.2196/48039] [Medline: 37768724]
14. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. JMIR Med Educ 2023;9:e48002 [FREE Full text] [doi: 10.2196/48002] [Medline: 37384388]
15. Lewandowski M, Łukowicz P, Świetlik D, Barańska-Rybak W. ChatGPT-3.5 and ChatGPT-4 dermatological knowledge level based on the specialty certificate examination in dermatology. Clin Exp Dermatol 2024;49(7):686-691. [doi: 10.1093/ced/llad255] [Medline: 37540015]
16. Brin D, Sorin V, Vaid A, Soroush A, Glicksberg BS, Charney AW, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep 2023;13(1):16492 [FREE Full text] [doi: 10.1038/s41598-023-43436-9] [Medline: 37779171]
17. Meyer A, Riese J, Streichert T. Comparison of the performance of GPT-3.5 and GPT-4 with that of medical students on the written German medical licensing examination: observational study. JMIR Med Educ 2024;10:e50965 [FREE Full text] [doi: 10.2196/50965] [Medline: 38329802]
18. Aljindan FK, Al Qurashi AA, Albalawi IAS, Alanazi AMM, Aljuhani HAM, Almutairi FF, et al. ChatGPT conquers the Saudi Medical Licensing Exam: exploring the accuracy of artificial intelligence in medical knowledge assessment and implications for modern medical education. Cureus 2023;15(9):e45043 [FREE Full text] [doi: 10.7759/cureus.45043] [Medline: 37829968]
19. Angel M, Rinehart J, Canneson M, Baldi P. Clinical knowledge and reasoning abilities of AI large language models in anesthesiology: a comparative study on the ABA exam. medRxiv Preprint posted online on May 16, 2023. [doi: 10.1101/2023.05.10.23289805]

20. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. Ophthalmol Sci 2023;3(4):100324 [FREE Full text] [doi: 10.1016/j.xops.2023.100324] [Medline: 37334036]
21. Bartoli A, May A, Al-Awadhi A, Schaller K. Probing artificial intelligence in neurosurgical training: ChatGPT takes a neurosurgical residents written exam. Brain Spine 2024;4:102715 [FREE Full text] [doi: 10.1016/j.bas.2023.102715] [Medline: 38163001]
22. Beaulieu-Jones BR, Berrigan MT, Shah S, Marwaha JS, Lai S, Brat GA. Evaluating capabilities of large language models: performance of GPT-4 on surgical knowledge assessments. Surgery 2024;175(4):936-942. [doi: 10.1016/j.surg.2023.12.014] [Medline: 38246839]
23. Benirschke R, Wodskow J, Prasai K, Freeman A, Lee J, Groth J. Assessment of a large language model's utility in helping pathology professionals answer general knowledge pathology questions. Am J Clin Pathol 2024;161(1):42-48. [doi: 10.1093/ajcp/aqad106] [Medline: 37658808]
24. Chen TC, Multala E, Kearns P, Delashaw J, Dumont A, Maraganore D, et al. Assessment of ChatGPT's performance on neurology written board examination questions. BMJ Neurol Open 2023;5(2):e000530 [FREE Full text] [doi: 10.1136/bmjno-2023-000530] [Medline: 37936648]
25. Cheung BHH, Lau GKK, Wong GTC, Lee EYP, Kulkarni D, Seow CS, et al. ChatGPT versus human in generating medical graduate exam multiple choice questions: a multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). PLoS One 2023;18(8):e0290691 [FREE Full text] [doi: 10.1371/journal.pone.0290691] [Medline: 37643186]
26. Ebrahimian M, Behnam B, Ghayebi N, Sobhrakhshankhah E. ChatGPT in Iranian medical licensing examination: evaluating the diagnostic accuracy and decision-making capabilities of an AI-based model. BMJ Health Care Inform 2023;30(1):e100815 [FREE Full text] [doi: 10.1136/bmjhci-2023-100815] [Medline: 38081765]
27. Fang C, Wu Y, Fu W, Ling J, Wang Y, Liu X, et al. How does ChatGPT-4 preform on non-English national medical licensing examination? An evaluation in Chinese language. PLOS Digit Health 2023;2(12):e0000397 [FREE Full text] [doi: 10.1371/journal.pdig.0000397] [Medline: 38039286]
28. Giannos P. Evaluating the limits of AI in medical specialisation: ChatGPT's performance on the UK neurology specialty certificate examination. BMJ Neurol Open 2023;5(1):e000451 [FREE Full text] [doi: 10.1136/bmjno-2023-000451] [Medline: 37337531]
29. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does chatGPT perform on the United States medical licensing examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
30. Gobira M, Nakayama LF, Moreira R, Andrade E, Regatieri CVS, Belfort R. Performance of ChatGPT-4 in answering questions from the Brazilian national examination for medical degree revalidation. Rev Assoc Med Bras (1992) 2023;69(10):e20230848 [FREE Full text] [doi: 10.1590/1806-9282.20230848] [Medline: 37792871]
31. Guillen-Grima F, Guillen-Aguinaga S, Guillen-Aguinaga L, Alas-Brun R, Onambele L, Ortega W, et al. Evaluating the efficacy of ChatGPT in navigating the Spanish medical residency entrance examination (MIR): promising horizons for AI in clinical medicine. Clin Pract 2023;13(6):1460-1487 [FREE Full text] [doi: 10.3390/clinpract13060130] [Medline: 37987431]
32. Haddad F, Saade JS. Performance of chatGPT on ophthalmology-related questions across various examination levels: observational study. JMIR Med Educ 2024;10:e50842 [FREE Full text] [doi: 10.2196/50842] [Medline: 38236632]
33. Hoch CC, Wollenberg B, Lüers JC, Knoedler S, Knoedler L, Frank K, et al. ChatGPT's quiz skills in different otolaryngology subspecialties: an analysis of 2576 single-choice and multiple-choice board certification preparation questions. Eur Arch Otorhinolaryngol 2023;280(9):4271-4278 [FREE Full text] [doi: 10.1007/s00405-023-08051-4] [Medline: 37285018]
34. Huang CH, Hsiao HJ, Yeh PC, Wu KC, Kao CH. Performance of ChatGPT on stage 1 of the Taiwanese medical licensing exam. Digit Health 2024;10:20552076241233144 [FREE Full text] [doi: 10.1177/20552076241233144] [Medline: 38371244]
35. Huang RS, Lu KJQ, Meaney C, Kemppainen J, Punnett A, Leung F. Assessment of resident and AI chatbot performance on the University of Toronto family medicine residency progress test: comparative study. JMIR Med Educ 2023;9:e50514 [FREE Full text] [doi: 10.2196/50514] [Medline: 37725411]
36. Jain N, Gottlich C, Fisher J, Campano D, Winston T. Assessing ChatGPT's orthopedic in-service training exam performance and applicability in the field. J Orthop Surg Res 2024;19(1):27 [FREE Full text] [doi: 10.1186/s13018-023-04467-0] [Medline: 38167093]
37. Jang D, Yun T, Lee C, Kwon Y, Kim C. GPT-4 can pass the Korean national licensing examination for Korean medicine doctors. PLOS Digit Health 2023;2(12):e0000416 [FREE Full text] [doi: 10.1371/journal.pdig.0000416] [Medline: 38100393]
38. Knoedler L, Alfertshofer M, Knoedler S, Hoch CC, Funk PF, Cotofana S, et al. Pure wisdom or potemkin villages? A comparison of ChatGPT 3.5 and ChatGPT 4 on USMLE step 3 style questions: quantitative analysis. JMIR Med Educ 2024;10:e51148 [FREE Full text] [doi: 10.2196/51148] [Medline: 38180782]
39. Kufel J, Paszkiewicz I, Bielówka M, Bartnikowska W, Janik M, Stencel M, Czogalik, et al. Will ChatGPT pass the Polish specialty exam in radiology and diagnostic imaging? Insights into strengths and limitations. Pol J Radiol 2023;88:e430-e434 [FREE Full text] [doi: 10.5114/pjr.2023.131215] [Medline: 37808173]

40. Yanagita Y, Yokokawa D, Uchida S, Tawara J, Ikusaka M. Accuracy of ChatGPT on medical questions in the National Medical Licensing Examination in Japan: evaluation study. JMIR Form Res 2023;7:e48023 [FREE Full text] [doi: 10.2196/48023] [Medline: 37831496]
41. Kung J, Marshall C, Gauthier C, Gonzalez T, Jackson IJ. Evaluating ChatGPT performance on the orthopaedic in-training examination. JB JS Open Access 2023;8(3):e23 [FREE Full text] [doi: 10.2106/JBJS.OA.23.00056] [Medline: 37693092]
42. Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT-3.5, ChatGPT-4, and orthopaedic resident performance on orthopaedic assessment examinations. J Am Acad Orthop Surg 2023;31(23):1173-1179 [FREE Full text] [doi: 10.5435/JAAOS-D-23-00396] [Medline: 37671415]
43. Moshirfar M, Altaf A, Stoakes I, Tuttle J, Hoopes P. Artificial intelligence in ophthalmology: a comparative analysis of GPT-3.5, GPT-4, and human expertise in answering StatPearls questions. Cureus 2023;15(6):e40822 [FREE Full text] [doi: 10.7759/cureus.40822] [Medline: 37485215]
44. Oh N, Choi G, Lee WY. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. Ann Surg Treat Res 2023;104(5):269-273 [FREE Full text] [doi: 10.4174/astr.2023.104.5.269] [Medline: 37179699]
45. Toyama Y, Harigai A, Abe M, Nagano M, Kawabata M, Seki Y, et al. Performance evaluation of ChatGPT, GPT-4, and Bard on the official board examination of the Japan Radiology Society. Jpn J Radiol 2024;42(2):201-207 [FREE Full text] [doi: 10.1007/s11604-023-01491-2] [Medline: 37792149]
46. Wang X, Gong Z, Wang G, Jia J, Xu Y, Zhao J, et al. ChatGPT performs on the Chinese national medical licensing examination. J Med Syst 2023;47(1):86. [doi: 10.1007/s10916-023-01961-0] [Medline: 37581690]
47. Wang AY, Lin S, Tran C, Homer RJ, Wilsdon D, Walsh JC, et al. Assessment of pathology domain-specific knowledge of ChatGPT and comparison to human performance. Arch Pathol Lab Med 2024;148(10):1152-1158 [FREE Full text] [doi: 10.5858/arpa.2023-0296-OA] [Medline: 38244054]
48. Gupta R, Park JB, Herzog I, Yosufi N, Mangan A, Firouzbakht PK, et al. Applying GPT-4 to the plastic surgery inservice training examination. J Plast Reconstr Aesthet Surg 2023 Dec;87:78-82. [doi: 10.1016/j.bjps.2023.09.027] [Medline: 37812847]
49. Ali R, Tang OY, Connolly ID, Fridley JS, Shin JH, Zadnik Sullivan PL, et al. Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. Neurosurgery 2023 Nov 01;93(5):1090-1098. [doi: 10.1227/neu.0000000000002551] [Medline: 37306460]
50. Khan AA, Yunus R, Sohail M, Rehman TA, Saeed S, Bu Y, et al. Artificial intelligence for anesthesiology board-style examination questions: role of large language models. J Cardiothorac Vasc Anesth 2024 May;38(5):1251-1259. [doi: 10.1053/j.jvca.2024.01.032] [Medline: 38423884]
51. Haze T, Kawano R, Takase H, Suzuki S, Hirawa N, Tamura K. Influence on the accuracy in ChatGPT: differences in the amount of information per medical field. Int J Med Inform 2023 Dec;180:105283. [doi: 10.1016/j.ijmedinf.2023.105283] [Medline: 37931432]
52. Rizzo MG, Cai N, Constantinescu D. The performance of ChatGPT on orthopaedic in-service training exams: a comparative study of the GPT-3.5 Turbo and GPT-4 models in orthopaedic education. J Orthop 2024 Apr;50:70-75. [doi: 10.1016/j.jor.2023.11.056] [Medline: 38173829]
53. Lin S, Chan PK, Hsu W, Kao C. Exploring the proficiency of ChatGPT-4: an evaluation of its performance in the Taiwan advanced medical licensing examination. Digit Health 2024;10:20552076241237678 [FREE Full text] [doi: 10.1177/20552076241237678] [Medline: 38449683]
54. Long C, Lowe K, Zhang J, Santos AD, Alanazi A, O'Brien D, et al. A novel evaluation model for assessing ChatGPT on otolaryngology-head and neck surgery certification examinations: performance study. JMIR Med Educ 2024;10:e49970 [FREE Full text] [doi: 10.2196/49970] [Medline: 38227351]
55. Lum ZC. Can artificial intelligence pass the American Board of Orthopaedic Surgery examination? Orthopaedic residents versus ChatGPT. Clin Orthop Relat Res 2023;481(8):1623-1630 [FREE Full text] [doi: 10.1097/CORR.0000000000002704] [Medline: 37220190]
56. Mackey BP, Garabet R, Maule L, Tadesse A, Cross J, Weingarten M. Evaluating ChatGPT-4 in medical education: an assessment of subject exam performance reveals limitations in clinical curriculum support for students. Discov Artif Intell 2024;4(1):1-5. [doi: 10.1007/s44163-024-00135-2]
57. Mahajan AP, Shabet CL, Smith J, Rudy SF, Kupfer RA, Bohm LA. Assessment of artificial intelligence performance on the otolaryngology residency in-service exam. OTO Open 2023;7(4):e98 [FREE Full text] [doi: 10.1002/oto2.98] [Medline: 38034065]
58. Maitland A, Fowkes R, Maitland S. Can ChatGPT pass the MRCP (UK) written examinations? Analysis of performance and errors using a clinical decision-reasoning framework. BMJ Open 2024;14(3):e080558 [FREE Full text] [doi: 10.1136/bmjopen-2023-080558] [Medline: 38490655]
59. Morjaria L, Burns L, Bracken K, Ngo QN, Lee M, Levinson AJ, et al. Examining the threat of ChatGPT to the validity of short answer assessments in an undergraduate medical program. J Med Educ Curric Dev 2023;10:23821205231204178 [FREE Full text] [doi: 10.1177/23821205231204178] [Medline: 37780034]

60. Nakao T, Miki S, Nakamura Y, Kikuchi T, Nomura Y, Hanaoka S, et al. Capability of GPT-4V(ision) in the Japanese national medical licensing examination: evaluation study. JMIR Med Educ 2024;10:e54393 [FREE Full text] [doi: 10.2196/54393] [Medline: 38470459]
61. Panthier C, Gatinel D. Success of chatGPT, an AI language model, in taking the French language version of the European board of ophthalmology examination: A novel approach to medical knowledge assessment. J Fr Ophtalmol 2023;46(7):706-711 [FREE Full text] [doi: 10.1016/j.jfo.2023.05.006] [Medline: 37537126]
62. Riedel M, Kaefinger K, Stuehrenberg A, Ritter V, Amann N, Graf A, et al. ChatGPT's performance in German OB/GYN exams – paving the way for AI-enhanced medical education and clinical practice. Front Med (Lausanne) 2023;10:1296615 [FREE Full text] [doi: 10.3389/fmed.2023.1296615] [Medline: 38155661]
63. Saad A, Iyengar KP, Kurisunkal V, Botchu R. Assessing ChatGPT's ability to pass the FRCS orthopaedic part A exam: a critical analysis. Surgeon 2023;21(5):263-266. [doi: 10.1016/j.surge.2023.07.001] [Medline: 37517980]
64. Sarangi PK, Narayan RK, Mohakud S, Vats A, Sahani D, Mondal H. Assessing the capability of ChatGPT, Google Bard, and Microsoft Bing in solving radiology case vignettes. Indian J Radiol Imaging 2024;34(2):276-282 [FREE Full text] [doi: 10.1055/s-0043-1777746] [Medline: 38549897]
65. Scaioli G, Lo MG, Conrado F, Rosset L, Bert F, Siliquini R. Exploring the potential of ChatGPT for clinical reasoning and decision-making: a cross-sectional study on the Italian medical residency exam. Ann Ist Super Sanita 2023;59(4):267-270. [doi: 10.1093/eurpub/ckad160.862]
66. Skalidis I, Cagnina A, Luangphiphat W, Mahendiran T, Muller O, Abbe E, et al. ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story? Eur Heart J Digit Health 2023;4(3):279-281 [FREE Full text] [doi: 10.1093/ehjdh/ztad029] [Medline: 37265864]
67. Surapaneni K. Assessing the performance of ChatGPT in medical biochemistry using clinical case vignettes: observational study. JMIR Medical Education 2023;9:e47191. [doi: 10.2196/preprints.47191]
68. Watari T, Takagi S, Sakaguchi K, Nishizaki Y, Shimizu T, Yamamoto Y, et al. Performance comparison of ChatGPT-4 and Japanese medical residents in the general medicine in-training examination: comparison study. JMIR Med Educ 2023;9:e52202 [FREE Full text] [doi: 10.2196/52202] [Medline: 38055323]
69. Wójcik S, Rulkiewicz A, Pruszczyk P, Lisik W, Poboży M, Domienik-Karłowicz J. Reshaping medical education: performance of ChatGPT on a PES medical examination. Cardiol J 2024;31(3):442-450 [FREE Full text] [doi: 10.5603/cj.97517] [Medline: 37830257]
70. Yu P, Fang C, Liu X, Fu W, Ling J, Yan Z, et al. Performance of ChatGPT on the Chinese postgraduate examination for clinical medicine: survey study. JMIR Med Educ 2024;10:e48514 [FREE Full text] [doi: 10.2196/48514] [Medline: 38335017]
71. Zong H, Li J, Wu E, Wu R, Lu J, Shen B. Performance of ChatGPT on Chinese national medical licensing examinations: a five-year examination evaluation study for physicians, pharmacists and nurses. BMC Med Educ 2024;24(1):143 [FREE Full text] [doi: 10.1186/s12909-024-05125-7] [Medline: 38355517]
72. Sahin MC, Sozer A, Kuzucu P, Turkmen T, Sahin MB, Sozer E, et al. Beyond human in neurosurgical exams: ChatGPT's success in the Turkish neurosurgical society proficiency board exams. Comput Biol Med 2024 Feb;169:107807. [doi: 10.1016/j.compbiomed.2023.107807] [Medline: 38091727]
73. Guerra GA, Hofmann H, Sobhani S, Hofmann G, Gomez D, Soroudi D, et al. GPT-4 artificial intelligence model outperforms ChatGPT, medical students, and neurosurgery residents on neurosurgery written board-like questions. World Neurosurg 2023 Nov;179:e160-e165. [doi: 10.1016/j.wneu.2023.08.042] [Medline: 37597659]
74. Isleem UN, Zaidat B, Ren R, Geng EA, Burapachaisri A, Tang JE, et al. Can generative artificial intelligence pass the orthopaedic board examination? J Orthop 2024 Jul;53:27-33. [doi: 10.1016/j.jor.2023.10.026] [Medline: 38450060]
75. Wang H, Wu WZ, Dou Z, He L, Yang L. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: pave the way for medical AI. Int J Med Inform 2023 Sep;177:105173. [doi: 10.1016/j.ijmedinf.2023.105173] [Medline: 37549499]
76. Gravina AG, Pellegrino R, Palladino G, Imperio G, Ventura A, Federico A. Charting new AI education in gastroenterology: cross-sectional evaluation of ChatGPT and Perplexity AI in medical residency exam. Dig Liver Dis 2024 Aug;56(8):1304-1311. [doi: 10.1016/j.dld.2024.02.019] [Medline: 38503659]
77. Franco D'Souza R, Amanullah S, Mathew M, Surapaneni KM. Appraising the performance of ChatGPT in psychiatry using 100 clinical case vignettes. Asian J Psychiatr 2023 Nov;89:103770. [doi: 10.1016/j.ajp.2023.103770] [Medline: 37812998]
78. Mannam SS, Subtirelu R, Chauhan D, Ahmad HS, Matache IM, Bryan K, et al. Large language model-based neurosurgical evaluation matrix: a novel scoring criteria to assess the efficacy of ChatGPT as an educational tool for neurosurgery board preparation. World Neurosurg 2023 Dec;180:e765-e773. [doi: 10.1016/j.wneu.2023.10.043] [Medline: 37839567]
79. Wilhelm TI, Roos J, Kaczmarczyk R. Large language models for therapy recommendations across 3 clinical specialties: comparative study. J Med Internet Res 2023;25:e49324 [FREE Full text] [doi: 10.2196/49324] [Medline: 37902826]
80. Jin H, Lee H, Kim E. Performance of ChatGPT-3.5 and GPT-4 in national licensing examinations for medicine, pharmacy, dentistry, and nursing: a systematic review and meta-analysis. BMC Medical Education 2024;24(1):1013.
81. Rosoł M, Gąsior J, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the medical final examination. medRxiv Preprint posted online on August 16, 2023. [doi: 10.1101/2023.06.04.23290939]

82. Farhat F, Chaudhry BM, Nadeem M, Sohail SS, Madsen. Evaluating large language models for the national premedical exam in India: comparative analysis of GPT-3.5, GPT-4, and Bard. JMIR Med Educ 2024;10:e51523 [FREE Full text] [doi: 10.2196/51523] [Medline: 38381486]
83. Newton P, Xiromeriti M. ChatGPT performance on multiple choice question examinations in higher education: A pragmatic scoping review. Assess Eval High Educ 2024;49(6):781-798. [doi: 10.1080/02602938.2023.2299059]
84. Zhu Q, Luo J. Toward artificial empathy for human-centered design: a framework. In: Proceedings of the ASME 2023 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. New York: American Society of Mechanical Engineers; 2023.
85. Yang Z, Yao Z, Tasmin M, Vashisht P, Jang WS, Ouyang F, et al. Unveiling GPT-4V's hidden challenges behind high accuracy on USMLE questions: observational study. J Med Internet Res 2025;27:e65146 [FREE Full text] [doi: 10.2196/65146] [Medline: 39919278]
86. Levin G, Horesh N, Brezinov Y, Meyer R. Performance of ChatGPT in medical examinations: a systematic review and a meta-analysis. BJOG 2024;131(3):378-380. [doi: 10.1111/1471-0528.17641] [Medline: 37604703]
87. Chaiban T, Nahle Z, Assi G, Cherfane M. The intent of ChatGPT usage and its robustness in medical proficiency exams: a systematic review. Discov Educ 2024;3(1):232. [doi: 10.1007/s44217-024-00332-2]
88. Kasai J, Kasai Y, Sakaguchi K, Yamada Y, Radev D. Evaluating GPT-4 and ChatGPT on Japanese medical licensing examinations. arXiv Preprint posted online on March 31, 2023. [doi: 10.48550/arXiv.2303.18027]
89. Akhtarshenas A, Dini A, Ayoobi N. ChatGPT or a silent everywhere helper: a survey of large language model. arXiv Preprint posted online on March 19, 2025. [doi: 10.48550/arXiv.2503.17403]
90. Liu M, Okuhara T, Dai Z, Huang W, Okada H, Emi F, et al. Performance of advanced large language models (GPT-4o, GPT-4, Gemini 1.5 Pro, Claude 3 Opus) on Japanese medical licensing examination: a comparative study. medRxiv Preprint posted online on July 9, 2024. [doi: 10.1101/2024.07.09.24310129]

Abbreviations

AI: artificial intelligence
LLM: large language model
MLEs: medical licensing examinations
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
ROBINS-I: Risk of Bias in Nonrandomized Interventional Studies
RR: risk ratio
USMLE: United States Medical Licensing Examination

Edited by B Lesselroth; submitted 28.10.24; peer-reviewed by J Riese, Z Yang, S Jaleel, W Shabbir; comments to author 08.04.25; revised version received 06.05.25; accepted 16.07.25; published 19.09.25.

Please cite as:
Jaleel A, Aziz U, Farid G, Zahid Bashir M, Mirza TR, Khizar Abbas SM, Aslam S, Sikander RMH
Evaluating the Potential and Accuracy of ChatGPT-3.5 and 4.0 in Medical Licensing and In-Training Examinations: Systematic Review and Meta-Analysis
JMIR Med Educ 2025;11:e68070
URL: https://mededu.jmir.org/2025/1/e68070
doi:10.2196/68070
PMID:

Review

Barriers and Enablers to the Production of Open Access Medical Education Platforms: Scoping Review

Ahmed Abdelfattah Eltomelhussein Ahmed[1], MBBS, MCh; Arushi Biswas[2], BA, BSE; Nefti Bempong-Ahun[3], BSc, MMedSci; Ines Perić[4], MA, PhD; Eric Patrick O'Flynn[4], BA, MA, MSc

[1]Department of Surgical Affairs, Royal College of Surgeons in Ireland, Dublin, Ireland

[2]Johns Hopkins University School of Medicine, Baltimore, MD, United States

[3]The Global Surgery Foundation, Geneva, Switzerland

[4]Institute of Global Surgery, Royal College of Surgeons in Ireland, Dublin, Ireland

Corresponding Author:
Eric Patrick O'Flynn, BA, MA, MSc
Institute of Global Surgery
Royal College of Surgeons in Ireland
118 St Stephen's Green, Dublin 2
Dublin, D02X0N1
Ireland
Phone: 353 851178005
Email: ericoflynn@rcsi.com

Abstract

Background: Free Open Access Medical Education has the potential to democratize access to medical knowledge globally; however, this potential remains largely unrealized, particularly in resource-limited settings. Content is increasingly concentrated on a small number of platforms, each hosting large volumes of material compiled from diverse sources.

Objective: This scoping review aimed to identify and synthesize reported barriers and enablers to the successful design, production, and operation of open access medical education platforms, with the goal of informing strategies to improve their impact, reach, and sustainability.

Methods: We conducted a scoping review using the Arksey and O'Malley framework. A structured search was carried out on April 17, 2023, in PubMed and EBSCOhost. Citation chaining with the SnowGlobe tool and manual reference checking supplemented the search. Studies were eligible for inclusion if they examined platforms that compile content from multiple sources and reported barriers and enablers. Two reviewers (AAEA and AB) independently screened records and extracted data, with discrepancies resolved by a third reviewer (EPOF). Beginning with an a priori framework of "barriers" and "enablers," coding was then developed inductively. Thematic synthesis categorized findings by stakeholder group.

Results: Of 1108 records identified, 1064 unique records were screened, and 64 full-text papers were assessed; 34 met the inclusion criteria. The most frequently reported barriers were concerns about content-quality control, incomplete or unstructured materials, and the resources needed to sustain platforms long-term. Key enablers included the use of validated tools to assess content quality and collaboration with existing content providers and platforms to enhance visibility and learner engagement. Findings were organized into 3 stakeholder groups: learners and training programs, content designers and creators, and platform managers.

Conclusions: Open access medical education platforms have significant untapped potential to enhance global medical training. Addressing these persistent challenges—particularly around quality assurance, content organization, and sustainability—will require more structured, collaborative, and internationally coordinated approaches.

(JMIR Med Educ 2025;11:e65306) doi:10.2196/65306

KEYWORDS

open access publishing; medical education; scoping review; educational technology; digital learning; barriers

Introduction

Advances in technology have revolutionized medical education and training. e-Learning) has become an accepted part of medical training worldwide, at all levels, in both high- and low-income environments [1,2], offering health care professionals and students the promise of accessible and flexible learning opportunities, regardless of their location [3,4]. e-Learning can be defined as "an approach to teaching and learning, representing all or part of the educational model applied, that is based on the use of electronic media and devices as tools for improving access to training, communication and interaction and that facilitates the adoption of new ways of understanding and developing learning" [5]. Included in this definition are a multitude of different forms including asynchronous (self-directed) courses, synchronous (live) tutorials and webinars, blogs, podcasts, social media, learning communities and communities of practice, online videos and images, and more. Learners may interact with this content as part of a formal training program, as formal in-service professional development, or at their own volition. e-Learning content is hosted on a variety of platforms of different types—including learning management systems, content management systems, and learning destination sites. In this review, we will refer to all online locations hosting e-learning content as "platforms."

Open access content is "digital, online, free of charge, and free of most copyright and licensing restrictions" [6]. This is of particular importance in resource-limited settings, where open access education has the potential to alleviate persistent educational disparities [7]. The acronym "FOAM" (or "FOAMed," Free Open Access Medical Education) is often used in open access medical education, describing "a dynamic collection of resources and tools for lifelong learning in medicine, as well as a community and an ethos" [8]. As FOAM covers many different types of content, and medical education technologies are quickly evolving, it is difficult to quantify the use of FOAM across the globe and across medical specialties. However, it is clear that there is a lot of FOAM content available, that it is increasingly accepted as a valid medium for medical education, and that it is increasingly used. Taking podcasts as one example, a study looking at podcasts related to a limited set of medical specialties found 169 different English language podcasts with over 6500 combined hours of educational content [9]. Acceptance of podcasts as a medical education tool is growing. While a 2007 study found that most of the US students and physicians surveyed felt podcasts had no role in medical education [10], a 2019 study reported that 71% of residents from various US training programs supported the utility of podcasts [11]. In total, 39% of US nephrology fellows report regularly learning from podcasts [12].

Open access does not, however, mean equal access. Use is higher in high-income countries than in low- and middle-income countries (LMICs) [13], and LMIC learners face additional challenges [14]. A systematic review of e-learning for medical education in LMICs finds that it has "not met its expected potential" in such settings [15]. Even within a single LMIC institution, some learners have greater access than others [16]. Thus, due to these disparities, open access platforms may not in fact fulfill their promise of decreasing pre-existing educational [17], health care, and economic inequalities [18]. This is a profound issue: those most in need of medical education and training information, such as professionals in resource-limited settings, often find it the most elusive [19].

The total number of FOAM sites now seems to be decreasing, as open access medical education resources are increasingly consolidated in a smaller number of platforms [20-22]. Due to the effort and expense involved in producing, maintaining, and hosting e-learning content, it can be expected that this trend will continue; the number of platforms will decrease, yet the volume of content on each platform will increase, compiled from multiple sources. Barriers and enablers to the production and sustainability of e-learning material of various kinds have been described in the literature [23]. This review aims, for the first time, to identify and synthesize from the literature the barriers and enablers to the successful design, production, and operation of e-learning platforms compiling content from multiple sources. In doing so, this review looks to inform efforts to optimize the impact and reach of open access medical education resources.

Methods

Study Design

A scoping review methodology was selected "to 'map' relevant literature in the field of interest," following the framework proposed by Arksey and O'Malley [24]. The protocol was publicly registered on Open Science Framework [25]. Results are reported in line with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) [26]. The PRISMA-ScR checklist is included as Multimedia Appendix 1.

Research Question

We defined our research question as "What is known in the existing literature about the barriers and corresponding enablers to developing, producing, and maintaining open access medical education platforms containing material from multiple sources?"

Search Strategy

Searches were conducted on April 17, 2023, using PubMed and EBSCOhost, which includes Academic Search Complete, ERIC, and CINAHL Plus with Full text. The search string used was ((resources) OR (material)) AND ((((medical education) OR (training)) AND (((((identify) OR (evaluate)) OR (integrate)) OR (compile)) OR (collect))) AND (open-access)). The SnowGlobe tool [27] was then used for citation chaining (snowballing), automatically retrieving references of included studies as well as studies citing them. Manual reference searching additionally contributed a number of papers. Search results were imported into Covidence (Veritas Health Innovation), and duplicates were removed. Gray literature, such as institutional websites, preprints, and reports, was not searched.

Study Selection

Studies were eligible for inclusion if they described open access medical education or training platforms that curate or compile content from multiple sources or institutions for use by health care providers. To be included, studies also needed to provide information on factors that facilitate or hinder the design, production, implementation, or ongoing operation of such platforms. All paper types, including reviews, were considered. No date restrictions were applied, given the emerging nature of this field. Studies were excluded if they focused solely on a single course or on materials developed by a single institution. Conference abstracts and non-English language studies were also excluded.

All identified papers underwent a 2-stage screening process. In stage 1, a total of 2 authors (AAEA and AB) independently reviewed the title and abstract of each paper against the predefined inclusion and exclusion criteria. Discrepancies were resolved with the input of EPOF. In stage 2, the full texts of potentially relevant papers were assessed for final inclusion by AAEA and AB, with EPOF adjudicating on discrepancies.

Charting the Data

Beginning with an a priori framework of "barriers" and "enablers," the data chart form was then developed through an inductive process, with input from all authors. Following initial immersion in the data, coding for each barrier and enabler was developed. Data from included papers were then independently extracted by AAEA and AB into this form. Any disagreements were resolved through discussion or with the input of EPOF. Finally, the coding for barriers and enablers was reviewed, and minor coding edits were made. To derive practical recommendations for the various stakeholder groups involved in the production, management, and use of open access educational platforms, barriers and enablers were then mapped to the appropriate stakeholder group: (1) learners and training programs, (2) content designers and creators, and (3) platform managers.

Data Synthesis and Reporting

Thematic analysis was undertaken. Once authors were familiar with the data, and coding agreed, themes were then proposed, explained, and clarified through iterative team discussion. Both barriers and enablers were then grouped under these themes, and the themes were reviewed for appropriateness. Descriptive statistics were used to analyze characteristics of the included studies.

Ethical Considerations

The review involved analysis of publicly available secondary data only. No human or animal participants were involved; therefore, ethics approval was not required.

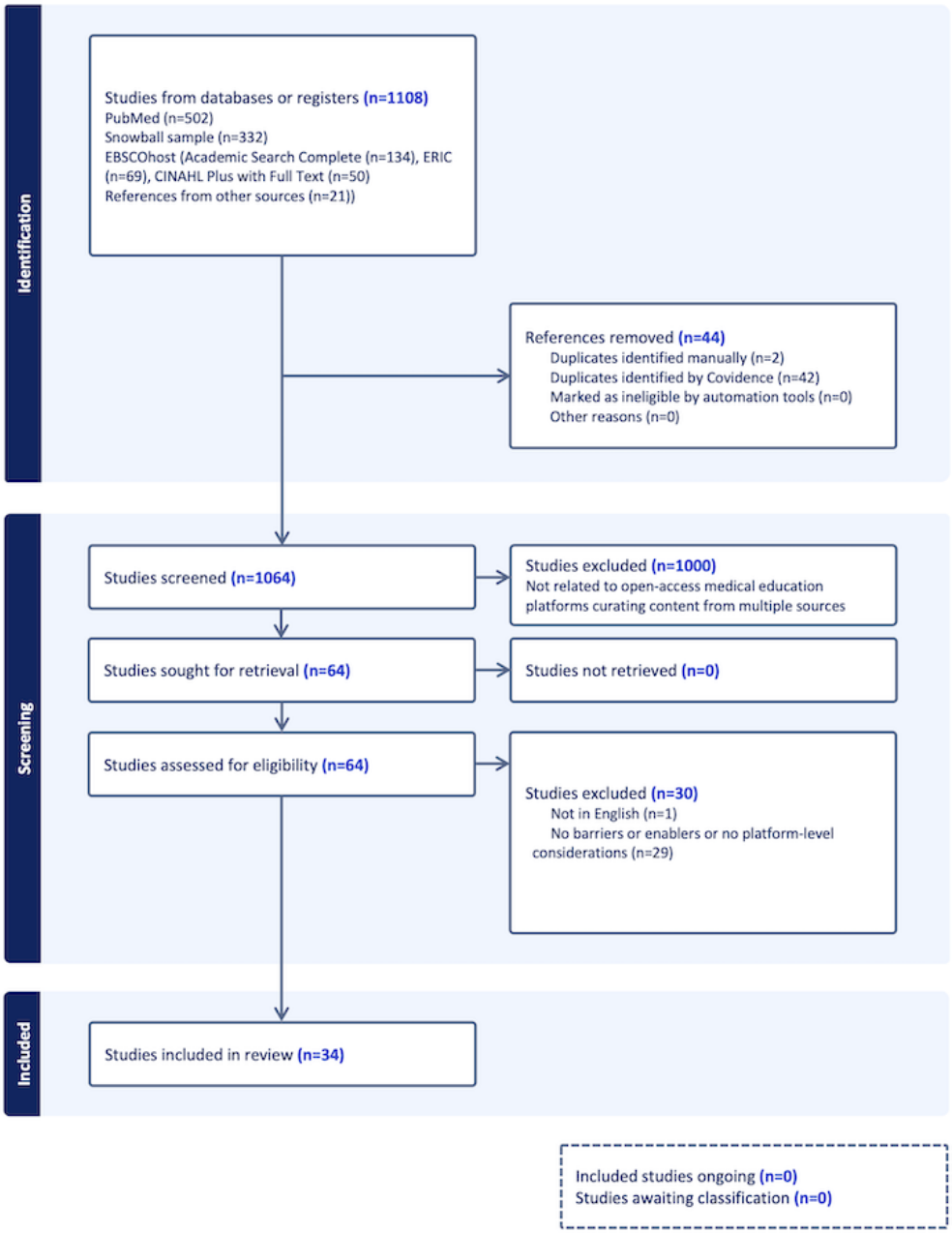
Results

Overview

Our search strategy identified a total of 1108 records from multiple sources. PubMed contributed 502 records, followed by 253 from EBSCOhost databases (including Academic Search Complete, ERIC, and CINAHL Plus with Full Text). Snowball sampling added 332 records, and 21 were identified through manual reference searching. After removing duplicates, 1064 records were screened by title and abstract. Of these, 64 full-text papers were assessed for eligibility.

A total of 29 full-text studies were excluded because, although they addressed open access content, they did not examine barriers or enablers or did not relate to platform-level considerations. Ultimately, 34 studies were included in the final review. The study selection process is shown in Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.



Characteristics of the Included Studies

All 34 studies described educational resources targeted at doctors or medical students. Nurses were also targeted in 14 (41%) studies, and allied health professionals in 11 (32%) studies. Most studies did not explicitly focus on a particular geography. Some focused on LMICs [2,28], while others were explicitly "global" [29] or "international" [30] in their scope. However, most first authors were based in high-income countries, with a majority from the United States. Other first authors were from Canada, Germany, Belgium, Italy, the United Arab Emirates, the United Kingdom, Australia, Botswana, Papua New Guinea, and South Africa. In total, 16 (47%) studies were categorized as reviews or expert opinion, 11 (32%) were primarily qualitative research, 5 (15%) were primarily quantitative research, and 2 (6%) were reports on educational innovations. While all included studies discussed open access medical education platforms, this was rarely their primary focus. Most included studies focused on e-learning in medical education in general, or in specific medical education and training contexts, or specific topics and specialties, or for specific cohorts of learners. Other studies again focused on the FOAM education movement. Characteristics of the included studies are given in Table 1.

Table 1. Characteristics of included studies.

Author name (year)	Country of first author	Type of study	Platform or context	Barriers	Enablers
Kleinpell et al (2011) [31]	United States	Qualitative research	Early open online nursing education resources (blogs, wikis, repositories); emphasis on discoverability and maintenance burdens.	• Navigability, searchability • Access • Effort and cost to maintain	• Crowdsourced feedback • Interactive content • Tools to help learners find resources • Student-centered, personalized • Content integrated with curricula • Networking of educators • Tools to make content creation easier
Nieder et al (2022) [2]	Germany	Scoping review	Global or LMIC[a]-friendly open resources for medical education (FOAM[b]-style blogs or podcasts, YouTube, mobile-first sites) with attention to bandwidth and language.	• Computer, phone, internet • Time • Language • Access • Digital literacy • Learning culture	• Collaboration for awareness • Social, instructor, and cognitive presence • Learner familiarity • Low bandwidth content • Local language • External pressure • Appropriate content structure • Interactive content • Quality tool • Regular content review
Pizzolato et al (2020) [32]	Belgium	Qualitative study	Open educational websites with structured navigation and mapped content areas to guide learners.	• Incompleteness, lack of structure • Interactivity, design	• Navigable, searchable • Interactive content • Map content areas
Grock et al (2019) [33]	United States	Qualitative research	FOAM resources in emergency medicine; blogs or podcasts and their quality or appropriateness concerns.	• Volume of content • Quality control • Appropriateness of content • Incompleteness, lack of structure	• Quality tool
Schettino and Capone (2022) [34]	Italy	Scoping review	Open platforms used in undergraduate medical curricula; focus on training content creators and interactive, learner-centered design.	• Language • Digital literacy • Learning culture • Motivation • Time	• Local language • Student-centered, personalized • Appropriate content structure • Training of content creators • Interactive content
Bansal (2021) [35]	Australia	Expert opinion	Broad FOAM ecosystem across blogs, podcasts, Twitter, YouTube, and online question banks; strategies to expand reach and impact.	• Navigability, searchability • Interactivity, design • Quality control • Motivation • Learning culture • Effort and cost to maintain	• Collaboration for awareness • Networking of educators • Training of content creators
Regmi and Jones (2020) [23]	United Kingdom	Systematic review	Systematic review of digital or open resources integrated with trainee experience and formal curricula; clarity and structure emphasized.	• Digital literacy • Motivation • Interactivity, design • Appropriateness of content • Incompleteness, lack of structure • Effort and cost to build • Effort and cost to maintain	• Content integrated with trainee experience • Content integrated with curricula • Clarity of content • Student-centered, personalized • Appropriate content structure • Interactive content
Culbert et al (2022) [36]	United States	Review paper	Review of open access educational resources in medicine; integration into curricula and institutional adoption.	• Access • Quality control • Quality tool issues • Effort and cost to maintain	• Collaboration for awareness • Content integrated with curricula

Author name (year)	Country of first author	Type of study	Platform or context	Barriers	Enablers
Rodman et al (2021) [28]	United States	Needs assessment	Needs assessment for OERs[c], preference for locally authored content, and regular review cycles.	• Language • Effort and cost to maintain • Appropriateness of content	• Technological advancement • Local language • Local content creators • Regular content review
Cevik et al (2021) [30]	United Arab Emirates	Expert opinion	Expert guidance on leveraging FOAM or OER with curation and Creative Commons licensing; curated compilations to aid discovery.	• Computer, phone, internet • Awareness • Language • Interactivity, design	• Collaboration for awareness • Curated compilations • Open Access and Creative Commons
Knopf et al (2020) [37]	United States	Descriptive analytical study	Descriptive analysis of learner engagement with open platforms; emphasis on social or instructor presence and tool quality.	• Computer, phone, internet • Language • Learning culture • Quality control • Incompleteness, lack of structure	• Social, instructor, and cognitive presence • Quality tool
Chan et al (2019) [38]	Canada	Expert opinion	Critical appraisal of FOAM resources; encourages structured evaluation frameworks for blogs or podcasts and social media posts.	• Volume of content • Trust • Quality control • Incompleteness, lack of structure	• Learner familiarity • Collaboration for awareness • Peer review, editorial process • Quality tool • Crowdsourced feedback • Map content areas
Cameron and Schofield (2017) [29]	Canada	Review paper	Review of open online medical education with emphasis on ethics, intellectual property, and learner-centered design; recommends low-bandwidth delivery.	• Computer, phone, internet • Learning culture • Motivation • Effort and cost to build • Ethics • Intellectual property	• Low bandwidth content • Content integrated with trainee experience • Clarity of content • Student-centered, personalized • Appropriate content structure • Immediate feedback • Social, instructor, and cognitive presence • Existing ethical guidelines • Open Access and Creative Commons
Lin et al (2016) [39]	United States	Educational innovation report	Approved Instructional Resources initiative; curated, peer-reviewed emergency medicine FOAM blogs or podcasts aligned to curricula.	• Quality control • Quality tool issues • Incompleteness, lack of structure	• Collaboration for awareness • Curated compilations • Free tools for hosting content
Thurtle et al (2016) [40]	Australia	Qualitative research	Qualitative exploration of FOAM awareness and trust; learners relying on local creators and site familiarity.	• Computer, phone, internet • Awareness • Trust • Navigability, searchability	• Learner familiarity • Local content creators
Thoma et al (2015) [41]	Canada	Educational innovation report	Implementation of peer review on a major FOAM blog to improve quality assurance and transparency.	• Expertise of authors • Quality control	• —[d]
Thoma et al (2014) [42]	Canada	Perspective paper		• Trust • Navigability, searchability • Volume of content • Quality control	

Author name (year)	Country of first author	Type of study	Platform or context	Barriers	Enablers
			Perspective on navigating FOAM: curated compilations, mentor-recommended resources, and residency-endorsed lists.		• Tools to help learners find resources • Resources recommended by colleagues, mentors • Curated compilations • Resources recommended by training programs
Alexiou and Falagas (2008) [43]	Greece	Expert opinion	Early expert commentary on open online resources in medical education; calls for better organization and curation.	• Volume of content • Incompleteness, lack of structure • Effort and cost to maintain	• Tools to help learners find resources • Interactive content • Curated compilations
Grock et al (2021) [44]	United States	Observational study	Analysis of FOAM coverage of core curricular topics; identifies gaps relative to competency frameworks.	• Navigability, searchability • Quality control • Incompleteness, lack of structure	• —
Grock et al (2021) [45]	United States	Usability study	Revised Approved Instructional Resources initiative; curated, peer-reviewed emergency medicine FOAM blogs or podcasts aligned to curricula.	• Quality control • Quality tool issues	• Quality tool
Ghiathi et al (2020) [46]	United States	Evaluation study	Evaluation of engagement strategies and tooling that make creating and maintaining open content easier.	• Quality control • Quality tool issues	• Collaboration for awareness • Engagement strategy • Tools to make content creation easier
Misra and Lawson (2019) [47]	United States	Technical report	Technical report on operationalizing open resources; stresses maintenance effort and need for periodic review.	• Incompleteness, lack of structure • Effort and cost to maintain	• Quality tool • Curated compilations • Regular content review
Hsiao et al (2021) [48]	United States	Review paper	Review of topic-focused open resources; supports curated lists and quality tools for better navigation.	• Navigability, searchability • Volume of content • Quality control • Incompleteness, lack of structure	• Quality tool • Curated compilations
Lin et al (2022) [21]	United States	Cross-sectional study	Census of active emergency medicine or critical care FOAM blogs or podcasts; sustainability and trend analysis.	• Dated content • Effort and cost to build • Effort and cost to maintain	• Collaboration for awareness • Content integrated with curricula
Mncube and Mthethwa (2022) [49]	South Africa	Qualitative research	Qualitative study of FOAM in South Africa; emphasizes open licensing and social or instructor presence.	• Awareness • Access • Quality control • Effort and cost to maintain • Ethics	• Collaboration for awareness • Social, instructor, and cognitive presence • Open Access and Creative Commons
Zhang et al (2019) [50]	Canada	Observational study	Observational work on open platform structure and interactivity; highlights the need for quality tools.	• Navigability, searchability • Interactivity, design • Quality control • Incompleteness, lack of structure • Effort and cost to build	• Collaboration for awareness • Technological advancement • Appropriate content structure • Interactive content • Quality tool

Author name (year)	Country of first author	Type of study	Platform or context	Barriers	Enablers
Grabow Moore et al (2021) [51]	United States	Descriptive study	Descriptive account of learner-centered open content design and routine review processes.	• Computer, phone, internet • Time • Quality control • Incompleteness, lack of structure • Effort and cost to maintain	• Student-centered content design • Crowdsourced feedback • Regular content review
Lee et al (2022) [52]	Canada	Economic evaluation	Economic evaluation of creating or maintaining OERs; discusses funding or recognition mechanisms.	• Effort and cost to build • Effort and cost to maintain	• Learner familiarity • Funding and recognition for content creation
Wolbrink et al (2019) [53]	United States	Analytic review	Review of selected open resources for critical care.	• Access • Interactivity, design	• Navigable, searchable • Appropriate content structure • Interactive content • Peer review, editorial process • Quality tool • Regular content review • Map content areas
Shappell et al (2017) [54]	United States	Review paper	Review of FOAM ecosystems; maps content areas, curated compilations, and crowdsourced feedback models.	• Interactivity, design • Quality control • Incompleteness, lack of structure	• Collaboration for awareness • Appropriate content structure • Interactive content • Quality tool • Crowdsourced feedback • Curated compilations • Map content areas
Evans et al (2022) [55]	United States	Cross-sectional study	Cross-sectional study of open resources, processes, and tooling (peer review, quality instruments) across sites.	• Computer, phone, internet • Access • Interactivity, design • Quality control • Quality tool issues • Effort and cost to build • Effort and cost to maintain	• Collaboration for awareness • Interactive content • Peer review, editorial process • Quality tool
Lo et al (2018) [56]	Canada	Expert opinion	Expert commentary on integrating FOAM into formal education; highlights peer-reviewed tools and mentor-curated lists.	• Volume of content	• Resources recommended by colleagues, mentors • Interactive content • Quality tool
Ting et al (2020) [57]	Canada	Literature review	Quality appraisal or assurance techniques for FOAM; proposes methods to evaluate blogs or podcasts and social media content.	• Ethics • Quality control • Ethics	• Content integrated with curricula • Quality tool • Funding and recognition for content creation
Marée (2019) [58]	Belgium	Mini review paper	Mini review on OER or FOAM with focus on maintenance costs and intellectual property; advocates collaboration for awareness.	• Computer, phone, internet • Incompleteness, lack of structure • Effort and cost to maintain • Intellectual property	• Collaboration for awareness

[a]LMIC: low- and middle-income country.

[b]FOAM: Free Open Access Medical Education.

[c]OER: open educational resource.

[d]Not available.

Across the 34 included studies, we identified 127 instances of barriers and 122 instances of enablers. These were coded into 22 unique barriers and 33 unique enablers, and then, mapped to 1 of 3 stakeholder-based themes. A full list of the coded barriers and enablers, and the frequency with which they were identified, is presented in Tables 2 and 3.

Table 2. Frequency and detail of barriers identified.

Barrier name	Barrier detail	Values, n
Theme: learners and training programs		
Learning culture	Lack of learner culture of e-learning, and lack of culture of or dislike of interactive learning	5
Motivation	General lack of motivation and adherence	4
Digital literacy	Lack of learner digital literacy	3
Time	Learners' limited time availability or conflict of priorities	3
Trust	Learners' lack of trust in content, lack of connection to content creators	3
Theme: content designers and creators		
Quality control	Lack of quality control	18
Incompleteness, lack of structure	Material does not equally cover all required learner knowledge areas and is not presented in a logical structure	14
Interactivity, design	Lack of interactivity of resources and poor design	8
Computer, phone, internet	Learners' lack of access to adequate computer, phone, or internet	8
Navigability, searchability	Resources are difficult to find and navigate	7
Volume of content	Overwhelming volume of content	6
Language	Lack of learner competence or comfort in the language of resources	5
Quality tool issues	Issues with quality tools used	5
Appropriateness of content	Content not appropriate for learners' context and level	3
Dated content	Dated content	1
Expertise of authors	Lack of author subject area expertise	1
Theme: platform managers		
Effort and cost to maintain	Human and financial resources required to maintain and update content	13
Access	Restrictions on learner access to material (requirement to register or be a member of certain organizations)	6
Effort and cost to build	Human and financial resources required to develop content	6
Awareness	Learners' lack of awareness of resources available	3
Ethics	Ethical issues	3
Intellectual property	Intellectual property issues	2

Table 3. Frequency and detail of enablers identified.

Enabler name	Enabler detail	Values, n
Theme: learners and training programs		
Curated compilations	Use of curated compilations of resources	7
Learner familiarity	Growth of learner familiarity with and awareness of e-learning and FOAM[a] in general	4
Resources recommended by colleagues and mentors	Recommendations of resources by trusted colleagues and mentors	2
External pressure	External circumstances and pressures	1
Engagement strategy	Production of a learner engagement strategy	1
Resources recommended by training programs	Curation or recommendation of resources by training programs	1
Theme: content designers and creators		
Quality tool	Use of a tool to rate quality of online resources or framework development	13
Interactive content	Interactive content	11
Appropriate content structure	Appropriate content structure (eg, iterative, with learning objectives, using multimedia)	7
Content integrated with curricula	Integration of resources with established curricula	5
Regular content review	Regular content review	5
Student-centered, personalized	Student-centered, personalized learning approach	5
Crowdsourced feedback	Crowdsourced feedback and curriculum development	4
Social, instructor, and cognitive presence	Establishment of a social presence (peers), instructor presence, and cognitive presence (material and assignments)	4
Map content areas	Mapping and categorization of content areas, aiming for comprehensiveness	4
Local language	Content translation into local language	3
Peer review, editorial process	Peer review, editorial process	3
Navigable, searchable	Systematic characterization of resources for navigability and searchability	2
Networking of educators	Networking of educators, sharing best practice	2
Training of content creators	Training of content creators	2
Clarity of content	Clarity of content	2
Content integrated with trainee experience	Integration of content with trainee experience	2
Local content creators	Recruitment of content creators from the same context as learners or launch LMIC[b] platform	2
Technological advancement	General global technological advancement	2
Low bandwidth content	Use of low-bandwidth content where the internet is poor or expensive	2
Immediate feedback	Provision of immediate feedback to learner	1
Theme: platform managers		
Collaboration for awareness	Collaborate and partner with existing content providers or hosts to increase awareness	13
Tools to help learners find resources	Tools to help learners find resources (eg, automatically queued resources and custom search engines)	3
Open Access and Creative Commons	Availability of Open Access and Creative Commons resources	3
Funding and recognition for content creation	Funding and recognition for content creation	2
Tools to make content creation easier	Use of tools to make content creation easier	2
Free tools for hosting content	Use of free tools for hosting content	1
Existing ethical guidelines	Use of existing ethical guidelines	1

[a]FOAM: Free Open Access Medical Education.

[b]LMIC: low- and middle-income country.

Learners and Training Programs

Barriers

Most barriers under this theme related to learner engagement: the lack of an interactive learning culture conducive to self-directed e-learning, particularly in resource-limited settings [2,29,34,35,37]; a lack of trust in content or content creators [38,40,41]; and a general lack of learner motivation and adherence [23,29,34,35]. Practical barriers were a lack of learner digital literacy [2,23,34] and learners' limited available time due to competing priorities [2,34,51]. Barriers under this theme were particularly reported by studies in resource-limited contexts.

Enablers

Some of the barriers found can be expected to be partially addressed by global macrotrends, such as increasing learner familiarity with, and acceptance of, e-learning and interactive learning approaches [2,38,40,52]. Several enablers were reported, which may address challenges of learner trust and motivation—peer-to-peer recommendation of content [42,56], training bodies recommending specific content to their trainees [42], and the use of trusted curated compilations of resources [30,39,42,43,47,48,54]. Other enablers can be leveraged by training programs to increase engagement, such as the development of a formal engagement strategy [46]. External pressure, such as when completion of certain e-learning content is determined to be a mandatory part of a training program, with appropriate deadlines and reminders, was also reported as an enabler [2].

Content Designers and Creators

Barriers

Barriers reported relating to content creation and design included a lack of resource interactivity [23,30,32,35,50,53-55], resources that are difficult to find and navigate [31,35,40,42,44,48,50], and an overwhelming volume of content [33,38,42,43,48,56]. However, the major challenge under this theme—and the most cited barrier in this study—was a lack of quality control, which was identified in the majority of included studies (n=18, 53%) [33,35-39,41,42,44-46,48-51,54,55,57]. Other quality-related barriers reported were problems with quality evaluation tools [36,39,45,46,55], a lack of author subject area expertise [41], and dated content [21].

Incompleteness of content or lack of structure—where material does not appropriately cover all required learner knowledge areas or is not presented in a logical structure—was repeatedly reported [23,32,33,37-39,43,44,47,48,50,51,54,58]. Some topics appear to be covered in detail on multiple platforms, while other topics are seemingly absent. Grock et al [44] describe the "uneven distribution" of open access content as "holes in the FOAM." Content in some cases was also reported not to be at the appropriate level for target learners or not appropriate for the learners' context [23,28,33].

Finally, practical barriers found that can be taken into consideration by content designers and creators were learners' lack of access to adequate equipment and infrastructure such as computers, phones, and reliable internet connections [2,29,30,37,40,51,55,58] and lack of learner competence in the language in which content was provided [2,28,30,34,37].

Enablers

Many of the enablers reported under this theme can be considered general good practice in adult education, such as interactive learning approaches [2,23,31,32,34,43,50,53-56], clarity of content [23,29], the provision of immediate feedback to the learner [29], and student- or learner-centered design [23,29,31,34,51]. The "flipped classroom" was cited as an example of "an interactive learner-centric approach particularly well-suited to the needs of contemporary ... learners" [51]. Gamification was reported as another effective student-centered design approach, as it "represents a way to increase the attractiveness of learning content among students and foster their motivation to participate in the proposed activities" [34]. Other enablers centered on e-learning design, such as the use of an appropriate structure with learning objectives, multimedia, and the employment of an iterative design process [2,23,29,34,50,53,54]. Another reported enabler was the establishment of 3 forms of "presence": social presence (peers), instructor presence, and cognitive presence (material and assignments) [2,29,37,49].

Two enablers reported focused on the development of content creators: the networking and sharing of best practices among educators [31,35] and the training of content creators [34,35]. Bansal [35] proposes faculty development programs to "train a cohort of clinical educators who can drive the creation of FOAM resources." The final enabler under this theme was the systematic characterization of resources for navigability and searchability [32,53].

The most commonly reported enabler to address the issues of content quality was the use of objective, validated content quality evaluation tools to assess the quality of content before publication [2,33,37,38,45,47,48,50,53-57]. Examples given include the "Approved Instructional Resources" tool [45] and the revised "Medical Education Translational Resources: Impact and Quality" tool [59]. Other enablers addressing issues of content quality were the implementation of regular content review [2,28,47,51,53], content peer review as part of a transparent selection and evaluation process [38,53,55], and crowdsourced feedback [31,38,51,54].

Two enablers under this theme focused on ensuring that content was appropriate for the target learner's level and context: the recruitment of content creators from the same context as targeted learners [28,40] and the integration of learners' experience into content [23,29]. Integration of resources with established curricula [21,23,31,36,57] was reported as a means to ensure both content appropriateness and comprehensiveness. Mapping and categorizing content areas [32,38,53,54] were reported as a way to identify content area gaps and thus enable the production of more complete and comprehensive learning resources.

Where the internet connection of the target learner population is likely to be expensive or of poor quality, ensuring content is not bandwidth-intensive was an identified enabler [2,29]. In such contexts, Nieder et al [2] propose "[p]roviding access to

course media such as video content in lower resolution and downloading course materials for offline usage in places with a better connection." The global spread of technology and improvements in internet speed can also be expected to enable greater access over time [28,50]. Learner participation was reported to be facilitated by translation of content into local languages [2,28,34].

Platform Managers

Barriers

Barriers found, which can be addressed by platform managers, include learners' inability to access content due to requirements to register or be a member of an organization [2,31,36,49,53,55]. Lack of learner awareness of available resources [30,40,49] can also be addressed by platform managers.

A number of barriers to the long-term sustainability of open access platforms were reported—most commonly, the resources and efforts required to maintain and update content [21,23,28,31,35,36,43,47,49,51,52,55,58]. Lin et al [21] note that open access medical education platforms originated from "volunteers providing free education to all who wish to learn. This volunteerism comes at the expense of opportunity costs and may have been unsustainable for many sites that no longer exist." The direct costs of content development are also a notable barrier [21,23,29,50,52,55]. Barriers that were reported related to intellectual property [29,58], included questions over the legality of the common practice of reproducing content from journal papers [29]. Ethical concerns, such as the deliberate sharing of incorrect information, bullying on social learning spaces, and the unethical appropriation of material, were also raised [29,49,57].

Enablers

Collaboration between content providers and platforms was repeatedly reported as an enabler, in order to make resources available on highly visible platforms, thus addressing a lack of learner awareness of these resources [2,21,30,35,36,38,39,46,49,50,54,55,58]. Several relatively simple technological enablers were also found. Where cost is a barrier, the use of free tools [39], such as Google Drive for content hosting, was reported as an enabler. User-friendly content creation tools were reported to reduce the effort required to produce resources [31,46]. Incorporation of tools into platforms, such as custom search engines to help learners find resources [31,42,43], was proposed to address barriers around awareness, searchability, and navigability.

Financial support and academic recognition for content creation were seen to enable the long-term effort required of content creators [52,57]. The use of Creative Commons licenses to clarify the legal status of content was reported as an enabler to address intellectual property–related barriers [29,30,49]. Users are free to retain, reuse, revise, remix, and redistribute the content (the 5Rs) under Creative Commons licenses, which have a range of openness depending on how many of the 5Rs are allowed to the user. Mncube and Mthethwa [49] consider that "Once such [Open Access Educational Resource] common laws are well established in faculty ... collaboration ... will continue to grow unviolated." One study reported the use of existing ethical guidelines as an enabler [29].

Discussion

Principal Findings

As has already been seen in the specialties of emergency medicine and critical care [21], it seems likely that open access medical education resources in other areas and specialties are becoming increasingly consolidated in a smaller number of larger online platforms. Such compilations of open access medical resources have the potential to improve the learner experience by offering a "1-stop shop," where learners are confident that they can find easy-to-use, quality-assured, up-to-date, comprehensive learning resources appropriate for their context.

Our review finds numerous barriers reported to the design, production, and operation of such platforms. The literature also shows that there is much that open access content designers, authors, and platform administrators can do to address these barriers; and indeed, there is much that can be done by training programs and learners themselves. We can reasonably expect that some of these barriers will diminish in severity over time. Some of the practical barriers identified, particularly in resource-limited settings [2], will partially be addressed by improving internet connectivity, while growing learner familiarity with interactive online learning methodologies can be expected to partially address some of the cultural and motivational barriers.

The 3 most cited barriers in this review were a perceived lack of content quality control, incompleteness of content, and concerns that the demand on human and financial resources required to maintain and update content are unsustainable.

While informal peer and expert recommendations of resources may be helpful, neither learners nor experts are consistently able to make an accurate "gut feeling" appraisal of the quality of online educational resources [60]; therefore, informal recommendation of resources alone may not suffice as a quality control measure. Objective, validated content quality evaluation tools such as the "Approved Instructional Resources" tool [45] or the revised "Medical Education Translational Resources: Impact and Quality" tool [59] should be routinely used by content creators and open access platform curators.

Where e-learning content is targeted at students or trainees in formal training programs, such content should be integrated with learners' other educational activities, rather than existing apart from them. The connection between content and curricula should be made clear and explicit through a mapping of content to appropriate established curricula. Training programs should identify appropriate open access resources, which fulfill the requirements of their curricula and should guide their students and trainees toward them. The curation of content into platforms, and the linking of content and curricula, will make clear which training needs are met and which are not met by currently available open access content. Identification of content area gaps will reduce duplication of effort and allow "holes" [44] in

open access medical education content to be filled by appropriate content.

The barrier presented by the human and financial resources required to maintain and update content is the least amenable to the solutions found in this review. Enablers addressing the sustainability of human resources, such as paying content creators, may negatively impact the financial sustainability of such platforms. Cost and effort may be contained to some degree by the enablers identified in this review; however, creating, hosting, and updating high-quality content will continue to cost money and require significant effort. Hosting content on curated platforms, such as the United Nations Global Surgery Learning Hub [61] and OpenCriticalCare [62], which are free for both content providers and learners, may address some of the issues associated with long-term hosting of content as well as partially address issues of quality control and trust. Ultimately, the value of material and platforms will have to be acknowledged for the associated costs to be met.

Comparison With Prior Work

This review differs from prior published work by focusing on open access e-learning platforms compiling content from multiple sources, whereas previously published literature has predominantly focused on specific courses and cohorts. Many of the barriers and enablers identified relating to platforms have previously been found in studies of individual courses, and meta-analyses relating to e-learning in health sciences in general [23] and e-learning in medical education in LMICs [15].

These include learner motivation and familiarity with e-learning, incompleteness and inappropriateness of content, and challenges in maintaining and updating content. In LMICs, basic challenges in getting online were highlighted [15]. Enablers related to all of these barriers were similar to those found in this study. Some challenges related to courses are not applicable to platforms. Barteit et al [15] found a preponderance of pilot or small-scale e-learning interventions for resource-limited settings that never produced content at scale. De facto compilations of medical education contain significant volumes of content.

This study found the financial sustainability of open access medical education compilations, and the sustainability of the level of effort required, to be challenges with few straightforward solutions. This concern is echoed throughout the broader literature on open access medical education. Lee et al [52] calculated the median overall value of a FOAM website to be US $22,815 and concluded that there is “substantial value being generated from these resources and ... this should be recognized by academia.” Platform costs will either have to be met by funding from academic institutions and funding bodies or by charging fees—to content providers, to learners in high-income settings, to training programs, or for learners to access “premium” content. Lin et al [21] found that the majority of the 109 FOAM sites that they analyzed generated income, with 18% embedding advertisements; 27.5% asking for donations or payment for merchandise, continuing medical education credits, books, or web-based courses; and 44% being affiliated with a sponsoring institution, such as a professional society, hospital, or journal. Lee et al [52] draw parallels between the FOAM movement and the open access publication movement, noting that in making publications available open access, publishers shifted the cost from the content consumer to the content producer. They note that grant funding mechanisms exist to fund these publication costs and call for similar support for medical education content creation. Mncube and Mthethwa [49] quote an unnamed academic, “When we write and publish the research articles or book chapters we are incentivized, however in the development of [open access resources], there are no incentives.” Where it is not possible to fund content creators, and until more sustainable models are developed, greater acknowledgment and recognition of content creators and others involved in this work may be a simple way to help sustain their motivation.

Two important barriers found relating to compilations of medical education resources, that were either not discussed or not given prominence in the literature on individual courses, are issues of trust and language. It may simply be that analyses at a course level exclude potential learners who either do not trust the content or are not comfortable learning in the course language of instruction—they would never take the courses in the first place. We found that managers of such platforms aspiring to make content sourced from multiple different institutions available to a global audience must give much higher priority to issues of trust and language than content creators may have previously done. We found several means by which platform managers can build trust—most prominently, the transparent use of objective, validated content quality evaluation tools. Translation of content into appropriate languages was found to be a key enabler.

Recommendations for Future Research

In addition to mapping existing literature, future empirical research should include systematic surveys of open access platform websites and interviews with their developers to provide deeper insight into their design, governance, and real-world implementation. There is a notable lack of longitudinal research evaluating the lasting impact of open access medical education platforms. Additionally, limited evidence exists on the structured integration of these resources into formal medical training programs. Sustainable models—particularly those detailing financial and operational strategies for the long-term maintenance of such platforms—are also rarely published.

Implication of Findings

This review identifies from the literature actions that each stakeholder group can take to improve open access medical platforms and increase their impact. Recommendations are summarized in Table 4.

Table 4. Recommended actions for stakeholder groups.

Stakeholder group	Recommended actions
Learners	• Use trusted curated compilations of resources • Peer-to-peer recommendation of content
Training programs	• Recommend specific content on curated compilations of resources to learners • Require learner completion of certain content, with deadlines • Develop a formal learner engagement strategy
Content designers and creators	• Evaluate the quality of educational content using validated evaluation tools • Share best practices among educators and content creators • Recruit content creators from the same context as targeted learners and integrate learners' experience into content • Provide low-bandwidth and downloadable resources for learners with poor or expensive internet connectivity • Map and categorize new and existing content to ensure new content fills gaps rather than duplicates existing content • Integrate new resources into established curricula • Combine self-directed cognitive learning with peer interaction and instructor presence • Adhere to best practices in adult education, including interactive learning and learner-centered approaches, and the provision of immediate feedback • Translate content into appropriate languages
Platform managers	• Use Creative Commons licenses to clarify the legal status of content • Host content on open access platforms and use free tools to reduce costs • Enhance navigability and searchability with custom search tools and other tools • Formally recognize the contribution of content creators

Limitations

Only studies published in English were included, which may introduce language bias and limit the representation of perspectives from non-English–speaking regions. Furthermore, the majority of included studies originated from high-income countries, potentially narrowing the cultural and economic contexts considered. We recognize the possibility of funding bias influencing the evidence base. The exclusion of gray literature may also have led to the omission of relevant nonindexed reports or institutional documents. Changes in medical education technology are fast-paced, and the literature necessarily lags behind changes in learner behavior. Nevertheless, this review establishes a foundation for future research, implementation, and policy development in the field of open access medical education.

Conclusions

Open access medical education has the potential to add significant value to preservice and in-service medical education worldwide. However, numerous challenges are preventing open access medical education platforms from achieving their potential, particularly a lack of content quality control, incompleteness of content, and the challenge of sustaining the necessary human and financial resources.

Ultimately, if open access medical education is to achieve its potential, much greater coordination and collaboration on a global scale is required—between learners, educational institutions, content creators, and platform managers.

Authors' Contributions

All authors contributed to the design of the study. AAEA conducted the search, and AAEA and AB independently undertook each stage of the screening and extraction process, with discrepancies resolved by EPOF. All authors contributed to interpreting the data as well as drafting and reviewing the manuscript. All authors have approved the submitted manuscript and agree to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR checklist.
[PDF File (Adobe PDF File), 84 KB - mededu_v11i1e65306_app1.pdf]

References

1. Longhini J, Rossettini G, Palese A. Massive open online courses for nurses' and healthcare professionals' continuous education: a scoping review. Int Nurs Rev 2021;68(1):108-121. [doi: 10.1111/inr.12649] [Medline: 33855697]
2. Nieder J, Nayna Schwerdtle P, Sauerborn R, Barteit S. Massive open online courses for health worker education in low- and middle-income countries: a scoping review. Front Public Health 2022;10:891987 [FREE Full text] [doi: 10.3389/fpubh.2022.891987] [Medline: 35903395]
3. Vaona A, Banzi R, Kwag KH, Rigon G, Cereda D, Pecoraro V, et al. e-Learning for health professionals. Cochrane Database Syst Rev 2018;1(1):CD011736 [FREE Full text] [doi: 10.1002/14651858.CD011736.pub2] [Medline: 29355907]
4. George P, Papachristou N, Belisario JM, Wang W, Wark PA, Cotic Z, et al. Online eLearning for undergraduates in health professions: a systematic review of the impact on knowledge, skills, attitudes and satisfaction. J Glob Health 2014;4(1):010406 [FREE Full text] [doi: 10.7189/jogh.04.010406] [Medline: 24976965]
5. Sangrà A, Vlachopoulos D, Cabrera N. Building an inclusive definition of e-learning: an approach to the conceptual framework. Int Rev Res Open Dis 2012;13(2):145-159 [FREE Full text]
6. Suber P. Open Access. Cambridge, MA: MIT Press; 2012.
7. Doyle DJ. Web-based education in anesthesiology: a critical overview. Curr Opin Anaesthesiol 2008;21(6):766-771. [doi: 10.1097/aco.0b013e3283173e44] [Medline: 19009692]
8. Nickson C, Cadogan MD. Free Open Access Medical Education (FOAM) for the emergency physician. Emerg Med Australas 2014;26(1):76-83. [doi: 10.1111/1742-6723.12191] [Medline: 24495067]
9. Little A, Hampton Z, Gronowski T, Meyer C, Kalnow A. Podcasting in medicine: a review of the current content by specialty. Cureus 2020;12(1):e6726 [FREE Full text] [doi: 10.7759/cureus.6726] [Medline: 32104642]
10. Sandars J, Schroter S. Web 2.0 technologies for undergraduate and postgraduate medical education: an online survey. Postgrad Med J 2007;83(986):759-762 [FREE Full text] [doi: 10.1136/pgmj.2007.063123] [Medline: 18057175]
11. Block J, Lerwick P. 1064: Educational preferences among residents in the ICU. Crit Care Med 2019;47(1):509 [FREE Full text] [doi: 10.1097/01.ccm.0000551809.99859.12]
12. Larsen D, Boscardin C, Sparks MA. Engagement in Free Open Access Medical Education by US nephrology fellows. Clin J Am Soc Nephrol 2023;18(5):573-580 [FREE Full text] [doi: 10.2215/CJN.0000000000000123] [Medline: 36800537]
13. Burkholder T, Bellows J, King RA. Free Open Access Medical Education (FOAM) in emergency medicine: the global distribution of users in 2016. West J Emerg Med 2018;19(3):600-605 [FREE Full text] [doi: 10.5811/westjem.2018.3.36825] [Medline: 29760862]
14. O'Flynn E, Ahmed AAE, Biswas A, Bempong-Ahun N, Perić I, Puyana JC. e-Learning supporting surgical training in low-resource settings. Curr Surg Rep 2024;12(6):1-9 [FREE Full text] [doi: 10.1007/s40137-024-00399-8]
15. Barteit S, Guzek D, Jahn A, Bärnighausen T, Jorge MM, Neuhann F. Evaluation of e-learning for medical education in low- and middle-income countries: a systematic review. Comput Educ 2020;145:103726 [FREE Full text] [doi: 10.1016/j.compedu.2019.103726] [Medline: 32565611]
16. Takavarasha S, Cilliers L, Chinyamurindi W. Assessing ICT Access Disparities Between the Institutional and Home Front: A Case of University Students in South Africa's Eastern Cape. 2018 Presented at: 13th IFIP TC 9 International Conference on Human Choice and Computers; September 19-21, 2018; Poznan, Poland URL: https://link.springer.com/book/10.1007/978-3-319-99605-9#bibliographic-information [doi: 10.1007/978-3-319-99605-9_4]
17. Hadjiat Y. Healthcare inequity and digital health—a bridge for the divide, or further erosion of the chasm? PLOS Digit Health 2023;2(6):e0000268 [FREE Full text] [doi: 10.1371/journal.pdig.0000268] [Medline: 37267232]
18. García-Escribano M. Low internet access driving inequality. IMF BLOG. 2020. URL: https://www.imf.org/en/Blogs/Articles/2020/06/29/low-internet-access-driving-inequality [accessed 2025-10-10]
19. Yamey G. Open access to medical literature can boost global public health. Virtual Mentor 2009;11(7):546-550 [FREE Full text] [doi: 10.1001/virtualmentor.2009.11.7.oped1-0907] [Medline: 23199390]
20. Chan T, Stehman C, Gottlieb M, Thoma B. A short history of Free Open Access Medical Education. The past, present, and future. ATS Sch 2020;1(2):87-100 [FREE Full text] [doi: 10.34197/ats-scholar.2020-0014PS] [Medline: 33870273]
21. Lin M, Phipps M, Yilmaz Y, Nash CJ, Gisondi MA, Chan TM. A fork in the road for emergency medicine and critical care blogs and podcasts: cross-sectional study. JMIR Med Educ 2022;8(4):e39946 [FREE Full text] [doi: 10.2196/39946] [Medline: 36306167]
22. Cadogan M, Thoma B, Chan T, Lin M. Free Open Access Meducation (FOAM): the rise of emergency medicine and critical care blogs and podcasts (2002-2013). Emerg Med J 2014;31(e1):e76-e77 [FREE Full text] [doi: 10.1136/emermed-2013-203502] [Medline: 24554447]
23. Regmi K, Jones L. A systematic review of the factors—enablers and barriers—affecting e-learning in health sciences education. BMC Med Educ 2020;20(1):91 [FREE Full text] [doi: 10.1186/s12909-020-02007-6] [Medline: 32228560]
24. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005;8(1):19-32 [FREE Full text] [doi: 10.1080/1364557032000119616]
25. Ahmed A, Biswas A, Peric I, Bempong-Ahun N, O'Flynn E. Barriers and enablers to producing compilations of open-access medical education and training material: a scoping review protocol. Open Science Framework. 2023. URL: https://osf.io/mtqyf/overview [accessed 2025-10-23]

26. Tricco A, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
27. McWeeny S, Choe J, Norton E. SnowGlobe: an iterative search tool for systematic reviews and meta-analyses. Open Science Framework. 2021. URL: https://osf.io/u25rn/overview [accessed 2025-10-23]
28. Rodman A, Abrams H, Watto M, Trivedi S, Barbee J, Meraz-Munoz A, et al. Medical podcasting in low- and middle-income countries: a needs assessment and vision for the future. Teach Learn Med 2021;33(4):416-422. [doi: 10.1080/10401334.2021.1875834] [Medline: 33587858]
29. Cameron BH, Schofield S. e-Learning in global surgery. In: Park A, Price A, editors. Global Surgery: The Essentials. Cham: Springer International Publishing; 2017:127-144.
30. Cevik A, Cakal E, Kwan J. From the pandemic's front lines: a social responsibility initiative to develop an international free online emergency medicine course for medical students. Afr J Emerg Med 2021;11(1):1-2 [FREE Full text] [doi: 10.1016/j.afjem.2020.11.005] [Medline: 33304802]
31. Kleinpell R, Ely E, Williams G, Liolios A, Ward N, Tisherman SA. Web-based resources for critical care education. Crit Care Med 2011;39(3):541-553. [doi: 10.1097/CCM.0b013e318206b5b5] [Medline: 21169819]
32. Pizzolato D, Abdi S, Dierickx K. Collecting and characterizing existing and freely accessible research integrity educational resources. Account Res 2020;27(4):195-211. [doi: 10.1080/08989621.2020.1736571] [Medline: 32122167]
33. Grock A, Bhalerao A, Chan TM, Thoma B, Wescott AB, Trueger NS. Systematic Online Academic Resource (SOAR) review: renal and genitourinary. AEM Educ Train 2019;3(4):375-386 [FREE Full text] [doi: 10.1002/aet2.10351] [Medline: 31637355]
34. Schettino G, Capone V. Learning design strategies in MOOCs for physicians' training: a scoping review. Int J Environ Res Public Health 2022;19(21):14247 [FREE Full text] [doi: 10.3390/ijerph192114247] [Medline: 36361125]
35. Bansal A. Expanding Free Open-Access Medical Education. Front Med (Lausanne) 2021;8:794667 [FREE Full text] [doi: 10.3389/fmed.2021.794667] [Medline: 35004769]
36. Culbert M, Brisson R, Oladeru OT. The landscape of digital resources in radiation oncology. Tech Innov Patient Support Radiat Oncol 2022;24:19-24 [FREE Full text] [doi: 10.1016/j.tipsro.2022.08.006] [Medline: 36133932]
37. Knopf J, Kumar R, Barats M, Klimo P, Boop FA, Michael LM, et al. Neurosurgical operative videos: an analysis of an increasingly popular educational resource. World Neurosurg 2020;144:e428-e437 [FREE Full text] [doi: 10.1016/j.wneu.2020.08.187] [Medline: 32889185]
38. Chan T, Bhalerao A, Thoma B, Trueger NS, Grock A. Thinking critically about appraising FOAM. AEM Educ Train 2019;3(4):398-402 [FREE Full text] [doi: 10.1002/aet2.10352] [Medline: 31637359]
39. Lin M, Joshi N, Grock A, Swaminathan A, Morley EJ, Branzetti J, et al. Approved instructional resources series: a national initiative to identify quality emergency medicine blog and podcast content for resident education. J Grad Med Educ 2016;8(2):219-225 [FREE Full text] [doi: 10.4300/JGME-D-15-00388.1] [Medline: 27168891]
40. Thurtle N, Banks C, Cox M, Pain T, Furyk J. Free Open Access Medical Education resource knowledge and utilisation amongst emergency medicine trainees: a survey in four countries. Afr J Emerg Med 2016;6(1):12-17 [FREE Full text] [doi: 10.1016/j.afjem.2015.10.005] [Medline: 30456058]
41. Thoma B, Chan T, Desouza N, Lin M. Implementing peer review at an emergency medicine blog: bridging the gap between educators and clinical experts. CJEM 2015;17(2):188-191 [FREE Full text] [doi: 10.2310/8000.2014.141393] [Medline: 25927262]
42. Thoma B, Joshi N, Trueger NS, Chan TM, Lin M. Five strategies to effectively use online resources in emergency medicine. Ann Emerg Med 2014;64(4):392-395 [FREE Full text] [doi: 10.1016/j.annemergmed.2014.05.029] [Medline: 24962889]
43. Alexiou V, Falagas ME. E-meducation.org: an open access medical education web portal. BMC Med Educ 2008;8:6 [FREE Full text] [doi: 10.1186/1472-6920-8-6] [Medline: 18218119]
44. Grock A, Chan W, Aluisio AR, Alsup C, Huang D, Joshi N. Holes in the FOAM: an analysis of curricular comprehensiveness in online educational resources. AEM Educ Train 2021;5(3):e10556 [FREE Full text] [doi: 10.1002/aet2.10556] [Medline: 34124504]
45. Grock A, Jordan J, Zaver F, Colmers-Gray IN, Krishnan K, Chan T, et al. The revised Approved Instructional Resources score: an improved quality evaluation tool for online educational resources. AEM Educ Train 2021;5(3):e10601 [FREE Full text] [doi: 10.1002/aet2.10601] [Medline: 34141997]
46. Ghiathi C, Seitz K, Kritek P. How to create and evaluate a resident-led audio program: six clinical podcasts for medicine house staff. MedEdPORTAL 2020;16:11062 [FREE Full text] [doi: 10.15766/mep_2374-8265.11062] [Medline: 33409359]
47. Misra A, Lawson C. Use of technology to identify and cover underweighted topics in resident conference curriculum. Cureus 2019;11(1):e3967 [FREE Full text] [doi: 10.7759/cureus.3967] [Medline: 30956919]
48. Hsiao J, Pedigo R, Bae SW, Jung J, Zhao L, Trueger NS, et al. Systematic Online Academic Resource (SOAR) review: endocrine, metabolic, and nutritional disorders. AEM Educ Train 2021;5(4):e10716 [FREE Full text] [doi: 10.1002/aet2.10716] [Medline: 34966884]
49. Mncube L, Mthethwa LC. Potential ethical problems in the creation of open educational resources through virtual spaces in academia. Heliyon 2022;8(6):e09623 [FREE Full text] [doi: 10.1016/j.heliyon.2022.e09623] [Medline: 35706951]

50. Zhang X, Holbrook A, Nguyen L, Lee J, Al Qahtani S, Garcia MC, et al. Evaluation of online clinical pharmacology curriculum resources for medical students. Br J Clin Pharmacol 2019;85(11):2599-2604 [FREE Full text] [doi: 10.1111/bcp.14085] [Medline: 31385322]
51. Grabow Moore K, Ketterer A, Wheaton N, Weygandt PL, Caretta-Weyer HA, Berberian J, et al. Development, implementation, and evaluation of an open access, level-specific, core content curriculum for emergency medicine residents. J Grad Med Educ 2021;13(5):699-710 [FREE Full text] [doi: 10.4300/JGME-D-21-00067.1] [Medline: 34721800]
52. Lee M, Hamilton D, Chan TM. Cost of Free Open-Access Medical education (FOAM): an economic analysis of the top 20 FOAM sites. AEM Educ Train 2022;6(5):e10795 [FREE Full text] [doi: 10.1002/aet2.10795] [Medline: 36189455]
53. Wolbrink T, Rubin L, Burns J, Markovitz B. The top ten websites in critical care medicine education today. J Intensive Care Med 2019;34(1):3-16. [doi: 10.1177/0885066618759287] [Medline: 29519206]
54. Shappell E, Chan T, Thoma B, Trueger NS, Stuntz B, Cooney R, et al. Crowdsourced curriculum development for online medical education. Cureus 2017;9(12):e1925 [FREE Full text] [doi: 10.7759/cureus.1925] [Medline: 29464134]
55. Evans F, Krotinger A, Lilaonitkul M, Khaled HF, Pereira GA, Staffa SJ, et al. Evaluation of open access websites for anesthesia education. Anesth Analg 2022;135(6):1233-1244. [doi: 10.1213/ANE.0000000000006183] [Medline: 35983999]
56. Lo A, Shappell E, Rosenberg H, Thoma B, Ahn J, Trueger NS, et al. Four strategies to find, evaluate, and engage with online resources in emergency medicine. CJEM 2018;20(2):293-299. [doi: 10.1017/cem.2017.387] [Medline: 28893344]
57. Ting D, Boreskie P, Luckett-Gatopoulos S, Gysel L, Lanktree MB, Chan TM. Quality appraisal and assurance techniques for Free Open Access Medical Education (FOAM) resources: a rapid review. Semin Nephrol 2020;40(3):309-319. [doi: 10.1016/j.semnephrol.2020.04.011] [Medline: 32560781]
58. Marée R. Open practices and resources for collaborative digital pathology. Front Med (Lausanne) 2019;6:255 [FREE Full text] [doi: 10.3389/fmed.2019.00255] [Medline: 31799253]
59. Colmers-Gray I, Krishnan K, Chan TM, Seth Trueger N, Paddock M, Grock A, et al. The revised METRIQ score: a quality evaluation tool for online educational resources. AEM Educ Train 2019;3(4):387-392 [FREE Full text] [doi: 10.1002/aet2.10376] [Medline: 31637356]
60. Krishnan K, Thoma B, Trueger NS, Lin M, Chan TM. Gestalt assessment of online educational resources may not be sufficiently reliable and consistent. Perspect Med Educ 2017;6(2):91-98 [FREE Full text] [doi: 10.1007/s40037-017-0343-3] [Medline: 28243948]
61. O'Flynn E, Tolani M, Joos E, O'Sullivan J, Sharma D, Wren SM, et al. Advancing open-access education for the surgical team worldwide: the development and rollout of the United Nations Global Surgery Learning Hub (SURGhub). World J Surg 2025;49(7):1700-1707. [doi: 10.1002/wjs.12661] [Medline: 40542462]
62. The Open Critical Care Project. Open Critical Care. URL: https://opencriticalcare.org/ [accessed 2025-10-10]

Abbreviations

FOAM: Free Open Access Medical Education
LMIC: low- and middle-income country
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

Edited by B Lesselroth; submitted 12.08.24; peer-reviewed by T Wolbrink, SH Sung, M Bordonaro, I Araujo-Filho, N Komasawa; comments to author 17.06.25; revised version received 31.08.25; accepted 07.10.25; published 07.11.25.

Please cite as:
Ahmed AAE, Biswas A, Bempong-Ahun N, Perić I, O'Flynn EP
Barriers and Enablers to the Production of Open Access Medical Education Platforms: Scoping Review
JMIR Med Educ 2025;11:e65306
URL: https://mededu.jmir.org/2025/1/e65306
doi:10.2196/65306
PMID:

What Are the Opportunities and Challenges of Using AI in Medical Education in Vietnam?

Trung Anh Nguyen[1], MD; Thanh Binh Nguyen[1], MD, PhD; Duy Cuong Nguyen[1], MD, PhD; Anh Dung Vu[1], MD, PhD; Khanh Linh Dang[1]; Nhu Quynh Le[1]; Duy Anh Ngo[2]; Dang Kien Nguyen[1], MD, PhD; Van Thuan Hoang[1], MD, PhD; Thanh Binh Ngo[1], MD, PhD

[1]Thai Binh University of Medicine and Pharmacy, 373 Ly Bon Street, Hung Yen, Vietnam

[2]Hanoi University of Science and Technology, Ha Noi, Vietnam

Corresponding Author:
Thanh Binh Ngo, MD, PhD
Thai Binh University of Medicine and Pharmacy, 373 Ly Bon Street, Hung Yen, Vietnam

Abstract

Artificial intelligence (AI) has the potential to transform medical training through adaptive learning, immersive simulations, automated assessments, and data-driven insights, offering solutions to persistent issues such as high student-to-faculty ratios, overcrowded classrooms, and limited clinical exposure. Globally, many universities have already embedded AI literacy and competencies into undergraduate, postgraduate, and continuing education programs, while in Vietnam, the use of AI in medical education remains limited and fragmented. Most students have little formal exposure to AI, and empirical evidence on faculty or institutional readiness is scarce. Experiences from other countries, including Malaysia, Palestine, and Oman, demonstrate that incremental adoption and faculty development can facilitate cultural acceptance and curricular innovation, providing useful lessons for Vietnam. At the same time, significant barriers remain. These include inadequate infrastructure in provincial universities, low levels of AI literacy among both students and educators, underdeveloped regulatory and ethical frameworks, and resistance to pedagogical change. Cost-effectiveness and sustainability are additional concerns in a middle-income context, where upfront investments must be balanced against long-term benefits and equitable access. Advancing AI in Vietnamese medical education will therefore require a coordinated national strategy that prioritizes infrastructure, AI literacy, faculty development, quality assurance, and sustainable funding models, alongside ethical and legal safeguards. By addressing these key foundations, Vietnam can harness AI not only to modernize medical education but also to strengthen preparedness for a digitally enabled health workforce.

(JMIR Med Educ 2025;11:e77817) doi:10.2196/77817

KEYWORDS

artificial intelligence; medical education; Vietnam; digital transformation; adaptive learning; virtual simulation; curriculum innovation; educational technology

Introduction

The rapid development of artificial intelligence (AI) technologies is transforming numerous industries, including the field of education. The last few years have witnessed significant potential for AI to enhance the way knowledge is delivered, tested, and learned at every level of education, from common learning spaces to highly technical fields such as medical training. AI applications in medical education range from personalized learning systems, automated evaluations, and virtual simulated patients to complex decision-support systems that enable clinical thinking and learning for diagnosis [1-3]. These developments align with broader patterns across the health care industry, where innovations backed by AI technology are increasingly being integrated into clinical practice workflows and diagnostic systems [4,5].

Globally, several countries possessing robust health and education systems have actively incorporated AI into their medicine curricula. These cross-disciplinary programs embed AI competence and proficiency within undergraduate, postgraduate, and continuing education courses [6,7]. It is seen through scoping reviews that learning enriched by AI can potentially maximize learning efficiency, clinical preparedness, and adaptability to real-world medical problems [8,9]. A significant case is that of the Icahn School of Medicine at Mount Sinai, which in 2025 announced a full roll-out of ChatGPT Edu, a partnership with OpenAI for use in advancing education, research, and innovation for its programs [10]. All of this can be enabled only through robust infrastructure, long-term funding, and an innovative culture that supports transformational curricula and technology infusion.

In comparison, Vietnamese medical training is confronted by unique context-related issues. The pedagogical model remains

strongly traditional, and hospital-based clinical attachments and didactic lectures dominate the learning environment. Limited digital infrastructure, uneven readiness among faculty members, and the shortage of national guidelines for integrating AI further limit innovation in teaching [11,12]. In addition, the disparity between AI development and pedagogical implementation is increasing across most low- and middle-income countries, and there are fears over learning inequity and readiness for the digital health workforce [13,14]. These inequalities explain the necessity to study the use, perception, and regulation of AI in Vietnamese medical education today.

Despite the significant focus on AI within international medical education, empirical research on preparedness, acceptance, and implementation of AI in the Vietnamese context is scarce. A study using a survey among pharmacy and medical students from the south region of Vietnam found that nearly all students had no idea what AI is and what it can do for health, while almost 80% thought it would be useful for their professional development [12]. There is a striking shortage when it comes to research on faculty members, institutional decision-makers, and policy-level interactions, but preparedness evaluations for AI within Vietnam's health care system have identified issues ranging from limited sociopolitical support to inadequate information infrastructure [15]. This lack of data makes it difficult for evidence-based planning and for congruence between technological progress and reality within education (Textbox 1). In addition, although the global discourse on AI in medical education is expanding rapidly, Vietnam has received little scholarly attention despite its unique challenges of limited infrastructure, high student-to-faculty ratios, and uneven policy development. This viewpoint aims to address this gap by synthesizing existing literature on the use of AI in medical education, with a specific focus on the opportunities and challenges of integrating AI into Vietnamese medical education. This is intended for medical educators, institutional leaders, and policymakers who are navigating digital transformation in health professions training. We suggest that AI could play a key role in solving persistent issues such as overcrowded classrooms, limited clinical training opportunities, and unequal access to learning materials. However, several barriers stand in the way, including underdeveloped infrastructure, a general lack of digital and AI literacy, unclear ethical and legal frameworks, and hesitation around changing traditional teaching methods. Drawing on global developments, current research, and our own professional experience, we believe Vietnam needs a strategic, inclusive plan to integrate AI. This should focus on upgrading infrastructure, equipping both teachers and students with the necessary skills, reforming curricula, and establishing national-level guidelines. With the right approach, AI could not only bring Vietnamese medical education into the digital age but also help build a future-ready health workforce.

Textbox 1. Gaps in the literature on artificial intelligence (AI) in medical education in Vietnam.

Although artificial intelligence (AI) is gaining momentum in global medical education, current literature reveals critical gaps in the Vietnamese context that hinder evidence-based planning and risk misalignment between technological advances and educational realities.

Lack of empirical research in medical education

- Most existing studies on AI adoption focus on general higher education institutions rather than health professional training [11,13].

Absence of national data on institutional readiness

- There is limited insight into how medical schools and policy stakeholders perceive or plan for AI integration [15,16].

Low AI literacy among health care students

- A cross-sectional study found that 92.2% of students in southern Vietnam had no understanding of AI in health care, and 70.6% had never received formal instruction on the topic [12].

Fragmented and small-scale initiatives

- Although programs such as the Fulbright-Google collaboration [17] show promise, they are isolated efforts and are not embedded in national medical curricula.

The gap between clinical AI use and educational application

- While AI is being introduced in Vietnamese health care [14], there is a lack of studies on its application in teaching, assessment, or simulation-based learning in medical education.

Characteristics of Medical Education in Vietnam

The Vietnamese system for training doctors has distinctive features at the structural and organizational level that are significantly different from high-income countries. There are about 30 universities across the country that offer undergraduate and postgraduate training for health sciences. A 6-year undergraduate program is typically completed by medical students to obtain a general medical doctor degree (MD, General Doctor).

Vietnamese teaching hospitals host, apart from undergraduate trainees, numerous other groups of learners simultaneously, including residents (Resident), master's students (Postgraduate student, Master), and candidates for Specialty Level I and II (Specialist Level I Doctor, Specialist Level II Doctor), all of whom require clinical experience in the same health care facilities.

This diversity and volume of learners place significant pressure on the teaching infrastructure. Hospital clinical sites are often stretched not only by patient care demands but also by educational requirements. The student-to-patient ratio is generally low, which reduces opportunities for hands-on practice, case-based discussions, and guided clinical reasoning [18].

Moreover, many faculty members manage multiple roles, including classroom instruction, clinical supervision, and full-time medical duties, which results in fragmented teaching time and limits the capacity for educational innovation [11,12].

Opportunities for AI in Vietnamese Medical Education

Lessons from global best practices highlight how AI can be adapted to resource-constrained settings similar to Vietnam. In Malaysia, pilot programs for AI literacy in medical schools have emphasized curriculum integration at the undergraduate level while addressing cultural resistance through faculty workshops [19]. A recent study conducted in Palestine reported high levels of AI adoption among medical students, with frequent use of tools such as ChatGPT (OpenAI) and virtual simulators. Students indicated that these technologies supported their academic performance and research productivity, particularly in tasks such as literature review and data analysis. Notably, students highlighted AI's usefulness in improving time management by automating repetitive academic tasks, thereby enhancing overall efficiency [20]. A study from Oman reported that medical students demonstrated a moderate level of readiness to adopt AI as they transitioned into their clinical training years. ChatGPT emerged as the most frequently used AI tool, and notable differences in attitudes were observed between students with higher technological proficiency and those with limited digital skills. The authors emphasized that medical schools should integrate AI into their curricula to better prepare students for future clinical practice [21]. These experiences demonstrate the value of incremental implementation, regional collaboration, and leveraging domestic innovation ecosystems. For Vietnam, adapting such strategies could mean initiating AI pilots at a small number of universities, gradually expanding to provincial schools, and fostering partnerships with both local startups and global technology firms. This phased approach would ensure that AI adoption is context-sensitive, financially feasible, and culturally acceptable.

Personalized Learning Experiences

AI facilitates the shift from 1-size-fits-all instruction to personalized teaching that adapts to the unique needs of each student. AI-powered adaptive learning systems continuously evaluate students' progress using diagnostic tests and real-time performance monitoring. These systems can then dynamically adjust how content is delivered, providing remedial modules for those who need extra help with specific topics or accelerating the pace for advanced learners [22-24].

In Vietnam, where medical schools often struggle with high student-to-teacher ratios, such systems can play a vital role in improving educational equity. For example, though not yet widely adopted, AI-based platforms have demonstrated how they can deliver tailored feedback and interactive learning content for medical students [1]. In a recent survey of health care students in Vietnam, over 80% said they were willing to use AI tools to aid their studies, yet more than 90% had never actually used such technologies [12].

In addition, AI can be a valuable asset for students learning in English by offering translation, summarizing complex medical texts, or providing culturally relevant explanations. Chatbots such as ChatGPT or the Medical Pathways Language Model can act as around-the-clock learning companions, helping students understand difficult topics, review for exams, or practice clinical thinking through interactive questions-and-answers simulations [25].

Enhanced Clinical Simulations

One of the most significant applications of AI in medical training is the use of clinical simulations combined with virtual reality, augmented reality, and natural language processing. These technologies can recreate high-stakes medical scenarios such as cardiac arrest, trauma response, or neonatal resuscitation, allowing students to make diagnostic and treatment decisions in a risk-free environment [26,27].

For example, AI-powered virtual patient platforms such as SimX (XR technology) and Body Interact (Take the Wind) offer students the opportunity to repeatedly practice, experience different case scenarios, and receive real-time feedback. These tools are designed to sharpen both psychomotor and cognitive skills, which are crucial for developing clinical competency [28,29]. In Vietnam, where access to real patient cases, particularly in rural training institutions, may be limited, such simulation-based learning can help fill crucial gaps in clinical exposure.

In addition, AI that uses the power of large language models (LLMs; advanced AI systems trained on vast amounts of text data, capable of generating human-like responses in natural language) can serve as a private coach to improve medical students' soft skills. For example, the Artificial Intelligence Medical History Evaluation Instrument (the University of Arizona), an AI coaching tool developed by the Arizona Simulation Technology and Education Center, has been designed to enhance the communication skills of medical students. This system analyzes data from medical interviews conducted between students and patients, offering comprehensive evaluations of their interpersonal communication and medical competency skills. Assessment and feedback are provided according to guidelines from various organizations, including the Liaison Committee on Medical Education, the American College of Surgeons, and the World Health Organization. Owing to its automated nature, this technology provides significant opportunities for personalized coaching and can transform medical education in terms of time efficiency and cost-effectiveness. This technology may also be applicable in terms of scalability and broader access, which can be adapted to various medical education settings, including those in low- and middle-income countries.

AI also makes it possible to adjust the complexity of simulations based on a student's performance. For instance, a student who successfully manages a simulated sepsis case might be presented with a more challenging case next time, one involving comorbid conditions or rare complications, following a mastery learning approach supported by AI [30].

Efficient Administrative Processes

AI's role in medical education goes beyond classroom instruction—it can also help resolve administrative challenges that are common in Vietnamese medical schools. For example, machine learning–based scheduling tools can streamline the creation of rotation timetables, classroom assignments, and faculty schedules. Software such as TimeTabler (October ReSolutions Limited) uses smart algorithms to handle logistical tasks that would otherwise demand extensive manual effort [31]. International experiences show that such systems can reduce scheduling conflicts dramatically and allow faculty to dedicate more time to teaching and mentoring.

AI also plays a role in assessment. It can be used to grade objective structured clinical examinations, multiple-choice questions (MCQs), and even written answers using natural language processing. Automated grading not only delivers faster feedback but also helps reduce human bias. Research shows that LLMs can support educators by generating standardized, objective structured clinical examination cases and evaluation rubrics. These AI tools can even benchmark student performance against expert evaluations [32,33]. In Vietnam, where faculty members often juggle teaching responsibilities across multiple institutions, such automation could reduce workload and improve consistency in student assessment.

Language processing capabilities allow AI tools to use LLMs to address issues related to medical education assessment. One particular domain that could significantly benefit from the application of LLM-based AI tools is the creation of MCQs for medical examinations. A multinational prospective study assessed the ability of ChatGPT to generate high-quality MCQs compared to university professors using standard medical textbooks [34]. The study showed that for the task of generating 50 MCQs for the graduate medical exam, ChatGPT took approximately 20 minutes, while professors took over 200 minutes. The analyses showed no significant differences in overall quality between MCQs generated by ChatGPT and by humans. This finding suggests that ChatGPT can produce MCQs of comparable quality to those generated by humans in a significantly shorter amount of time. It is important to note that since the publication of this study, more sophisticated LLM models with advanced reasoning capacity, including GPT-4 or GPT-5 (advanced LLMs developed by OpenAI), have been released by OpenAI, potentially outperforming the older models in such language processing tasks. Given that these models can be accessed via the application programming interface, tools that allow different software systems to communicate with each other—for example, enabling an educational platform to integrate AI functions—can automate this process without the need for manual input. Several tools using this approach have been developed, including Questgen (QuestgenAI Inc) or NoteGPT (NoteGPT AI Inc).

In Vietnam, where faculty members often juggle teaching responsibilities across various institutions or departments, automation like this can significantly lighten their load, giving them more time to focus on teaching and mentoring students.

In addition, predictive analytics powered by AI can help universities identify students who are at risk of failing based on their behavior and academic performance. This allows for timely intervention and support. Such tools are especially helpful in the competency-based education model that is gaining ground in Vietnamese medical training. International evidence supports the use of machine learning techniques such as k-nearest neighbors and exam performance predictors—to flag students who may need extra help [35,36]. In Vietnam, where competency-based education is being introduced, these systems could be integrated into student tracking to ensure timely support and reduce attrition. By combining automation with predictive tools, AI has the potential to not only ease administrative burdens but also strengthen educational outcomes through early intervention.

Data-Driven Decision-Making

AI can analyze large sets of educational data to support quality improvements in medical education. For instance, when integrated into learning management systems, AI can monitor things such as student engagement, quiz scores, and participation in discussion forums. This type of data can help educators identify which parts of the curriculum are not working well, pinpointing specific modules or concepts that consistently challenge students [37]. In other countries, AI-powered dashboards have been used to redesign underperforming modules, improve exam blueprints, and assess the impact of new teaching methods, demonstrating clear benefits for curriculum quality assurance [38,39].

At the broader institutional level, AI can aid in accreditation and quality assurance by generating detailed dashboards that track metrics such as student learning outcomes, faculty performance, and the success of graduates. These insights can help academic leaders—such as deans and curriculum planners—make well-informed decisions about curriculum updates, faculty development, and student support systems [40,41].

In Vietnam, where a national competency-based medical education framework is in development, AI could play a crucial role. It can help map student progress against national competency standards and produce standardized performance reports across various medical schools [12].

An additional opportunity emerges from Vietnam's health care digitization initiatives. With all hospitals nationwide mandated to implement electronic medical records by September 30, 2025 [42], medical schools will soon be connected to richer clinical data environments. This creates the possibility of linking students' performance in clinical rotations with electronic medical records-based outcomes, allowing education to be evaluated alongside real-world patient care. Such integration

could establish a feedback loop that improves both medical training and health care delivery.

Finally, AI-powered decision support systems can also assist policymakers by aggregating anonymized student and institutional data to forecast workforce needs. Predictive models could estimate shortages in specific specialties or geographic areas, ensuring that Vietnam's medical education system aligns more closely with national health priorities. In a context where disparities in health care access remain significant, the use of data-driven decision-making could guide not only academic reform but also broader strategies for building an equitable health workforce.

Challenges in Implementing AI in Vietnamese Medical Education

Table 1 summarizes the contrast between international adoption of AI in medical education and the Vietnamese context, highlighting the structural, educational, and policy gaps that shape the challenges discussed (Table 1).

Table . Comparison of artificial intelligence (AI) integration in medical education: international versus Vietnam.

Dimension	International context	Vietnamese context
Curriculum integration	Many universities embed AI[a] literacy and competencies into undergraduate, postgraduate, and continuing medical education programs [43].	No national AI curriculum, minimal exposure, and most students report no formal training in AI in health care [12].
Infrastructure	Robust digital platforms, simulation centers, and cloud-based tools are widely available, with strong institutional investment [29,43].	Limited infrastructure, particularly in provincial schools and outdated labs, and poor internet continue to hinder AI adoption [11,16].
Faculty readiness	Faculty development programs in AI pedagogy and interdisciplinary teaching are increasingly common [6].	Faculty often lack AI knowledge, have limited training opportunities, and have a heavy clinical workload that further limits innovation [12,14].
Policy and regulation	Clearer data protection and AI ethics frameworks in regions such as the European Union and the United States [44].	A national AI strategy exists, but there are no specific laws for AI in education or student data protection.
Student exposure	Widespread use of AI tools for adaptive learning, automated assessment, and clinical simulation [1,22-25].	Over 90% of health care students have never used AI in their studies, despite high interest [12].
Cultural acceptance	Increasingly normalized as part of digital transformation in education [45].	Skepticism and resistance among faculty and uneven acceptance across institutions.

[a]AI: artificial intelligence.

Limited Awareness and Understanding

A major obstacle to integrating AI effectively into Vietnamese medical education is the widespread lack of knowledge and basic understanding among both students and educators [12]. While discussions about AI in health care are growing globally, many medical trainees in Vietnam are still unfamiliar with its core concepts and practical applications. According to a 2023 cross-sectional study, 92.2% of health care students in southern Vietnam reported having no understanding of how AI works in health care, and 70.6% had never received any formal education on AI-related topics [12]. These figures highlight a significant educational gap and point to an urgent need to introduce AI literacy into the medical curriculum. The urgent call for this integration has also been sparked in Malaysia, a country located in the same geographic area as Vietnam, where unreadiness for AI has been reported widely among medical students [46].

Moreover, most medical schools in Vietnam still rely heavily on traditional, lecture-based teaching methods and have limited integration of digital tools. Without structured exposure to AI, whether through dedicated coursework, hands-on workshops, or interdisciplinary collaboration with departments such as data science, future health care professionals may lack the skills needed to use AI tools responsibly or interpret AI-generated clinical data. This lack of preparation not only hampers the adoption of AI in training but also raises long-term concerns about the quality of health care delivery as AI becomes increasingly embedded in global medical practice.

Infrastructural Constraints

The effective integration of AI in medical education relies heavily on a strong technological foundation. This includes dependable high-speed internet, access to cloud computing, up-to-date hardware, and specialized software. Unfortunately, many medical schools in Vietnam, especially those in rural areas or linked to lower-tier universities, face significant infrastructure challenges. Common issues such as limited budgets, outdated computer labs, and a lack of technical support continue to obstruct the adoption of AI-powered learning tools [11,15,16].

As of now, no medical university in Vietnam has officially developed or launched an AI-based curriculum tailored to medical training. The digital divide between urban and rural institutions further exacerbates the issue, highlighting both a lack of preparedness and a deep educational disparity.

Globally, research shows that while AI is making strides in medical education in well-funded regions, many low- and middle-income countries lag due to insufficient infrastructure, a shortage of trained faculty, and limited institutional readiness

[2,47]. These international gaps are mirrored within Vietnam, where centrally located universities may have the means and partnerships to explore AI, whereas many provincial medical schools simply lack the resources and expertise to do so.

Without unified national investment and clear policy direction to support digital transformation across all levels of medical education, AI advantages risk being confined to a few elite institutions, ultimately undermining efforts to build a fair and forward-looking health care workforce for the entire country.

Quality Assurance, Ethical, Legal, and Policy Challenges

Ensuring the quality, validity, and reliability of AI-assisted assessment is critical in medical education, where evaluation outcomes have direct implications for patient safety and professional competency [48]. Although AI shows promise in automating grading and generating exam content, risks remain regarding algorithmic bias, variability, and transparency [49]. To safeguard assessment integrity, AI-generated items and evaluations should undergo external expert review and be pilot-tested before use in high-stakes examinations. Institutions should adopt national or regional standards for AI-assisted evaluation, ensuring consistency across medical schools. Moreover, regular audits of AI performance against human benchmarks are needed to detect errors or unintended biases. Establishing clear oversight mechanisms will be essential for building trust in AI-enhanced assessments and protecting both educational and professional standards [50].

The use of AI in education also brings forward a range of ethical challenges, especially around data privacy, informed consent, fairness in algorithms, and transparency. Many AI systems rely on collecting and analyzing extensive data such as student performance, learning habits, and, in some cases, biometric or voice information. AI-powered proctoring tools used during online exams can record facial movements, keystrokes, and background activity. While these tools are meant to prevent cheating, universities have faced student protests and lawsuits due to concerns about surveillance and data misuse. Without strong legal or institutional safeguards, there is a real risk of similar issues in Vietnam, where institutional data governance remains underdeveloped and student privacy could easily be compromised [51].

A significant concern is algorithmic bias. If an AI system is trained on unrepresentative data or lacks clear decision-making criteria, it may unfairly evaluate students. This is especially troubling in high-pressure fields such as medical education, where skewed assessments could influence a student's confidence, academic progress, or even future career opportunities [2,52]. The lack of transparency in how these algorithms operate makes it even harder to hold them accountable.

As of 2024, Vietnam does not yet have a solid legal framework to govern the use of AI in higher education; it is important to situate this gap within the country's broader digital transformation and health policy agenda. In 2020, the government launched the National Digital Transformation Program to 2025 with orientation to 2030, which identifies health care and education as 2 priority sectors for digital innovation [53]. The program calls for the adoption of digital platforms, cloud-based services, and data-driven management tools in universities and hospitals, laying a foundation that could facilitate the integration of AI into medical training.

Similarly, the Ministry of Health has advanced health care digitization initiatives under the "Smart Health" and National Health Digital Transformation plans [54], focusing on electronic medical records, telemedicine, and health information systems. According to recent regulations, all hospitals nationwide are mandated to implement electronic medical records by September 30, 2025 [42]. While these initiatives primarily target clinical service delivery, they create data ecosystems and digital infrastructure that could be leveraged for AI-driven medical education, especially in areas such as simulation, case-based training, and competency tracking.

Although early-stage policies have been suggested under the national AI development strategy, no specific laws are in place to address data protection, algorithmic responsibility, or the ethical role of AI in classrooms [16]. This gap raises pressing concerns about how institutions and regulators should manage these technologies responsibly.

Finally, while AI can enhance efficiency and scale learning, it risks sidelining the personal, human aspects of medical training. Skills such as empathy, ethical judgment, and effective communication are vital in health care, and they develop best through real human interaction and mentorship. AI may support education, but it cannot replace the essential human touch in shaping compassionate medical professionals [1,55].

Cultural and Acceptance Issues

One of the key obstacles to integrating AI in Vietnam's medical education system is the cultural and institutional resistance to change. Many faculty members and administrators, having been trained in conventional teaching methods, may approach AI with skepticism or even see it as a threat to their professional roles. Concerns about job security, diminished academic authority, or simply a lack of familiarity with emerging technologies can fuel this reluctance [56,57].

These attitudes are deeply rooted in Vietnamese pedagogical traditions, which remain strongly influenced by Confucian values emphasizing hierarchy, respect for authority, and teacher-centered learning. The traditional lecture-based approach positions the teacher as the primary source of knowledge, while students are expected to learn through memorization and obedience. In this context, the idea of AI systems providing personalized tutoring or automated feedback may be perceived as undermining the authority of faculty or disrupting established classroom dynamics.

Similarly, the relationship between faculty and students in Vietnamese medical schools often emphasizes formality and deference, which may affect how students engage with AI-driven tools. While younger generations may be more open to experimentation, students might hesitate to adopt AI-based tutoring systems if they sense disapproval from faculty or if

doing so appears to challenge established norms of learning. Conversely, if faculty members actively endorse and model the use of AI, acceptance among students is likely to grow more rapidly.

Moreover, Vietnamese academic culture tends to value collective conformity over individual experimentation. Students often prioritize standardized examination performance, and learning strategies are adapted toward achieving success in high-stakes tests. Because AI tools often encourage exploratory, self-paced, and competency-based learning, there may be a mismatch between the innovative potential of AI and the exam-driven educational ethos. Unless AI is aligned with assessment systems and institutional incentives, it risks being perceived as peripheral or even irrelevant.

Addressing these cultural barriers requires meaningful investment in faculty development. Training programs that introduce educators to AI tools, illustrate their value in teaching, and offer hands-on learning opportunities can help build trust and competence. Promising signs of progress can be seen in local efforts such as Fulbright University's AI education grant and VinBrain's nationwide campaigns promoting AI in health care, both of which highlight the power of sustained institutional backing in shifting academic mindsets [17,58].

Leadership also plays a crucial role in fostering a culture of innovation. Universities need to actively support digital experimentation, offer incentives for early adopters, and ensure that AI initiatives align with broader goals such as competency-based education and health care system reform. Without genuine buy-in from faculty, there is a risk that AI will be viewed as a top-down mandate, poorly integrated into everyday teaching, and ultimately ineffective [2,24,59].

Cost-Effectiveness and Sustainability

Cost-effectiveness and sustainability are central concerns for Vietnam as a middle-income country [60-62]. While AI technologies promise long-term benefits, including reduced faculty workload, standardized assessments, and improved student retention, their implementation requires upfront investment in infrastructure, training, and software. To ensure sustainability, policymakers and universities must consider the return on investment and adopt funding models that extend beyond 1-time grants. Public-private partnerships, collaborations with technology companies, and integration with national digital transformation and health care digitization programs represent potential pathways for sustainable financing [62]. Without deliberate strategies, there is a risk that only urban and better-resourced institutions will benefit, thereby deepening existing inequities with respect to provincial universities. Embedding cost-effectiveness evaluations into policy design and monitoring will be essential to ensure scalability and equitable distribution of AI benefits across the entire medical education system.

Conclusion and Implications

Integrating AI into the Vietnamese medical education system offers a powerful opportunity to modernize how future health care professionals are trained. AI can individualize learning, simulate complex clinical scenarios, automate assessment, and guide responsive curriculum reform—advantages that are particularly relevant in Vietnam, where faculty shortages and unequal resource distribution persist—yet realizing these benefits requires more than enthusiasm. Low levels of AI literacy, uneven technological capacity, concerns about ethics and data protection, and resistance within academic culture remain significant barriers. The recommendations outlined in Textbox 2 provide a practical roadmap for short-, medium-, and long-term action. With deliberate planning, sustained investment, and national coordination, Vietnam can ensure that AI integration is not only feasible but also equitable and sustainable, ultimately strengthening the country's preparedness for a digitally enabled health workforce.

Textbox 2. Recommendations for the integration of artificial intelligence (AI) in Vietnamese medical education.

Short-term

- Introduce artificial intelligence (AI) literacy modules in undergraduate and postgraduate curricula, covering:
- Fundamentals of machine learning and large language models.
- Principles of data ethics, bias awareness, and privacy.
- Clinical applications of AI in diagnostics, simulation, and decision support.
- Require medical universities to conduct faculty development workshops on AI tools for teaching, simulation, and assessment.
- Pilot the use of AI-assisted exam generation and automated grading (eg, multiple-choice questions and objective structured clinical examination checklists) at selected universities.
- The responsible parties include individual universities, with oversight provided by the Ministry of Education and Training (MOET) and the Ministry of Health (MOH).

Medium-term

- Establish minimum infrastructure standards, including:
- Reliable high-speed internet access.
- Cloud-based learning management systems.
- Secure institutional data storage and access to national health data platforms.
- Develop national ethical guidelines for AI in education, coordinated by MOET and MOH, with input from medical universities, professional associations, and legal experts.
- Expand pilot projects to provincial universities, accompanied by rigorous outcome evaluations.
- The responsible parties include MOET, MOH, national medical councils, and professional associations.

Long-term

- Integrate AI literacy as a core requirement across all medical training programs nationwide.
- Mandate external quality assurance mechanisms for AI-generated educational content and assessments to ensure validity and reliability.
- Create a sustainable funding framework through public-private partnerships and government-supported digital transformation budgets.
- Align AI-based educational analytics with national health workforce planning systems, ensuring that training outputs meet the needs of the health care system.
- The responsible parties include MOET, MOH, the Ministry of Science and Technology, and national accreditation bodies.

The integration of AI has broad implications for policy, practice, and research. Policymakers must establish clear ethical and regulatory frameworks and invest in robust digital infrastructure to prevent widening inequities between urban and provincial institutions. For educators and universities, AI provides practical solutions to faculty shortages, overcrowded classrooms, and limited clinical exposure, yet these benefits can only be realized if faculty development and support systems are prioritized. For researchers, there is an urgent need for Vietnam-specific studies evaluating the effectiveness, acceptability, and long-term outcomes of AI-based educational tools. Interdisciplinary collaborations between medical and technical institutions will be crucial in generating locally relevant evidence.

Drawing from the recommendations outlined in this paper, Vietnam should prioritize: (1) including AI literacy modules in undergraduate and postgraduate curricula, (2) investing in IT infrastructure and cloud-based educational platforms, particularly in provincial schools, (3) conducting faculty development programs focused on digital teaching tools, (4) establishing ethical guidelines for AI use in education to protect student data and ensure fairness, and (5) promoting interdisciplinary collaboration between medical and technology faculties.

By addressing these policy, practice, and research priorities, Vietnam can ensure that AI adoption not only modernizes medical education but also contributes to a resilient, equitable, and future-ready health workforce.

Acknowledgments

During the preparation of this work, the authors used ChatGPT-4o to check and improve grammar and clarity during manuscript development. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding

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

Authors' Contributions

TAN contributed to the conceptualization and methodology of the study, conducted the investigation, performed validation, and prepared the original draft of the manuscript. TBN contributed to conceptualization, methodology, investigation, and validation, in addition to participating in data collection. TBN also contributed to writing through original drafting, review, and editing of the manuscript. DCN, ADV, KLD, and NQL contributed to data collection and validation and assisted with the review and editing of the manuscript. DAN participated in data collection, validation, and manuscript review and editing and provided supervision throughout the project. DKN and VTH contributed to conceptualization, methodology, investigation, validation, and preparation of the original manuscript draft.

Conflicts of Interest

None declared.

References

1. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. JMIR Med Educ 2019 Jun 15;5(1):e13930. [doi: 10.2196/13930] [Medline: 31199295]
2. Tozsin A, Ucmak H, Soyturk S, et al. The role of artificial intelligence in medical education: a systematic review. Surg Innov 2024 Aug;31(4):415-423. [doi: 10.1177/15533506241248239] [Medline: 38632898]
3. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019 Dec 3;5(2):e16048. [doi: 10.2196/16048] [Medline: 31793895]
4. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023 Sep 22;23(1):689. [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
5. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019 Jan;25(1):30-36. [doi: 10.1038/s41591-018-0307-0] [Medline: 30617336]
6. Charow R, Jeyakumar T, Younus S, et al. Artificial intelligence education programs for health care professionals: scoping review. JMIR Med Educ 2021 Dec 13;7(4):e31043. [doi: 10.2196/31043] [Medline: 34898458]
7. Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Med Teach 2024 Apr 2;46(4):446-470. [doi: 10.1080/0142159X.2024.2314198]
8. Feigerlova E, Hani H, Hothersall-Davies E. A systematic review of the impact of artificial intelligence on educational outcomes in health professions education. BMC Med Educ 2025 Jan 27;25(1):129. [doi: 10.1186/s12909-025-06719-5] [Medline: 39871336]
9. Shaw K, Henning MA, Webster CS. Artificial intelligence in medical education: a scoping review of the evidence for efficacy and future directions. MedSciEduc 2025;35(3):1803-1816. [doi: 10.1007/s40670-025-02373-0]
10. Icahn School of Medicine at Mount Sinai expands AI innovation with OpenAI's ChatGPT Edu rollout. Mount Sinai. URL: https://www.mountsinai.org/about/newsroom/2025/icahn-school-of-medicine-at-mount-sinai-expands-ai-innovation-with-openais-chatgpt-edu-rollout [accessed 2025-05-20]
11. Loan TTT, Th y NTH. Vietnamese University lecturers apply AI in teaching. a case study in Thu Dau Mot University. ejtas 2024;2(6):643-650 [FREE Full text] [doi: 10.59324/ejtas.2024.2(6).57]
12. Truong NM, Vo TQ, Tran HTB, Nguyen HT, Pham VNH. Healthcare students' knowledge, attitudes, and perspectives toward artificial intelligence in the southern Vietnam. Heliyon 2023 Dec;9(12):e22653. [doi: 10.1016/j.heliyon.2023.e22653] [Medline: 38107295]
13. Quy VK, Thanh BT, Chehri A, Linh DM, Tuan DA. AI and digital transformation in higher education: vision and approach of a specific university in Vietnam. Sustainability 2023;15(14):11093. [doi: 10.3390/su151411093]
14. Doan Thu TN, Nguyen QK, Taylor-Robinson AW. Healthcare in Vietnam: harnessing artificial intelligence and robotics to improve patient care outcomes. Cureus 2023 Sep;15(9):e45006. [doi: 10.7759/cureus.45006] [Medline: 37829937]
15. Vuong QH, Ho MT, Vuong TT, et al. Artificial intelligence vs. natural stupidity: evaluating AI readiness for the Vietnamese medical information system. J Clin Med 2019 Feb 1;8(2):168. [doi: 10.3390/jcm8020168] [Medline: 30717268]
16. Nguyen TL, Nguyen VP, Dang TVD. Critical factors affecting the adoption of artificial intelligence: an empirical study in Vietnam. J Asian Finance Econ Bus 2022;9:225-237 [FREE Full text] [doi: 10.13106/jafeb.2022.vol9.no5.0225]

17. Thao T. Fulbright University Vietnam receives US1.5 million grant from Google to advance AI research and education in Vietnam. Fulbright University Vietnam. 2024. URL: https://fulbright.edu.vn/fulbright-university-vietnam-receives-us1-5-million-grant-from-google-to-advance-ai-research-and-education-in-vietnam [accessed 2025-05-14]
18. Densen P. Challenges and opportunities facing medical education. Trans Am Clin Climatol Assoc 2011;122:48-58. [Medline: 21686208]
19. Mat Yusoff S, Mohamad Marzaini AF, Hao L, Zainuddin Z, Basal MH. Understanding the role of AI in Malaysian higher education curricula: an analysis of student perceptions. Discov Computing 2025;28(1):62. [doi: 10.1007/s10791-025-09567-5]
20. Yousef M, Deeb S, Alhashlamon K. AI usage among medical students in Palestine: a cross-sectional study and demonstration of AI-assisted research workflows. BMC Med Educ 2025 May 12;25(1):693. [doi: 10.1186/s12909-025-07272-x] [Medline: 40355851]
21. AlZaabi A, Masters K. Assessing medical students' readiness for artificial intelligence after pre-clinical training. BMC Med Educ 2025 Jun 2;25(1):824. [doi: 10.1186/s12909-025-07008-x] [Medline: 40457325]
22. Das S, Mutsuddi I, Ray N. Advancing Adaptive Education: Technological Innovations for Disability Support: IGI Global Scientific Publishing; 2025. [doi: 10.4018/979-8-3693-8227-1.ch002]
23. du Plooy E, Casteleijn D, Franzsen D. Personalized adaptive learning in higher education: a scoping review of key characteristics and impact on academic performance and engagement. Heliyon 2024 Nov 15;10(21):e39630. [doi: 10.1016/j.heliyon.2024.e39630] [Medline: 39524879]
24. Merino-Campos C. The impact of artificial intelligence on personalized learning in higher education: a systematic review. Trends High Educ 2025;4(2):17. [doi: 10.3390/higheredu4020017]
25. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. Nature New Biol 2023 Apr;616(7956):259-265. [doi: 10.1038/s41586-023-05881-4] [Medline: 37045921]
26. Elendu C, Amaechi DC, Okatta AU, et al. The impact of simulation-based training in medical education: a review. Medicine (Baltimore) 2024 Jul 5;103(27):e38813. [doi: 10.1097/MD.0000000000038813] [Medline: 38968472]
27. Tene T, Vique López DF, Valverde Aguirre PE, Orna Puente LM, Vacacela Gomez C. Virtual reality and augmented reality in medical education: an umbrella review. Front Digit Health 2024;6:1365345. [doi: 10.3389/fdgth.2024.1365345] [Medline: 38550715]
28. Kononowicz AA, Zary N, Edelbring S, Corral J, Hege I. Virtual patients--what are we talking about? A framework to classify the meanings of the term in healthcare education. BMC Med Educ 2015 Feb 1;15:11. [doi: 10.1186/s12909-015-0296-3] [Medline: 25638167]
29. Kyaw BM, Saxena N, Posadzki P, et al. Virtual reality for health professions education: systematic review and meta-analysis by the Digital Health Education Collaboration. J Med Internet Res 2019 Jan 22;21(1):e12959. [doi: 10.2196/12959] [Medline: 30668519]
30. Basu K, Sinha R, Ong A, Basu T. Artificial intelligence: how is it changing medical sciences and its future? Indian J Dermatol 2020;65(5):365-370. [doi: 10.4103/ijd.IJD_421_20] [Medline: 33165420]
31. Nwile CB, Edo BL. Artificial intelligence and robotic tools for effective educational management and administration in the state-owned universities in Rivers State, Nigeria. FNAS-JMSE 2023;4(1):28-36 [FREE Full text]
32. Misra SM, Suresh S. Artificial intelligence and objective structured clinical examinations: using ChatGPT to revolutionize clinical skills assessment in medical education. J Med Educ Curric Dev 2024;11:23821205241263475. [doi: 10.1177/23821205241263475] [Medline: 39070287]
33. Geathers J, Hicke Y, Chan C, Rajashekar N, Sewell J, Cornes S, et al. Benchmarking generative AI for scoring medical student interviews in Objective Structured Clinical Examinations (OSCES). arXiv. Preprint posted online on Jul 22, 2025 URL: https://arxiv.org/abs/2501.13957 [accessed 2025-11-20] [doi: 10.1007/978-3-031-98420-4_17]
34. Cheung BHH, Lau GKK, Wong GTC, Lee EYP, Kulkarni D, Seow CS, et al. ChatGPT versus human in generating medical graduate exam multiple choice questions—a multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). PLOS One 2023;18(8):e0290691. [doi: 10.1371/journal.pone.0290691] [Medline: 37643186]
35. Kumar A, DiJohnson T, Edwards RA, Walker L. The application of adaptive minimum match k-nearest neighbors to identify at-risk students in health professions education. J Physician Assist Educ 2023 Sep 1;34(3):171-177. [doi: 10.1097/JPA.0000000000000513] [Medline: 37548617]
36. Shioiri T, Nakashima M, Tsunekawa K. When can we identify the students at risk of failure in the National Medical Licensure Examination in Japan using the predictive pass rate? BMC Med Educ 2024 Aug 27;24(1):930. [doi: 10.1186/s12909-024-05948-4] [Medline: 39192215]
37. Kolachalama VB, Garg PS. Machine learning and medical education. npj Digital Med 2018;1(1):1-3. [doi: 10.1038/s41746-018-0061-1]
38. AI-assisted item grouping and test blueprint development. The e-Assessment Association. 2024. URL: https://www.e-assessment.com/news/ai-assisted-item-grouping-and-test-blueprint-development [accessed 2025-09-12]
39. Ahmed A, Kerr E, O'Malley A. Quality assurance and validity of AI-generated single best answer questions. BMC Med Educ 2025 Feb 25;25(1):300. [doi: 10.1186/s12909-025-06881-w] [Medline: 40001164]

40. Meskó B, Hetényi G, Győrffy Z. Will artificial intelligence solve the human resource crisis in healthcare? BMC Health Serv Res 2018 Jul 13;18(1):545. [doi: 10.1186/s12913-018-3359-4] [Medline: 30001717]
41. Wartman SA, Combs CD. Medical education must move from the information age to the age of artificial intelligence. Acad Med 2018 Aug;93(8):1107-1109. [doi: 10.1097/ACM.0000000000002044] [Medline: 29095704]
42. The Ministry of Health requires the completion of the implementation of electronic medical records before September 30, 2025. Vietnam.vn. 2024. URL: https://www.vietnam.vn/en/bo-y-te-yeu-cau-hoan-tat-trien-khai-ho-so-benh-an-dien-tu-truoc-30-9-2025 [accessed 2025-09-12]
43. Rincón EHH, Jimenez D, Aguilar LAC, Flórez JMP, Tapia Á, Peñuela CLJ. Mapping the use of artificial intelligence in medical education: a scoping review. BMC Med Educ 2025 Apr 12;25(1):526. [doi: 10.1186/s12909-025-07089-8] [Medline: 40221725]
44. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell 2019;1(9):389-399. [doi: 10.1038/s42256-019-0088-2]
45. Zawacki-Richter O, Marín VI, Bond M, Gouverneur F. Systematic review of research on artificial intelligence applications in higher education – where are the educators? Int J Educ Technol High Educ 2019 Dec;16(1):39. [doi: 10.1186/s41239-019-0171-0]
46. Ong QC, Ang CS, Lai NM, Neves AL, Car J. Dearth of digital health education: the need for an accelerated medical curriculum reform in Malaysia. Lancet Reg Health West Pac 2025 Feb;55(101476):101476. [doi: 10.1016/j.lanwpc.2025.101476] [Medline: 39902151]
47. Busch F, Hoffmann L, Truhn D, et al. Global cross-sectional student survey on AI in medical, dental, and veterinary education and practice at 192 faculties. BMC Med Educ 2024 Sep 28;24(1):1066. [doi: 10.1186/s12909-024-06035-4] [Medline: 39342231]
48. Mir MM, Mir GM, Raina NT, et al. Application of artificial intelligence in medical education: current scenario and future perspectives. J Adv Med Educ Prof 2023 Jul;11(3):133-140 [FREE Full text] [doi: 10.30476/JAMP.2023.98655.1803] [Medline: 37469385]
49. Hirosawa T, Yokose M, Sakamoto T, et al. Utility of generative artificial intelligence for Japanese medical interview training: randomized crossover pilot study. JMIR Med Educ 2025 Aug 1;11:e77332. [doi: 10.2196/77332] [Medline: 40749190]
50. Chen HL, Lee CW, Chang CW, Chiu YC, Hung TY. Evaluating tailored learning experiences in emergency residency training through a comparative analysis of mobile-based programs versus paper- and web-based approaches: feasibility cross-sectional questionnaire study. JMIR Med Educ 2025 Jul 24;11:e57216. [doi: 10.2196/57216] [Medline: 40705816]
51. Regan PM, Jesse J. Ethical challenges of edtech, big data and personalized learning: twenty-first century student sorting and tracking. Ethics Inf Technol 2019 Sep;21(3):167-179. [doi: 10.1007/s10676-018-9492-2]
52. Holstein K, Wortman Vaughan J, Daumé H, Dudik M, Wallach H. Improving fairness in machine learning systems: what do industry practitioners need? 2019 Presented at: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems; May 4-9, 2019; Glasgow, UK p. 1-16. [doi: 10.1145/3290605.3300830]
53. Decision no.749/QD-ttg 2020 national digital transformation program through 2025. LuatVietnam. URL: https://english.luatvietnam.vn/decision-no-749-qd-ttg-on-approving-the-national-digital-transformation-program-until-2025-with-a-vision-184241-doc1.html [accessed 2025-09-12]
54. Tran DM, Thwaites CL, Van Nuil JI, et al. Digital health policy and programs for hospital care in Vietnam: scoping review. J Med Internet Res 2022 Feb 9;24(2):e32392. [doi: 10.2196/32392] [Medline: 35138264]
55. Masters K, Ellaway R. e-Learning in medical education Guide 32 Part 2: technology, management and design. Med Teach 2008 Jun;30(5):474-489. [doi: 10.1080/01421590802108349] [Medline: 18576186]
56. Hoffman J, Wenke R, Angus RL, Shinners L, Richards B, Hattingh L. Overcoming barriers and enabling artificial intelligence adoption in allied health clinical practice: a qualitative study. Digit Health 2025;11:20552076241311144. [doi: 10.1177/20552076241311144] [Medline: 39906878]
57. Rosenheck M. Chapter 3 - AI in medical education: challenges and opportunities. In: Digit Health: Academic Press; 2025:27-40. [doi: 10.1016/B978-0-443-23901-4.00003-9]
58. "AI's superior speed & learning capability make it a natural fit for healthcare tasks demanding precision & early detection". BioSpectrum Asia Edition. URL: https://www.biospectrumasia.com/opinion/94/23657/ais-superior-speed-learning-capability-make-it-a-natural-fit-for-healthcare-tasks-demanding-precision-early-detection.html [accessed 2025-05-14]
59. Akinwalere SN, Ivanov V. Artificial intelligence in higher education: challenges and opportunities. Border Crossing 2022;12(1):1-15. [doi: 10.33182/bc.v12i1.2015]
60. Ferk Savec V, Jedrinović S. The role of AI implementation in higher education in achieving the sustainable development goals: a case study from Slovenia. Sustainability 2025;17(1):183. [doi: 10.3390/su17010183]
61. Ciecierski-Holmes T, Singh R, Axt M, Brenner S, Barteit S. Artificial intelligence for strengthening healthcare systems in low- and middle-income countries: a systematic scoping review. NPJ Digit Med 2022 Oct 28;5(1):162. [doi: 10.1038/s41746-022-00700-y] [Medline: 36307479]

62. El Arab RA, Al Moosa OA. Systematic review of cost effectiveness and budget impact of artificial intelligence in healthcare. NPJ Digit Med 2025 Aug 26;8(1):548. [doi: 10.1038/s41746-025-01722-y] [Medline: 40858882]

Abbreviations

AI: artificial intelligence
LLM: large language model
MCQ: multiple-choice question

Edited by T Leung; submitted 20.05.25; peer-reviewed by BT George, C Anyaegbuna, FA Baloch; revised version received 15.09.25; accepted 28.10.25; published 02.12.25.

Please cite as:
Nguyen TA, Nguyen TB, Nguyen DC, Vu AD, Dang KL, Le NQ, Ngo DA, Nguyen DK, Hoang VT, Ngo TB
What Are the Opportunities and Challenges of Using AI in Medical Education in Vietnam?
JMIR Med Educ 2025;11:e77817
URL: https://mededu.jmir.org/2025/1/e77817
doi:10.2196/77817

The Need for Health Care Innovation Training in Medical Education

Lily Zhu[1], BS; Jeffrey Khong[1*], BS; Oren Wei[1*], MSE; Katherine C Chretien[1], MD; Youseph Yazdi[2], MBA, PhD

[1]School of Medicine, Johns Hopkins University, 733 N Broadway, Baltimore, MD, United States

[2]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, United States

[*]these authors contributed equally

Corresponding Author:
Lily Zhu, BS
School of Medicine, Johns Hopkins University, 733 N Broadway, Baltimore, MD, United States

Abstract

The rapid transformation of the health care landscape requires physicians to not only be skilled clinically but also navigate and lead a highly dynamic, innovation-driven environment. This also provides an avenue for physicians to significantly enhance their ability to help their patients, through participation in health innovation projects. Despite this growing need and opportunity, few medical schools provide formal training in innovation and entrepreneurship (I&E). In this perspective, we examine the need for I&E education in medical curricula by exploring student interest, effective program models, and implementation strategies. To better understand medical student interest in innovation and willingness to participate in I&E programs during medical school, we surveyed 480 medical students at our institution, the Johns Hopkins University School of Medicine, and received 90 responses with a 19% response rate. We observed a strong interest in health care I&E, with 97% (87/90) of respondents valuing knowledge or experience in I&E and 63% (56/90) expressing intent to incorporate I&E into their careers. To assess the real-world impact of I&E education on medical professionals, we surveyed 12 alumni of the Johns Hopkins Center for Bioengineering Innovation and Design (CBID) Master's program who had also completed medical school. Graduates reported that their experiences cultivated transferable skills—design thinking, interdisciplinary collaboration, and leadership—that shaped their professional trajectories. We propose three models for incorporating I&E education into existing medical curricula—short-term workshops, one-year gap programs, and longitudinal tracks—and discuss their advantages and trade-offs. Early and structured exposure to I&E education in medical school empowers students to identify unmet clinical needs, collaborate across disciplines, and develop real-world solutions. As the pace of innovation continues to accelerate, integration of I&E education into medical curricula offers a timely opportunity for medical schools to cultivate physician leaders in this space.

(JMIR Med Educ 2025;11:e79489) doi:10.2196/79489

KEYWORDS

health care innovation; medical education; curriculum development; bioengineering; leadership

Introduction

Medical students traditionally receive extensive training in biological sciences, clinical reasoning, and patient care, but minimal exposure to innovation methodologies, design thinking, and entrepreneurial concepts [1]. As the health care landscape undergoes unprecedented transformation through technological advances, changing information and data landscapes, and evolving patient expectations, traditional medical school curricula that primarily emphasize disease diagnosis and treatment protocols may no longer fully prepare physicians for the opportunities and challenges they will encounter [2,3]. As medical students, educators, and administrators, we believe that health care innovation and entrepreneurship (I&E) represents a critical area of education that should be integrated into modern medical school curricula. These concepts provide an important framework for how to identify real-world health care challenges, characterize these challenges in detail, and design effective solutions. They mix both theory and practice, spanning four essential domains: Medical, Business, Technical, and Entrepreneurial. From a medical student perspective, we uniquely recognize the critical gap between the rapidly evolving health care environment we are entering and the limited exposure to I&E concepts during medical education. Physicians are uniquely positioned to identify clinical problems worth solving, yet most lack the formal training to translate these insights into viable solutions, representing a missed opportunity to significantly expand their impact on the care of patients around the world.

Current Landscape and Needs

Although there has been increasing recognition of the importance of health care innovation, few medical schools offer structured I&E programs [4]. Examples include Harvard's Health Sciences and Technology (HST) program, Stanford's Biodesign Innovation coursework for medical students, University of Michigan's Biomedical Innovation & Entrepreneurship Certificate Program, George Washington's

Clinical Practice Innovation and Entrepreneurship Track, and at Johns Hopkins, the Center for Bioengineering Innovation & Design (CBID)'s one-year full time Master's program. Since its inception in 2009, 24 medical students and physicians have completed the CBID program [5].

A 2021 study found that only 15.2% (26/171) of American and Canadian allopathic medical schools offered I&E-oriented medical education programs [6]. Among these 28 programs, 57% (n=16) had been started within the past four years, 75% (n=21) required a selective application process, and 79% (n=22) required students to complete a capstone project [6]. The most common program structure is a four-year track or concentration (15/28, 54%) [6]. Other formats vary in duration, ranging from weeks-long courses to a five-year dual-degree program. Teaching strategies in these I&E programs include lecture series, progress meetings, problem-based learning, workshops, mentorship, and guest lectures.

Survey of Student Interest

Given this growing but still limited landscape of opportunities, we sought to assess medical student interest in health care I&E programs at our institution, the Johns Hopkins University School of Medicine. During the 2023 - 2024 academic year, we distributed an electronic survey via email to 480 students across all years and received 90 responses, an approximately 19% response rate. The survey assessed student perceptions of innovation importance and career intentions and was approved by the Johns Hopkins Medicine Institutional Review Board (ID: IRB00381640).

The results suggest strong interest among the student body in integrating health care I&E into medical education and a recognition of its importance. Most respondents (n=87, 97%) expressed that a physician's knowledge and experience with health care innovation was at least somewhat important, including 39 students (44.8%) who thought it was very important or essential. Additionally, 63% of students expressed that they were likely or very likely to incorporate health care innovation into their future practice. A total of 36% percent of students indicated that they were likely to devote more than one quarter of their future career time to health care innovation. Importantly, 43% of students reported that the availability of opportunities to learn about innovation was a positive factor in their choice of medical school (Table 1).

Table . Medical student survey.

Survey questions	Respondents, n (%)
How important is it for a physician to have knowledge about and experience with health care innovation?	
Not important	3 (3)
Somewhat important	48 (54)
Very important	37 (41)
Essential	2 (2)
Thinking about your vision for your medical career long term, what percentage of your time do you anticipate devoting to health care innovation?	
0% - 25%	58 (64)
25% - 50%	26 (29)
50% - 75%	4 (5)
75% - 100%	2 (2)
When you were applying to medical school, how did the availability of opportunities for exposure to healthcare innovation affect your choice of medical school?	
It was a negative factor	0 (0)
It was not a factor	51 (57)
It was a minor positive factor	21 (23)
It was a moderate positive factor	13 (14)
It was a major positive factor	5 (6)

We acknowledge several limitations of our student survey. The relatively low response rate may introduce selection bias, as students with a preexisting interest in I&E may have been more motivated to respond. Second, the survey reflects the perspectives of students at a single institution, and these findings may not be generalizable to other institutions. Nevertheless, with nearly all respondents considering familiarity with I&E somewhat important, this represents a sizable number of students, even factoring in the response rate.

Impact of Health Care Innovation and Entrepreneurship Training

Our institution offers a one-year immersive Master's degree program managed by CBID that provides a structured framework to allow students to study medical device innovation in a hands-on, team-based format, with attention to all four domains mentioned previously. To better understand the impact of early exposure to health care I&E, we distributed an electronic

survey via email to 12 medical school graduates who completed the program between 2018 and 2024.

Five graduates had completed the CBID program during medical school, and seven after medical school. The respondents' subsequent career paths demonstrate a strong commitment to I&E. Two practicing physicians currently dedicate 90% of their time to clinical practice and 10% to entrepreneurship, while ten current residents anticipate distributing their time across clinical practice (54%), research (19%), industry (9%), entrepreneurship (16%), and other roles (2%), including teaching programs like CBID (Table 2).

We also sought to assess the impact of skills that alumni had developed through the CBID program. When asked about which skills have been most applicable to their medical career, the most common responses were design thinking and ideation (n=11, 91.7%), leadership and team management (n=8, 67.7%), knowledge of health care markets (n=7, 58.3%), and interdisciplinary collaboration (n=6, 50%) (Table 2).

Table . CBID graduate survey.

Survey questions	Respondents
When did you complete the CBID program? n (%)	
During medical school	5 (42)
After medical school	7 (58)
How do you currently allocate your time to the following? n=2, (% of time)	
Clinical practice	90
Entrepreneurship	10
How do you hope to allocate your time to the following in your future career? n=10, (% of time)	
Clinical practice	54
Research	19
Entrepreneurship	16
Industry	9
Other	2
Which skills or experiences gained from CBID have been most applicable in your medical career? n=12, n (%)	
Design thinking and ideation	11 (92)
Leadership and team management	8 (68)
Knowledge of health care markets	7 (58)
Interdisciplinary collaboration	6 (50)
Project management	3 (25)
Other	1 (8)

[a]CBID: Center for Bioengineering Innovation and Design

Participants highlighted the program's transformative impact, with one stating, "The full CBID experience made a big difference in my career trajectory and the exposure alone to the space was invaluable as this is something most medical students do not have access to and gives a unique perspective on health care." Similarly, another student noted, "CBID offers a truly unique, collaborative, and hands-on experience that transformed my medical and surgical training. Participants learn to tackle clinical challenges comprehensively and without bias, enabling the development of clinically, commercially, and technically sound innovations." These findings suggest that structured I&E education through programs like CBID can equip students with skills that complement existing clinical training and experiences that potentially shape students' career trajectories.

In our opinion, design thinking skills acquired through such programs have the potential to augment a trainee's development across multiple aspects of medical practice and professional growth. Design thinking entails an organized, iterative process in which one repeatedly interrogates user needs, defines aspects of a problem, and explores potential solutions to address those aspects. This framework trains students to approach clinical scenarios with a structured, problem-solving mindset which can improve diagnostic reasoning and clinical judgment [7-9]. Moreover, exposure to design thinking can empower clinicians to reimagine their career trajectories and even contribute to innovations such as the development of new medical devices and therapies [10,11].

Interestingly, while 50% (6/12) of alumni identified interdisciplinary collaboration as valuable, it ranked lower than expected given the inherently collaborative nature of health care. In clinical practice, physicians routinely coordinate with a wide range of professionals and specialists from other disciplines. Likewise, successful physician innovators depend on interdisciplinary teams including engineers, entrepreneurs, and business professionals to translate ideas into viable solutions. Navigating this ecosystem is a critical "soft skill"

that I&E programs are uniquely positioned to cultivate through hands-on, team-based experiences. One possible explanation for these findings is that many medical students may have already developed a foundation of interdisciplinary skills by the time they joined the CBID program, which was applied to the innovation space with relative ease. Nonetheless, these findings do not detract the importance of supporting interdisciplinary collaboration through I&E curricula, as developing these skills early may better prepare future physicians to lead and thrive in complex health care environments.

Our alumni survey has limitations including a small sample size (n=12). Graduates of the CBID program are also likely to report that they plan to use this training in their future careers. However, our findings may actually underreport the true impact of such programs, as the long-term effects of such programs have not yet been fully realized. It may take time for graduates to assume leadership roles and make visible impacts in their respective fields.

Critical Timing and Developmental Considerations

In our opinion, exposure to health care I&E is important early in medical education. When students are exposed to I&E training before clinical routines are developed, they are more likely to look for opportunities for improvement over existing practices and have the tools to act effectively on these opportunities. This timing allows students to see clinical situations and identify clinical challenges through a critical lens. Rather than being solely acclimated to current standards, these students are more likely to question assumptions, recognize unmet needs, and envision improved solutions. As current students ourselves who are deeply involved in I&E programs at Johns Hopkins, we can attest to the ways in which these opportunities have complemented our development of clinical knowledge and helped us better tackle patient care obstacles.

Additionally, we believe that early exposure to health care I&E can help to cultivate physician leaders in the health care innovation space. This is evident in the current and anticipated future professional activities of the CBID alumni. While many respondents are still in residency training, most remain actively involved in I&E through various pursuits. Seven graduates serve as chief executive officers or cofounders of startups, while others mentor engineering students (n=1) or lead research labs (n=1). These career trajectories suggest that early engagement with innovation during medical education not only enhances students' ability to identify and solve complex clinical problems but also fosters leadership and sustained involvement in health care I&E.

Structural Considerations

We believe that the goal of a medical school I&E program should be to expose students to the principles of design thinking, to provide them with practical experience navigating the design process, and to enable them to become conversant with a wide range of experts in this field. By the end of the program, students should be equipped with the skills necessary to apply the design process to real-world clinical problems. While we are aware that not all physicians will be interested in pursuing a career in health care I&E, these skills are applicable to all aspects of problem-solving in medical practice and will help students build sufficient fluency in order to collaborate with peers in related fields [8,9]. There are several models for incorporating health care I&E training into medical education, each with its own advantages and disadvantages (Table 3).

Table . Structural models for innovation and entrepreneurship programs.

Model	Duration	Advantages	Disadvantages
Short-term programs	Weeks to months	Broad accessibility Easy curriculum integration Low time commitment	Limited hands-on experience Theoretical focus
Gap year programs	1 year	Comprehensive immersion Complete project cycles Dedicated focus time	Financial barriers Limited accessibility
Longitudinal tracks	4 years	Integrated learning Extended project development	Challenges for sustained engagement Requires curriculum flexibility Resource intensive

Short-Term Programs

Week- or month-long programs provide an excellent structure to introduce foundational I&E principles such as design thinking, problem identification, and prototyping to a broad range of students. Their shorter time commitment may be appealing to students who are interested but hesitant to dedicate significant time to I&E. Additionally, these programs are often easier to implement into the existing medical school curricula.

However, the limited duration may not provide students sufficient time for meaningful engagement in a hands-on project—an essential component for understanding the nuances of the design process. Without having practical experience, graduates of these curricula may only gain theoretical knowledge. Therefore, they may not be fully prepared to tackle and address more practical challenges that emerge along the journey of solving a clinical problem.

Gap Year Programs (Eg, One-Year Master's Degrees Such as the CBID Program)

Gap year programs provide comprehensive, hands-on experience working through the entire innovation process from identifying a clinical need to developing a prototype. By stepping away from traditional coursework and clerkships, students can fully immerse themselves in their projects without competing academic responsibilities, augmenting the learning experience.

However, financial or academic constraints may deter students from pursuing an additional year of education. Some students may be reluctant to delay their clinical training and career progression, while others interested in I&E may prefer earlier engagement rather than waiting several years to begin formal education.

Longitudinal Tracks (Eg, Integrated Four-Year Programs Alongside the MD Curriculum)

A four-year longitudinal track can seamlessly integrate innovation-focused classes alongside traditional medical curriculum, allowing students to receive both theoretical and practical exposure in innovation. By introducing I&E training at the beginning of medical school, students have more time for hands-on exposure allowing for deeper engagement in the project, time for iterative problem-solving, and a greater chance of seeing projects through to implementation. Additionally, the extended time frame allows students to develop meaningful relationships with mentors in both academia and industry. This structure enriches the learning experience and helps prepare students with the skills needed to succeed as future innovators.

However, sustained engagement over four years can be difficult especially as competing academic demands or student interest change. Furthermore, not all medical school curricula allot the time and flexibility necessary to incorporate this coursework. Additionally, implementing a longitudinal program requires substantial and ongoing faculty, mentorship, and institutional support to ensure strong educational quality across multiple years.

Challenges and Implementation Considerations

There are several challenges to implementing innovation-focused programs within a medical school curriculum. Notably, the highly structured and densely packed schedule of the medical school curriculum may leave limited room for additional courses or activities. However, the I&E program's duration and content can be tailored around the existing schedule. Many medical schools allocate decided time for optional independent scholarly activities and research [6]. Portions of this time can be allotted to core classes in the I&E track or independent work. In the clinical years, students often have the option to take research electives, which can also provide additional time to engage with the I&E curriculum.

A phased implementation approach may help effectively introduce a health care I&E track, especially in institutions with limited resources. By initially offering short elective courses that require minimal infrastructure, institutions can assess student engagement and interest [12]. For example, during the 2024 - 2025 academic year, we designed and led a three-week long elective course for first year medical students focused on teaching the principles of the design process and applying them to simple real-world problems. Fifteen (out of 120) first-year medical students registered for and then completed the course, reflecting high student interest. As faculty observe the value students place on I&E training, they may become more receptive to integrating these concepts into the broader curriculum. The positive outcomes from a pilot program can also build a compelling case for increased institutional support and may also help create a collaborative environment in which faculty and administration can collectively support the evolution of the medical curriculum.

Another challenge for the implementation of a health care I&E program is securing sustainable funding. One strategy is to pursue external funding sources dedicated to improving medical education. For example, the American Medical Association's ChangeMedEd initiative offers Innovation Grants to support transformation educational projects. Another option is forming partnerships with health care organizations and technology companies. For example, collaborations with venture capital or medical device companies can provide students with projects to address real-world clinical needs. This offers students practical experience while also advancing the partners' objectives. These partnerships also provide students with expertise and mentorship. In fact, the aforementioned CBID program successfully uses such a partnership model for many of the project opportunities offered to their students. In addition, at Johns Hopkins, pre- and postdoctoral training grants focused on health care I&E have provided support for some medical students and residents in pursuing yearlong Masters programs and projects.

Future Directions and Expected Trends

We anticipate that interest in I&E education among medical students will continue to increase. First, growing student awareness of the importance of health care I&E is evident in our survey findings, where 43% of students considered innovation opportunities when choosing their medical school. Second, external forces, including health care system pressures, technological advancement, and changing care delivery models create increasing demand for physician innovators. Third, the increase in dual-degree and specialized programs reflects students' desire to differentiate themselves for residency applications and career development [4,6]. Notably, 57% of the 28 identified programs were launched within the past four years, underscoring the recent surge in innovation-focused medical education.

Future studies should be conducted to further understand the role of I&E education in medical curricula. These studies should

be larger and involve geographically diverse institutions across different tiers and program structures to validate the observations from our study. Additionally, longitudinal studies with larger cohorts that follow students for multiple years through medical training and into their careers will be important to capture the full effect of structured I&E training on career trajectories and professional contributions in health care.

Conclusion

As medicine evolves at an ever-accelerating pace, it is increasingly important for the next generation of physicians to have a proper foundation in clinical problem solving and develop the design thinking principles and effective problem-solving skills that will allow them to integrate emerging technologies into their solutions and contribute to the creation new ones. Early exposure to health care I&E allows students to nurture both of these aspects of their academic development simultaneously, augmenting the quality of their overall training and promoting their ability to serve as physician leaders of health care I&E in the future. Our findings suggest substantial interest in I&E training among medical students at our institution, consistent with national trends, and we expect this trend to increase rather than diminish. Therefore, following in the footsteps of prior successfully implemented programs, we encourage all medical schools to investigate ways in which they can integrate I&E programs into their curriculum. These efforts will better prepare students to collaborate across disciplines and lead future innovation initiatives that will ultimately improve the care of the patients they will serve.

Acknowledgments

No generative AI tools were used in the writing, editing, data analysis, or figure creation for this manuscript.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Authors' Contributions

LZ: Conceptualization, Methodology, Data curation, Formal analysis, Writing - original draft
JK: Conceptualization, Methodology, Data curation, Formal analysis, Writing - original draft
OW: Conceptualization, Methodology, Data curation, Formal analysis, Writing - original draft
KCC: Methodology, Writing - review & editing
YY: Supervision, Methodology, Writing - review & editing

Conflicts of Interest

None declared.

References

1. Hindin DI, Mazzei M, Chandragiri S, et al. A National Study on Training Innovation in US Medical Education. Cureus 2023 Oct;15(10):e46433. [doi: 10.7759/cureus.46433] [Medline: 37927762]
2. Frenk J, Chen L, Bhutta ZA, et al. Health professionals for a new century: transforming education to strengthen health systems in an interdependent world. Lancet 2010 Dec 4;376(9756):1923-1958. [doi: 10.1016/S0140-6736(10)61854-5] [Medline: 21112623]
3. Kocher R, Emanuel EJ, DeParle NAM. The Affordable Care Act and the future of clinical medicine: the opportunities and challenges. Ann Intern Med 2010 Oct 19;153(8):536-539. [doi: 10.7326/0003-4819-153-8-201010190-00274] [Medline: 20733178]
4. Niccum BA, Sarker A, Wolf SJ, Trowbridge MJ. Innovation and entrepreneurship programs in US medical education: a landscape review and thematic analysis. Med Educ Online 2017;22(1):1360722. [doi: 10.1080/10872981.2017.1360722] [Medline: 28789602]
5. Yazdi Y, Acharya S. A new model for graduate education and innovation in medical technology. Ann Biomed Eng 2013 Sep;41(9):1822-1833. [doi: 10.1007/s10439-013-0869-4] [Medline: 23943068]
6. Arias J, Scott KW, Zaldivar JR, et al. Innovation-oriented medical school curricula: review of the literature. Cureus 2021 Oct;13(10):e18498. [doi: 10.7759/cureus.18498] [Medline: 34754659]
7. Health care providers can use design thinking to improve patient experiences. Harvard Business Review. URL: https://hbr.org/2017/08/health-care-providers-can-use-design-thinking-to-improve-patient-experiences [accessed 2025-10-11]
8. Sandars J, Goh PS. Design thinking in medical education: the key features and practical application. J Med Educ Curric Dev 2020;7:2382120520926518. [doi: 10.1177/2382120520926518] [Medline: 32548307]
9. McLaughlin JE, Wolcott MD, Hubbard D, Umstead K, Rider TR. A qualitative review of the design thinking framework in health professions education. BMC Med Educ 2019 Apr 4;19(1):98. [doi: 10.1186/s12909-019-1528-8] [Medline: 30947748]

10. Roberts JP, Fisher TR, Trowbridge MJ, Bent C. A design thinking framework for healthcare management and innovation. Healthcare (Basel) 2016 Mar;4(1):11-14. [doi: 10.1016/j.hjdsi.2015.12.002]
11. Sorelle R. News: real-world health challenges solved by design. Emergency Medicine News 2017;39(2):28. [doi: 10.1097/01.EEM.0000512783.55588.82]
12. Ahrari A, Sandhu P, Morra D, McClennan S, Freeland A. Creating a healthcare entrepreneurship teaching program for medical students. JRMC 2021 Jan 28;4(1):1. [doi: 10.24926/jrmc.v4i1.3564]

Abbreviations

CBID: Center for Bioengineering Innovation and Design
HST: Health Sciences and Technology
I&E: Innovation and Entrepreneurship

Edited by A Stone; submitted 22.06.25; peer-reviewed by R Buckley, S Kolla; revised version received 19.11.25; accepted 19.11.25; published 19.12.25.

Please cite as:
Zhu L, Khong J, Wei O, Chretien KC, Yazdi Y
The Need for Health Care Innovation Training in Medical Education
JMIR Med Educ 2025;11:e79489
URL: https://mededu.jmir.org/2025/1/e79489
doi:10.2196/79489

Leveraging Generative Artificial Intelligence to Improve Motivation and Retrieval in Higher Education Learners

Noahlana Monzon, BS; Franklin Alan Hays, BS, PhD

Department of Nutritional Sciences, University of Oklahoma Health Sciences, 1200 N Stonewall Ave, 3064 Allied Health Building, Oklahoma City, OK, United States

Corresponding Author:
Franklin Alan Hays, BS, PhD
Department of Nutritional Sciences, University of Oklahoma Health Sciences, 1200 N Stonewall Ave, 3064 Allied Health Building, Oklahoma City, OK, United States

Abstract

Generative artificial intelligence (GenAI) presents novel approaches to enhance motivation, curriculum structure and development, and learning and retrieval processes for both learners and instructors. Though a focus for this emerging technology is academic misconduct, we sought to leverage GenAI in curriculum structure to facilitate educational outcomes. For instructors, GenAI offers new opportunities in course design and management while reducing time requirements to evaluate outcomes and personalizing learner feedback. These include innovative instructional designs such as flipped classrooms and gamification, enriching teaching methodologies with focused and interactive approaches, and team-based exercise development among others. For learners, GenAI offers unprecedented self-directed learning opportunities, improved cognitive engagement, and effective retrieval practices, leading to enhanced autonomy, motivation, and knowledge retention. Though empowering, this evolving landscape has integration challenges and ethical considerations, including accuracy, technological evolution, loss of learner's voice, and socioeconomic disparities. Our experience demonstrates that the responsible application of GenAI's in educational settings will revolutionize learning practices, making education more accessible and tailored, producing positive motivational outcomes for both learners and instructors. Thus, we argue that leveraging GenAI in educational settings will improve outcomes with implications extending from primary through higher and continuing education paradigms.

(JMIR Med Educ 2025;11:e59210) doi:10.2196/59210

KEYWORDS

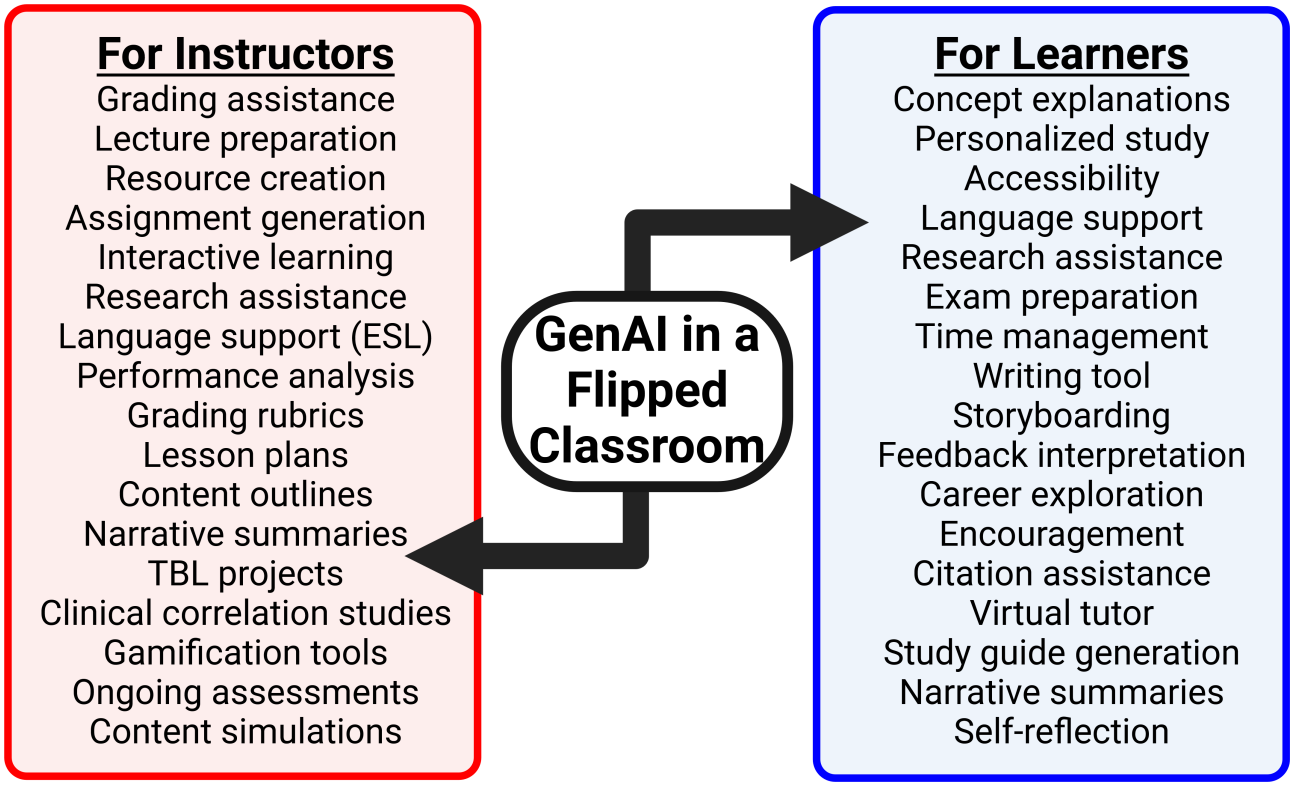
educational technology; retrieval practice; flipped classroom; cognitive engagement; personalized learning; generative artificial intelligence; higher education; university education; learners; instructors; curriculum structure; learning; technologies; innovation; academic misconduct; gamification; self-directed; socio-economic disparities; interactive approach; medical education; chatGPT; machine learning; AI; large language models

Introduction

Generative artificial intelligence (GenAI) is impacting educational spaces in ways that few technologies have since the personal computer and calculator [1-3]. Though GenAI is not a new concept, its inroads into education and pedagogy started in earnest following the release of "ChatGPT" (November 30, 2022). We observed learners using ChatGPT within weeks of its release. GenAI continues to rapidly evolve with new "GPTs," models, websites, application programming interfaces, and GenAI-enabled hardware [3-6]. GenAI is now "mainstream" with low activation barriers for use. This new reality is sending shockwaves through educational institutions and districts, including higher and clinical education. Learners, instructors, and administration work to understand and define implications while either leveraging or obstructing GenAI implementation. Indeed, banning and blocking GenAI use in some educational settings remains with no clear consensus on what role GenAI can, or should, play going forward. With this rapidly evolving space, it's challenging to differentiate inflated expectations and hype from productivity and enlightenment, to borrow from the Gartner Hype Cycle. One could argue that the GenAI "peak of inflated expectations" has yet to be reached. However, the new reality in clinical and higher education is one where GenAI will play a role going forward, both within classrooms and clinical practice [3,6-8]. What that looks like remains variable and will change depending on the knowledge of those involved, personal perspectives on GenAI, learner and programmatic needs, and accreditation standards and expectations. Our immediate approach was to embrace GenAI, like ChatGPT and other tools as they have come online (eg, Claude, Bard, and CoPilot), as a new tool to facilitate learning, retrieval, and motivation in a higher education, clinically focused, instructional environment. The rationale is that modern GenAI can generate diverse outputs (eg, text, images, videos, and language) derived from data-centric training sets [9] using narrative style prompts or data inputs (eg, content outlines, documents, PDFs, and images). Thus, there is a pragmatic reality that new GenAI tools have a

low activation barrier for use while being capable of generating high-quality output focused on user needs. It's evident that modern GenAIs ability to generate extensive, coherent, responses is fundamental to increasing engagement, communication, and motivation in educational settings. Though not seamless, especially when considering potential for malfeasance or hallucinations [10], GenAI can integrate throughout curricula to reduce overhead and improve outcomes (Figure 1). This integration fosters environments that inspire and empower learners, promote motivation and collaboration, and facilitates the creation of dynamic and individualized curricula.

Figure 1. Generative artificial intelligence application in a flipped classroom model enhances both learner (blue) and instructor (red) experiences. Shown are examples of generative artificial intelligence benefit and impact within respective domains. Bidirectional arrows indicate reciprocal enhancement of generative artificial intelligence applications, demonstrating improvements in instructor-driven activities inherently enrich learner experience, thereby reinforcing a flipped classroom learning environment. GenAI: generative artificial intelligence.



The objective of this viewpoint is to advance a positive perspective on leveraging GenAI tools in modern medical education environments while presenting examples and methods tested in our hands since ChatGPT's initial release. This viewpoint is presented from both learner (Mrs Monzon) and instructor (Dr Hays) perspectives as, in our experience, both offer unique opportunities on GenAI use. Learners are focused on knowledge acquisition and retrieval from an individualized perspective. Not all learners have the same motivation, hidden curriculum, previous knowledge, and experience, or ability to learn and retain learning objectives defined by instructors. Likewise, instructors are unable to individualize curricula across multiple learners or sections while ensuring productive exposure to core learning objectives defined by accreditation and program standards. It's a conundrum of modern higher education, learners seeking individualized instruction amidst information overload while instructors are bandwidth-limited and hamstrung by program and accreditation demands. GenAI tools will directly impact this reality in a positive manner and empower both learners and instructors. The current challenge is what does that look like? How can GenAI be integrated into learning environments to facilitate learning and retrieval, drive motivation, and improve outcomes while avoiding pitfalls such as loss of voice, data ownership and use, academic misconduct or malfeasance, and incorrect information? This future must balance innovation and GenAI integration with established guidelines, integrity and safety guardrails, and equity. By presenting a nuanced perspective of the interplay between GenAI and learning theories from both learner and instructor perspectives, this viewpoint intends to inform GenAI integration that is inclusive, forward-thinking, and collaborative while not ignoring tangible GenAI benefits for all stakeholders in the learning ecosystem. Integration should not overshadow essential human elements of teaching and learning but rather complement and enhance both, thereby creating a dynamic and inclusive educational environment that is responsive. Finally, we argue that GenAI should not be ignored but embraced. It's imperative that learners are exposed to new technologies that will increasingly impact workforce dynamics going forward. Instructors are innovators and our learners are digital natives surrounded by AI technologies. We implemented and evolved methods described in this viewpoint within graduate (PhD and MS), clinical (dietetics and RD), and undergraduate curricula. Leveraging GenAI in courses does require initial effort, yet subsequent improvements in effort needed, instructional quality, and learner feedback justify the initial cost. GenAI has proven, in our hands, to positively impact every pedagogical niche. It should be noted that we acknowledge significant ethical

concerns regarding GenAI use in educational settings. This has been covered extensively elsewhere [10,11] and the current viewpoint starts with the perspective that GenAI can, and should, play a role in educational settings.

Learner Perspective

Technology is a powerful means to facilitate collaboration between learners and instructors. Learning management systems (eg, Canvas or D2L) are an example of this point, leveraging technology to facilitate learning and retention. In this sense, bringing GenAI into classroom settings is an evolutionary step with clear emerging data that it enhances learner engagement, motivation, and personalized learning in a self-directed manner. The pragmatic meaning is that interactive collaboration can be extended from instructor-learner or learner-learner to include learner-GenAI where the scope and implementation of learner-GenAI interactions is defined by tools being used, prompt design (Figure 2), and personalized needs (Figures 1 and 3). This approach fosters learner motivation as a key driver for positive outcomes [12]. In addition, learning is more effective when it's relevant, engaging, and contextualized to real-life scenarios (eg, team-based learning or clinical applications) in accordance with adult learning theory. Cognitive load is reached when germane, intrinsic, and extraneous factors become unmanageable [13]. Incorporating GenAI into the educational framework can simplify the intrinsic load, reduce the extraneous load and in turn maximize the germane load. This is consistent with our observations using GenAI to foster collaborative interactions in clinical courses. To maintain learner motivation, one must account for both intrinsic and extrinsic factors [13]. Intrinsic factors include self-efficacy, self-determination, curiosity, cognitive engagement, emotional well-being, professional well-being, and innate interest in the material presented. Extrinsic factors include pedagogical approach, peer interactions, assessment methods, learning environment, curriculum design, and quality and scope of feedback (both peers and instructors). A unique aspect to the learner-GenAI interaction is that it impacts both intrinsic and extrinsic motivational elements for learners. For instance, GenAI can be implemented as a personal tutor or study partner that encourages conversations and positive feedback in a low activation barrier environment (eg, compared with instructor office hours). Engaging GenAI "chatbots" like ChatGPT can also be conversational for learners, like interacting with a human counterpart (see "Current Limitations and Future Hurdles" section below). Thus, leveraging the learner-GenAI interaction provides agency to learners which increases autonomy and motivation [14].

Figure 2. Prompt design. General overview for developing prompts with clearly defined role (red), bounds (green), and input (blue). Each component is essential for effective generative artificial intelligence use toward minimizing refinement while helping ensure output is optimized toward specific needs.

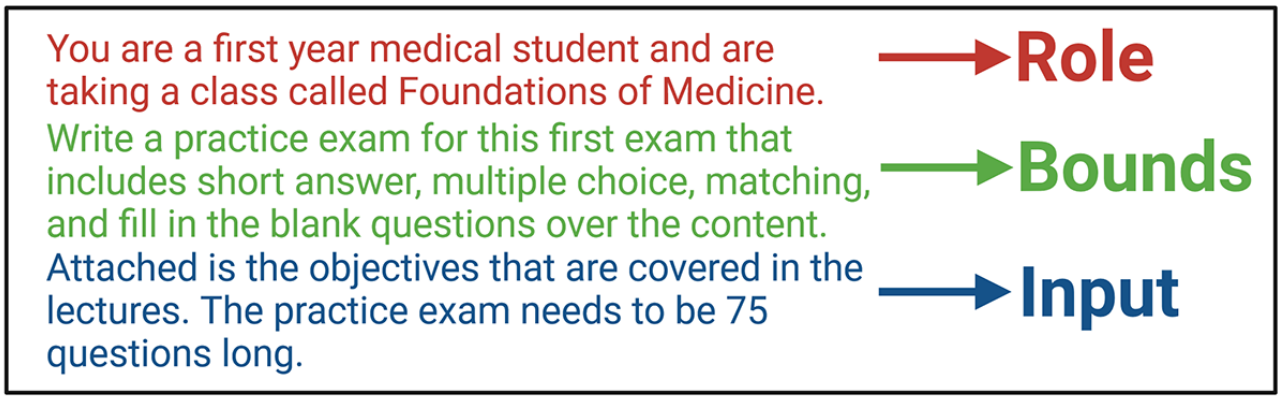
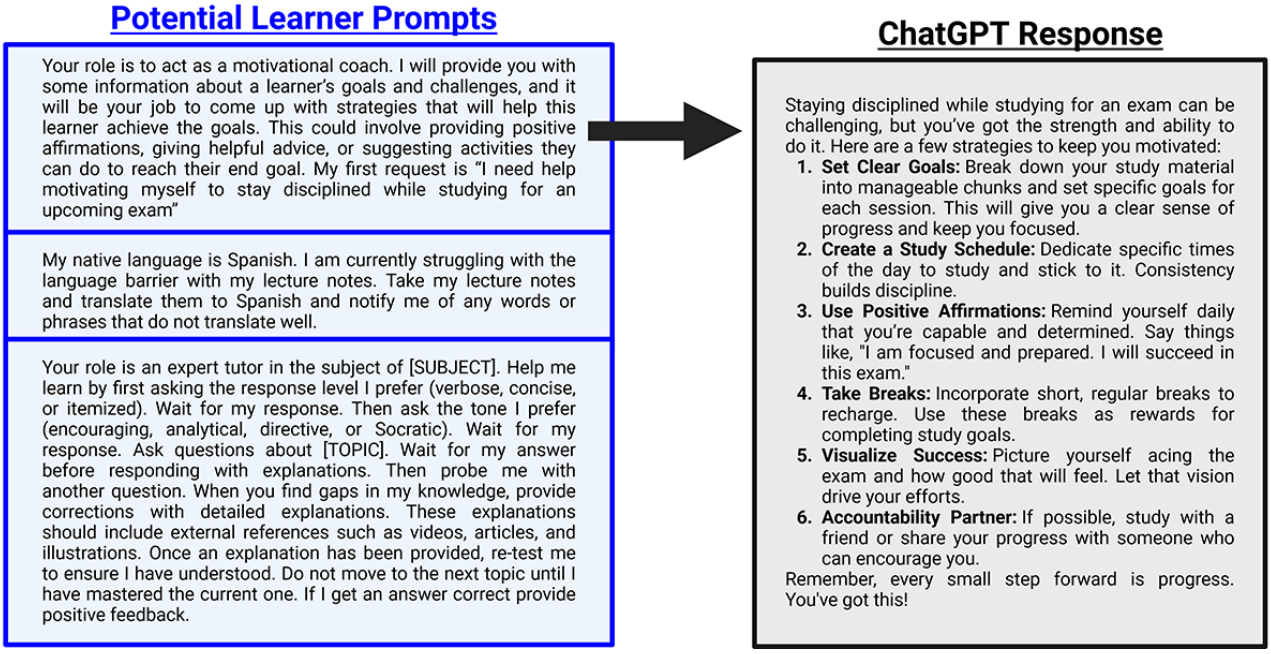


Figure 3. Example learner prompts for leveraging generatice artificial intelligence. Three example learner input prompts are shown (blue box) with representative ChatGPT (accessed on December 10, 2024) output (black box) for the top prompt (black arrow). In these instances, the resultant output can change significantly with minor changes in the input prompt provided. This is also true when using the same input prompt across different generatice artificial intelligence tools. Thus, specificity and clear instructions are key to effect desired output.



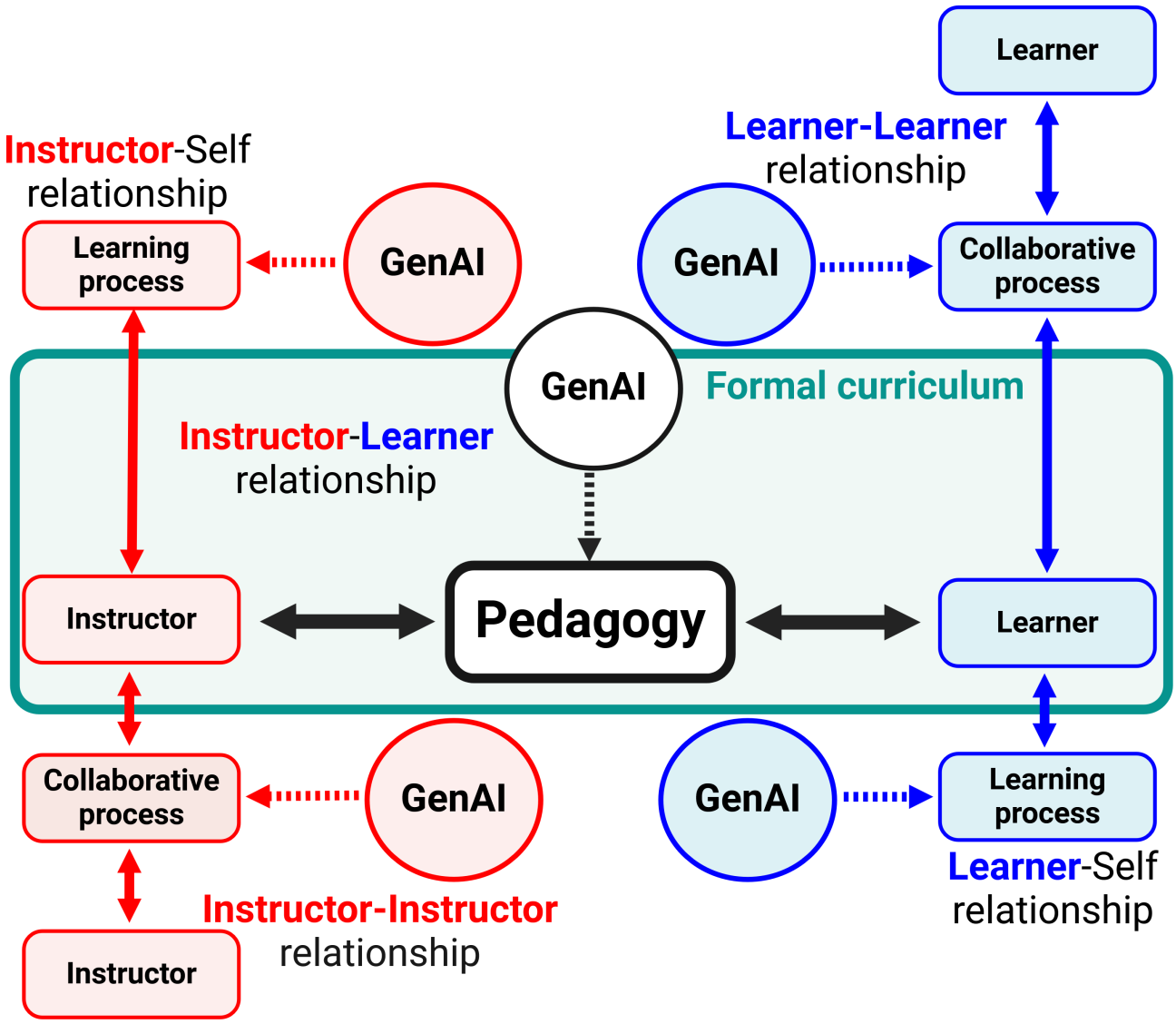
Beyond motivation, retrieval practice is an essential component for learner ownership over new material and, for example, applying knowledge in a clinical setting where integration and rapid access is often required [13]. Indeed, learned information is rapidly forgotten without reinforcement [15]. This is a core consideration of the "Desirable Difficulties" theory by Bjork and Bjork [15] suggesting that one should introduce challenges (eg, spacing or testing effects and varied practice) to enhance long-term memory retention of new information. Thus, retrieval practice requires active effort by learners to bolster information recall. This active engagement promotes deeper processing and understanding to facilitate ownership. A common example of retrieval practice in medical learner training is leveraging the Socratic method in clinical rounds, case discussions, simulations, journal clubs, team-based learning, and even mortality and morbidity conferences. In these scenarios, clinicians are pushed to understand, integrate, and verbalize knowledge under immediate critique and assessment. This moves beyond simple passive recall or reading to test true understanding and identify areas where learners assume ownership of knowledge but fail accurate retrieval or application [16]. In simplistic terms, retrieval practice is a common element to most curricula through formative (common in direct clinical training) and summative (common in formalized classroom instruction) assessments. Incorporating learner-GenAI methods into the curriculum provides a dynamic, ongoing, personalized, and iterative method to facilitate retrieval practice for learners outside of formal, instructor-based, course design. The learner-GenAI axis is instructor-independent in this instance. GenAI can generate adaptive quizzes and assessments while customizing difficulty level and content based on learner proficiency (eg, Figure 3). As learners progress and improve in retrieval practice, GenAI can dynamically adjust question complexity, ensuring continued adaptive learning. These tools analyze users learning patterns, preferences, performance data, and needs to personalize content. In this instance, GenAI recommends specific retrieval or practice exercises and intervals to drive memory consolidation. Learner-GenAI natural language interactions can efficiently manage spaced repetition schedules based upon individual learning patterns and needs to adapt timing and frequency of review sessions, ensuring learners revisit information at optimal intervals for memory retention. Finally, it's important for instructors to consider that learners do not enter courses on equitable footing in knowing how to access, use, and leverage GenAI tools. Initial training with pragmatic examples, discussion of prompt engineering, setting up accounts if needed, and reviewing available tools and associated strengths and weaknesses is strongly advised for courses that allow GenAI use.

Instructor Perspective

Instructors, faculty, and programs within higher education and clinical training settings are primary determinants of motivational factors for learners [12-14]. These include accreditation and departmental oversight (meaning "static" curriculum), designing and structuring assessment, determining feedback mechanisms, managing learning environment, defining expectations, and even implementing recognition and reward systems for positive outcomes or performance. Thus, in our experience, learner motivation is impacted in significant ways before the first day of class. With this backdrop in place, what role can GenAI play in higher education and clinical training from the instructors' perspective? This question has interesting parallels to the "great calculator debate." These include questions of access and equity, learners using these tools outside class regardless of policies set by schools and instructors, learners not gaining essential skills, or knowledge, through their use, modifications required for existing curricula with a major

shift away from algorithms and "rules" toward meaning, concepts, and applications, and ability to trust accuracy of answers produced from novel technologies. Yet, even with the rapidly evolving current landscape surrounding GenAI, we posit that GenAI has significant benefits for instructors. Thus, the instructors' role in harnessing GenAI as an educational tool is multifaceted and includes instructional design, creating dynamic learning content, and even streamlining administrative tasks (Figures 1 and 4), all of which are predicated on training and learning about a rapidly evolving field with new tools appearing almost daily.

Figure 4. Role of generative artificial intelligence in educational relationships and processes. Generative artificial intelligence intersects with and supports relational dynamics between instructors (red) and learners (blue). Four primary interactions are shown, instructor-self, instructor-learner, learner-self, and learner-learner, where generative artificial intelligence serves as a pivotal tool for enhancing learning processes. Generative artificial intelligence's contribution to the pedagogical framework is central, mediating and enriching explicit curriculum delivery and assimilation. Bidirectional arrows between actors and generative artificial intelligence signify a feedback loop allowing for continuous improvement of educational strategies. It underscores generative artificial intelligence's potential to facilitate collaborative processes as well as promoting self-directed learning and peer-to-peer engagement. GenAI: generative artificial intelligence.



GenAI integration by instructors can lower activation barriers to create dynamic, engaging, personalized, and efficient learning environments that optimize learner outcomes (inclusive of motivation and retrieval). We approach this using a flipped classroom (Figure 1), content gamification, streamlined workflows, team-based learning, knowledge gap analysis, and consistent feedback using "exit tickets," all facilitated using GenAI tools. The concept of classroom "flipping" has gained attention in recent years as an approach to instructional delivery. Flipping involves inverting traditional curriculum structure with learners acquiring, or at least engaging, new content outside of structured class time and using active learning methods during class to reinforce, expand upon, and use retrieval practice to reinforce and learn content [17]. All of which is instructor-guided as part of the instructor-learner core axis (Figure 4) [18]. If done well, this approach allows instructors to focus on providing targeted and personalized feedback, requires higher-order critical assembly and thinking skills, and facilitates meaningful discussion or reinforcement during class time for learners. Core tenants of this approach are engaging motivation for learners and expanded retrieval practice outside of high-stress graded assessments like quizzes or exams. Though

our experience with the above approach involves class sizes ranging from 5‐40 learners, others have successfully implemented flipped classroom methods with more than 200‐300 learners [19]. Great examples of this dynamic approach are “metabolic melodies” in which the instructor, Dr Kevin Ahern from Oregon State University, uses complex biochemistry content to generate songs set to popular music such as “Yellow Submarine” by the Beatles. In this instance, Dr Ahern is extremely creative and a musician with an affinity for the Beatles. GenAI empowers instructors, even those lacking creative brilliance, to turn complex content into interactive and dynamic content such as games, clinical case scenarios, creative narratives, or even music and images. Thus, leveraging GenAI in a flipped classroom environment can reduce instructor workload while improving learning outcomes.

Gamification of course content is one mechanism toward merging GenAI tools with positive learner outcomes. GenAI can quickly generate games (eg, Kahoot!, bingo, Jeopardy!, crossword puzzles, or quiz show questions) using content outlines, slides, or even lecture as input (eg, Figure 5). Games are a low stress means of retrieval practice while promoting an interactive and engaged classroom experience [20]. GenAI can also generate complex clinical practice scenarios with rich hypothetical patient details. These scenarios require learners to use critical thinking and diverse knowledge outputs in a similar method to group-based socratic questioning used in medical education. One area we have had great success in using GenAI is developing gap analysis surveys for learners to assess knowledge levels upon course entry, midcourse for progression, and end of course for effectiveness relative to course objectives. GenAI can quickly generate gap assessments using course learning objectives, prerequisite course content outlines and turnkey instructor needs to provide immediate input on needed content modifications when introducing new content. This is an effective approach to identify knowledge gaps or needs that an instructor may assume are covered in prior courses, or were covered but not retained. Finally, GenAI can be leveraged to streamline workflows before, during, and after content delivery. This includes using templates to generate individualized self-assessment tools for learners, providing narrative feedback for learners on correct and incorrect responses, turning content outlines into focused assessments, integrating lecture modalities under core and redundant learning objectives, and statistical analysis on batched learner responses to identify learning gaps post-content delivery (Figures 1 and 6) [21]. Thus, improving learner outcomes using GenAI tools shows significant promise even with clear limitations, discussed below, and rapidly evolving tools. A central theme for instructors when considering GenAI integration into preplanning, execution, and postanalysis is to balance promoting motivation with opportunities for knowledge retrieval. Course guidelines and instructor expectations must be very clearly defined as to acceptable GenAI use during a given course. This is important considering the current environment where broad institutional or district policies may be lacking, or nonexistent, and variability in what is acceptable between different instructors and courses. Effective communication and clear policies and procedures remain the most important means to avoid academic misconduct or malfeasance. Adapting retrieval strategies to accommodate different learning styles, while ensuring inclusive and personalized learning experiences, is important yet challenging to implement in practice. GenAI holds promise for instructors as these tools provide opportunities to reduce activation barriers (eg, time constraints) toward delivering more effective content and meaningful assessments.

Figure 5. Example instructor prompts for leveraging generative artificial intelligence. Three example instructor input prompts are shown (red box) with representative ChatGPT (accessed on December 10, 2024) output (black box) for the middle prompt (black arrow). A key aspect for instructors is to clearly define the level of instruction being provided, type of learner being instructed, with narrowly defined content scope relative to learning objectives. The latter can be accomplished by inputting content outlines, lecture slides, or narrative summaries.

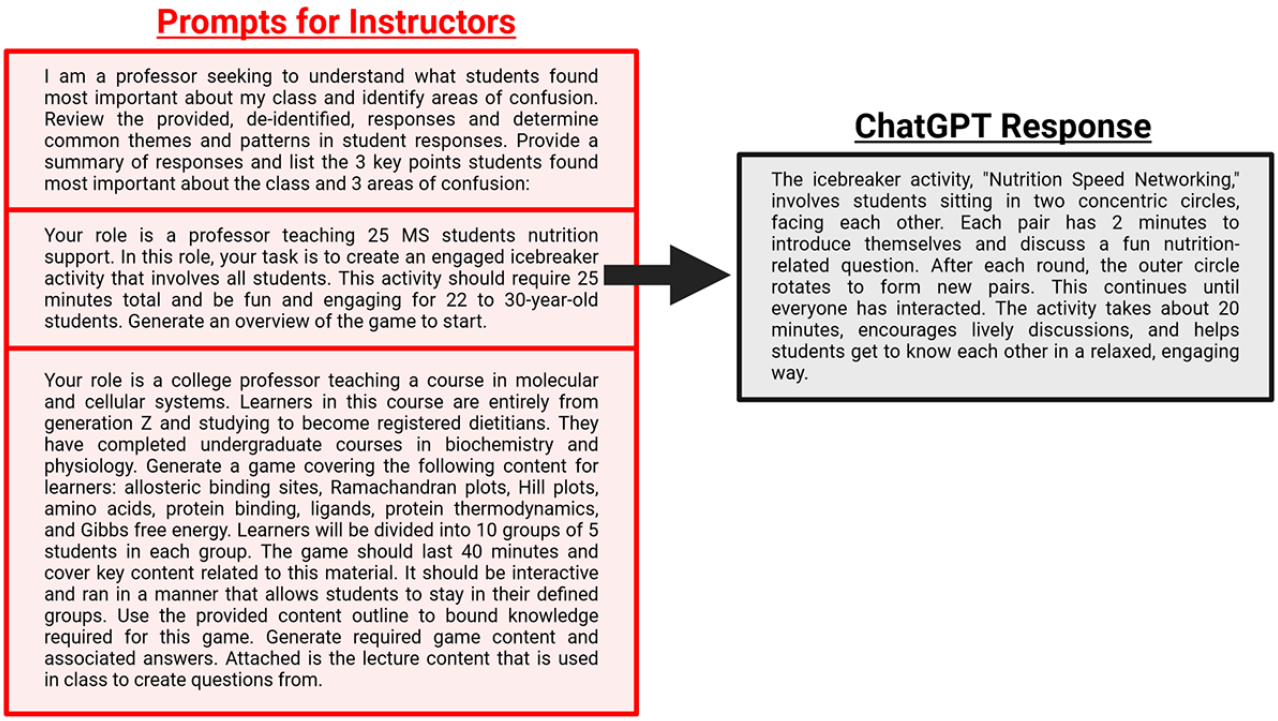
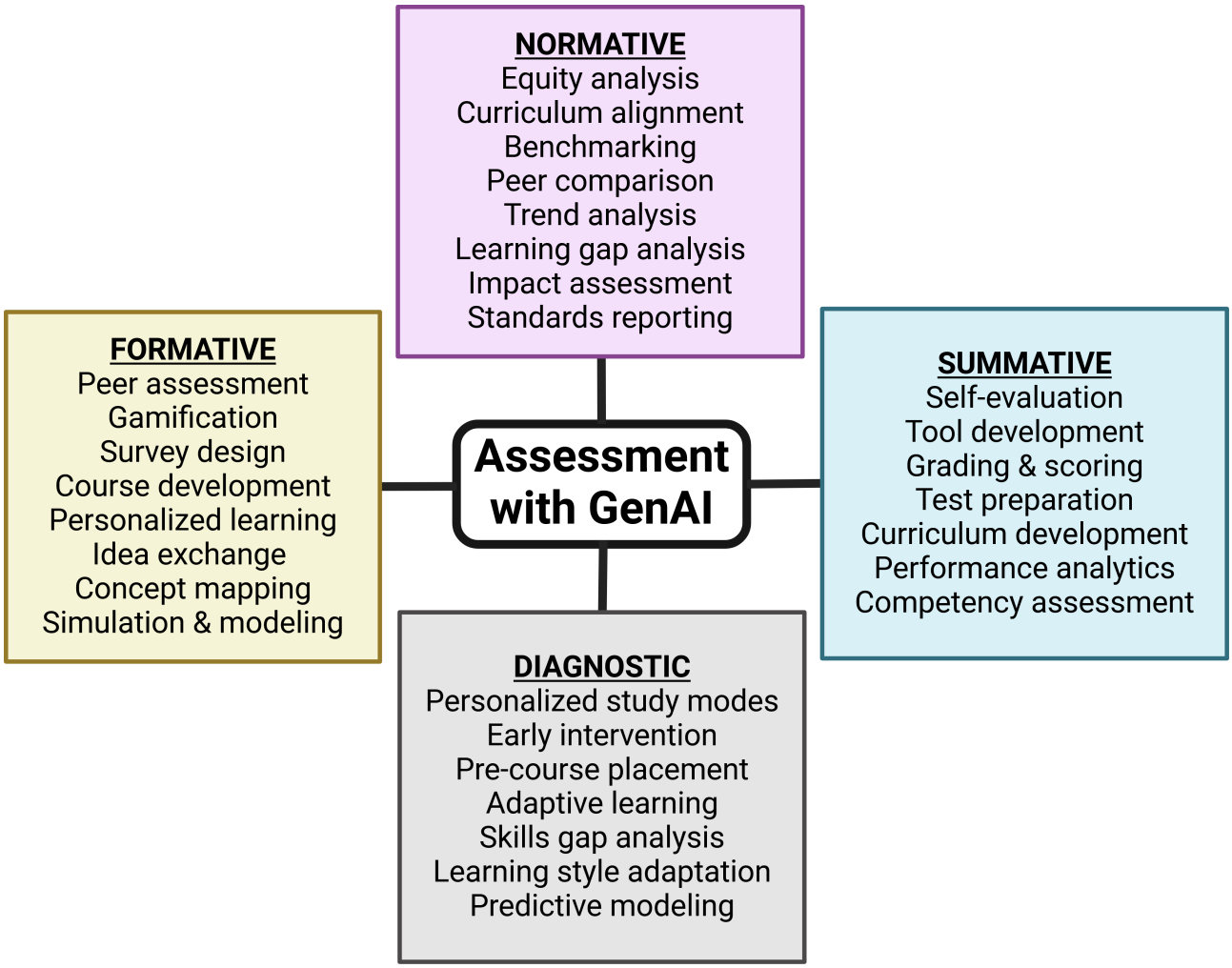


Figure 6. Generative artificial intelligence integration across assessment types. Generative artificial intelligence use across assessment modalities is shown. In formative assessment, generative artificial intelligence aids in creating interactive content and personalized feedback mechanisms. For summative assessment, generative artificial intelligence capabilities extend to evaluating overall learning achievements and generating comprehensive exams. Normative assessment with generative artificial intelligence focuses on establishing benchmarks and evaluating learning outcomes against standards, while diagnostic assessment leverages generative artificial intelligence for early detection of learning gaps and customization of learning experiences. The central position of "Assessment with GAI" emphasizes its role as a centralized tool, facilitating a comprehensive and integrated approach to educational assessment. GenAI: generative artificial intelligence.



Current Limitations and Future Hurdles

Embracing GenAI tools holds promise for both learners and instructors, yet significant hurdles remain, and user caution is warranted in a rapidly evolving environment of GenAI capabilities, tools, ethics, and acceptable use policies and procedures. Accuracy of GenAI output is a critical aspect that requires careful consideration and diligence, especially when used as a training tool for future clinicians and scientists. LLMs are trained by large datasets and leverage analytics to produce predictions, not logic-derived integrations linking informed input to informed output [22]. Thus, a randomness can exist between prompt design and output obtained. Learners and instructors must both carefully assess and evaluate output from GenAI tools to detect "hallucinations" (ie, incorrect GenAI output that presents as correct). Outside of local application programming interface iterations, general releases of these GenAI models are trained on large datasets that may include unvalidated data from the internet. Thus, even if effort is made to include reliable and authoritative sources, these large training datasets may contain misinformation, biases, or outdated information from uncurated data. We have observed this on several occasions when implementing GenAI as an educational tool, where output sounds entirely factual and even referenced only to be completely incorrect with nonexistent references (this improved substantially in GPT-4o and Claude-3.5 Sonnet, accessed on April 11, 2024). As of this submission, caution is still warranted even with significant improvements in model quality. One effective strategy to minimize accuracy issues is uploading content, such as lecture outlines or even slides, and designing focused prompts working from provided content. If possible, use an institutionally firewalled GenAI tool where content is not shared beyond immediate use. This works well to develop focused learner assessments. This leads to another limitation and hurdle: the rapid pace of GenAI tool development, improvement, and deployment has created an environment where time-limited instructors and digital native learners are increasingly overwhelmed with determining best practices, tools, methods, or even workflows. We have responded to this reality by developing open training courses (eg, on Canvas Learning Management System from Instructure) with frequent

updates and ongoing informational seminars, often targeting instructors, to raise awareness of AI changes as they relate to clinical practice and pedagogy. While this rapidly evolving landscape of tools and capabilities is a challenge, it's also an opportunity to leverage new features and expand impact.

In addition, a caveat to implementing "chatbot" style GenAI like ChatGPT in educational frameworks is the input prompt. The relationship between input prompt and output produced is so integrated that "prompt engineering" is a growing career emerging alongside GenAI [23]. Prompt engineering is the careful construction of input prompts or instructions for GenAI models to influence content, style, voice, depth, and even accuracy of resultant output. How input prompts are constructed is a primary determinant of what models produce, even down to small changes like omitting single words, changing adverbs, or using commas versus numbers for a list [24]. These small changes can produce vastly different results with biased, misleading, or inaccurate information [24]. Truly effective prompts are often complex and descriptive, striking a balance between specificity and openness [25]. Meaning, a prompt that is too specific may limit the ability to generate diverse or creative responses, and a prompt that is too open-ended may result in ambiguous or irrelevant output. The approach we use is constructing "Role+ Bounds+Inputs" style input prompts. In this formula, "Role" involves assigning the chatbot a specific job or identity for the analysis (eg, college professor teaching a specific course and learner type), "Bounds" will establish limitations and constraints for model operation (eg, academic context, subject matter, and level of expected answer), and "Inputs" includes relevant contextual information (eg, content outlines, rubrics, or slide summaries). Using a structured prompt guides GenAI models to produce more focused, accurate, and relevant responses. In addition, one can use a scaffolding approach with iterative prompts building toward a common theme or objective. The complexities of prompt engineering have been discussed elsewhere and leveraging the resources provides a deeper discussion of the benefits of a productive prompt while leveraging GenAI [26,27].

GenAI tools such as ChatGPT are proving transformative, impactful, and hold immense potential for enhancing the educational experience for both learners and instructors. Yet, this new reality is problematic considering the lack of transparency on training data content; ethical and socioeconomic implications; quality control and model accuracy; and the potential to perpetuate bias, loss of voice, and agency for both learners and instructors [21]. Learners should be empowered to make informed decisions regarding their participation in GenAI-related or driven activities, and instructors should encourage learners to ask questions, express concerns, and participate in shaping ethical guidelines related to GenAI use in curricula [11,21]. Furthermore, instructors and administrators have an expectation to proactively define clear policies and procedures for GenAI use in educational settings that provide flexibility for both learners and instructors. A major benefit to GenAI use in educational settings is its collaborative potential to rapidly generate personalized and dynamic content, yet this requires equity in understanding and use. Considering GenAI's rapid evolution, this equity has always been absent in courses we have started since ChatGPT was released in November 2022. It's incumbent on instructors to ensure learners are familiar with GenAI tools being used and available. Data-driven insights from GenAI analytics enable instructors to provide targeted support to individual learners as skills develop and courses progress. This can facilitate multimodal learning experiences, incorporating various media formats such as videos, audio, and interactive simulations mentioned above. Significant hurdles remain regarding GenAI use in higher education. These include data ownership and privacy, output accuracy, linking learner needs with instructor resources, and ensuring sufficient training to avoid equity and GenAI skills gaps when being used. Academic institutions are increasingly looking at internal, data-protected and firewalled, GenAI resources (eg, Microsoft CoPilot) yet there remain limited options for analyzing course content and metrics. Finally, ethical concerns associated with leveraging GenAI by both learners and instructors is a major (probably primary at most institutions) topic of discussion.

Conclusions

Our experiences in leveraging GenAI across 5 academic semesters have, overall, been very positive. This implementation is from the perspective of informing and training both learners and instructors; establishing clear policies and procedures relating to academic misconduct at the department, college, and institutional levels; ensuring equity and ability in use; and constant vigilance regarding content accuracy and limitations. LLMs and GenAI have evolved through decades of iterative research across multiple disciplines including statistics, mathematics, and computer science data science [28]. They are not new technologies. Development of the transformer architecture in 2017 was a key transition point for the emergence of current "chatbots" gaining momentum in popular media and use [29]. This technology continues to evolve with diffusion, attention mechanism variants, and retrieval-enhanced transformer mechanisms being new examples of how GenAI technology is rapidly evolving [30]. Through leveraging large datasets with high-demand computational needs, current GenAI models show significant promise. The important point being, these models (eg, ChatGPT or Claude) excel at pattern recognition yet struggle with defining logical connections between training data and outputs produced ("reasoning"). This is an important caveat for application in higher educational settings focused on critical thinking, developing advanced knowledge and skills within specific disciplines, clinical training, and scientific discovery. As such, it's essential for instructors, administrators, policymakers, institutions, districts, and learners to collaborate and communicate toward what this future will look like as GenAI models evolve. Through this collaboration, GenAI use in educational settings can be leveraged while minimizing negative aspects like potential misconduct, data privacy, algorithmic bias, accuracy, and equity concerns. We argue that GenAI can play a valuable role in higher education settings to improve learner motivation and knowledge retrieval while facilitating workflows and content generation for instructors. This viewpoint explores GenAI's potential as an educational tool including alignment with learning theories (eg, behaviorism and cognitive load theory),

implications for learners and instructors (eg, flipped classrooms and self-directed assessments), responsible implementation (eg, bias and equity), and evolving challenges (eg, hallucinations and misconduct).

Acknowledgments

The authors would like to extend appreciation to professional and graduate students from the Department of Nutritional Sciences and College of Medicine Graduate Program in Biological Sciences for providing feedback and insights during spring and fall 2023 courses as we developed material presented.

Conflicts of Interest

None declared.

References

1. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887. [doi: 10.3390/healthcare11060887] [Medline: 36981544]
2. Arif TB, Munaf U, Ul-Haque I. The future of medical education and research: Is ChatGPT a blessing or blight in disguise? Med Educ Online 2023 Dec;28(1):2181052. [doi: 10.1080/10872981.2023.2181052] [Medline: 36809073]
3. Stretton B, Kovoor J, Arnold M, Bacchi S. ChatGPT-based learning: generative artificial intelligence in medical education. Med Sci Educ 2024 Feb;34(1):215-217. [doi: 10.1007/s40670-023-01934-5] [Medline: 38510403]
4. Gilbert TK. Generative AI and generative education. Ann N Y Acad Sci 2024 Apr;1534(1):11-14. [doi: 10.1111/nyas.15129] [Medline: 38512308]
5. Rodriguez DV, Lawrence K, Gonzalez J, et al. Leveraging generative AI tools to support the development of digital solutions in health care research: case study. JMIR Hum Factors 2024 Mar 6;11:e52885. [doi: 10.2196/52885] [Medline: 38446539]
6. Raza MM, Venkatesh KP, Kvedar JC. Generative AI and large language models in health care: pathways to implementation. NPJ Digit Med 2024 Mar 7;7(1):62. [doi: 10.1038/s41746-023-00988-4] [Medline: 38454007]
7. Montazeri M, Galavi Z, Ahmadian L. What are the applications of ChatGPT in healthcare: Gain or loss? Health Sci Rep 2024 Feb;7(2):e1878. [doi: 10.1002/hsr2.1878] [Medline: 38361810]
8. Tessler I, Wolfovitz A, Livneh N, et al. Advancing medical practice with artificial intelligence: ChatGPT in healthcare. Isr Med Assoc J 2024 Feb;26(2):80-85. [Medline: 38420977]
9. Aydın Ö, Karaarslan E. Is ChatGPT leading generative AI? What is beyond expectations? SSRN Journal 2023. [doi: 10.2139/ssrn.4341500]
10. Weidener L, Fischer M. Teaching AI ethics in medical education: a scoping review of current literature and practices. Perspect Med Educ 2023;12(1):399-410. [doi: 10.5334/pme.954] [Medline: 37868075]
11. Adams C, Pente P, Lemermeyer G, Rockwell G. Ethical principles for artificial intelligence in K-12 education. Comput Artif Intell 2023;4:100131. [doi: 10.1016/j.caeai.2023.100131]
12. Meece JL, Anderman EM, Anderman LH. Classroom goal structure, student motivation, and academic achievement. Annu Rev Psychol 2006;57:487-503. [doi: 10.1146/annurev.psych.56.091103.070258] [Medline: 16318604]
13. Morris LS, Grehl MM, Rutter SB, Mehta M, Westwater ML. On what motivates us: a detailed review of intrinsic v. extrinsic motivation. Psychol Med 2022 Jul;52(10):1801-1816. [doi: 10.1017/S0033291722001611] [Medline: 35796023]
14. Han K. Fostering students' autonomy and engagement in EFL classroom through proximal classroom factors: autonomy-supportive behaviors and student-teacher relationships. Front Psychol 2021;12:767079. [doi: 10.3389/fpsyg.2021.767079] [Medline: 34744946]
15. Bjork RA, Bjork EL. Desirable difficulties in theory and practice. J Appl Res Mem Cogn 2020;9(4):475-479. [doi: 10.1016/j.jarmac.2020.09.003]
16. Roediger HL, Agarwal PK, McDaniel MA, McDermott KB. Test-enhanced learning in the classroom: long-term improvements from quizzing. J Exp Psychol Appl 2011 Dec;17(4):382-395. [doi: 10.1037/a0026252] [Medline: 22082095]
17. Uzunboylu H, Karagozlu D. Flipped classroom: a review of recent literature. WJET 2015;7:142-147 [FREE Full text] [doi: 10.18844/wjet.v7i2.46]
18. Schiff D. Out of the laboratory and into the classroom: the future of artificial intelligence in education. AI Soc 2021;36(1):331-348. [doi: 10.1007/s00146-020-01033-8] [Medline: 32836908]
19. Ahern K. Teaching biochemistry online at Oregon State University. Biochem Mol Biol Educ 2017 Jan 2;45(1):25-30. [doi: 10.1002/bmb.20979] [Medline: 27228905]
20. Lee JJ, andHammer J. Gamification in education: What, how, why bother? Academic exchange quarterly 2011;15(2):1-5.
21. Dave M, Patel N. Artificial intelligence in healthcare and education. Br Dent J 2023 May;234(10):761-764. [doi: 10.1038/s41415-023-5845-2] [Medline: 37237212]
22. Kassab J, El Dahdah J, Chedid El Helou M, et al. Assessing the accuracy of an online chat-based artificial intelligence model in providing recommendations on hypertension management in accordance with the 2017 American College of

Cardiology/American Heart Association and 2018 European Society of Cardiology/European Society of Hypertension Guidelines. Hypertension 2023 Jul;80(7):e125-e127. [doi: 10.1161/HYPERTENSIONAHA.123.21183] [Medline: 37190998]
23. Heston TF, Khun C. Prompt engineering in medical education. Medicine and Pharmacology Preprint posted online on 2023. [doi: 10.20944/preprints202307.0813.v1]
24. Shunyu Yao DY, Zhao J, Shafran I, Griffiths TL, Cao Y, Narasimhan K. Tree of thoughts: deliberate problem solving with large language models. arXiv. Preprint posted online on 2023. [doi: 10.48550/arXiv.2305.10601]
25. Ray PP. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems 2023;3:121-154. [doi: 10.1016/j.iotcps.2023.04.003]
26. Zaghir J, Naguib M, Bjelogrlic M, Névéol A, Tannier X, Lovis C. Prompt engineering paradigms for medical applications: scoping review. J Med Internet Res 2024 Sep 10;26:e60501. [doi: 10.2196/60501] [Medline: 39255030]
27. Abhari S, Fatahi S, Saragadam A, Chumachenko D, Pelegrini Morita P. A road map of prompt engineering for ChatGPT in healthcare: a perspective study. Stud Health Technol Inform 2024 Aug 22;316:998-1002. [doi: 10.3233/SHTI240578] [Medline: 39176959]
28. Kaul V, Enslin S, Gross SA. History of artificial intelligence in medicine. Gastrointest Endosc 2020 Oct;92(4):807-812. [doi: 10.1016/j.gie.2020.06.040] [Medline: 32565184]
29. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomes AN, et al. Attention is all you need computation and language. arXiv. URL: https://arxiv.org/abs/1706.03762v7
30. Shamshad F, Khan S, Zamir SW, et al. Transformers in medical imaging: a survey. Med Image Anal 2023 Aug;88:102802. [doi: 10.1016/j.media.2023.102802] [Medline: 37315483]

Abbreviations

AI: artificial intelligence
GAI: generative artificial intelligence

Edited by B Lesselroth; submitted 05.04.24; peer-reviewed by A Das, C Lian; revised version received 11.11.24; accepted 02.01.25; published 11.03.25.

Please cite as:
Monzon N, Hays FA
Leveraging Generative Artificial Intelligence to Improve Motivation and Retrieval in Higher Education Learners
JMIR Med Educ 2025;11:e59210
URL: https://mededu.jmir.org/2025/1/e59210
doi:10.2196/59210

Quo Vadis, AI-Empowered Doctor?

Gary Takahashi*, MS, MD; Laurentius von Liechti*, BS; Ebrahim Tarshizi*, PhD

Shiley-Marcos School of Engineering, University of San Diego, 5998 Alcalá Park, San Diego, CA, United States

*all authors contributed equally

Corresponding Author:
Gary Takahashi, MS, MD
Shiley-Marcos School of Engineering, University of San Diego, 5998 Alcalá Park, San Diego, CA, United States

Abstract

In the first decade of this century, physicians maintained considerable professional autonomy, enabling discretionary evaluation and implementation of new technologies according to individual practice requirements. The past decade, however, has witnessed significant restructuring of medical practice patterns in the United States, with most physicians transitioning to employed status. Concurrently, technological advances and other incentives drove the implementation of electronic systems into the clinic, which these physicians were compelled to integrate. Health care practitioners have now been introduced to applications based on large language models, largely driven by artificial intelligence (AI) developers as well as established electronic health record vendors eager to incorporate these innovations. Although generative AI assistance promises enhanced clinical efficiency and diagnostic precision, its rapid advancement may potentially redefine clinical provider roles and transform workflows, as it has already altered expectations of physician productivity, as well as introduced unprecedented liability considerations. Recognition of the input of physicians and other clinical stakeholders in this nascent stage of AI integration is essential. This requires a more comprehensive understanding of AI as a sophisticated clinical tool. Accordingly, we advocate for its systematic incorporation into standard medical curricula.

(JMIR Med Educ 2025;11:e70079) doi:10.2196/70079

KEYWORDS

clinical medicine; artificial intelligence; large language models; decision support; AI; LLM; AI in medicine

Introduction

Artificial intelligence (AI) has demonstrated long-standing potential to fundamentally transform health care delivery. Prior to the emergence of large language models (LLMs) in the modern era, the implementation and advancement of AI applications were predominantly concentrated in domains such as diagnostic imaging and predictive analytics. These early efforts endeavored to provide decision support for clinicians in critical clinical contexts, such as sepsis identification and management. These implementations were not patient-facing, and these benefits were generally perceived as natural extensions of broader technological progress.

In contrast, today's interactive chat apps, showcasing advances in LLMs, are able to simulate sentient conversational speech, which has prompted a reconceptualization of AI capabilities. The proficiency of these systems to rapidly process and summarize relevant information from a vast collection of stored knowledge has sparked debates as to the potential of these models to exceed human cognitive performance in tasks requiring sophisticated clinical decision-making and interpretative analysis [1].

Heralded for its transformative potential, AI in medicine has promised to enhance administrative efficiency through the automation of repetitive and time-intensive processes, support doctors through improved diagnostic accuracy, meticulously reduce iatrogenic errors, facilitate personalized medicine tailored to individual patient characteristics, and enable clinicians to navigate the continually expanding corpus of medical research advances and evolving practice guidelines [2,3]. However, earlier initiatives to integrate AI into health care frameworks saw limited adoption, as clinicians remained unconvinced as to the technology's capacity to add substantive value in the clinical setting [4-6]. Technological constraints in computer vision and natural language processing impeded widespread clinical adoption of nascent AI applications, while evolving regulatory frameworks constituted significant barriers to commercialization [5].

Another significant factor impacting the trajectory of health care AI implementation was a shift in professional autonomy. Prior to the preceding decade, the medical profession within the United States operated with greater practitioner independence. Physicians unfamiliar with AI technology, or unconvinced of its practical advantages, had little incentive to incorporate the new technology into their workflow [7]. Notably, they were able to determine for themselves when and how best to invest in and implement AI into their medical practice. The contemporary practice landscape has since undergone significant transformation, as the majority of physicians have transitioned from autonomous ownership to employment relationships with

hospitals or other corporate health care systems [8]. This structural shift has profound implications for the implementation and governance of AI technologies in clinical settings, as employed health care professionals, unable to keep pace with these developments, risk marginalization as key stakeholders [9]. Their essential perspectives may be overlooked in critical decisions that will shape clinical workflows, promote work-life balance, and address professional burnout, ultimately redefining their intrinsic role in the health care system [10].

What Practicing Physicians Need to Understand Regarding the Role of LLMs

In the past year, multiple reports have highlighted the remarkable achievements of LLMs on medical knowledge tasks, often claiming accuracy near 100%, which surpasses human capability [11]. The benchmark testing panels used to evaluate these models have included datasets of clinical vignettes, urgent care encounters, and medical licensing or board exam datasets [12]. Such impressive results, widely publicized in both the general and industry media, have significantly influenced perceptions of medical AI capabilities compared with human practitioners [13].

The inadequacy of standard LLM evaluation metrics as grounds for physician workforce reduction has been comprehensively examined previously [14-18]. For example, the performance of medical LLMs is still dependent on the provision of pertinent clinical history information and salient features of the physical examination, and it is still not clear that this critical initial step in successfully identifying the nature of a medical condition can be adequately performed by an LLM. Automated techniques to acquire the clinical history by requiring that the user select from a predetermined menu of symptoms and descriptors may fail to capture nuanced empathetic human interaction, such as a sense of advocacy, caring, comfort, and dedication that emerges during genuine patient-provider encounters [19,20].

Although LLMs can demonstrate proficiency in tasks involving logic, reasoning, and assimilating large volumes of structured data, these models still lack essential clinical skills such as observation of a patient's demeanor, interpretation of nuanced nonverbal clues, and establishing rapport—competencies instinctively performed by a seasoned physician. Such limitations in basic sensorimotor and perceptual processing represent a manifestation of Moravec's paradox, a theoretical conundrum that poses formidable challenges to researchers investigating generative AI [21]. Simulated expressions of empathy and clinical judgment can still be perceived as superficial and scripted, precisely because their responses rely on predicted or pretrained responses, rather than authentic and experiential understanding of a patient's lived reality.

Limitations of LLM Capabilities

Physicians should understand that inference on LLMs is highly dependent on the data on which they have been trained. Details on specific dataset selection for model pretraining are proprietary knowledge, but many have been trained on datasets such as PubMed Central, MIMIC-III clinical notes, sanitized data from electronic health record interactions, and clinical practice guidelines [22,23]. These models undergo further fine-tuning on additional medical knowledge datasets as well as physician-patient dialog datasets [24]. As with any commercial deployments, medical LLMs must adhere to "continuous integration/continuous deployment" principles in machine learning operations, with monitoring to assure that the application dataset does not drift too far from the training dataset and that regular maintenance fine-tuning and dataset updating are performed [25].

Physicians should also be aware that LLMs, functioning as statistical pattern generators rather than verified information arbiters, generate outputs based on probabilistic distributions within their training data rather than through systematic verification of factual accuracy. Hallucinations remain problematic, afflicting even the latest reasoning models [26,27]. These confabulatory responses can be difficult for the clinician user to detect, creating a risk for their use in the clinic. Compounding this issue, it has been noted that references cited by LLMs to support their claims may themselves be hallucinatory [28].

Bayesian inference plays a significant role in the clinical application of LLMs in medical decision support. Despite having been trained on extensive medical corpora encompassing comprehensive clinicopathological knowledge, these models may exhibit deficiencies in appropriately weighting disease prevalence. The adage "when you hear hoofbeats, think horses, not zebras," reflects the experience of physicians that more common etiologies may present atypically and should still be prioritized. Current LLMs may still struggle in providing reasonable estimates of pretest disease probability, a skill that physicians acquire after years of clinical experience [29]. As a consequence, LLMs may disproportionately elevate rare conditions with close symptom concordance over more common diseases with partial clinical alignment [30]. LLMs may also fail to understand that the diagnostic process is dynamic and iterative, requiring ongoing refinement in response to emerging patient data revealed in subsequent encounters.

The Importance of Prompting

The role of system prompt customization in the efficacy of the physician-LLM interaction has been largely unexplored. Physicians may find benefit in interacting with an LLM that behaves like a trusted colleague, rather than a chatbot. Being able to manage the tone of an LLM might encourage a more exploratory and conversational interaction that lowers anxiety and stress, rather than isolated zero-shot querying as with a search engine. Strategic modifications to the system prompt can significantly influence model output, potentially resulting in divergent clinical management recommendations [31]. A demonstration of the efficacy of engineered prompting is the use of Medprompt and AutoMedPrompt, which invoke advanced techniques, such as chain-of-thought reasoning, *k*-nearest-neighbor–selected few-shot prompting, ensemble voting, and textual gradients, to extract high performance from generalist foundation models in standardized question-answer benchmarks, surpassing that of specialist models [32,33]. These

prompt enhancement techniques can yield impressive scores on multiple-choice question-answer datasets, such as MedQA-USMLE or PubMedQA. However, it is important to recognize that zero-shot (unassisted) performance on unstructured input is the more clinically relevant paradigm, an area where there is a comparative paucity of empirical performance data. A comprehensive study of various open-source models, including several that were fine-tuned on medical corpora, demonstrated that 1- to 3-shot prompting was requisite for optimal clinical language comprehension. The investigators concluded that while LLMs demonstrate proficiency in exam-style question-answer tasks with provided options, they exhibit significant limitations in open-ended scenarios [34].

Public LLMs typically restrict access to system prompting, but domain-specific consultative LLMs should offer this as a customization option. Currently, certain industry stakeholders regard proficiency in prompt engineering as "simply an expected skill," exemplifying a troublesome paradigm in which the vast majority of physicians, inadequately trained in this regard, are dependent on software developers to craft tools that physicians poorly understand [35]. Physicians should seek training to develop expertise in crafting suitable prompts to obtain the most relevant and suitably formatted information, while minimizing the likelihood of hallucinatory outputs [36,37].

LLM Performance Compared With Physician Performance

In addition to reports describing expert-level performance in question-answer multiple-choice testing, LLMs have been touted as being superior in the generation of differential diagnoses when presented with clinical vignettes [38]. These capabilities may stem from the models' capacity to recall factual information from their training corpora, rather than from any inherent ability to synthesize insight from a panoply of clinical indicators, as with human clinical reasoning [38]. For example, the performance of GPT-4 in identifying the diagnosis of published internal medicine cases was significantly decreased when challenged with unpublished clinical vignettes [39].

A recent systematic review and meta-analysis encompassing 83 studies across diverse models (including GPT-4, GPT-4o, PaLM2, and Perplexity, as well as open-source models fine-tuned in the medical domain) found that the pooled accuracy of the generative AI models was 52.1%, demonstrating no overall advantage over physician performance. The models were tested against a variety of clinical vignette datasets, as well as challenges posed in prominent medical journals. Notably, the performance of expert physicians was 15.8% higher, while nonexpert (resident) physicians maintained a marginal 0.6% advantage over LLMs [40].

A counterpoint to these observations, in a commentary highlighting 6 selected studies that examined the effectiveness of LLMs as diagnostic adjuncts, concluded that LLM assistance failed to enhance clinicians' diagnostic accuracy, with the models purportedly demonstrating superior performance on various assessment metrics [41]. We concur with the contention that claims of physician inferiority in these studies remain inconclusive, given methodological limitations, including an insufficient number of valid datapoints for robust comparison [42]. Nevertheless, it is readily apparent that physicians unaccustomed to AI-augmented workflows found LLM assistance unhelpful or counterproductive, especially when resolving discordant or ambiguous model outputs, which consumed valuable clinical time [43].

Physicians should also be cognizant of special legal ramifications regarding the use of AI for clinical decision support. The use of LLMs in patient care potentially exposes a clinician to novel vulnerabilities, broadly including model overreliance, inadequate appreciation of performance limitations, informed consent challenges, and potential bias with ethical ramifications [44,45]. These risks highlight the need for the robust regulatory oversight of LLM-based technology [46]. In litigation, physicians have been required to demonstrate adherence to a reasonable standard of care; however, these norms may evolve in response to transformative technologies [47]. In the event of an adverse outcome, physicians also risk penalization by juries whether or not an AI recommendation is accepted or overruled [48,49]. A rigorous discussion of the legal ramifications of using AI in clinical decision-making is beyond the scope of this viewpoint, but in light of the above considerations, the most prudent use of medical AI may be to confirm an existing medical decision, rather than as a means to augment care [50].

The Need for Active Physician Involvement in Shaping the Future of Generative AI in Health Care

Machine learning and generative AI will undoubtedly catalyze remarkable advancements in health care delivery, especially in clinic settings. These technological advances will undoubtedly exert differential impacts across medical specialties as advances in machine learning are increasingly leveraged to assist in image-processing tasks; however, they are unlikely to wholly replace the clinical expertise of physicians [51]. Indeed, Geoffrey Hinton, the "godfather of AI," was notoriously inaccurate as to his predictions regarding the demise of diagnostic radiology as a career [52]. We feel that health care providers will continue to play essential roles and that AI technology has the potential to augment the capabilities of physicians, nurses, pharmacists, and clinical researchers through the identification of more effective therapeutics and facilitation of novel technological innovations.

We also wish to emphasize, however, that the notion that a physician empowered by AI may outperform a doctor without this advantage may obscure deeper issues [53]. Near-term enhancements in AI-driven productivity gains may ultimately lead to its commoditization and may not necessarily translate to increased compensation, decreased burnout, or even job security [54]. In the early stages, physicians may even see an increased demand for their services (Jevons paradox) [55]; however, some warn that the augmentation or empowered role of health care providers may ultimately lead to a restructuring

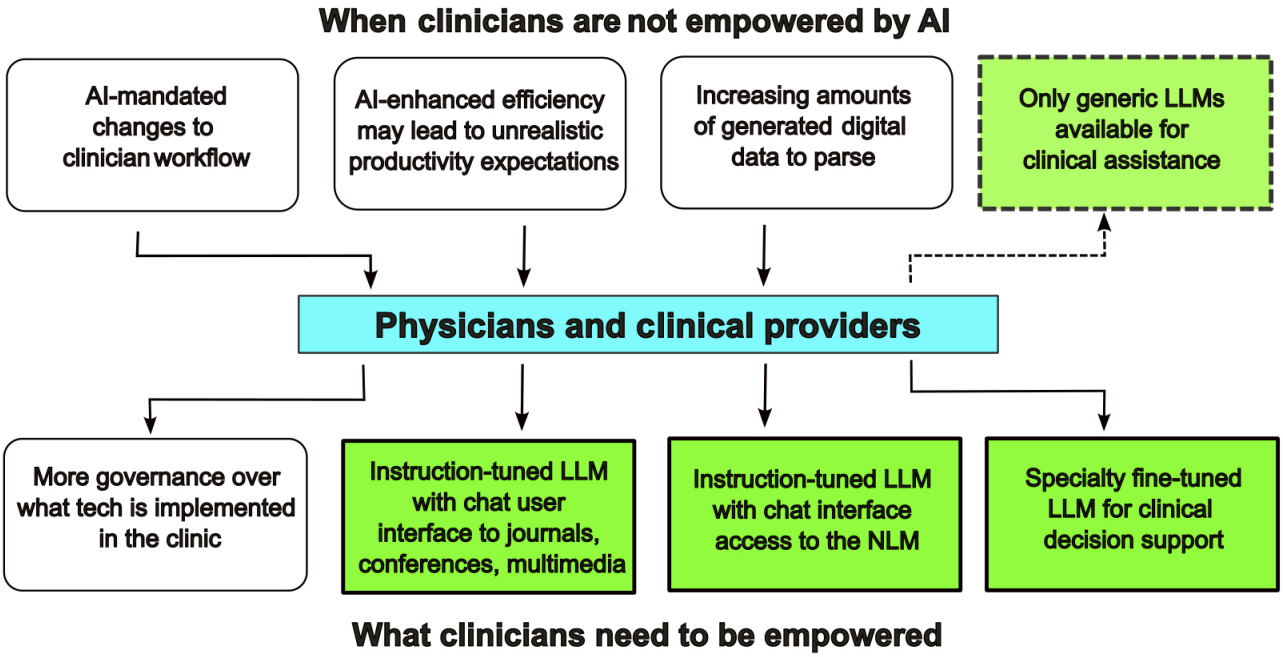
of the health care system. Patient intake and flow structures may be eventually redirected to meet the needs of third parties, such as insurers or hospital administrators, to prioritize revenue cycle management, or even to interface with other AI systems, such as those that seek to leverage actionable insights from outcomes data to guide evidence-based treatment recommendations. The adaptation of AI integration may reconfigure key decision-making in health care systems away from the employed physician to those whose priorities put greater weight on economic or political factors.

Physician input remains critically important in this process, especially in the transformative stages of AI integration into the clinic. We posit that the aforementioned structural shift in the physician employment landscape has significantly attenuated their influence as essential stakeholders and arbiters regarding technological implementation decisions [56]. Clinical practitioners should avoid defaulting to passive acceptance as institutionally procured software systems integrate AI technologies into their established clinical workflows.

Generative AI applications in medicine are still early in development, necessitating an approach that balances technological promotion with the practice-refined workflow of the clinical diagnostic process. The complexities of medical decision-making transcend simplistic evaluation through multiple-choice question-answering from medical datasets. Concern has already been raised that AI-based applications are being adopted too rapidly by hospitals eager to offer the latest in technological innovation, but without the necessary continuous oversight. Relying on the Food and Drug Administration to develop and regulate safeguards is not feasible [57]. A different approach, centered on the physician and accommodating the workflow requirements of the practitioner, will better foster physician-AI synergy [58,59]. Achieving this will require that physicians develop a deeper understanding of the workings of AI technology, comparable to their understanding of more traditional medical tools (Figure 1). We advocate for research initiatives exploring optimal physician-AI collaboration, potentially including practitioner proficiency in customizing LLM tools to address specific needs. Physicians with such expertise will be better able to advise regulatory bodies on establishing appropriate guardrails against potentially deleterious applications, privacy violations, and the perpetuation of bias and misinformation in health care contexts [60].

Furthermore, clinicians who are well-versed in the limitations of LLMs and related AI applications can provide essential expertise in medicolegal proceedings involving adverse clinical outcomes associated with AI utilization. Enhanced training in AI methodologies will equip physicians to critically evaluate medical research, which increasingly applies advanced data analytics in clinical settings. Such training will also enable physicians to contribute experiential insights and conduct rigorous critiques of machine learning applications designed to enhance predictive analytics. Actualization of these objectives necessitates comprehensive integration of AI education within the pathways of standard medical curricula [60].

Figure 1. Arrows indicate the direction of cause and effect or action initiated to its effect. Green shaded boxes indicate factors where the involvement of AI is direct. AI: artificial intelligence; LLM: large language model; NLM: National Library of Medicine.



Proposals for Physician Engagement in AI

As AI increasingly transforms health care delivery, physicians must proactively expand their expertise to include the following principles, ensuring responsible and effective integration of these technologies into clinical practice:

- Physicians should have some understanding of how deep learning models are trained and be aware of factors that can impact accuracy, such as dataset bias, covariate shift, out-of-distribution generalization, and concept drift.

- Physicians should understand how deep learning models are evaluated and, when possible, demand from software vendors the provenance of the datasets used for model training as well as performance metrics before they are introduced into the clinic.
- Physicians should understand the mechanism underlying LLMs; their intrinsic limitations and vulnerabilities; the impact of prompt engineering on output quality; and how to reduce hallucinatory behavior. Physicians should understand how to evaluate the capabilities of LLM models, as well as whether the information they generate will be exported and used for training purposes. Physicians should understand the ramifications of ambient LLM listening, for example, the custody and retention issues regarding the source recordings generated by AI scribes. These issues pertain to data privacy and confidentiality.
- Physicians should understand the potential ethical concerns intrinsic to how LLMs are trained, so as to minimize their perpetuation.
- Physicians should understand the legal ramifications of using LLMs as clinical diagnostic support. Physicians should recognize that medical LLMs function best when used adjunctively to validate evidence-based practice, rather than to generate novel treatments or be allowed to operate autonomously.
- Physicians should understand how privacy and confidentiality may be breached by incautious use of public LLM models.
- Physicians should develop sufficient understanding of clinical AI to be able to critique commercial software.
- Physicians should be able to educate and help train ancillary health care staff as to the proper use of AI technology, as well as to instill confidence in patients that such technology will be responsibly deployed.
- There should be greater physician participation in the development, validation, and implementation of clinical AI systems, tailored to local deployments.
- Physicians should collaborate with clinical informaticians throughout clinical AI implementation to ensure regulatory preparedness and compliance.

By embracing these essential AI competencies, physicians can maintain their central role in patient care while leveraging this technology to enhance clinical outcomes and preserve the integrity of the medical profession.

Conflicts of Interest

None declared.

References

1. Castagno S, Khalifa M. Perceptions of artificial intelligence among healthcare staff: a qualitative survey study. Front Artif Intell 2020;3(578983):578983. [doi: 10.3389/frai.2020.578983] [Medline: 33733219]
2. Bekbolatova M, Mayer J, Ong CW, Toma M. Transformative potential of AI in healthcare: definitions, applications, and navigating the ethical landscape and public perspectives. Healthcare (Basel) 2024 Jan 5;12(2):125. [doi: 10.3390/healthcare12020125] [Medline: 38255014]
3. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J 2021 Jul;8(2):e188-e194. [doi: 10.7861/fhj.2021-0095] [Medline: 34286183]
4. Hirani R, Noruzi K, Khuram H, et al. Artificial intelligence and healthcare: a journey through history, present innovations, and future possibilities. Life (Basel) 2024 Apr 26;14(5):557. [doi: 10.3390/life14050557] [Medline: 38792579]
5. Goldfarb A, Teodoridis F. Why is AI adoption in health care lagging? Brookings. 2022 Mar 9. URL: https://www.brookings.edu/articles/why-is-ai-adoption-in-health-care-lagging/ [accessed 2025-06-13]
6. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019 Oct 29;17(1):195. [doi: 10.1186/s12916-019-1426-2] [Medline: 31665002]
7. Arvai N, Katonai G, Mesko B. Health care professionals' concerns about medical AI and psychological barriers and strategies for successful implementation: scoping review. J Med Internet Res 2025 Apr 23;27(1):e66986. [doi: 10.2196/66986] [Medline: 40267462]
8. PAI-Avalere study on physician employment-practice ownership trends 2019-2023. Physicians Advocacy Institute. URL: https://www.physiciansadvocacyinstitute.org/PAI-Research/PAI-Avalere-Study-on-Physician-Employment-Practice-Ownership-Trends-2019-2023 [accessed 2025-05-14]
9. Hoffman J, Wenke R, Angus RL, Shinners L, Richards B, Hattingh L. Overcoming barriers and enabling artificial intelligence adoption in allied health clinical practice: a qualitative study. Digit Health 2025;11:20552076241311144. [doi: 10.1177/20552076241311144] [Medline: 39906878]
10. Wolfgruber DM. AI's healthcare revolution needs a human touch in 2025. Future Healthcare Today. 2025 Feb 18. URL: https://futurehealthcaretoday.com/ais-healthcare-revolution-needs-a-human-touch-in-2025/ [accessed 2025-05-14]
11. Wu K, Wu E, Wei K, et al. An automated framework for assessing how well LLMs cite relevant medical references. Nat Commun 2025 Apr 16;16(1):3615. [doi: 10.1038/s41467-025-58551-6] [Medline: 40240349]
12. Open Medical-LLM leaderboard – a Hugging Face space by openlifescienceai. Hugging Face. URL: https://huggingface.co/spaces/openlifescienceai/open_medical_llm_leaderboard [accessed 2025-05-14]
13. Rajpurkar P, Topol EJ. Opinion | the robot doctor will see you now. The New York Times. 2025 Feb 2. URL: https://www.nytimes.com/2025/02/02/opinion/ai-doctors-medicine.html [accessed 2025-05-14]

14. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. JAMA 2023 Sep 5;330(9):866-869. [doi: 10.1001/jama.2023.14217] [Medline: 37548965]
15. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA 2025 Jan 28;333(4):319-328. [doi: 10.1001/jama.2024.21700] [Medline: 39405325]
16. Raji ID, Daneshjou R, Alsentzer E. It's time to bench the medical exam benchmark. NEJM AI 2025 Jan 23;2(2):AIe2401235. [doi: 10.1056/AIe2401235]
17. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med 2024 Sep;30(9):2613-2622. [doi: 10.1038/s41591-024-03097-1] [Medline: 38965432]
18. Liu F, Zhou H, Hua Y, Rohanian O, Clifton L, Clifton DA. Large language models in healthcare: a comprehensive benchmark. medRxiv. Preprint posted online on Apr 25, 2024. [doi: 10.1101/2024.04.24.24306315]
19. Zakim D. Development and significance of automated history-taking software for clinical medicine, clinical research and basic medical science. J Intern Med 2016 Sep;280(3):287-299. [doi: 10.1111/joim.12509] [Medline: 27071980]
20. AI Patient Actor app – Thesen Laboratory. Dartmouth Geisel School of Medicine. URL: https://geiselmed.dartmouth.edu/thesen/patient-actor-app/ [accessed 2025-05-14]
21. LoAlza-Bonilla A. Moravec's paradox comes to the clinic. LinkedIn. 2024 Dec 31. URL: https://www.linkedin.com/pulse/moravecs-paradox-comes-clinic-arturo-loaiza-bonilla-md-lgvee [accessed 2025-05-18]
22. Zhou H, Liu F, Gu B, et al. A survey of large language models in medicine: progress, application, and challenge. arXiv. Preprint posted online on Nov 9, 2023. [doi: 10.48550/arXiv.2311.05112]
23. Zhang D, Xue X, Gao P, et al. A Survey of Datasets in Medicine for Large Language Models: Intell Robot OAE Publishing Inc; 2024, Vol. 4:457-478. [doi: 10.20517/ir.2024.27]
24. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature 2023 Aug;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
25. Wornow M, Xu Y, Thapa R, et al. The shaky foundations of large language models and foundation models for electronic health records. NPJ Digit Med 2023 Jul 29;6(1):135. [doi: 10.1038/s41746-023-00879-8] [Medline: 37516790]
26. Kim Y, Jeong H, Chen S, et al. Medical hallucinations in foundation models and their impact on healthcare. arXiv. Preprint posted online on Feb 26, 2025. [doi: 10.48550/arXiv.2503.05777]
27. OpenAI o3 and o4-mini system card. OpenAI. 2025 Apr 16. URL: https://openai.com/index/o3-o4-mini-system-card/ [accessed 2025-05-15]
28. Jaźwińska K, Chandrasekar A. AI search has a citation problem. Columbia Journalism Review. 2025 Mar 6. URL: https://www.cjr.org/tow_center/we-compared-eight-ai-search-engines-theyre-all-bad-at-citing-news.php [accessed 2025-05-15]
29. Gao Y, Myers S, Chen S, et al. Position paper on diagnostic uncertainty estimation from large language models: next-word probability is not pre-test probability. arXiv. Preprint posted online on Nov 7, 2024. [doi: 10.48550/arXiv.2411.04962]
30. A follow up on o1's medical capabilities + major concern about it's utility in medical diagnosis. Substack - Artificial Intelligence Made Simple. 2024 Sep 24. URL: https://artificialintelligencemadesimple.substack.com/p/a-follow-up-on-o-1s-medical-capabilities [accessed 2025-05-15]
31. Wang L, Chen X, Deng X, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. NPJ Digit Med 2024 Feb 20;7(1):41. [doi: 10.1038/s41746-024-01029-4] [Medline: 38378899]
32. Nori H, Lee YT, Zhang S, et al. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. arXiv. Preprint posted online on Nov 28, 2023. [doi: 10.48550/arXiv.2311.16452]
33. Wu S, Koo M, Scalzo F, Kurtz I. AutoMedPrompt: a new framework for optimizing LLM medical prompts using textual gradients. arXiv. Preprint posted online on Feb 21, 2025. [doi: 10.48550/arXiv.2502.15944]
34. Liu F, Li Z, Zhou H, et al. Large language models are poor clinical decision-makers: a comprehensive benchmark. In: Al-Onaizan Y, Bansal M, Chen YN, editors. Presented at: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; Nov 12-16, 2024; Miami, FL p. 13696-13710. [doi: 10.18653/v1/2024.emnlp-main.759]
35. Chandonnet H. "AI is already eating its own": prompt engineering is quickly going extinct. Fast Company. 2025 Jun 5. URL: https://www.fastcompany.com/91327911/prompt-engineering-going-extinct [accessed 2025-08-08]
36. Zaghir J, Naguib M, Bjelogrlic M, Névéol A, Tannier X, Lovis C. Prompt engineering paradigms for medical applications: scoping review. J Med Internet Res 2024 Sep 10;26:e60501. [doi: 10.2196/60501] [Medline: 39255030]
37. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res 2023 Oct 4;25:e50638. [doi: 10.2196/50638] [Medline: 37792434]
38. McDuff D, Schaekermann M, Tu T, et al. Towards accurate differential diagnosis with large language models. arXiv. Preprint posted online on Nov 30, 2023. [doi: 10.48550/arXiv.2312.00164]
39. Rutledge GW. Diagnostic accuracy of GPT-4 on common clinical scenarios and challenging cases. Learn Health Syst 2024 Jul;8(3):e10438. [doi: 10.1002/lrh2.10438] [Medline: 39036534]
40. Takita H, Kabata D, Walston SL, et al. A systematic review and meta-analysis of diagnostic performance comparison between generative AI and physicians. NPJ Digit Med 2025 Mar 22;8(1):175. [doi: 10.1038/s41746-025-01543-z] [Medline: 40121370]
41. Topol E, Rajpurkar P. When doctors with A.I. are outperformed by A.I. Substack - Ground Truths. 2025 Feb 2. URL: https://erictopol.substack.com/p/when-doctors-with-ai-are-outperformed [accessed 2025-05-15]

42. Polevikov S. The “AI outperforms doctors” claim is false, despite NYT story - a rebuttal (part 2 of 6). Substack - AI Health Uncut. 2024 Nov 21. URL: https://sergeiai.substack.com/p/the-ai-outperforms-doctors-claim [accessed 2025-05-15]
43. Agarwal N, Moehring A, Rajpurkar P, Salz T. Combining human expertise with artificial intelligence: experimental evidence from radiology. National Bureau of Economic Research. 2023 Jul. URL: https://www.nber.org/papers/w31422 [accessed 2025-08-08]
44. Arvai N, Katonai G, Mesko B. Health care professionals’ concerns about medical AI and psychological barriers and strategies for successful implementation: scoping review. J Med Internet Res 2025 Apr 23;27:e66986. [doi: 10.2196/66986] [Medline: 40267462]
45. Jones C, Thornton J, Wyatt JC. Artificial intelligence and clinical decision support: clinicians’ perspectives on trust, trustworthiness, and liability. Med Law Rev 2023 Nov 27;31(4):501-520. [doi: 10.1093/medlaw/fwad013] [Medline: 37218368]
46. Weissman GE, Mankowitz T, Kanter GP. Unregulated large language models produce medical device-like output. NPJ Digit Med 2025 Mar 7;8(1):148. [doi: 10.1038/s41746-025-01544-y] [Medline: 40055537]
47. FSMB releases recommendations on the responsible and ethical incorporation of AI into clinical practice. Federation of State Medical Boards. 2024 May 2. URL: https://www.fsmb.org/advocacy/news-releases/fsmb-releases-recommendations-on-the-responsible-and-ethical-incorporation-of-ai-into-clinical-practice/ [accessed 2025-05-17]
48. Appel JM. Artificial intelligence in medicine and the negative outcome penalty paradox. J Med Ethics 2024 Dec 23;51(1):34-36. [doi: 10.1136/jme-2023-109848] [Medline: 38290853]
49. Patil SV, Myers CG, Lu-Myers Y. Calibrating AI reliance-a physician’s superhuman dilemma. JAMA Health Forum 2025 Mar 7;6(3):e250106. [doi: 10.1001/jamahealthforum.2025.0106] [Medline: 40116804]
50. Price WN, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. JAMA 2019 Nov 12;322(18):1765-1766. [doi: 10.1001/jama.2019.15064] [Medline: 31584609]
51. Wolfe D. How physicians are vulnerable to AI. Healthcare Recruiting. 2025 Apr 29. URL: https://www.npnow.com/how-physicians-are-vulnerable-to-ai/ [accessed 2025-06-15]
52. Stempniak M. NY times revisits nobel prize winner’s prediction AI will render radiologists obsolete. Radiology Business. 2025 May 15. URL: https://radiologybusiness.com/topics/artificial-intelligence/ny-times-revisits-nobel-prize-winners-prediction-ai-will-render-radiologists-obsolete [accessed 2025-08-08]
53. Choudary SP. The many fallacies of “AI won’t take your job, but someone using AI will”. Substack - Platforms, AI, and the Economics of BigTech. 2025 Apr 13. URL: https://platforms.substack.com/p/the-many-fallacies-of-ai-wont-take [accessed 2025-08-08]
54. Kim BJ, Lee J. The mental health implications of artificial intelligence adoption: the crucial role of self-efficacy. Humanit Soc Sci Commun 2024 Nov 17;11(1):1-15. [doi: 10.1057/s41599-024-04018-w]
55. Nguyen B. Will AI really lighten the load in allied health? navigating the jevons paradox. LinkedIn. 2025 Jan 15. URL: https://www.linkedin.com/pulse/ai-really-lighten-load-allied-health-navigating-jevons-nguyen-pvjnc [accessed 2025-05-19]
56. Five key trends driving purchasing decisions in healthcare IT. Signify Research. 2023 Mar 13. URL: https://www.signifyresearch.net/insights/five-key-trends-driving-purchasing-decisions-in-healthcare-it/ [accessed 2025-05-15]
57. Lenharo M. Medicine’s rapid adoption of AI has researchers concerned. Nature New Biol 2025 Jun 9. [doi: 10.1038/d41586-025-01748-y] [Medline: 40490519]
58. Henry T. Physicians’ greatest use for AI? Cutting administrative burdens. American Medical Association. 2025 Mar 20. URL: https://www.ama-assn.org/practice-management/digital-health/physicians-greatest-use-ai-cutting-administrative-burdens [accessed 2025-08-08]
59. Lohr S. A.i. was coming for radiologists’ jobs. So far, they’re just more efficient. The New York Times. 2025 May 14. URL: https://www.nytimes.com/2025/05/14/technology/ai-jobs-radiologists-mayo-clinic.html [accessed 2025-05-16]
60. Schuitmaker L, Drogt J, Benders M, Jongsma K. Physicians’ required competencies in AI-assisted clinical settings: a systematic review. Br Med Bull 2025 Jan 16;153(1):ldae025. [doi: 10.1093/bmb/ldae025] [Medline: 39821209]

Abbreviations

AI: artificial intelligence
LLM: large language model

Edited by B Lesselroth; submitted 14.12.24; peer-reviewed by A Benitez, E Markus, MV Ebrahim; revised version received 17.06.25; accepted 25.07.25; published 15.08.25.

Please cite as:
Takahashi G, von Liechti L, Tarshizi E
Quo Vadis, AI-Empowered Doctor?
JMIR Med Educ 2025;11:e70079
URL: https://mededu.jmir.org/2025/1/e70079
doi:10.2196/70079

Shaping the Future of Digital Health Education in Canada: Prioritizing Competencies for Health Care Professionals Using the Quintuple Aim

Glynda Rees[1,2*], MSN; Lorelli Nowell[3*], PhD; Tracie Risling[3*], PhD

[1]School of Health Sciences, British Columbia Institute of Technology, 3700 Willingdon Avenue, Burnaby, BC, Canada

[2]Doctoral Student, Faculty of Nursing, University of Calgary, Calgary, AB, Canada

[3]Faculty of Nursing, University of Calgary, Calgary, AB, Canada

[*]all authors contributed equally

Corresponding Author:
Glynda Rees, MSN
School of Health Sciences, British Columbia Institute of Technology, 3700 Willingdon Avenue, Burnaby, BC, Canada

Abstract

The integration of digital health and informatics competencies into health care education in Canada is essential for preparing a workforce capable of leveraging health care technologies to enhance care delivery and patient outcomes. Despite significant advancements, the current educational landscape in digital health remains inconsistent, characterized by fragmented curricula and uneven competency attainment. Addressing these gaps requires an innovative reframing of digital health competencies guided by a robust, outcomes-oriented framework. These authors propose the Quintuple Aim as an effective framework for outlining and organizing digital health and informatics competencies, focusing simultaneously on improving patient experience, enhancing population health, reducing health care costs, improving health care provider experience, and advancing health equity. Each dimension of the Quintuple Aim provides a critical lens for identifying, prioritizing, and contextualizing core competencies. Within the "patient experience" aim, competencies prioritize patient-centered technology use, including digital literacy, privacy awareness, and the ability to empower patients through technology. "Healthcare provider experience" competencies prioritize usability, workflow integration, and strategies to mitigate technology-related burnout. Under "population health," competencies emphasize data-driven decision-making, analytics, and health informatics to support effective public health interventions. Competencies associated with "cost reduction" focus on operational efficiency, resource optimization, and economic evaluation of digital health solutions. Finally, "health equity" competencies emphasize inclusivity, cultural safety, and the elimination of digital divides, ensuring equitable access to digital health technologies. Potential assessment strategies aligned with each competency area are highlighted, emphasizing formative and summative evaluations that include simulation-based assessments, real-world technology integration projects, and reflective practice portfolios. By applying the Quintuple Aim as a guiding structure, digital health education can achieve greater standardization, clarity, and alignment with health care system needs, while simultaneously allowing for tailored adaptations responsive to specific regional and institutional priorities. This paper introduces the Quintuple Aim as a guiding framework to comprehensively identify and organize core digital health and informatics competencies for health professional education.

(JMIR Med Educ 2025;11:e75904) doi:10.2196/75904

KEYWORDS

digital health; health informatics; competency-based education; Quintuple Aim

Introduction

The rapid advancement of digital health technologies necessitates a health care workforce equipped with the requisite knowledge, skills, and abilities to effectively integrate and use these technologies [1-3]. Educating the Canadian health care workforce in digital health is essential to enhance patient care and safety, improve health care accessibility, and ensure the sustainability of health care systems. The International Council of Nurses, the World Health Organization (WHO), and the Organisation for Economic Co-operation and Development (OECD) emphasize the importance of digital literacy for health care providers to effectively leverage technology and adapt to evolving health care challenges [4-7]. Despite the increasing importance of digital health and health informatics, there is a significant gap in the number of educational programs offered across Canada that prepare health care professionals for these challenges [2,8].

There is a recognized and acute need for more digital health professionals who bring informatics knowledge, digital health understanding, and data analytics skills to health care settings

[2,5,8-12]. Terms essential to this discussion include digital health, health informatics, and competency. Digital Health refers to the use of technologies for improving the health and well-being of people at individual and population levels, as well as enhancing the care of patients through intelligent processing of clinical and genetic data and leveraging digital health technologies and services to transform care delivery [5,7,13,14]. The Canadian Institutes of Health Research cite Canada Health Infoway's definition of Digital Health as "the use of information technology or electronic communication tools, services and processes to deliver health care services or to facilitate better health" [15]. Health Informatics is a focused area within digital health and is defined by the American Medical Informatics Association as "the science of how to use data, information and knowledge to improve human health and the delivery of health care services" [16]. A competency refers to defined knowledge, skills, and qualities expected to competently carry out a function [17]. In the context of health care, the term competency is defined as "an observable ability of a health professional, integrating multiple components such as knowledge, skills and attitudes" [18]. Informatics competencies are integral to the broader concept of digital health and should be integrated into an educational digital health program [1,2,8]. For the purposes of this paper, digital health is used as an umbrella term that includes health informatics competencies alongside broader digital technologies and services that enhance health system performance. These competencies are designed for health care professionals seeking digital health education to develop the knowledge and skills necessary for effectively integrating digital technologies into health care practice.

Current State of Digital Health Education

Despite multiple organizations highlighting the need for the health care workforce to acquire and maintain the knowledge and skills to navigate technologically advancing health care systems [2,5,9-12], current Canadian postsecondary educational programs often lack the comprehensive curriculum, faculty expertise, and resources necessary to equip graduates with these competencies [2,8]. This gap between the evolving digital health landscape and the preparedness of health care professionals can potentially lead to inefficiencies, reduced quality of care, increased clinician workload, and missed opportunities for innovation within health care settings [9,10].

If the gap in digital health education remains unresolved, the health care workforce will continue to struggle with the integration and use of digital health technologies, which are increasingly vital for modern health care delivery. The International Council of Nurses highlights that digital health is integral to enhancing care quality and patient outcomes, making it essential for nurses to be proficient in these technologies [5]. The WHO also stresses that a well-prepared workforce is crucial to achieving the full potential of digital health innovations globally [7]. Without adequate education, the health care sector risks failing to capitalize on digital evolutions as noted by Socha-Dietrich [6], whose OECD report emphasizes the importance of empowering health care professionals to adapt to these advancements. Furthermore, the 2023 Canadian Survey of Nurses by Canada Health Infoway reveals that gaps in digital health proficiency can directly impact the quality of care provided [19].

Prioritizing Digital Health and Informatics Competencies

Digital health technologies are transforming health care delivery, requiring a shift in education to equip health care professionals with essential digital health and informatics competencies. Effective digital health educational programs typically include components such as data analytics, health information system interoperability, privacy and ethical considerations, and health informatics theory [20]. These curricula also emphasize skills such as leadership, communication, project management, and change management, all of which are critical in implementing and managing digital health initiatives [21]. Studies have shown that graduates of digital health programs often demonstrate improved competencies in using digital health technologies, which translates to better patient outcomes and more efficient health care delivery [10,20-22].

The Quintuple Aim as a Framework

The Quintuple Aim is a framework for health care improvement that aims to make health care more equitable, effective, and efficient [23]. In the evolving landscape of health care, the integration of digital health and informatics competencies plays a critical role in achieving the goals outlined in the Quintuple Aim: enhancing patient experience, improving care team well-being, reducing costs, advancing population health, and promoting health equity [23,24]. These competencies empower providers to harness technology and data-driven insights, ultimately enhancing patient care and health outcomes. By aligning digital health and informatics competencies with the Quintuple Aim, health care professionals can better navigate and contribute to a system that prioritizes both efficiency and equity.

Many researchers and organizations have compiled lists of informatics competencies for health care professionals in various roles in the health care system [20-22,25-29,30,31]. In addition, several researchers have developed frameworks to guide the organization, prioritization, and context of the competencies [21,22,32-36]. These authors suggest organizing and combining these competencies into the 5 categories outlined in the Quintuple Aim (patient experience, provider experience, reduced cost, population health, and health equity). This competency framework supports clinicians, health care administrators, informatics specialists, and other professionals preparing for roles in digital health. These competencies are not only confined to single objectives but intersect across multiple areas of the Quintuple Aim. As health care systems increasingly adopt advanced digital solutions, it becomes evident that these competencies do not function in isolation; rather, they create synergies that benefit both individual and population-level outcomes. Understanding these overlaps is essential for designing education programs and initiatives that prepare health care providers to meet the complex and interconnected needs of today's health care environments. Table 1 maps each of the

Quintuple Aims to digital health competencies as well as potential assessment strategies.

Table . Mapping digital health competencies to the Quintuple Aim Framework.

Quintuple Aim and competency	Assessment strategies
Patient experience	
Integrating digital health technologies into care	Practical simulations, case studies, and reflective assignments around patient management, communication, and decision-making, ethical, privacy, and interoperability concerns.
Applying and understanding user-centered design	Project-based evaluations to design, test, and refine digital health solutions based on patient needs, usability, and accessibility.
Patient-centered digital literacy and virtual care skills	Role play scenarios and patient simulations to communicate via digital platform and ensure patients can navigate virtual care services.
Ethical practice	Case-based discussions, ethical dilemma scenarios, and reflective essays around privacy, data security, informed consent, and responsible use of technology.
Communication skills	Virtual patient consultations and collaborative team simulations to explain digital health tools, address patient concerns, and facilitate effective communication through digital platforms.
Genomics competencies	Case studies, quizzes, or practical exercises to interpret genetic data, apply genomic information in clinical decision-making, and use digital tools for personalized patient care.
Provider experience	
Interoperability principles	Scenario-based assignments and systems integration projects to navigate and connect various digital health platforms, ensuring seamless data exchange and collaboration across health care settings.
User-centered design and workflow optimization	Hands-on projects to design digital health solutions that prioritize user experience, streamline clinical processes, and demonstrate the ability to enhance efficiency and usability within health care environments.
Critical thinking and evaluation	Analytical case studies to critically assess the effectiveness of digital health tools, identify potential areas for improvement, and evaluate their impact on patient care and health care workflows.
Clinical decision support and evidence-based practice	Simulations and practical examinations to use digital tools to make data-driven decisions, incorporate the latest clinical guidelines, and evaluate the outcomes of their decisions in patient care scenarios.
Reduced cost	
Efficient use of digital health technologies	Real-world simulations and timed exercises to quickly and accurately navigate digital health systems, manage patient data, and complete tasks with minimal errors and maximum efficiency.
Cybersecurity, privacy, and ethical use of digital health technologies	Written examinations, case studies, and scenario-based discussions to test knowledge of data protection laws, ethical guidelines, and strategies for safeguarding patient information in digital health environments.
Digital health ecosystems and interoperability	Group projects to analyze and design solutions that demonstrate how different digital health platforms work together.
Data analysis	Data-driven case studies and analytical projects to interpret health care data, identify trends, and make recommendations.
Population health	
Data management and analysis	Projects to analyze large data sets to identify trends and evaluate strategies.
Adaptability to emerging technologies	Hands-on exercises and simulations to learn and integrate new tools into clinical practice.
Leadership, change management, and project management skills	Team-based projects to plan, execute, and evaluate implementation of digital health solutions.
Health equity	
Addressing digital health disparities	Case studies and community-based projects to identify barriers to digital health access, propose strategies for increasing inclusivity, and evaluate the impact of their solutions on underserved populations.

Quintuple Aim and competency	Assessment strategies
Health data access	Practical exercises and written assignments to analyze the ethical, legal, and technical aspects of accessing and sharing health data, ensuring compliance with regulations while considering patient privacy and data security.
Indigenous data sovereignty	Case studies and reflective essays that analyze the importance of respecting Indigenous peoples' control over their health data, explore ethical considerations, and demonstrate how to apply culturally appropriate practices in digital health.
Health literacy and patient education	Role-playing exercises and patient simulation scenarios to effectively communicate digital health information, assess patient comprehension, and tailor education to diverse literacy levels and needs.

Patient Experience

Overview

The patient experience category of the Quintuple Aim focuses on improving the quality, accessibility, and personalization of care to enhance patient satisfaction and outcomes. Digital health technologies play a crucial role in advancing this goal by enabling virtual care, mobile health apps, wearable devices, and patient portals that facilitate seamless communication between patients and providers. These tools empower individuals to actively engage in their care, access real-time health information, and receive timely interventions, ultimately leading to more patient-centered, efficient, and responsive health care experiences.

Integrating Digital Health Technologies Into Care

The ability to effectively translate digital health knowledge and theory into practical applications within patient care settings is essential to enhancing the patient experience with digital health technologies. Health care students should be supported in developing proficiency in using digital health technologies and applications for patient care, identifying appropriate patient education materials, and acting as advocates for system users [2,37]. Competencies in using electronic health records for comprehensive documentation, effective communication among care providers, and making evidence-informed decisions are needed in today's health care workforce [4,5,28]. Health care students should be adept at leveraging virtual care platforms for remote consultations and patient monitoring, expanding access to care and addressing health care disparities [10,25,26]. In addition, understanding how digital tools can enhance patient engagement and empower individuals to actively participate in their care and self-management of chronic conditions is essential [11]. Understanding the functionalities and interconnectedness of health information systems, virtual care platforms, mobile health apps, and emerging technologies like artificial intelligence (AI) and robotics is fundamental [1,3,38-40]. Furthermore, understanding the capabilities and limitations of these technologies is essential for their appropriate and effective application in providing quality patient care. If digital health technologies are not effectively integrated into patient care, patients may face reduced access to timely information, fragmented communication with providers, and lower engagement in their own health management, ultimately compromising their overall health care experience. The competency of integrating digital health technologies into care could be assessed through practical simulations, case studies, and reflective assignments that evaluate health care students' competency in using digital tools for patient management, communication, and decision-making, while also considering their understanding of ethical, privacy, and interoperability concerns (Table 1).

Applying and Understanding User-Centered Design

Informatics competencies that enhance patient-centered care focus on human-centered design, usability principles, and patient engagement strategies within digital health solutions. Students should learn to critique digital health technologies, such as patient portals, telehealth and virtual care systems, and mobile health apps, to ensure they are user-friendly, accessible, and culturally sensitive [41,42]. Training in health literacy principles and user-centered design methodologies can foster empathy by deepening health care students' understanding of patient needs, ensuring that digital health solutions are accessible, inclusive, and tailored to diverse populations. If health care professionals understand and apply user-centered design, it will make it easier for patients to navigate and engage with health technologies, ultimately leading to improved patient satisfaction [43]. If user-centered design is not applied when integrating digital health technologies, patients may struggle with inaccessible, confusing, or nonintuitive systems, leading to frustration, decreased engagement, and potential disparities in health care access and outcomes. Health care students' abilities to apply and understand user-centered design in relation to digital health technologies could be assessed through project-based evaluations where students design, test, and refine digital health solutions, considering patient needs, usability, and accessibility to ensure the technology meets the diverse requirements of end users (Table 1).

Patient-Centered Digital Literacy and Virtual Care Skills

For patients and caregivers, digital health literacy is crucial for understanding and managing their own health conditions. With the advent of virtual care, online patient portals, and mobile health apps, patients are increasingly expected to engage with digital platforms [44]. The increased adoption of virtual care, accelerated by the COVID-19 pandemic, underscores the need for health care professionals to be proficient in patient-centered digital literacy [45,46]. This includes using virtual care platforms and mobile health apps effectively to engage with patients

remotely [47]. Health care professionals are often required to support patients in using digital health technologies to manage their health [2,48]. Patient-centered digital literacy and virtual care competencies are particularly important as virtual care has become a staple in health care delivery, requiring providers to ensure equitable access and appropriate care via digital means [42,45,46,49]. A lack of digital health literacy can lead to misunderstandings, nonadherence to treatment plans, and ultimately poorer health outcomes [50]. Competencies in patient-centered digital literacy and virtual care skills could be assessed through role-playing scenarios, patient simulations, and assessments that measure health care students' ability to effectively communicate with patients via digital platforms, educate them about digital tools, and ensure patients can navigate virtual care services with confidence and ease (Table 1).

Ethical Practice

Ethical training on the appropriate use of digital tools is necessary to navigate complex privacy issues and ensuring health care professionals are equipped with these skills fosters secure and ethical practices in digital health environments [27,51]. In addition, competencies in data privacy, confidentiality, and security are essential, as safeguarding patient data contributes to clinicians' confidence and comfort in using digital health systems, thereby supporting a safer work environment [52]. Furthermore, integrating ethical frameworks into digital health education can help health care professionals recognize and address biases, promote equitable access to technology, and uphold patient autonomy in an increasingly digital health care landscape. If ethical practices are not incorporated into digital health education, there is a risk of compromising patient privacy, consent, and autonomy, which can undermine trust in health care systems and negatively impact patient experiences [52]. Ethical practice competencies may be assessed through case-based discussions, ethical dilemma scenarios, and reflective essays that evaluate health care students' understanding of privacy, data security, informed consent, and the responsible use of technology in patient care (Table 1).

Communication Skills

Strong communication skills are critical for the effective use of digital health technologies to interact with patients, explain complex information, and provide clear instructions [51]. For instance, explaining the use of health information technologies, providing educational resources on conditions, treatments, procedures, or guiding patients on using digital health apps requires excellent communication skills. Health care professionals must also be trained in communication competencies, such as digital compassion and patient advocacy, so as to better facilitate positive experiences for patients [20,21,53,54]. These competencies are particularly crucial in virtual care, where fostering patient trust and engagement without face-to-face interaction demands a specialized skill set [45,46]. If communication skills are not incorporated into digital health education, patients may experience a lack of emotional connection with health care providers, leading to diminished patient satisfaction and outcomes. Communication skills may be assessed through interactive exercises, such as virtual patient consultations or collaborative team simulations, where health care students are evaluated on their ability to clearly explain digital health tools, address patient concerns, and facilitate effective communication through digital platforms (Table 1).

Genomics Competencies

Genomic competencies empower health care providers to deliver more precise and personalized care, identifying genetic factors that influence disease risk, medication responses, and treatment options [41]. By integrating genomic data into clinical decision-making, practitioners can tailor interventions to each patient's unique genetic profile, potentially improving the efficacy of treatments and reducing adverse reactions to medications [41,55,56]. As genomic technology continues to evolve, training in this area not only supports patient outcomes but also contributes to more efficient, cost-effective, and equitable health care by reducing trial-and-error in treatments and focusing on individualized approaches [55-58]. Without adequate training in genomics, health care professionals may struggle to interpret complex genetic data, limiting their ability to leverage these advancements in clinical practice and potentially widening disparities in access to precision medicine. Genomics competencies may be assessed through case studies, quizzes, and practical exercises that test health care students' ability to interpret genetic data, apply genomic information in clinical decision-making, and use digital tools for personalized patient care (Table 1).

Provider Experience

Overview

The provider experience category of the Quintuple Aim emphasizes the well-being, satisfaction, and efficiency of health care professionals, recognizing that a supported workforce is essential for high-quality patient care. Digital health technologies contribute to this goal by streamlining workflows, reducing administrative burdens, and enhancing communication through electronic health records, clinical decision support systems, and telehealth platforms. These tools can help mitigate burnout by improving efficiency, reducing repetitive tasks, and allowing providers to focus more on patient care. When effectively integrated, digital health solutions potentially foster a more sustainable and rewarding work environment for health care professionals.

Interoperability Principles

Understanding interoperability standards to facilitate enhanced care coordination can help leverage the full potential of digital health technologies [9]. This encompasses an understanding of interoperability and data standards, enabling the seamless exchange of health information between different systems and care settings [59]. As scopes of practice shift and we are faced with workforce shortages and the emergence of new roles, having access to comparable, sharable data derived from the use of clinical data standards has the potential to provide support for human resource and health policy decisions at local, regional, and national levels [60,61]. As we plan for a future where AI is foundational to facilitate evidence-informed decision-making, it is vital that we have standardized clinical data to inform care

decisions and resource allocation. If interoperability principles are not incorporated into digital health education, health care providers may face fragmented systems, difficulty accessing patient information across platforms, and increased administrative burden, all of which can hinder clinical decision-making and job satisfaction [52]. Health care students' understanding and use of interoperability principles could be assessed through scenario-based assignments and system integration projects where they demonstrate their ability to navigate and connect various digital health platforms, ensuring seamless data exchange and collaboration across health care settings (Table 1).

User-Centered Design and Workflow Optimization

Designing user-friendly digital health systems that minimize cognitive load and enhance workflow efficiency can significantly improve provider well-being [34]. Digital health education programs should include principles of human-computer interaction, usability testing, and ergonomics to ensure systems are intuitive and supportive of provider workflows [4,5,28,34]. Digital health systems generate numerous alerts and notifications, potentially contributing to provider burnout. Training can address strategies for managing alert fatigue, including customization, prioritization, and effective communication protocols to minimize unnecessary interruptions and streamline information flow [62].

Key competencies also include ergonomics, cognitive load theory, and human factors in technology design. These informatics skills empower health care providers to design and implement systems that minimize administrative burden, optimize usability, and reduce unnecessary interruptions during clinical workflows [63]. For example, well-designed electronic health record interfaces that prioritize intuitive navigation and minimize excessive documentation can significantly reduce cognitive overload, allowing providers to focus more on direct patient care.

Training in workflow optimization and data automation is also vital, as these competencies enable professionals to implement time-saving measures within electronic health record systems and other digital health tools. Reducing the time spent on repetitive or low-value tasks allows clinicians to dedicate more time to patient care, which can enhance job satisfaction and reduce burnout [63]. In addition, an understanding of implementation science, which guides the successful integration of new technologies into health care systems, may help optimize workflow by identifying barriers to technology adoption, refining processes to enhance efficiency, and ensuring digital tools seamlessly integrate into clinical practice [10]. If user-centered design and workflow optimization are not incorporated into digital health education, health care providers may encounter inefficient, cumbersome systems that increase cognitive load, disrupt clinical workflows, and contribute to burnout and frustration. User-centered design and workflow optimization could be assessed through hands-on projects where health care students design digital health solutions that prioritize user experience, streamline clinical processes, and demonstrate the ability to enhance efficiency and usability within health care environments (Table 1).

Critical Thinking and Evaluation

The rapid evolution of the digital health landscape necessitates the ability to critically evaluate new tools, discern their potential benefits and limitations, and assess their impact on health care delivery and outcomes [48,64]. This competency involves building a strong theoretical understanding of health informatics, staying abreast of emerging technologies, recognizing their strengths and weaknesses, and developing the ability to translate data and research findings into evidence-informed practices. A strong understanding of data science and bioinformatics can further enhance the ability to critically analyze health data and contribute to the development of novel digital health solutions [42]. If critical thinking and evaluation are not incorporated into digital health education, health care providers may struggle to assess the effectiveness of digital tools, leading to suboptimal decision-making, reduced confidence in technology, and potential negative impacts on patient care. Critical thinking and evaluation skills could be assessed through analytical case studies, where health care students critically assess the effectiveness of digital health tools, identify potential areas for improvement, and evaluate their impact on patient care and health care workflows (Table 1).

Clinical Decision Support and Evidence-Based Practice

Digital technologies that enhance evidence-informed practice, such as clinical decision support systems, are becoming central to health care delivery. Educating clinicians on the use and optimization of these tools helps improve diagnostic accuracy, decision-making, and treatment planning [20]. Furthermore, familiarity with decision support technologies enables health care professionals to access real-time evidence, which is crucial for safe, high-quality care [20]. Failing to advance the use of clinical decision support and evidence-based practice using digital technologies may lead to outdated treatments, increased medical errors, and missed opportunities for improving patient outcomes through data-driven insights. Clinical decision support and evidence-based practice skills may be assessed through simulations or practical examinations where health care students use digital tools to make data-driven decisions, incorporate the latest clinical guidelines, and evaluate the outcomes of their decisions in patient care scenarios (Table 1).

Reduced Cost

Overview

The reduced costs category of the Quintuple Aim focuses on lowering health care expenses while maintaining or improving the quality of care. Digital health technologies contribute significantly to cost reduction by streamlining administrative tasks, reducing hospital readmissions, and enabling remote monitoring through telehealth and mHealth apps. By automating routine processes, improving care coordination, and providing patients with more proactive management of chronic conditions, these technologies can minimize the need for costly in-person visits and prevent unnecessary hospitalizations. Ultimately, the integration of digital tools could create a more efficient and financially sustainable health care system.

Efficient Use of Digital Health Technologies

Proficiency in using electronic health records to streamline documentation, reduce redundancy, and improve care coordination can lead to significant cost savings. Training should go beyond basic data entry and focus on leveraging decision support systems, expert systems, and other features to optimize clinical workflows and resource usage [37,48]. Telehealth adoption is rapidly increasing, offering cost-effective virtual care delivery models. Digital health education programs should incorporate telehealth and virtual care training, equipping providers with the skills to conduct virtual visits, remotely monitor patients, and use virtual care platforms effectively [2,45-47]. To help empower informatics professionals to advocate for cost-effective technology adoption and usage, they should also be aware of the economic implications of digital health technologies, including cost-benefit analysis, return on investment, and value-based health care models [34]. If the efficient use of digital health technologies is not incorporated into digital health education, health care systems may face increased operational costs due to inefficiencies, higher rates of preventable hospitalizations, and missed opportunities for cost-saving innovations such as telehealth and automation. Efficient use of digital health technologies could be assessed through timed exercises or real-world simulations where health care students demonstrate their ability to quickly and accurately navigate digital health systems, manage patient data, and complete tasks with minimal errors and maximum efficiency (Table 1).

Cybersecurity, Privacy, and Ethical Use of Digital Health Technologies

With the growing reliance on digital tools comes the increased risk of data breaches and cyberattacks. Competency in cybersecurity, including understanding data protection principles and mitigating security risks, is essential for maintaining patient trust and ensuring the safety of health information systems [65]. Cybersecurity, privacy, and ethical use of digital health technologies reduce health care costs by safeguarding sensitive information, preventing data breaches, and ensuring trust in digital systems. Effective cybersecurity measures prevent costly breaches that can lead to financial penalties, legal liabilities, and expensive recovery processes [65]. In addition, ethical use of digital health technologies ensures transparency, fostering patient trust, which encourages technology adoption and self-management [27]. If cybersecurity, privacy, and the ethical use of digital health technologies are not incorporated into digital health education, health care systems may face costly data breaches, legal penalties, and loss of patient trust, ultimately leading to increased financial and reputational risks. Cybersecurity, privacy, and the ethical use of digital health technologies may be assessed through written examinations, case studies, or scenario-based discussions that test health care students' knowledge of data protection laws, ethical guidelines, and strategies for safeguarding patient information in digital health environments (Table 1).

Digital Health Ecosystems and Interoperability

Promoting interoperability and connected care in health care reduces costs by improving coordination and access to complete patient information, which decreases unnecessary tests, procedures, and medication errors [52]. These systems streamline administrative tasks, lowering overhead, and support value-based care models focused on quality outcomes. In addition, by enabling population health management and enhancing patient engagement, interoperability fosters preventive care and self-management, further lowering expenditures associated with chronic conditions and emergency care. Overall, interoperability creates a more efficient, cohesive, and cost-effective health care environment [52,66]. If digital health ecosystems and interoperability are not incorporated into digital health education, health care organizations may incur higher costs due to redundant testing, inefficiencies in data sharing, and fragmented care coordination, leading to suboptimal resource use. Health care students' understanding of digital health ecosystems and interoperability could be assessed through group projects where they analyze and design solutions that demonstrate how different digital health platforms can communicate and work together to improve patient care and health care outcomes (Table 1).

Data Analysis

The reduction of health care costs can be achieved through the application of informatics competencies such as data analysis. Cost analysis skills help professionals make data-informed decisions that balance cost and quality in patient care, while health economics education emphasizes understanding economic drivers within health care [34]. Cost-benefit analysis tools within digital health platforms allow for optimized resource allocation and reduced waste, ultimately reducing costs while maintaining quality. Through these competencies, digital health practitioners can advocate for financial strategies that lower costs without compromising care quality. If data analysis is not incorporated into digital health education, health care systems may struggle to identify cost-saving opportunities, leading to inefficient resource allocation, increased waste, and missed chances for data-driven decision-making that could enhance financial sustainability. The data analysis competency may be assessed through data-driven case studies or analytical projects where health care students interpret health care data, identify trends, and make evidence-based recommendations for improving patient outcomes or optimizing health care processes (Table 1).

Population Health

Overview

The population health category of the Quintuple Aim focuses on improving health outcomes across communities by addressing preventive care, chronic disease management, and health disparities. Digital health technologies support this goal by enabling large-scale data collection, predictive analytics, and remote monitoring, allowing health care providers to identify trends, allocate resources efficiently, and implement targeted interventions. Tools such as electronic health records, virtual care platforms, and mobile health apps enhance access to care, especially for underserved populations, while data-driven insights help inform public health strategies. By leveraging digital health, health care systems can promote proactive,

equitable, and efficient approaches to population health management.

Data Management and Analysis

To improve population health, health care professionals must be skilled in using data analytics to assess and address health trends and disparities. Health informatics education programs should thus prioritize competencies in population health informatics, which includes training in data collection, analysis, and interpretation related to population health metrics [20,21]. In addition, familiarity with tools for epidemiological analysis and population health management, including geospatial mapping and predictive analytics, can equip graduates with the knowledge to identify at-risk populations and forecast public health trends [21,25].

Understanding social determinants of health and leveraging big data are also central to enhancing population health, as they enable health care providers to integrate nonclinical factors impacting patient outcomes into their care approaches [42,49]. Analysis of social, behavioral, and environmental determinants of health supports the delivery of personalized care. The ability to collect, store, manage, analyze, and interpret health data is indispensable in the digital health era.

This competency also encompasses data literacy and the ability to critically appraise data quality and reliability [59]. Students should develop proficiency in basic statistics, data visualization techniques, and gain an understanding of data governance and privacy principles, which are particularly critical given the sensitive nature of health information and the need to maintain patient confidentiality [28,34]. As digital health technologies become increasingly sophisticated in capturing social, behavioral, and environmental determinants of health, health care professionals must be able to analyze and use this data to deliver personalized and holistic care, as well as appreciate the extent that digital health technologies impact the planet [2,67-69]. If data management and analysis are not incorporated into digital health education, health care systems may struggle to track disease patterns, allocate resources effectively, and develop targeted public health interventions, ultimately hindering efforts to improve population health outcomes. Health care students' understanding of data management and analysis in relation to digital health and population health could be assessed through projects where they analyze large datasets to identify health trends, evaluate interventions, and propose strategies for improving population health outcomes using digital tools and evidence-based insights (Table 1).

Adaptability to Emerging Technologies

AI is being increasingly used to enhance population health strategies through predictive analytics. Health care leaders must stay abreast of emerging technologies such as AI, robotics, and big data analytics [21,40,70]. AI marks a pivotal advancement in the digital transformation of health care, and traditional health care education and training often do not sufficiently address the digital competencies required for its effective use [40]. To ensure the safe and efficient adoption of AI, health care professionals need foundational knowledge in machine learning and neural networks, skills for critical evaluation of datasets, and the ability to integrate AI into clinical workflows while managing bias and ensuring smooth human-machine interaction in clinical settings. Furthermore, an understanding of the legal and ethical implications of digital health and the broader impact of AI adoption is essential [3,12,33,39,49,70]. If adaptability to emerging technologies is not incorporated into digital health education, health care systems may fall behind in implementing innovative solutions, limiting their ability to address evolving public health challenges and widening disparities in access to effective care. Adaptability to emerging technologies may be assessed through hands-on exercises or simulations where health care students are tasked with quickly learning and integrating new digital health tools into clinical practice, demonstrating flexibility and problem-solving skills in response to evolving health care technologies (Table 1).

Leadership, Change Management, and Project Management Skills

Health care professionals should be equipped to champion digital health transformation and spearhead initiatives that promote the adoption and integration of digital technologies within health care organizations and systems [27,42,59]. This competency includes a grasp of change management principles, enabling graduates to navigate the complexities and challenges associated with implementing new technologies and workflows [20,21]. Proficiency in project management is also vital, ensuring the organized planning, execution, and evaluation of digital health initiatives [20,21]. If leadership, change management, and project management skills are not incorporated into digital health education, health care organizations may struggle to implement and scale digital health initiatives effectively, leading to fragmented care, inefficient resource use, and missed opportunities to improve health outcomes. Leadership, change management, and project management skills in relation to digital health could be assessed through team-based projects where health care students plan, execute, and evaluate the implementation of digital health solutions, demonstrating their ability to lead, manage stakeholder expectations, and navigate challenges in health care technology integration (Table 1).

Health Equity

Overview

The health equity category of the Quintuple Aim focuses on reducing disparities in health care access, quality, and outcomes across diverse populations. Digital health technologies support this goal by expanding access to care through telemedicine, mobile health apps, and remote monitoring, particularly for underserved and rural communities. In addition, data analytics and AI can help identify health disparities, enabling targeted interventions and resource allocation. However, to fully realize these benefits, digital health solutions must be designed with inclusivity in mind, ensuring accessibility for individuals with varying levels of digital literacy, socioeconomic backgrounds, and health care needs.

Addressing Digital Health Disparities

Digital health interventions must be designed to avoid widening existing health inequities. Training should highlight the role of social determinants and the need for equitable technology use, as well as the role that data standards play in ensuring clinical documentation accurately reflects representation of social determinants of health [58]. Culturally sensitive and accessible digital tools are essential, and education programs must stress the importance of cultural considerations to ensure fair access [42,45,49]. Health care professionals also need advocacy skills for the ethical deployment of digital health solutions to prevent these technologies from exacerbating disparities [37]. If addressing digital health disparities is not incorporated into digital health education, existing health care inequities may worsen, as underserved populations could face barriers to accessing digital tools, leading to gaps in care and poorer health outcomes. Health care students' ability to address digital health disparities may be assessed through case studies or community-based projects where they identify barriers to digital health access, propose strategies for increasing inclusivity, and evaluate the impact of their solutions on underserved populations (Table 1).

Health Data Access

Limited access to personal health data leads to fragmented care, delays in diagnosis, and increased health care costs [52,66]. Ensuring patients have control over their health data is vital for improving outcomes, especially for marginalized communities [42,71]. Limited access disproportionately affects vulnerable groups, reinforcing health disparities and increasing the economic burden on both individuals and the health care system. Equitable access to health data is crucial for high-quality care and patient autonomy [42,66]. Importantly, "health data access" encompasses both individual-level access, where patients control, view, and share their personal health records, and population-level access, where aggregated data systems support public health surveillance, population health planning, and policy development. Competent digital health professionals must understand and ethically navigate both dimensions to effectively address inequities and improve community health outcomes [72]. If health data access is not incorporated into digital health education, marginalized communities may face continued barriers to care, as health care providers could lack the tools and knowledge to leverage data effectively, exacerbating disparities in health outcomes and access to services. Understanding of health data access could be assessed through practical exercises or written assignments where health care students analyze the ethical, legal, and technical aspects of accessing and sharing health data, ensuring compliance with regulations while considering patient privacy and data security (Table 1).

Indigenous Data Sovereignty

Understanding Indigenous Data Sovereignty helps professionals create data systems respectful of Indigenous cultural and ethical values. Principles like OCAP (Ownership, Control, Access, Possession) emphasize self-determination, advocating for ethical guidelines that honor Indigenous perspectives on health data [73,74]. Indigenous community involvement in research phases is essential to uphold ethical standards and enhance trust and relevancy in findings [75]. However, the scope of Indigenous Data Sovereignty extends beyond research. It must also be applied to public health surveillance, health system planning, and digital infrastructures to ensure that data about Indigenous populations is governed in ways that align with Indigenous laws, protocols, and priorities aligned with specific Nations. Indigenous communities should have equitable access to their own data to support local decision-making, health planning, and service delivery. Embedding Indigenous data governance at community, institutional, and policy levels enables Indigenous-led responses to health challenges and advances broader goals of health equity and reconciliation [76]. If Indigenous data sovereignty is not incorporated into digital health education, there is a risk of exploiting or misusing Indigenous health data without proper consent or oversight, potentially undermining trust and perpetuating historical injustices in health care. Health care students' understanding of Indigenous data sovereignty may be assessed through case studies or reflective essays where they analyze the importance of respecting Indigenous peoples' control over their health data, explore ethical considerations, and demonstrate how to apply culturally appropriate practices in digital health initiatives (Table 1).

Health Literacy and Patient Education

Digital health literacy is vital for patients, caregivers, and health care professionals alike. Patients need digital health literacy to manage their care effectively. Providers, leaders, and researchers must understand and use digital tools, as literacy across these groups supports quality care, informed decision-making, and research innovation. Digital health literacy fosters improved understanding, adherence, and outcomes across the health care system [50]. If health literacy and patient education are not incorporated into digital health education, patients may struggle to understand and effectively use digital health tools, exacerbating health disparities and hindering efforts to promote equitable access to care and tailor education to diverse literacy levels and needs. Health literacy and patient education in relation to digital health could be assessed through role-playing exercises or patient simulation scenarios where health care students demonstrate their ability to effectively communicate digital health information, assess patient comprehension, and tailor education to diverse literacy levels and needs (Table 1).

Integrated Digital Health Competencies for the Quintuple Aim

Digital health competencies, such as interoperability, data literacy, and person-centered care principles, overlap across the 5 domains of the Quintuple Aim, enhancing both patient and provider experiences. These competencies promote equitable patient-centered care by improving access, streamlining care processes, and supporting population health efforts. A comprehensive approach to digital health education enables health care professionals to meet the diverse, interconnected demands of modern health care, supporting both individual and population-level outcomes. Failing to create a comprehensive approach to digital health education in relation to the Quintuple Aim could lead to fragmented health care delivery, where

inefficiencies persist, patient experiences, health outcomes remain unequal, and costs rise, ultimately hindering the achievement of improved health for all.

Standardized Yet Customizable Digital Health Education

Standardized informatics competency-based curricula, as described by organizations such as the Canadian Association of Schools of Nursing, the International Medical Informatics Association and the American Medical Informatics Association, provide a crucial foundation for health informatics education [20,21,77]. However, the health care landscape varies significantly across regions, influenced by factors like technological infrastructure, health care systems, cultural norms, and local health priorities. These differences require that standardized curricula also provide additional opportunities for curricula customization, versatility, and variety of delivery methods [78-80], bridging academic learning and real-world application, and aligning curricula with evolving health care demands [20,78].

To align with the Quintuple Aim framework, educational delivery strategies should intentionally match the competency emphasis to each aim. For instance, competencies that address the “Improving Population Health” aim should be emphasized in programs targeting public health professionals, while “Reducing Costs” and “Enhancing Efficiency” may be prioritized in executive and administrative education. This alignment ensures that the education of health care professionals is not only technically robust but also strategically positioned to support each domain of the Quintuple Aim.

The integration of case-based learning (CBL) with competency-based education potentially offers an effective pedagogical strategy, especially in digital health education, where learners need to master both theoretical concepts and practical skills in an ever-evolving field. CBL engages students with practical scenarios that simulate real-world challenges, encouraging critical thinking, collaborative problem-solving, and decision-making. In complex and rapidly evolving environments like health care, competency-based education can provide context-rich scenarios that mimic real-life challenges, while CBE ensures that students meet predefined learning outcomes essential for professional competence [78-81].

Achieving a balance between standardized and customized education could also involve adopting a modular approach. Core competencies can be standardized across programs, while additional modules can be tailored to meet local needs. For instance, health care educational programs can integrate standardized content on digital ethics and health data infrastructure but offer elective modules or microcredentials on specific areas such as local legislative and regulatory policies, electronic health records design, or virtual care workflows relevant to regional health care priorities [80]. Digital competencies are also not uniformly applicable across all health professions. Direct care providers, such as nurses, physicians, and allied health professionals, require competencies focused on clinical decision support, patient communication, and digital documentation. Health administrators need competencies focused on systems-level data analytics, resource management, and technology implementation. Public health practitioners, in contrast, may require competencies related to population surveillance, data ethics, and cross-sectoral data integration. These variations highlight the need for educational frameworks that are both discipline-sensitive and practice-informed, ensuring that digital health education is relevant to specific roles while maintaining shared foundational competencies.

Collaborations between academia and industry can ensure that curricula stay relevant. Industry insights help shape competencies that meet real-world demands, while partnerships provide students with practical experiences through internships and projects. This alignment fosters a skilled workforce ready to integrate digital health technologies effectively [34,40].

Together with maintaining strong academic and health care system partnerships, a hybrid approach, combining core standardization with customization through CBL and modular content delivery, potentially offers a sustainable model for developing digital health curricula that align with both global standards and local needs, and ensures health care professionals are prepared to navigate the complexities of digital health care, ultimately contributing to more effective, equitable, and patient-centered care delivery.

Conclusion

The development of a standardized digital health curriculum that integrates both core competencies and context-specific customization is essential for equipping health care professionals to thrive in a rapidly evolving technological landscape. As health care systems increasingly rely on digital health technologies to enhance patient care and improve outcomes, health care professionals should at minimum be competent in skills such as data management, interoperability, digital literacy, and virtual care delivery. Standardization ensures consistency and fosters interprofessional collaboration, while context-specific case studies and modular customization enable education programs to address specific health care priorities, such as rural health care or emerging technologies. Achieving a balance between these approaches will ensure that digital health education remains relevant, responsive, and capable of meeting both global standards and local needs. Ultimately, a digitally competent workforce will be crucial to advancing health care delivery, supporting innovation, and achieving sustainable health outcomes in individuals, communities, and populations.

Conflicts of Interest

None declared.

References

1. Booth RG, Strudwick G, McBride S, O'Connor S, Solano López AL. How the nursing profession should adapt for a digital future. BMJ 2021;373:n1190. [doi: 10.1136/bmj.n1190]
2. Kleib M, Arnaert A, Nagle LM, et al. Digital health education and training for undergraduate and graduate nursing students: scoping review. JMIR Nurs 2024 Jul 17;7(1):e58170. [doi: 10.2196/58170] [Medline: 39018092]
3. Risling T. Educating the nurses of 2025: technology trends of the next decade. Nurse Educ Pract 2017 Jan;22:89-92. [doi: 10.1016/j.nepr.2016.12.007] [Medline: 28049072]
4. Davies A, Davies A, Abdulhussein H, et al. Educating the healthcare workforce to support digital transformation. In: MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation: IOS Press; 2022:934-936. [doi: 10.3233/SHTI220217]
5. International Council of Nurses. The future of nursing and digital health: new ICN position statement highlights opportunities and risks. ICN - International Council of Nurses. 2023. URL: https://www.icn.ch/news/future-nursing-and-digital-health-new-icn-position-statement-highlights-opportunities-and [accessed 2024-08-11]
6. Socha-Dietrich K. Empowering the health workforce to make the most of the digital revolution. In: IDEAS Working Paper Series from RePEc 2021, Vol. 129:1-67. [doi: 10.1787/37ff0eaa-en]
7. Global strategy on digital health 2020-2025. World Health Organization. 2021. URL: https://apps.who.int/iris/bitstream/handle/10665/344249/9789240020924-eng.pdf [accessed 2023-03-18]
8. Nagle L, Kleib M, Furlong K. Digital health in Canadian schools of nursing part A: nurse educators' perspectives. QANE-AFI 2020;6(1). [doi: 10.17483/2368-6669.1229]
9. Abernethy A, Adams L, Barrett M, et al. The Promise of Digital Health: Then, Now, and the Future: NAM Perspectives; 2022. [doi: 10.31478/202206e]
10. Lee KH, Kim MG, Lee JH, et al. Empowering healthcare through comprehensive informatics education: the status and future of biomedical and health informatics education. Healthc Inform Res 2024 Apr;30(2):113-126. [doi: 10.4258/hir.2024.30.2.113] [Medline: 38755102]
11. Nagle L, Kleib M, Furlong K. Digital health in Canadian schools of nursing—part B: academic nurse administrators' perspectives. QANE-AFI 2020;6(3). [doi: 10.17483/2368-6669.1256]
12. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
13. Snowdon A. HIMSS defines digital health for the global healthcare industry. HIMSS. 2020. URL: https://www.himss.org/news/himss-defines-digital-health-global-healthcare-industry
14. How health information systems are driving digital transformation in healthcare. World Economic Forum. 2022. URL: https://www.weforum.org/agenda/2022/09/health-information-system-digital-transformation-healthcare/
15. Canada Health Infoway. Definition of digital health. Canadian Institutes of Health Research. URL: https://cihr-irsc.gc.ca/e/47345.html [accessed 2024-03-18]
16. Informatics: research and practice. American Medical Informatics Association. URL: https://www.amia.org/about-amia/why-informatics/informatics-research-and-practice [accessed 2025-10-25]
17. Giddens J. Demystifying concept-based and competency-based approaches. J Nurs Educ 2020 Mar 1;59(3):123-124. [doi: 10.3928/01484834-20200220-01] [Medline: 32130412]
18. Englander R, Cameron T, Ballard AJ, Dodge J, Bull J, Aschenbrener CA. Toward a common taxonomy of competency domains for the health professions and competencies for physicians. Acad Med 2013 Aug;88(8):1088-1094. [doi: 10.1097/ACM.0b013e31829a3b2b] [Medline: 23807109]
19. 2023 Canadian Survey of Nurses Understanding the utilization and impact of digital health technologies on nursing practice. Canada Health Infoway. 2023. URL: https://insights.infoway-inforoute.ca/2023-nursing-survey [accessed 2024-10-25]
20. Bichel-Findlay J, Koch S, Mantas J, et al. Recommendations of the International Medical Informatics Association (IMIA) on education in biomedical and health informatics: second revision. Int J Med Inform 2023 Feb;170:104908. [doi: 10.1016/j.ijmedinf.2022.104908] [Medline: 36502741]
21. Gadd CS, Steen EB, Caro CM, Greenberg S, Williamson JJ, Fridsma DB. Domains, tasks, and knowledge for health informatics practice: results of a practice analysis. J Am Med Inform Assoc 2020 Jun 1;27(6):845-852. [doi: 10.1093/jamia/ocaa018] [Medline: 32421829]
22. Monkman H, Mir S, Bond J, Borycki EM, Courtney KL, Kushniruk AW. Canadian employers' perspectives on a new framework for health informatics competencies. Int J Med Inform 2024 Mar;183:105324. [doi: 10.1016/j.ijmedinf.2023.105324] [Medline: 38218130]
23. Nundy S, Cooper LA, Mate KS. The Quintuple Aim for health care improvement: a new imperative to advance health equity. JAMA 2022 Feb 8;327(6):521-522. [doi: 10.1001/jama.2021.25181] [Medline: 35061006]
24. Shah YB, Goldberg ZN, Harness ED, Nash DB. Charting a path to the Quintuple Aim: harnessing AI to address social determinants of health. Int J Environ Res Public Health 2024 May 31;21(6):718. [doi: 10.3390/ijerph21060718] [Medline: 38928964]
25. Mainz A, Nitsche J, Weirauch V, Meister S. Measuring the digital competence of health professionals: scoping review. JMIR Med Educ 2024 Mar 29;10(1):e55737. [doi: 10.2196/55737] [Medline: 38551628]

26. Moore RA, Berner ES. Comparison of health/medical informatics curricula against multiple sets of professional criteria. AMIA Annu Symp Proc 2003;2003:942. [Medline: 14728447]
27. Poncette AS, Glauert DL, Mosch L, Braune K, Balzer F, Back DA. Undergraduate medical competencies in digital health and curricular module development: mixed methods study. J Med Internet Res 2020 Oct 29;22(10):e22161. [doi: 10.2196/22161] [Medline: 33118935]
28. Davies A, Mueller J, Moulton G. Core competencies for clinical informaticians: a systematic review. Int J Med Inform 2020 Sep;141:104237. [doi: 10.1016/j.ijmedinf.2020.104237] [Medline: 32771960]
29. Fenton SH, Gongora-Ferraez MJ, Joost E. Health information technology knowledge and skills needed by HIT employers. Appl Clin Inform 2012;3(4):448-461. [doi: 10.4338/ACI-2012-09-RA-0035] [Medline: 23646090]
30. Garde S, Harrison D, Huque M, Hovenga EJS. Building health informatics skills for health professionals: results from the Australian Health Informatics Skill Needs Survey. Aust Health Rev 2006 Feb;30(1):34-45. [Medline: 16448376]
31. Monkman H, Mir S, Borycki EM, Courtney KL, Bond J, Kushniruk AW. Updating professional competencies in health informatics: a scoping review and consultation with subject matter experts. Int J Med Inform 2023 Feb;170:104969. [doi: 10.1016/j.ijmedinf.2022.104969] [Medline: 36572000]
32. Brice S, Almond H. Health professional digital capabilities frameworks: a scoping review. J Multidiscip Healthc 2020;13:1375-1390. [doi: 10.2147/JMDH.S269412] [Medline: 33173300]
33. Davies A, Hassey A, Williams J, Moulton G. Creation of a core competency framework for clinical informatics: from genesis to maintaining relevance. Int J Med Inform 2022 Dec;168:104905. [doi: 10.1016/j.ijmedinf.2022.104905] [Medline: 36332519]
34. Khairat S, Sandefer R, Marc D, Pyles L. A review of biomedical and health informatics education: a workforce training framework. JHA 2016;5(5):10. [doi: 10.5430/jha.v5n5p10]
35. Nazeha N, Pavagadhi D, Kyaw BM, Car J, Jimenez G, Tudor Car L. A digitally competent health workforce: scoping review of educational frameworks. J Med Internet Res 2020 Nov 5;22(11):e22706. [doi: 10.2196/22706] [Medline: 33151152]
36. Ramirez-Montoya AM, Fernández-Luque MS. Health Professionals' Competencies in the Framework of Complexity: Digital Training Model for Education 40. 2024. URL: https://click.endnote.com/viewer?doi=10.3989%2Fredc.2024.2.1470&token=WzQzMDA3MDQsIjEwLjM5ODkvcmVkYy4yMDI0LjIuMTQ3MCJd.5s4e8d3TZSHQdsocI2UxaWj3wno [accessed 2024-10-30]
37. Choi J, Woo S, Tarte V. Informatics competencies of students in a doctor of nursing practice program: a descriptive study. Healthc Inform Res 2024 Apr;30(2):147-153. [doi: 10.4258/hir.2024.30.2.147] [Medline: 38755105]
38. Moore JB. From personalised nutrition to precision medicine: the rise of consumer genomics and digital health. Proc Nutr Soc 2020 Aug;79(3):300-310. [doi: 10.1017/S0029665120006977]
39. Buchanan C, Howitt ML, Wilson R, Booth RG, Risling T, Bamford M. Predicted influences of artificial intelligence on nursing education: scoping review. JMIR Nurs 2021;4(1):e23933. [doi: 10.2196/23933] [Medline: 34345794]
40. Malerbi FK, Nakayama LF, Gayle Dychiao R, et al. Digital education for the deployment of artificial intelligence in health care. J Med Internet Res 2023 Jun 22;25(1):e43333. [doi: 10.2196/43333] [Medline: 37347537]
41. Assamad D, Majeed S, Aguda V, et al. Digital health tools in genomics: advancing diversity, equity, and inclusion. Public Health Genomics 2023;26(1):194-200. [doi: 10.1159/000534804] [Medline: 37883926]
42. Brewer LC, Fortuna KL, Jones C, et al. Back to the future: achieving health equity through health informatics and digital health. JMIR Mhealth Uhealth 2020 Jan 14;8(1):e14512. [doi: 10.2196/14512] [Medline: 31934874]
43. Lennox-Chhugani N. A user-centred design approach to integrated information systems - a perspective. Int J Integr Care 2018 May 22;18(2):15. [doi: 10.5334/ijic.4182] [Medline: 30127699]
44. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun 16;8(2):e9. [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
45. Clement David-Olawade A, Olawade DB, Ojo IO, Famujimi ME, Olawumi TT, Esan DT. Nursing in the digital age: harnessing telemedicine for enhanced patient care. Informatics and Health 2024 Sep;1(2):100-110. [doi: 10.1016/j.infoh.2024.07.003]
46. Ochs N, Franco H, Gallegos B, Baba D, Crossland J. Telehealth in nursing education: navigating the new normal. Nursing (Auckl) 2022 Mar 1;52(3):12-14. [doi: 10.1097/01.NURSE.0000820044.72197.f9] [Medline: 35196275]
47. Nowell L, Dolan S, Johnston S, Jacobsen M, Lorenzetti D, Oddone Paolucci E. Exploring student perspectives and experiences of online opportunities for virtual care skills development: sequential explanatory mixed methods study. JMIR Nurs 2024 Aug 21;7(1):e53777. [doi: 10.2196/53777] [Medline: 39167789]
48. Scott IA, Shaw T, Slade C, et al. Digital health competencies for the next generation of physicians. Intern Med J 2023 Jun;53(6):1042-1049. [doi: 10.1111/imj.16122] [Medline: 37323107]
49. Ronquillo CE, Mitchell J, Alhuwail D, Peltonen LM, Topaz M, Block LJ. The untapped potential of nursing and allied health data for improved representation of social determinants of health and intersectionality in artificial intelligence applications: a rapid review. Yearb Med Inform 2022 Aug;31(1):94-99. [doi: 10.1055/s-0042-1742504] [Medline: 35654435]
50. Mackert M, Mabry-Flynn A, Champlin S, Donovan EE, Pounders K. Health literacy and health information technology adoption: the potential for a new digital divide. J Med Internet Res 2016 Oct 4;18(10):e264. [doi: 10.2196/jmir.6349] [Medline: 27702738]

51. Mannevaara P, Kinnunen UM, Egbert N, et al. Discovering the importance of health informatics education competencies in healthcare practice. A focus group interview. Int J Med Inform 2024 Jul;187:105463. [doi: 10.1016/j.ijmedinf.2024.105463] [Medline: 38643700]
52. Affleck E, Murphy T, Williamson T, et al. Interoperability Saves Lives. 2023. URL: www.albertavirtualcare.org
53. Fitzpatrick PJ. Improving health literacy using the power of digital communications to achieve better health outcomes for patients and practitioners. Front Digit Health 2023;5:1264780. [doi: 10.3389/fdgth.2023.1264780] [Medline: 38046643]
54. Wiljer D, Charow R, Costin H, et al. Defining compassion in the digital health age: protocol for a scoping review. BMJ Open 2019 Feb 15;9(2):e026338. [doi: 10.1136/bmjopen-2018-026338] [Medline: 30772865]
55. Bombard Y, Ginsburg GS, Sturm AC, Zhou AY, Lemke AA. Digital health-enabled genomics: opportunities and challenges. Am J Hum Genet 2022 Jul 7;109(7):1190-1198. [doi: 10.1016/j.ajhg.2022.05.001] [Medline: 35803232]
56. Halbert CH. Equity in genomic medicine. Annu Rev Genomics Hum Genet 2022 Aug 31;23:613-625. [doi: 10.1146/annurev-genom-112921-022635] [Medline: 35363547]
57. Carter AB, Abruzzo LV, Hirschhorn JW, et al. Electronic health records and genomics. J Mol Diagn 2022 Jan;24(1):1-17. [doi: 10.1016/j.jmoldx.2021.09.009]
58. Fein R. Innovate or die!: Genomic data and the electronic health record (EHR). Appl Transl Genom 2014 Dec 1;3(4):130-131. [doi: 10.1016/j.atg.2014.09.007] [Medline: 27294027]
59. Jarva E, Mikkonen K, Andersson J, et al. Aspects associated with health care professionals' digital health competence development – a qualitative study. FinJeHeW 2022;14(1):79-91. [doi: 10.23996/fjhw.111771]
60. Hannah KJ, White PA, Nagle LM, Pringle DM. Standardizing nursing information in Canada for inclusion in electronic health records: C-HOBIC. J Am Med Inform Assoc 2009;16(4):524-530. [doi: 10.1197/jamia.M2974] [Medline: 19261936]
61. White P, Nagle LM, Hannah K. Adopting national nursing data standards in Canada. CAN NURSE 2017;113(3):18-22. [Medline: 29235786]
62. Scott PJ, Dunscombe R, Evans D, Mukherjee M, Wyatt JC. Learning health systems need to bridge the 'two cultures' of clinical informatics and data science. BMJ Health Care Inform 2018 Apr;25(2):126-131. [doi: 10.14236/jhi.v25i2.1062]
63. Büssing A, Falkenberg Z, Schoppe C, Recchia DR, Poier D. Work stress associated cool down reactions among nurses and hospital physicians and their relation to burnout symptoms. BMC Health Serv Res 2017 Aug 10;17(1):551. [doi: 10.1186/s12913-017-2445-3] [Medline: 28797258]
64. Longhini J, Rossettini G, Palese A. Digital health competencies among health care professionals: systematic review. J Med Internet Res 2022 Aug 18;24(8):e36414. [doi: 10.2196/36414] [Medline: 35980735]
65. Wells JSG. Preparing for hybrid warfare and cyberattacks on health services' digital infrastructure: What nurse managers need to know. J Nurs Manag 2022 Sep;30(6):2000-2004. [doi: 10.1111/jonm.13633] [Medline: 35419846]
66. Brend Y. Canada's health-care system has a data problem, experts say. and it puts patients at risk. CBC. 2022. URL: https://www.cbc.ca/news/canada/health-data-canada-sharing-information-1.6652770 [accessed 2023-10-01]
67. Austin RR, Alexander S, Jantraporn R, Rajamani S, Potter T. Planetary health and nursing informatics. Comput Inform Nurs 2023 Dec 1;41(12):931-936. [doi: 10.1097/CIN.0000000000001085] [Medline: 38062545]
68. Lokmic-Tomkins Z, Davies S, Block LJ, et al. Assessing the carbon footprint of digital health interventions: a scoping review. J Am Med Inform Assoc 2022 Nov 14;29(12):2128-2139. [doi: 10.1093/jamia/ocac196] [Medline: 36314391]
69. Lokmic-Tomkins Z, Borda A, Humphrey K. Designing digital health applications for climate change mitigation and adaptation. Med J Aust 2023 Feb 20;218(3):106-110. [doi: 10.5694/mja2.51826] [Medline: 36625463]
70. Risling T. Understanding the foundations of artificial intelligence: data, math and machine learning. In: Nursing and Informatics for the 21st Century - Embracing a Digital World, 3rd edition 2022, Vol. 3:95-112.
71. Fort MP, Manson SM, Glasgow RE. Applying an equity lens to assess context and implementation in public health and health services research and practice using the PRISM framework. Front Health Serv 2023;3:1139788. [doi: 10.3389/frhs.2023.1139788] [Medline: 37125222]
72. Richardson S, Lawrence K, Schoenthaler AM, Mann D. A framework for digital health equity. NPJ Digit Med 2022 Aug 18;5(1):119. [doi: 10.1038/s41746-022-00663-0] [Medline: 35982146]
73. Information Governance Centre FN. Barriers and Levers for the Implementation of OCAP(TM). Int Indig Policy J 2014;5(2). [doi: 10.18584/iipj.2014.5.2.3]
74. Pringle A, Pavagadhi K. Using OCAP and IQ as frameworks to address a history of trauma in indigenous health research. AMA J Ethics 2020;22(10):E868-E873. [doi: 10.1001/amajethics.2020.868]
75. Truth and Reconciliation Commission of Canada. Government of Canada. 2015 Dec 14. URL: https://www.rcaanc-cirnac.gc.ca/eng/1450124405592/1529106060525 [accessed 2025-08-20]
76. FNIGC. First Nations Data Governance Strategy: Relevance to Canada's Health Information and Statistical Systems 2023. [doi: 10.1001/amajethics.2020.868]
77. Nursing informatics entry-to-practice competencies for registered nurses | canadian association of schools of nursing / association canadienne des écoles de sciences infirmières (CASN / ACESI). Canadian Association of Schools of Nursing. URL: https://www.casn.ca/2014/12/nursing-informatics-entry-practice-competencies-registered-nurses-2/ [accessed 2025-08-20]

78. Sajjad T, Younas A, Abbasi LS. Case-based learning with a twist: testing the effectiveness of integrated case-based learning in an undergraduate dental curriculum. JSTMU 2023;6(2):65-71. [doi: 10.32593/jstmu/Vol6.Iss2.242]
79. Srinivasan M, Wilkes M, Stevenson F, Nguyen T, Slavin S. Comparing problem-based learning with case-based learning: effects of a major curricular shift at two institutions. Acad Med 2007 Jan;82(1):74-82. [doi: 10.1097/01.ACM.0000249963.93776.aa] [Medline: 17198294]
80. Wijnia L, Noordzij G, Arends LR, Rikers R, Loyens SMM. The effects of problem-based, project-based, and case-based learning on students' motivation: a meta-analysis. Educ Psychol Rev 2024 Mar;36(1):29. [doi: 10.1007/s10648-024-09864-3]
81. Jones S, Gerdtz M, Ukovich D, Marriott P, Merolli M. Case-Based Learning in a Simulated Electronic Medical Record: Digital Health Education for Nursing Students: IOS Press; 2024, Vol. 310:1181. [doi: 10.3233/SHTI231151]

Abbreviations

AI: artificial intelligence
CBL: Case-Based Learning
OCAP: Ownership, Control, Access, Possession
OECD: Organisation for Economic Co-operation and Development
WHO: World Health Organization

Edited by T Gladman; submitted 14.04.25; peer-reviewed by G McKee, J Bichel-Findlay, UM Kinnunen; revised version received 07.07.25; accepted 15.07.25; published 08.09.25.

Please cite as:
Rees G, Nowell L, Risling T
Shaping the Future of Digital Health Education in Canada: Prioritizing Competencies for Health Care Professionals Using the Quintuple Aim
JMIR Med Educ 2025;11:e75904
URL: https://mededu.jmir.org/2025/1/e75904
doi:10.2196/75904

An Ecosystem Approach to Developing and Implementing a Cocreated Bachelor's Degree in Digital Health and Biomedical Innovation

Patrícia Alves[1,2,3], PhD; Elisio Costa[2,4], PhD; Altamiro Costa-Pereira[2,5], MD, PhD; Inês Falcão-Pires[2,6], PhD; João Fonseca[1,2], MD, PhD; Adelino Leite-Moreira[2,6], MD, PhD; Bernardo Sousa-Pinto[1,2], MD, PhD; Nuno Vale[1,2], PhD

[1]MEDCIDS–Department of Community Medicine, Information and Health Decision Sciences, Faculty of Medicine, University of Porto, Rua Dr. Plácido da Costa, Porto, Portugal

[2]RISE-Health, Porto, Portugal

[3]Centre for Research and Intervention in Education, Porto, Portugal

[4]Department of Biological Sciences, Faculty of Pharmacy, University of Porto, Porto, Portugal

[5]Faculty of Medicine, University of Porto, Porto, Portugal

[6]Surgery and Physiology Department, Faculty of Medicine, University of Porto, Porto, Portugal

Corresponding Author:
João Fonseca, MD, PhD
MEDCIDS–Department of Community Medicine, Information and Health Decision Sciences, Faculty of Medicine, University of Porto, Rua Dr. Plácido da Costa, Porto, Portugal

Abstract

This paper aims to describe the cocreation and development processes of an educational ecosystem-centered Bachelor's degree in Digital Health and Biomedical Innovation (SauD InoB). This program is shaped by a multidisciplinary, intersectoral, and collaborative framework, involving more than 60 organizations in teaching activities, internship supervision, or hosting, most of which collaborated in needs assessment, curriculum development, and public promotion of the degree. In the context of health care digital transformation, this comprehensive Bachelor's degree will respond to unmet demands of the labor market by training students with technological, research, and management skills, as well as with basic clinical and biomedical concepts. Graduates will become transdisciplinary, creative professionals capable of understanding and integrating different "languages," reasoning, clinical processes, and scenarios.

(JMIR Med Educ 2025;11:e63903) doi:10.2196/63903

KEYWORDS

digital health; biomedical innovation; curriculum design; co-creation; health professionals' education; ecosystem-centered: biomedical; education; health care; graduates; teaching; learning; digital transformation; health information system; health education; information system; health management; project management; ecosystem

Introduction

Biomedical innovation may be defined as the process of creation and application of scientific and technological knowledge to improve health care and promote health and well-being [1]. It includes developing therapeutic and diagnostic health technologies, biotechnology, precision medicine, drug discovery, or health care digitalization [2], which may encompass developing and using technologies such as health information systems, artificial intelligence, and advanced health data analytics [3,4].

To address global health challenges, imposed by aging demographics and sustainability of health and care systems, biomedical innovation is expected to integrate digital health [1], which plays an increasingly important role in the organization and management of health systems and institutions, health research, and health interventions [5]. However, biomedical innovation and digital transformation processes are frequently long, costly, risky, and challenging [2], involving changes in management dynamics [6,7] of complex project ecologies in which knowledge creation, flow, and outputs depend on the communication between multidisciplinary areas and sectors of activity [1]. Such a demanding context is further rendered more complex by management challenges, such as interoperability problems, data security concerns, ethical and legal aspects, low digital and scientific literacy among patients [6], resistance to change [7], or a recognizably insufficient health professionals' workforce, simultaneously trained in digital health and biomedical innovation areas [8-10].

Overcoming these challenges and promoting successful change dynamics depends on multiple factors, namely a bold vision shared with different actors, engagement with the society

through the involvement of various stakeholders, infrastructures, or the involvement of professionals with technical and scientific expertise, capable of adjusting the changes of management processes to contextual and cultural specificities [6] and to create synergies between different disciplinary areas and sectors of activity.

Identifying the unmet need for these professionals led to the creation of an innovative and comprehensive Bachelor's degree in Digital Health and Biomedical Innovation (SauD InoB) at the University of Porto, one of the largest Portuguese universities.

In Portugal, the BSc degree leads to the degree of "licenciado," corresponding to the level 6 of the National Qualifications Framework and European Qualifications Framework. A BSc degree may range from 180 to 240 European Credit Transfer and Accumulation System credits (ECTS) and last from 6 to 8 curricular semesters [11].

The BSc degree was structured according to the requirements and milestones of the Portuguese Agency for Assessment and Accreditation of Higher Education (A3ES) and grounded on the first steps of Kern's approach to curriculum development: (1) identification and characterization of a problem, (2) identification of training needs, (3) definition of goals and objectives, and (4) definition of educational strategies [12]. This process comprised innovative aspects, engaging from the outset on a comprehensive and collaborative educational ecosystem–centered approach focused on the enhancement of students' experiences and outcomes and on the strengthening of their contributions to society through the development of creativity and innovation capacity [13].

The BSc degree was created within an educational ecosystem, involving academic, research, health care, business, and society (eg, patients' associations). An educational ecosystem refers to the interconnected network of factors and processes that influence students' learning and development. Just like a natural ecosystem where different elements, such as air, water, plants, and animals, interact to create a balanced environment, an educational ecosystem consists of various components that work together to shape the learning experience [14-16]. According to Bronfenbrenner's Ecological Systems Theory, learning and development are influenced by factors, processes, and actors situated in different contexts, some of them more directly involved in the learning process, such as family, teachers, higher education institutions, or internship sites (microsystem), others situated in a broader context, such as the course regulations (macrosystem), or the economic situation and cultural values that may influence learning (exosystem). Furthermore, learning may also be influenced by the interactions between contexts, for example, between the universities and industry (mesosystem) and by the dimension of time, for instance, by the effects of the technological development (chronosystem) [17].

Like natural ecosystems, educational ecosystems involve disturbances that may challenge their balance and require resilience, flexibility, and responsiveness to the needs and characteristics of different contexts and actors [14]. The concept of ecological university seems to entail this notion of disturbance and adaptive management. With the development of knowledge societies, higher education institutions no longer occupy a hegemonic position in the creation and legitimization of knowledge, and the actions of the individual and collective knowledge creators, users, legitimizers, and beneficiaries are informed by a widening diversity of perspectives and voices (eg, companies also have their own laboratories). In this context, higher education institutions must be able to interact with other contexts or ecological zones (eg, industry, patient associations, and government) with different rhythms, interests, epistemologies, and perspectives [18-20], but also to cross boundaries and build bridges between those contexts, not only to maintain their legitimacy, but also to be able to contribute to the health and functionality of the ecosystem [14].

Some authors concluded that the involvement of diverse stakeholders may promote an alignment with the needs of patients [21] or with the specific needs of diverse populations [22] and foster creativity, entrepreneurship, and innovation [13,23]. Curriculum development processes in transdisciplinary areas involving multidisciplinary and multisectoral teams tend to be complex processes in which tensions can arise, for example, related to different interests and rules [15]. Nevertheless, research on ecosystem-based curriculum development is still scarce, leaving the actors involved in these processes without a consistent theoretical basis to support their practices and decisions.

This viewpoint aims to contribute to the knowledge about curriculum development by describing the educational ecosystem–centered, multidisciplinary, and multisectoral cocreation and development process of the SauD Inob BSc degree, from needs assessment and curriculum design to the public promotion of the degree, aiming to attract its first group of students and the results of the first edition of the degree application.

The Educational Ecosystem

The creation of this BSc involved 3 faculties within 1 university, the collaboration of 3 polytechnic institutions, and partnerships with more than 60 organizations, including hospitals and other health care providers, pharmaceutical and health technologies companies, research and governmental institutions, and patients' associations. Cocreation processes involved more than 130 contributors, including faculty, researchers, master's and PhD students and graduates in health, health informatics and data science, professionals working in diverse functional areas and sectors of activity, and managers and policy makers. The recruitment of students and alumni was strategically carried out in relevant master's and doctoral programs, with some focus on digital health and biomedical innovation (medical informatics, clinical and health services research, and health data science), ensuring a highly relevant and involved participant base. Besides contributing to the BSc degree through involvement in teaching and internship supervision or hosting internships, most partners and collaborators were also integrated into needs assessment and curriculum development.

A list of the partners involved in the course is listed in [24].

The Need for a BSc Degree in Digital Health and Biomedical Innovation

Needs assessment was based on a narrative literature review, document analysis, and meetings with the stakeholders integrated into the educational ecosystem. Individual meetings with partner representatives were carried out to discuss their perspectives on the interplay between labor market needs and the curriculum of SauD InoB and to establish a collaborative commitment. The curriculum was also presented and discussed in the first meeting of partners of the BSc degree in Digital Health and Biomedical Innovation in September 2023.

A literature review revealed a compelling but still unmet need to develop digital health and biomedical innovation skills and knowledge among the different professional groups involved in health care provision [8-10,25]. This need has been enhanced by a remarkable acceleration of the development of information and communication technologies, with repercussions for the adoption of electronic health records, the development of apps for monitoring chronic diseases, or advances in data analysis [3,4]. This unmet need motivated the introduction of digital health and biomedical innovation contents in the curricula of health and information technologies degree-awarding courses, the development of continuous education courses in biomedical and health informatics [26], and the creation of a few degree-awarding courses, in digital health and biomedical innovation [27-31], such as the BSc in Digital Health (Politechnic Institute of Porto, Portugal), the BSc in Digital Technologies and Health (University Institute of Lisbon, Portugal) [27], the Master's Degree in Digital Health (Deggendorf Institute of Technology, Germany) [28], or the Master's Degree in Digitalization in the Health Sector (University of Oslo, Norway) [31]. Nevertheless, the stakeholders involved in SauD InoB creation considered that the training offered in these areas is still insufficient to overcome the challenges of this fast-paced development context, namely in the North of Portugal.

The stakeholders identified the need to develop new professional profiles of transdisciplinary health care professionals who may establish bridges between different disciplinary or professional areas, academic and research cultures, knowledge, languages, and epistemologies. Graduates from SauD InoB are expected to be able to hold careers in (1) data analytics and artificial intelligence in health, (2) biomedical and clinical research, (3) health information systems and telemedicine, and (4) health consulting and management, innovation management, and other health-related areas. The labor market for these professionals may include digital technology and health informatics, pharmaceutical, medical device, and biotechnology companies, hospitals and health care institutions, research and development organizations (biomedical and clinical research laboratories, academia, research and development service companies, and consulting firms), and health policy and regulatory organizations.

Aims and Learning Outcomes of SauD InoB

SauD InoB aims at training health professionals capable of (1) understanding concepts and "languages" in human biology, clinical medicine, and health care services and (2) mastering skills in health information systems, programming, data science, and research methodology. Therefore, these professionals will be able to identify, develop, and solve the main problems associated with health care services and with their digital transformation, offering them practical solutions that bridge different fields.

Upon completion of the study cycle, graduates should be able to apply critical knowledge about (1) human biology; (2) the pathophysiological basis of health and disease, different types of diseases, diagnostic tests, and therapeutic interventions; (3) clinical language and reasoning; (4) mathematics and statistics; (5) informatics, computing, and programming; (6) organization, pathways, and processes of health care provision in multiple contexts; (7) health care management and innovation; and (8) ethical, regulatory and security, data privacy in digital health, and health innovation processes. They should also be able to participate in the development, design and implementation, or management of databases, health care research and innovation projects, digital health systems and strategies, and apply data science approaches in health (processing, statistical analysis, and presentation of data).

SauD InoB goes beyond developing specific digital health and biomedical innovation competencies. It aims to create a new professional identity, which is defined as the individual sense of identification with a (new) profession and a sense of belonging to a community of professionals sharing specific knowledge, competencies, values, activities, and norms [32]. This is a distinctive characteristic of SauD InoB derived from being a BSc and not a subsequent degree, in which students' professional identities are more likely to be influenced by the roles and competencies that characterize the professional or disciplinary area of their initial academic training [32].

Further information on the objectives, expected learning outcomes, and competencies of SauD InoB can be found on the course web page [33].

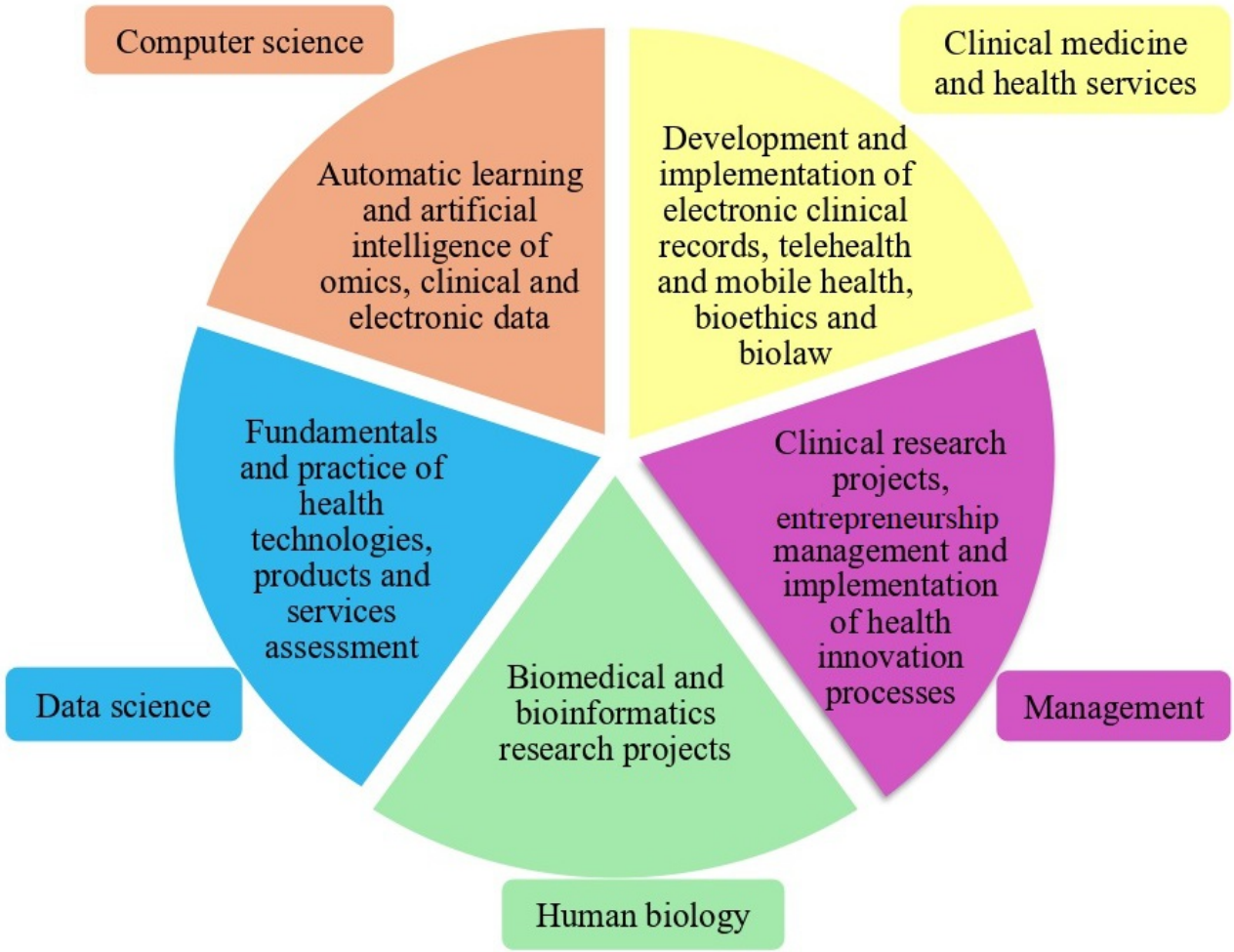
Curriculum and Teaching and Learning Methods

SauD InoB comprises 6 academic semesters and 180 ECTS, integrating standardized and pluralized curricular components [34]. ECTS are units of learning based on learning outcomes and their associated workload and were defined to promote transparency and mobility within the European Higher Education Area.

Multimedia Appendix 1 lists the curricular structure of SauD InoB. The first 4 semesters (120 ECTS) constitute a common core, including a set of curricular units that all students must complete.

These 120 ECTS are divided into the areas of Computer Science (29 ECTS), Data Science (26 ECTS), Human Biology (26 ECTS), Clinical Medicine and Health Services (25 ECTS), and Management (11 ECTS). Elective units correspond to 3 more ECTS. Figure 1 shows the fundamental areas of SauD InoB.

Figure 1. Fundamental areas of SauD InoB and main topics addressed by each area.



In the last 2 semesters (60 ECTS), students must choose 1 of 4 branches, each of them with a stronger focus in 1 of the fundamental areas of SauD InoB: (1) Data Analytics and Artificial Intelligence is more focused on Data Science, (2) Health Information Systems and Telemedicine is more focused on Computer Science, (3) Clinical Research and Health Innovation Management is more focused on Clinical Medicine and Health Services, and (4) Biomedical Research and Bioinformatics is more focused on topics related to Human Biology. The creation of these branches was aligned not only with the fundamental scientific areas of the course but also with the professional profiles defined earlier by the educational ecosystem stakeholders. Each branch comprises 3 internship curricular units in the first semester and a project curricular unit in the second semester, in addition to a set of core and elective curricular units related to the subject.

Thus, the course enables an early differentiation in these profiles while providing introductory training in the different areas of knowledge, even in the third year, which confers a valuable competitive advantage for the labor market and responds to a pressing need in the health sector for professionals who have an overview of the sector and can build bridges between different professions and areas of knowledge.

The BSc includes theoretical, theoretical-practical, laboratory practice, and seminar classes, in which active teaching-learning methods, which promote interactions, will be privileged (eg, discussion of fictitious or real case studies, problem- or project-based learning, and flipped classroom). All branches include a project curricular unit in collaboration with one of the partners. Professional practice will be promoted through short-term observation and immersion activities and 3 internships in different areas in the third year. Despite its transdisciplinary approach, as health care processes are the central focus of SauD InoB, all branches include an internship in clinical medicine and health services.

Horizontal and vertical articulations will be promoted, for example, through use cases discussion, assignments involving various curricular units, simulation of patients' pathway in the health care services, and assignments fostering progressively complex, transdisciplinary, and critical approaches.

Teaching and learning activities will be supported by education technologies (eg, e-learning platforms and medical simulation devices).

Preliminary Assessment, Accreditation, and Funding of Saud InoB

The creation of new cycles of studies in Portugal requires a preliminary assessment and accreditation process [35,36], comprising the digital submission of a detailed proposal. After a preliminary analysis of compliance with national regulations and quality guidelines, the proposal is assessed by an external assessment commission. Besides assessing the written proposal, the commission conducts site visits and consultations and issues an evaluation report, which is the basis for the accreditation decision. More detailed information about accreditation can be found on the A3ES website [35].

The process of cocreating this BSc degree began in August 2021 and entailed the elaboration of a brief and an extended creation and accreditation proposal. The proposals were approved by the Pedagogical and Scientific Councils of the faculties involved in the cycle of studies and by the dean and senate of the university and were submitted to A3ES approval in February 2022.

In March 2023, the A3ES External Assessment Commission, including 3 external experts, issued a preliminary assessment report, which configured a peer review process comprising advice on the curriculum design and other suggestions for optimizing the BSc degree (eg, changes in the name of the BSc degree and a decrease in the number of branches, requiring changes to the curricular structure and to the syllabi of some curricular units). The commission included an international bioinformatician with expertise in genome analysis, systems biology, and computational tools for biomedical research; a professor of pharmacology and medical doctor with a strong background in biomedical sciences and medical education; and a digital health and health care management expert, with leadership experience in eHealth systems and medical informatics.

The task force incorporated most of the suggestions in a final proposal, and in May of 2023, SauD InoB was approved. The conclusions of the final assessment report asserted that “this training offer is not only pertinent and very innovative, but also of the highest quality and demand. The professionals it will train will make an unrivaled contribution to biomedical innovation structuring” [37].

The assessment reports can be found at the A3ES website. More information about the composition of the External Assessment Commission can be found in its final report [38].

The first edition of the cycle of studies was held in the 2024-2025 school year.

This cycle of studies was also included in the application of the University of Porto to the funds of Portugal’s Recovery and Resilience Plan, a plan integrated into NextGenerationEU [39], aiming to restore sustained economic growth in the postpandemic period and to respond to the challenges of the dual climate and digital transition [40].

Public Promotion of SauD InoB

Public promotion of SauD InoB emphasized that the new course aimed to promote health innovation in Portugal by creating professionals with skills suited to the current and future needs of the health care market. Prospective students were invited to deepen their knowledge of human biology, experience the daily work of health care services alongside doctors and health professionals, develop programming skills, and gain expertise in cutting-edge methods for analyzing biological or clinical data.

In structural terms, a plan was defined to publicize the new degree in May 2023 (soon after the course was approved by A3ES). Public promotion of SauD InoB included posts in social media, institutional newsletter and webpage, news media publications, direct marketing actions (eg, email, SMS, course, and flyer), distribution of course gifts, dissemination of the course in academic institutional events, education fairs and targeted events, activities and presentations in high schools, and a newsletter (SauD InoB News).

Multimedia Appendix 2 shows the strategic plan for the promotion of SauD InoB.

Results of the Competition for Accessing the First Edition of SauD InoB

Admission to first cycles of studies in Portuguese public higher education institutions is limited to a maximum number of placements (numerus clausus) and requires an application through an annual competition held by the Directorate-General for Higher Education under the general regime, special conditions (eg, top-level athletes and permanent staff of the Portuguese Armed Forces), or special competitions (eg, applicants older than 23 years or holders of a previous BSc degree) [11]. The admission process under the general regime is based on a weighted average of internal high school and national exams (entry score), ranging between 0 and 200 points. Some courses may also require the fulfillment of specific prerequisites, such as physical or functional aptitude tests. In the first edition of SauD InoB, 40 vacancies were opened for the first phase: 35 for the general regime, one under special conditions (for top athletes), and 4 for a special competition (holders of a previous BSc degree).

SauD InoB received 284 applications for the general competition, of which 83 were first option (candidates may apply for up to 6 cycles of studies, indicating their order of preference). The entry scores of the admitted applicants ranged between 179 and 199, placing SauD InoB as the 27th (out of 1119) cycle of studies with the highest entry score in the country, and 36 of the admitted applicants chose the course as their first option.

Considering the special competition for holders of a previous BSc degree, SauD InoB received 14 applications and filled all

4 vacancies. Results from the competition under special conditions were not available when this article was written.

Final Remarks and Future Work

This paper describes the cocreation and the development process of a BSc degree in Digital Health and Biomedical Innovation, intricately developed within an educational multidisciplinary and multisectoral ecosystem, aiming to train transdisciplinary, highly skilled, and creative professionals able to respond to the needs and challenges of the health care sector in the context of fast-paced innovation and digital transformation. Besides following the requirements and milestones for the creation and accreditation of cycles of studies as determined by A3ES, this creation process was based on Kern's approach to curriculum development [12] and grounded on the concept of ecological university [18-20].

This ecosystem-centered cocreation of a bachelor's degree enabled the team to learn a few lessons, which may be helpful for other researchers and curriculum developers:

1. The integration of diverse stakeholders in the cocreation process seems to have fostered the development of a more adapted curriculum, flexible and responsive to the needs of diverse employers and to the needs of the users of the services that will be delivered by the professionals trained by Saud InoB;
2. The results of the competition for accessing the first edition of SauD InoB and subsequent admission of students with high classifications confirmed the relevance of this BSc course. Furthermore, the vacancies under special conditions and the special competition may increase the diversity of students and contribute to the democratization of access to SauD InoB;
3. More research is needed on curriculum development methods.

Further work will be needed to monitor the implementation of SauD InoB and promote continuous adaptation of the education processes to the needs of students and other actors. This involves pursuing collaboration with its current partners, namely through regular partner meetings, widening community involvement by joining new partners, and designing tools and strategies to involve students (and, in the future, graduates) in the processes of continuous assessment and improvement of SauD InoB. Course assessment will be pursued through meetings with faculty and students' representatives. The University of Porto also assesses all the courses through pedagogical questionnaires from the students. Furthermore, all the degree-awarding courses in Portugal are periodically assessed by A3ES.

This work may contribute to the knowledge about processes involved in creating new cycles of studies and informing academics and other stakeholders on the practices involved in such processes. Moreover, and even if the first edition of SauD InoB is still taking its first steps, this cocreation process and educational ecosystem–centered approach may be already considered a successful experience of productive interactions [41], which questions (even if not in an uncritical way) [18] a discourse of higher education in crisis [42,43] and reflects higher education's more than ever active role in promoting bridges between different actors who, in a knowledge society, join forces to promote the improvement of societies and the well-being of the population.

Acknowledgments

The authors would like to acknowledge the support and contributions to the creation of the Bachelor's degree in Digital Health and Biomedical Innovation (Saud InoB) from the Academic Management and Communication Management Units of the Faculty of Medicine of the University of Porto (FMUP), the Communication and Image Service of the University of Porto, the CHAIR in Onco-Innovation of FMUP, and all the partners (Parceiros [24]) and faculty (Docentes da Licenciatura SauD InoB [44]) of Saud InoB.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Curricular structure.
[DOCX File, 37 KB - mededu_v11i1e63903_app1.docx]

Multimedia Appendix 2
Strategic plan for the promotion of the new SauD InoB degree at the Faculty of Medicine of the University of Porto (FMUP).
[DOCX File, 17 KB - mededu_v11i1e63903_app2.docx]

References

1. Newell S, Goussevskaia A, Swan J, Bresnen M, Obembe A. Interdependencies in complex project ecologies: the case of biomedical innovation. Long Range Plann 2008 Feb;41(1):33-54. [doi: 10.1016/j.lrp.2007.10.005]

2. Swan J, Goussevskaia A, Newell S, Robertson M, Bresnen M, Obembe A. Modes of organizing biomedical innovation in the UK and US and the role of integrative and relational capabilities. Res Policy 2007 May;36(4):529-547. [doi: 10.1016/j.respol.2007.02.014]
3. Gopal G, Suter-Crazzolara C, Toldo L, Eberhardt W. Digital transformation in healthcare - architectures of present and future information technologies. Clin Chem Lab Med 2019 Feb 25;57(3):328-335. [doi: 10.1515/cclm-2018-0658] [Medline: 30530878]
4. Stoumpos AI, Kitsios F, Talias MA. Digital transformation in healthcare: technology acceptance and its applications. Int J Environ Res Public Health 2023 Feb 15;20(4):3407. [doi: 10.3390/ijerph20043407] [Medline: 36834105]
5. Agenda temática de investigação e inovação: saúde, investigação clínica e de translação [thematic research and innovation agenda: health, clinical and translational research], lisboa: fundação para a ciência e a tecnologia [Report in Portuguese]. : Fundação para a Ciência e a Tecnologia; 2019 URL: https://www.fct.pt/wp-content/uploads/2022/05/Agenda_Saude_Investigacao_Clinica_e_de_Translacao_Versao_Finalizacao.pdf [accessed 2025-08-21]
6. Hospodková P, Berežná J, Barták M, Rogalewicz V, Severová L, Svoboda R. Change management and digital innovations in hospitals of five European countries. Healthcare (Basel) 2021 Nov 5;9(11):1508. [doi: 10.3390/healthcare9111508] [Medline: 34828554]
7. Milella F, Minelli EA, Strozzi F, Croce D. Change and innovation in healthcare: findings from literature. Clinicoecon Outcomes Res 2021;13:395-408. [doi: 10.2147/CEOR.S301169] [Medline: 34040399]
8. Digital skills for health professionals. : European Health Parliament; 2016 URL: https://www.healthparliament.eu/wp-content/uploads/2017/09/Digital-skills-for-health-professionals.pdf [accessed 2025-08-21]
9. Global strategy on digital health 2020-2025. World Health Organization. 2021. URL: https://iris.who.int/handle/10665/344249 [accessed 2025-08-21]
10. Zainal H, Tan JK, Xiaohui X, Thumboo J, Yong FK. Clinical informatics training in medical school education curricula: a scoping review. J Am Med Inform Assoc 2023 Feb 16;30(3):604-616. [doi: 10.1093/jamia/ocac245] [Medline: 36545751]
11. Suplemento ao diploma [Web page in Portuguese]. Direção-Geral do ensino Superior. 2025. URL: https://www.dges.gov.pt/pt/pagina/suplemento-ao-diploma [accessed 2025-08-21]
12. Tomas PA, et al. Curriculum Development for Medical Education: A Six-Step Approach: Baltimore: Johns Hopkins University Press; 2009.
13. Crosling G, Nair M, Vaithilingam S. A creative learning ecosystem, quality of education and innovative capacity: a perspective from higher education. Studies in Higher Education 2015 Aug 9;40(7):1147-1163. [doi: 10.1080/03075079.2014.881342]
14. Hecht M, Crowley K. Unpacking the learning ecosystems framework: lessons from the adaptive management of biological ecosystems. Journal of the Learning Sciences 2020 Mar 14;29(2):264-284. [doi: 10.1080/10508406.2019.1693381]
15. Niemi H. Building partnerships in an educational ecosystem. CEPSj 2016;6(3):5-15 [FREE Full text] [doi: 10.26529/cepsj.62]
16. Väljataga T, Poom-Valickis K, Rumma K, Aus K. Transforming higher education learning ecosystem: teaching staff perspective. IxD&A 2021;46(46):47-69. [doi: 10.55612/s-5002-046-003]
17. Bronfenbrenner U, Morris PA. The bioecological model of human development. In: Handbook of Child Psychology, 6th edition: John Wiley & Sons, Inc; 2006, Vol. 1:793-828. [doi: 10.1002/9780470147658]
18. Barnett R. University knowledge in an age of supercomplexity. Higher Education 2000 Dec;40(4):409-422. [doi: 10.1023/A:1004159513741]
19. Barnett R. The Ecological University: A Feasible Utopia: Oxford: Routledge; 2018. [doi: 10.4324/9781315194899]
20. Barnett R, Bengtsen S. Universities and epistemology: from a dissolution of knowledge to the emergence of a new thinking. Education Sciences 2017;7(1):38. [doi: 10.3390/educsci7010038]
21. Belita E, Carter N, Bryant-Lukosius D. Stakeholder engagement in nursing curriculum development and renewal initiatives: a review of the literature. QANE-AFI 2020;6(1). [doi: 10.17483/2368-6669.1200]
22. Satterfield JM, Werder K, Reynolds S, Kryzhanovskaya I, Curtis AC. Transforming an educational ecosystem for substance use disorders: a multi-modal model for continuous curricular improvement and institutional change. Subst Abus 2022;43(1):1953-1962. [doi: 10.1080/08897077.2022.2116742] [Medline: 36053217]
23. Bischoff K, Volkmann CK, Audretsch DB. Stakeholder collaboration in entrepreneurship education: an analysis of the entrepreneurial ecosystems of European higher educational institutions. J Technol Transf 2018 Feb;43(1):20-46. [doi: 10.1007/s10961-017-9581-0]
24. List of partners. URL: https://saudinob.med.up.pt/parceiros/ [accessed 2025-08-21]
25. Mantas J, Ammenwerth E, Demiris G, et al. Recommendations of the International Medical Informatics Association (IMIA) on Education in Biomedical and Health Informatics. First Revision. Methods Inf Med 2010 Jan 7;49(2):105-120. [doi: 10.3414/ME5119] [Medline: 20054502]
26. Costa PD, Almeida J, Araujo SM, et al. Biomedical and health informatics teaching in Portugal: Current status. Heliyon 2023 Mar;9(3):e14163. [doi: 10.1016/j.heliyon.2023.e14163] [Medline: 36967900]
27. Pesquisa de cursos e instituições [Web page in Portuguese]. Direção-Geral do Ensino Superior. 2024. URL: https://www.dges.gov.pt/pt [accessed 2025-08-21]
28. Digital Health MSc. Deggendorf Institute of Technology. 2025. URL: https://th-deg.de/dh-m-en [accessed 2025-08-21]

29. Health and digital transformation. Maastricht University. 2025. URL: https://www.maastrichtuniversity.nl/education/master/programmes/health-and-digital-transformation [accessed 2025-08-21]
30. Master in digital healthcare. Barcelona Technology School. 2025. URL: https://barcelonatechnologyschool.com/master/master-in-digital-healthcare/ [accessed 2025-08-21]
31. IN4380 – digital transformation of healthcare [Web page in Norwegian]. Universitetet i oslo. 2025. URL: https://www.uio.no/studier/emner/matnat/ifi/IN4380/ [accessed 2025-08-21]
32. Ha Choi Y, Bouwma-Gearhart J, Ermis G. Doctoral students' identity development as scholars in the education sciences: literature review and implications. IJDS 2021;16:089-125. [doi: 10.28945/4687]
33. Digital health and biomedical innovation [Web page in Portuguese]. Faculdade de Medicina da Universidade do Porto. URL: https://sigarra.up.pt/fmup/en/CUR_GERAL.CUR_VIEW?pv_ano_lectivo=2024&pv_origem=CUR&pv_tipo_cur_sigla=L&pv_curso_id=30781 [accessed 2025-08-21]
34. González-Ocampo G, Kiley M, Lopes A, et al. The curriculum question in doctoral education. FLR 2015;3(3):23-38. [doi: 10.14786/flr.v3i3.191]
35. Prior accreditation of new study programmes [Web page in Portuguese]. Agência de avaliação e acreditação do ensino superior.: A3ES URL: https://a3es.pt/en/assessment-and-accreditation/study-programmes/new-study-programmes/ [accessed 2025-08-27]
36. Standards and guidelines for quality assurance in the european higher education area (ESG). European higher education area and bologna process. 2025. URL: https://tinyurl.com/yr5vuvsj [accessed 2025-08-21]
37. NCE/21/2100374: relatório final da CAE - novo ciclo de estudos [NCE/21/2100374: final report from CAE - new cycle of studies] [Report in Portuguese]. : External Assessment Commission; 2023 URL: https://si.a3es.pt/sia3es/page?stage=ConsultaPublicaProcesso&processID=25698 [accessed 2025-08-27]
38. Relatório final da CAE. Agência de Avaliação e Acreditação do Ensino Superior [Web page in Portuguese]. 2023. URL: https://si.a3es.pt/sia3es/page?stage=ConsultaPublicaProcesso&processID=25698 [accessed 2025-08-28]
39. NextGenerationEU. European Union. 2025. URL: https://next-generation-eu.europa.eu/index_en [accessed 2025-08-21]
40. Plano de recuperação e resiliência– desafios e oportunidades [Web page in Portuguese]. Recuperar Portugal. 2025. URL: https://recuperarportugal.gov.pt/plano-de-recuperacao-e-resiliencia/ [accessed 2025-08-21]
41. Akker W, Spaapen J. Productive interactions: societal impact of academic research in the knowledge society. : League of European Research; 2017 URL: https://www.leru.org/publications/productive-interactions-societal-impact-of-academic-research-in-the-knowledge-society [accessed 2025-08-21]
42. Mintz B. Neoliberalism and the crisis in higher education: the cost of ideology. American J Econ Sociol 2021 Jan;80(1):79-112 [FREE Full text] [doi: 10.1111/ajes.12370]
43. Readings B. The University in Ruins: Harvard University Press; 1996. URL: https://uchile.cl/dam/jcr:6bb0e38f-c349-43d5-92d4-7a358e025af5/Readings%20Bill%20University%20in%20Ruins.pdf [accessed 2025-09-12]
44. Docentes da licenciatura suad inob. Faculty of Medicina University of Porto [web page in Portuegese]. URL: https://saudinob.med.up.pt/ [accessed 2025-09-22]

Abbreviations

A3ES: Agency for Assessment and Accreditation of Higher Education
ECTS: European Credit Transfer and Accumulation System credits
InoB: Bachelor's degree in Digital Health and Biomedical Innovation

Edited by B Lesselroth; submitted 18.09.24; peer-reviewed by A Naghettini, M Friebe; revised version received 16.03.25; accepted 19.04.25; published 23.09.25.

Please cite as:
Alves P, Costa E, Costa-Pereira A, Falcão-Pires I, Fonseca J, Leite-Moreira A, Sousa-Pinto B, Vale N
An Ecosystem Approach to Developing and Implementing a Cocreated Bachelor's Degree in Digital Health and Biomedical Innovation
JMIR Med Educ 2025;11:e63903
URL: https://mededu.jmir.org/2025/1/e63903
doi:10.2196/63903

Beyond Chatbots: Moving Toward Multistep Modular AI Agents in Medical Education

Minyang Chow[1,2*], MBBS, MRCP, MS-HPEd; Olivia Ng[2*], PhD

[1]Group Clinical Education, National Healthcare Group, 1 Mandalay Rd, Singapore, Singapore

[2]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore

[*]all authors contributed equally

Corresponding Author:
Minyang Chow, MBBS, MRCP, MS-HPEd
Group Clinical Education, National Healthcare Group, 1 Mandalay Rd, Singapore, Singapore

Abstract

The integration of large language models into medical education has significantly increased, providing valuable assistance in single-turn, isolated educational tasks. However, their utility remains limited in complex, iterative instructional workflows characteristic of clinical education. Single-prompt AI chatbots lack the necessary contextual awareness and iterative capability required for nuanced educational tasks. This Viewpoint paper argues for a shift from conventional chatbot paradigms toward a modular, multistep artificial intelligence (AI) agent framework that aligns closely with the pedagogical needs of medical educators. We propose a modular framework composed of specialized AI agents, each responsible for distinct instructional subtasks. Furthermore, these agents operate within clearly defined boundaries and are equipped with tools and resources to accomplish their tasks and ensure pedagogical continuity and coherence. Specialized agents enhance accuracy by using models optimally tailored to specific cognitive tasks, increasing the quality of outputs compared to single-model workflows. Using a clinical scenario design as an illustrative example, we demonstrate how task specialization, iterative feedback, and tool integration in an agent-based pipeline can mirror expert-driven educational processes. The framework maintains a human-in-the-loop structure, with educators reviewing and refining each output before progression, ensuring pedagogical integrity, flexibility, and transparency. Our proposed shift toward modular AI agents offers significant promise for enhancing educational workflows by delegating routine tasks to specialized systems. We encourage educators to explore how these emerging AI ecosystems could transform medical education.

(JMIR Med Educ 2025;11:e76661) doi:10.2196/76661

KEYWORDS

artificial intelligence; agents; pedagogy; large language models; instructional design

Introduction

The use of large language models (LLMs), such as ChatGPT, has surged in medical education, largely due to their capacity to generate answers, summaries, or clinical scenarios on demand [1]. While these tools offer convenience, most implementations remain confined to single-turn interactions with minimal contextual awareness [2]. As such, they are ill-suited to the inherently complex, multistep processes that define both clinical reasoning and medical education instructional design.

Medical education is neither linear nor static. Learners must navigate evolving information, engage in iterative reflection, and integrate feedback across diverse domains. Educators, likewise, must design learning experiences that are both structured and flexible, adapting in real time to learner needs and contextual constraints. In this Viewpoint article, we argue that advancing the role of artificial intelligence (AI) in this domain requires a conceptual and architectural shift—from chatbots to modular agents.

Elements of the proposed multistep modular agent framework are already in pilot use at our institution. For example, a history-taking application uses an agentic patient alongside companion agents that surface summaries and nudge questions when learners stall. Similar workflows have been trialed for scenario design and rubric authoring across both undergraduate and postgraduate medical education settings. We hope this Viewpoint will be useful for educators and support staff (eg, instructional designers) who are seeking structured, pedagogically aligned ways to integrate AI into their teaching and assessment practices.

Why Single-Prompt AI Chatbots Fall Short in Complex Educational Workflows

LLMs have demonstrated strong performance in generating simulated patient responses, drafting multiple-choice questions with plausible distractors, and providing clear, conversational explanations to learners, making them valuable for isolated, language-driven tasks in medical education [1,3,4]. These interactions, however, remain fundamentally static and

transactional. The AI responds once to a prompt, with limited contextual memory and no intrinsic understanding of how its output fits into a broader pedagogical framework. This architecture is not suited to the realities of medical education, which demand systems that can manage evolving learner inputs, sustain contextual awareness, and scaffold multiple interconnected subtasks. Rather than seeking access to models' internal "reasoning," we emphasize transparent, reviewable intermediate artifacts. For example, tables of objectives, draft scenarios, and rubric levels. These should be open to educator review at frequently defined checkpoints.

Medical educational workflows combine structure with adaptability. On one hand, structured elements such as predefined curricula, learning objectives, and assessment frameworks provide learners with a clear roadmap [5]. These components are essential for building knowledge systematically, maintaining consistency across training programs, and aligning educational outcomes with professional standards [6]. At the same time, medical education demands flexibility. Learners must continuously adapt their understanding in response to new information, clinical contexts, and feedback [7]. This dynamic process mirrors the realities of clinical practice, where patient presentations are often unpredictable, and decisions evolve as new data emerge [8].

For instance, designing a simulation scenario or a performance rubric involves more than a single step [9]. It requires iterative alignment with learning objectives, incorporation of cultural and contextual considerations, and ongoing refinement based on feedback. These are complex and adaptive processes rather than linear tasks. Relying on a single, comprehensive prompt to guide such development can result in superficial or inconsistent outputs [2]. It also places an unrealistic expectation on the prompt designer to capture all instructional nuances at once.

Modular AI Agent Framework: A Pedagogically Aligned Alternative

To address the limitations of single-dimension chatbot systems in educational design, we propose a shift in consideration to a modular, multistep AI agent framework. An agent is a specialized AI process that operates with explicit goals, constraints, contextual inputs [10,11], and error-handling capacity. This contrasts with simple chained prompts, which merely sequence tasks without structured autonomy or memory (Table 1). Agents can operate in parallel as well as sequentially, whereas chained prompts are inherently linear.

Table . Comparison of key features of chatbots and multiagent frameworks [10,11].

Dimension	Chatbots (single or chained prompts)	Modular multiagent frameworks
Transparency	Outputs are all-in-one; intermediate reasoning often hidden	Intermediate artifacts (objectives, drafts, rubrics) explicit and reviewable
Model specialization	One model handles all subtasks; prompt engineering is the only lever	Each agent can call the model or tool best suited for its subtask
Error localization	Errors propagate; difficult to isolate or correct	Errors traceable to specific agents; easier to isolate and correct
Evaluation	Only final output assessable	Stepwise evaluation at each checkpoint
Cost	Lower per run but may require multiple retries	Higher per run (multiple agents) but more efficient for complex workflows
Pedagogical fit	Possible with skilled prompt design and oversight, but constrained by single system prompt	Agents can be designed to map systematically onto instructional frameworks
Instructional scope	Constrained by single system prompt (eg, 8000 characters in customGPT as of September 2025); must pack all instructions into one prompt; success depends on model's ability to follow and prioritize prompts	Each agent only receives the instructions it needs for its task, reducing ambiguity and improving compliance
Scalability	Harder to adapt; changing prompts may affect entire workflow	Extensible; new agents can be added or replaced independently
Workflow structure	Sequential, linear	Supports both series (one task after another) and parallel (simultaneous subtasks) workflows

Rather than tasking a single agent with the entire instructional-design workflow, responsibilities are distributed across a pipeline of specialized agents, each optimized for its step [10,11]. These agents can invoke function calls or other tool integrations so that every agent not only reasons in natural language but also acts with the precise utilities needed to fulfill its mandate.

An *AI agentic framework* can model the expert-led, multistep process educators use in scenario design by integrating task specialization, interactive user feedback, and tool integration. Each specialized agent hands its output to the next, preserving logical and pedagogical continuity while an overarching orchestrator agent monitors the entire pipeline, coordinates handoffs, and can route back to the relevant specialist agent for further iteration when needed [12]. This intuitive structure supports extensibility through tool integration and scalability

by allowing agents to be refined, replaced, or expanded based on evolving needs. In other words, pedagogy comes first: agents map intuitively onto how we already teach and assess. MEDCO (Medical Education Copilots), for example, optimizes learner-and-agent dialogue in simulated encounters [13]. In comparison, our educator-in-the-loop approach focuses sequentially on the lesson design objectives, story, tasks, and marking guide. Each draft is reviewed and approved before moving on. This enables educators to remain central to the process, reviewing and approving outputs at each stage to safeguard pedagogical integrity and contextual relevance.

An Example in Scenario Design

We present an example of a modular AI agentic framework that offers a structured yet flexible approach to developing clinical communication scenarios. This design allows the system to emulate the collaborative workflow of a multidisciplinary education team including educators, instructional designers, content experts, and assessment specialists by assigning each phase of scenario construction to an agent optimized for that purpose.

This example system is composed of 7 AI agents, each operating within clearly defined instructional boundaries. Other than the orchestrating agent, 6 other agents function sequentially, handing off their outputs to the next agent in the pipeline to ensure continuity, coherence, and pedagogical alignment. Each agent is responsible for a specific task that reflects real-world instructional design practices, where complex deliverables are broken into manageable, interdependent stages.

The orchestrating agent oversees the entire workflow, ensuring that each specialized agent performs its task effectively and that the outputs are cohesive and aligned with the pedagogical goals. The orchestrating agent can return generated outputs from specific agents for further iteration if required.

The scenario overview agent initiates the workflow by generating a scenario title and a structured table of learning or assessment objectives. Educators are prompted to review and refine objectives before proceeding.

The storyline agent builds on the approved objectives by producing 3 patient storylines. Each includes narrative elements, such as patient background, emotional and physical concerns, hidden agendas, and sociocultural context, with explicit localisation to the Singaporean setting.

The candidate instructions agent crafts clear, structured learner instructions, including task descriptions and background details.

The surrogate patient script agent generates a detailed script to guide the simulated patient's behavior. It incorporates behavioral cues, emotional tone, and plausible responses to candidate questions. The agent can retrieve example scripts via file searches and invoke web tools to integrate culturally relevant details.

The scoring rubric agent allows users to define their own rubric, including domains of competence and marking scales. Users can construct a table of performance criteria that aligns with the scenario's learning objectives. The rubric incorporates score levels, descriptive anchors, and exemplar behaviors. Additionally, tool-calling and web search functionality enable the retrieval of validated rubrics to ensure alignment with existing assessment frameworks.

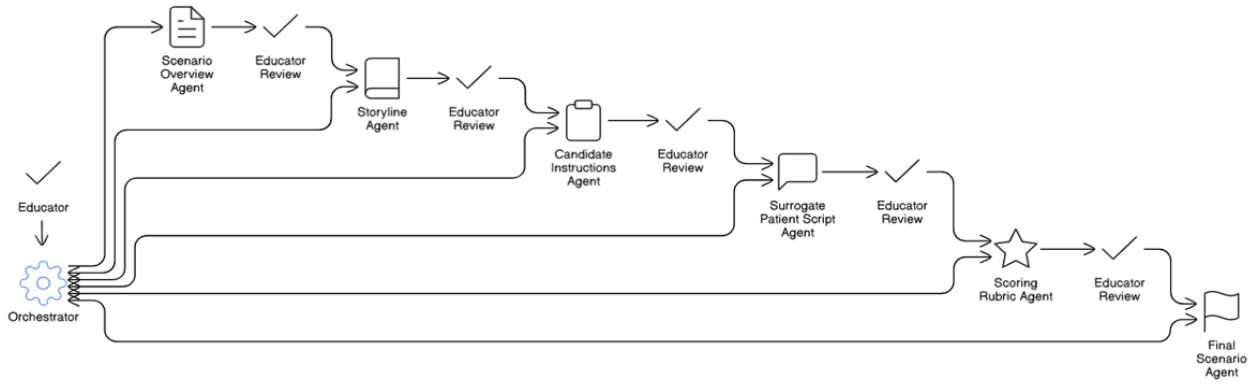
The final scenario agent compiles all approved components into a cohesive scenario document. It ensures consistency in formatting, terminology, and contextual accuracy, preparing the scenario for direct use or further adaptation by educators.

Figure 1 illustrates the 7-agent pipeline with educator checkpoints and the orchestrating agent. Each handoff (represented by arrows) is schema-checked; the educator can approve or request revision before progression.

In agent-based generative workflows, the concept of a discrete "error" is less relevant than in traditional rule-based systems. Output quality is influenced by multiple factors: the specificity and clarity of user prompts, the depth of reasoning elicited by the design structure, and the availability of exemplars or templates that guide generation. Agentic frameworks allow for improved output quality, as each agent is specifically trained and instructed to perform a certain task and provided with the necessary tools to execute it effectively. Specific model types can also be selected to best fit the nature of the task, for example, reasoning, long-context, tool-use, and vision tasks. The modular framework uses these elements to produce educational artifacts that are functional, pedagogically sound, and contextually sensitive.

By breaking down the scenario design process into agent-specific tasks, the system supports a structured and transparent development process.

Figure 1. Seven-agent pipeline for scenario design with educator checkpoints. The orchestrating agent routes tasks, monitors quality, and reruns only failing steps; educators review intermediate artifacts at each gate.



Educational Theory Alignment and Pedagogical Implications

Drawing from instructional design theory, each agent's role mirrors existing teaching and assessment practices—from objective setting to content creation to assessment alignment [14]. It also resonates with complexity theory in education, which acknowledges the nonlinear, adaptive, and emergent nature of curriculum design [8]. By assigning each AI agent a tailored reasoning strategy, such as generative models for creative storytelling or deterministic models for rubric precision, we preserve pedagogical coherence while leveraging the strengths of LLMs in context-specific ways. Moreover, the use of AI agents to build culturally localized cases (eg, context-specific language patient scripts) demonstrates how modularity can enable personalization without compromising scalability. The overall system thus acts not as a content generator, but as a structured collaborator in instructional design.

Human-in-the-Loop: Safeguarding Pedagogical Quality and Trust

Central to our framework is the conviction that AI agents must augment—not replace—human educators. Every stage of the agentic pipeline therefore includes a mandatory checkpoint where educators review, revise, and approve an agent's output before it is handed to the next specialist agent. Only when the reviewer is satisfied, and no further iterations are requested, does the workflow advance.

This checkpoint design yields 3 advantages. First, pedagogical assurance: continuous expert oversight guarantees that scenarios remain faithful to institutional standards, educator requirements, learner needs, and real-world clinical practice. Second, iterative flexibility: at any time, users can return to an earlier step, reopen an agent's draft, and trigger additional refinement cycles. This bidirectional flow prevents "lock-in" to premature decisions and supports rapid improvement when curricular goals evolve. Third, transparency and trust: because every AI-generated artifact is treated as a draft, it remains open to critique rather than being accepted as a "black box" verdict. Educators retain clear visibility into the rationale behind each decision. This collaborative, human-guided process reflects traditional instructional design workflows, ensuring that final materials are shaped by pedagogical judgment rather than relying solely on algorithmic output.

To provide readers with a familiar curricular reference point, we align the proposed agentic pipeline with Kern's [15] 6 steps of curriculum development. This alignment is not intended as a prescriptive template, but rather as an interpretive framework that illustrates how agent outputs and educator checkpoints correspond to established stages of curriculum design. In doing so, Kern's [15] model provides conceptual coherence, while the agentic workflow demonstrates a practical means of operationalizing each step. Table 2 summarizes this mapping.

This approach reflects the collaborative nature of instructional design, where materials are developed through iterative review and decisions are guided by pedagogical judgment rather than algorithmic output alone.

Table . Agent-to-pedagogy mapping: how each agent (and the educator) contributes across Kern's steps 1 - 6.

Kern's 6 steps	Agent and educator roles
Step 1: problem identification and general needs assessment	Scenario overview agent frames the educational problem and proposes objectives; orchestrator agent ensures consistency. Educator reviews problem framing and refines objectives.
Step 2: targeted needs assessment	Scenario overview agent structures learner-specific needs. Educator validates alignment with learner context and institutional requirements before progression.
Step 3: goals and objectives	Storyline agent translates objectives into case narratives with sociocultural context. Educator evaluates relevance, realism, and coherence of the narratives.
Step 4: educational strategies	Candidate instructions agent defines learner tasks; surrogate patient script agent operationalizes the interaction; storyline agent scaffolds narrative complexity. Educator provides iterative feedback, ensuring strategies remain pedagogically sound.
Step 5: implementation	Scoring rubric agent designs criteria for performance; final scenario agent compiles all components; orchestrator agent ensures alignment. Educator approves rubrics, validates final scenario quality, and ensures usability.
Step 6: evaluation and feedback	Scoring rubric agent embeds standards and feedback anchors; orchestrator agent supports routing to the educator for revision. Educator delivers final judgments, integrates feedback, and ensures final product is satisfactory.

Practical Considerations

Some may argue that carefully engineered single-prompt workflows within a large language model (eg, customGPTs), can produce comparable results. Yet empirical data show that single-prompt workflows still trail agentic teams in diagnostic accuracy and bias mitigation [16-18].

Agent modularity also offers benefits that chained prompts find difficult to replicate. First, it allows the educator end user to assign each subtask to the foundation model best suited to the cognitive demands of the task. For example, a reasoning-optimized model can be used for diagnostic planning, a vision-capable model for image annotation, and a fast, low-cost model for rubric scoring.

Second, specialized agents can issue narrowly focused tool or function calls, which helps ensure stricter compliance with domain-specific instructions. An analogy illustrates this: asking a single writer to produce an entire clinical simulation, including learning objectives, patient script, scoring rubric, and feedback, in one attempt is less reliable than assigning each section to professionals who have the right expertise and references. Modularity gives an AI pipeline that same division of labor advantage.

Third, modularity localizes failure and simplifies evaluations. Breaking the workflow into a series of small, purpose-built agents lets educators spot problems exactly where they occur, verify the quality of each step in isolation, and rerun or update only the affected agent without disturbing the rest of the pipeline.

While the modular-agent approach offers clear advantages, it is not without trade-offs. Stitching together multiple agents raises implementation complexity, especially in low-resource or nontechnical settings. Here, "vibe coding," a visual and block-based approach to wiring agents, tool calls, and routing rules, is helping to lower the barrier for educators who do not have coding experience [19]. Vibe coding allows educators to create educational apps by instructing AI in natural language. Any educator who hopes to pivot toward AI-augmented end-to-end educational workflows must begin with careful process decomposition. The instructional journey should be mapped into discrete steps, with the cognitive and technical requirements of each step clearly specified [20,21]. There are several practical implementation choices to consider. The input and output schemas between agents are standardized at each step with simple templates to ensure minimum quality standards. Educators are provided with "prompt cards" documenting instructions, data sources, tools, and approvals, with any changes logged for accountability. All drafts are automatically logged with source, model, prompt card, and a summary that enables traceability. Clear acceptance criteria guide fallback or rollback when a step underperforms, allowing repair or educator-initiated rewrites without restarting the pipeline. Versioning and update policies are also in place. Runs typically complete within minutes, with token use and costs managed by combining efficient models for routine steps with reasoning models for critical alignment.

Limitations

This Viewpoint advances a conceptual and architectural position supported by early pilot deployments; however, we have not yet conducted prospective controlled studies to compare modular multiagent pipelines with single-step chatbots in terms of their impact on learner outcomes.

Conclusion: Rethinking AI's Role in Medical Education Together

Moving beyond chatbot paradigms opens new possibilities for the use of AI in medical education. Modular, agent-based

systems can mirror the sophistication of human instructional design, support iterative feedback loops, and preserve important contextual nuance. Their true value lies not in replacing educator labor, but in amplifying educator intent. Realizing this potential requires investment in systems that prioritize transparency, alignment, and human AI collaboration.

One promising step is the shift toward purpose-built agent ecosystems. Rather than relying on single-prompt chatbots, educators can begin by decomposing the instructional process, mapping each step, and identifying where a specialized agent, equipped with the appropriate model, data access, and tool-calling privileges, might help relieve cognitive load or automate repetitive tasks. Delegating routine activities to autonomous agents allows human experts to reclaim valuable time for higher-order mentoring, empathy, and creative pedagogy. Embracing this agentic mindset can help AI feel less like a black box chatbot and more like a transparent, collaborative partner that transforms tedious processes into opportunities for deeper learning and innovation.

We invite educators who are curious and committed to innovation to join us in exploring and shaping these agentic ecosystems together.

Acknowledgments

The authors acknowledge the use of ChatGPT with GPT-4o and GPT-5 (OpenAI) in the editorial refinement of this manuscript. These tools were used exclusively to improve clarity of sentence structure and address grammatical nuances. The authors did not use generative AI for ideation, conceptual development, or drafting of content. At all stages, the authors' critical judgment and scholarly voice were preserved. All outputs were independently reviewed, revised, and verified to ensure scholarly integrity and intellectual authorship. There is no funder involvement in this Viewpoint.

Conflicts of Interest

None declared.

References

1. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ 2024 Nov;58(11):1276-1285. [doi: 10.1111/medu.15402] [Medline: 38639098]
2. Ray PP. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Int Things Cyber-Phys Syst 2023;3:121-154. [doi: 10.1016/j.iotcps.2023.04.003]
3. Ng O, Phua DH, Chu J, Wilding LVE, Mogali SR, Cleland J. Answering patterns in SBA items: students, GPT3.5, and Gemini. MedSciEduc 2024;35(2):629-632. [doi: 10.1007/s40670-024-02232-4]
4. Arun G, Perumal V, Urias F, et al. ChatGPT versus a customized AI chatbot (Anatbuddy) for anatomy education: a comparative pilot study. Anat Sci Educ 2024 Oct;17(7):1396-1405. [doi: 10.1002/ase.2502] [Medline: 39169464]
5. Harden RM. AMEE Guide No. 21: curriculum mapping: a tool for transparent and authentic teaching and learning. Med Teach 2001 Mar;23(2):123-137. [doi: 10.1080/01421590120036547] [Medline: 11371288]
6. Haji F, Morin MP, Parker K. Rethinking programme evaluation in health professions education: beyond "did it work?". Med Educ 2013 Apr;47(4):342-351. [doi: 10.1111/medu.12091] [Medline: 23488754]
7. Han SP, Wang X, Kiruparan P, et al. Preparation for practice: what are students learning for? Clin Teach 2024 Dec;21(6):e13796. [doi: 10.1111/tct.13796] [Medline: 39162347]
8. Mennin S. Self-organisation, integration and curriculum in the complex world of medical education. Med Educ 2010 Jan;44(1):20-30. [doi: 10.1111/j.1365-2923.2009.03548.x] [Medline: 20078753]
9. Shah C, Davtyan K, Nasrallah I, Bryan RN, Mohan S. Artificial intelligence-powered clinical decision support and simulation platform for radiology trainee education. J Digit Imaging 2023 Feb;36(1):11-16. [doi: 10.1007/s10278-022-00713-9] [Medline: 36279026]
10. Marik V, Stepankova O, Krautwurmova H, Luck M, editors. Multi-Agent-Systems and Applications II: 9th ECCAI-ACAI/EASSS 2001, AEMAS 2001, HoloMAS 2001. Selected Revised Papers: Springer; 2003, Vol. 2322.
11. Wooldridge M, Jennings NR. Intelligent agents: theory and practice. Knowl Eng Rev 1995 Jun;10(2):115-152. [doi: 10.1017/S0269888900008122]
12. Huang KA, Choudhary HK, Kuo PC. Artificial intelligent agent architecture and clinical decision-making in the healthcare sector. Cureus 2024 Jul;16(7):e64115. [doi: 10.7759/cureus.64115] [Medline: 39119387]
13. Wei H, Qiu J, Yu H, Yuan W. MEDCO: medical education copilots based on a multi-agent framework. . Preprint posted online on Aug 22, 2024. [doi: 10.48550/arXiv.2408.12496]
14. Frerejean J, Dolmans D, Merrienboer JJG. Research on instructional design in the health professions: from taxonomies of learning to whole-task models. Res Med Educ 2022:291-302. [doi: 10.1002/9781119839446]
15. Kern DE. A six-step approach to curriculum development. In: Thomas P, Kern D, M H, Chen B, editors. Curriculum Development for Medical Education: Hopkins Press; 2016:5-9.

16. Ke Y, Yang R, Lie SA, et al. Mitigating cognitive biases in clinical decision-making through multi-agent conversations using large language models: simulation study. J Med Internet Res 2024 Nov 19;26:e59439. [doi: 10.2196/59439] [Medline: 39561363]
17. Tang X, Zou A, Zhang Z, et al. MedAgents: large language models as collaborators for zero-shot medical reasoning. ArXiv. Preprint posted online on Jun 4, 2023. [doi: 10.48550/arXiv.2311.10537]
18. Chen X, Yi H, You M, et al. Enhancing diagnostic capability with multi-agents conversational large language models. NPJ Digit Med 2025 Mar 13;8(1):159. [doi: 10.1038/s41746-025-01550-0] [Medline: 40082662]
19. Chow M, Ng O. From technology adopters to creators: leveraging AI-assisted vibe coding to transform clinical teaching and learning. Med Teach 2025 Apr 9:1-3. [doi: 10.1080/0142159X.2025.2488353] [Medline: 40202513]
20. Bloom BS, Engelhart MD, Furst EJ, Hill WH, Krathwohl DR. Taxonomy of Educational Objectives: The Classification of Educational Goals Handbook I: Cognitive Domain: David McKay Company; 1956.
21. Tuma F, Nassar AK. Applying Bloom's taxonomy in clinical surgery: practical examples. Ann Med Surg (Lond) 2021 Sep;69:102656. [doi: 10.1016/j.amsu.2021.102656] [Medline: 34429945]

Abbreviations

AI: artificial intelligence
LLM: large language model
MEDCO: Medical Education Copilots

Edited by A Stone, T Leung; submitted 28.04.25; peer-reviewed by C Li, M Chatzimina, P Sharma; revised version received 17.09.25; accepted 18.09.25; published 02.10.25.

Please cite as:
Chow M, Ng O
Beyond Chatbots: Moving Toward Multistep Modular AI Agents in Medical Education
JMIR Med Educ 2025;11:e76661
URL: https://mededu.jmir.org/2025/1/e76661
doi:10.2196/76661

From Hype to Implementation: Embedding GPT-4o in Medical Education

Sumaia Sabouni[1], PhD; Mohammad-Adel Moufti[2,3], PhD; Mohamed Hassan Taha[3,4], PhD

[1]Department of Data Science and Artificial Intelligence, Faculty of Information Technology, Al-Ahliyya Amman University, Amman, Jordan

[2]College of Dental Medicine, University of Sharjah, Sharjah, United Arab Emirates

[3]Medical Education Center, University of Sharjah, Sharjah, United Arab Emirates

[4]College of Medicine, University of Sharjah, Sharjah, United Arab Emirates

Corresponding Author:
Mohammad-Adel Moufti, PhD
College of Dental Medicine, University of Sharjah, Sharjah, United Arab Emirates

Abstract

The release of GPT-4 Omni (GPT-4o), an advanced multimodal generative artificial intelligence (AI) model, generated substantial enthusiasm in the field of higher education. However, one year later, medical education continues to face significant challenges, demonstrating the need to move from initial experimentation with the integration of multimodal AIs in medical education toward meaningful integration. In this Viewpoint, we argue that GPT-4o's true value lies not in novelty, but in its potential to enhance training in communication skills, clinical reasoning, and procedural skills by offering real-time simulations and adaptive learning experiences using text, audio, and visual inputs in a safe, immersive, and cost-effective environment. We explore how this innovation has made it possible to address key medical educational challenges by simulating realistic patient interactions, offering personalized feedback, and reducing educator workloads and costs, where traditional teaching methods struggle to replicate the complexity and dynamism of real-world clinical scenarios. However, we also address the critical challenges of this approach, including data accuracy, bias, and ethical decision-making. Rather than seeing GPT-4o as a replacement, we propose its use as a strategic supplement, scaffolded into curriculum frameworks and evaluated through ongoing research. As the focus shifts from AI novelty to sustainable implementation, we call on educators, policymakers, and curriculum designers to establish governance mechanisms, pilot evaluation strategies, and develop faculty training. The future of AI in medical education depends not on the next breakthrough, but on how we integrate today's tools with intention and rigor.

(JMIR Med Educ 2025;11:e79309) doi:10.2196/79309

KEYWORDS

multimodal generative pretraining transformers; GPT-4o; artificial intelligence; interactive learning; medical education

Introduction

The impact of artificial intelligence (AI) within health care continues to expand, ranging from diagnostic imaging to medical training [1]. Its integration into medical education offers promising solutions to numerous challenges, particularly the need for substantial resources to effectively train students in competencies like communication, clinical reasoning, and procedural skills. AI tools offer scalable, low-resource alternatives that can help bridge these gaps.

In 2024, OpenAI released GPT-4 Omni (GPT-4o), a multimodal model combining text, image, and audio processing. Unlike text-only systems, GPT-4o supports more dynamic interactions, such as engaging in real-time, human-like conversations, interpreting visual content, and performing realistic voice-based simulations. GPT-4o received widespread attention within medical education due to its multimodal capabilities enabling realistic simulations that extend to procedural skill assessment and interactive case analyses. Early demonstrations highlight GPT-4o's ability to simulate patient interactions [2], visually identify medications [3], and undertake medical examinations with higher accuracy than GPT-4 and GPT-3.5 [4,5].

However, one year after its release, there is clearly a slow pace of adoption and implementation remains limited [6]. Although some institutions have experimented with GPT-4o integration [7] and others analyzed medical students' attitudes toward this integration [8], sustained adoption within formal curricula has lagged.

This Viewpoint highlights how GPT-4o shows promise as a relevant tool to enhance medical education and address many of its challenges, despite the initial hype fading. We examine its strengths in simulating patient interactions, increasing equity due to low-resource deployment, and reducing educator workload. We also address concerns including accuracy, bias, ethical reasoning, and digital access. Our aim is not to present GPT-4o as a cure-all, but to advocate for a structured, research-informed pathway for integrating such tools meaningfully into medical training. Although this discussion

focuses on general medical education, specific considerations for dental medicine and other health disciplines warrant future exploration.

Educational Challenges Addressed by GPT-4o

Here are some of the educational challenges addressed by GPT-4o:

1. Communication skills: Developing interpersonal communication is vital for effective clinical practice. Traditional role-playing methods, reliant on patients or actors, are resource-intensive and challenging to standardize and scale. GPT-4o–generated virtual patients simulate diverse and adaptive interactions, fostering essential skills such as delivering bad news and providing culturally sensitive care. By creating realistic, immersive environments, GPT-4o aligns with situated learning theory, enabling learners to acquire practical skills in context.
2. Clinical reasoning: Diagnostic and intervention decision-making requires iterative exposure to varied scenarios. Static case-based learning often fails to engage learners dynamically. GPT-4o generates interactive, context-rich cases tailored to learner actions, promoting experiential learning [9] and hypotheticodeductive reasoning [10]. Feedback is immediate and adaptive, allowing users to refine diagnostic strategies in real time. These features advance cognitive load management, enhancing learning efficiency.
3. Essay-based assessments: Essay writing fosters critical thinking and provides a platform for expressing individual ideas. Despite these benefits, educators often avoid using essay writing due to the time required for grading. GPT-4o automates the assessment of essay responses, providing detailed feedback on reasoning, structure, and language [11]. This automation reduces educator workloads, encourages higher-order assessments, and supports personalized student development. However, overreliance on automated grading may overlook nuanced clinical reasoning or professional reflection, underscoring the need for educator oversight [12] Additionally, some research suggests that students may perceive AI-generated feedback as impersonal, leading to reduced engagement or trust [11].
4. Procedural skills: Skill acquisition in procedures relies on hands-on practice, detailed feedback, and in-person, one-on-one observation and assessment. GPT-4o's ability to produce realistic simulations and analyze video inputs offers scalable solutions for performance evaluation. Future integration with augmented reality and virtual reality technologies could facilitate immersive training experiences, replicating clinical environments for enhanced procedural competence. However, the concept of fidelity, including how realistic and believable simulations appear to learners, is critical. Low-fidelity simulations risk invoking the "uncanny valley" effect, where interactions feel unnatural and may reduce engagement [13].

Potential Benefits

For Learners

GPT-4o supports ongoing skill development by offering, personalized instant performance evaluation and feedback. Due to its realistic clinical simulations, students are exposed to a wide range of medical cases and can practice clinical reasoning in a realistic and practical way. Learners can practice breaking bad news to virtual patients, interpret multimodal data such as radiological images or heart sounds, and receive adaptive feedback tailored to their level. As a result, students can develop their confidence in a secure, flexible, and nonjudgmental learning environment.

For Educators

GPT-4o helps ease the workload by performing time-consuming tasks such as writing patient scenarios, setting assessments, marking essays, and providing personalized feedback on student performance. This gives educators more time to focus on tasks such as student support and course development. Additionally, seeing as GPT-4o is already pretrained, there is no need for heavy investment in custom AI tools, making it easier and quicker for institutions to adopt. Nevertheless, institutions should address faculty resistance by providing hands-on workshops, co-designing AI use cases with educators, and sharing evidence from pilot studies to build confidence in its implementation.

Global Accessibility

Due to GPT-4o being cloud-based, it does not require costly infrastructure, making it accessible to both learners and educators in low-resource settings. This helps democratize medical education and training and results in a more even distribution of health care expertise globally [14].

Challenges and Considerations

Although AI holds promise in medical education, there are also some challenges and considerations:

1. Accuracy: AI programs like GPT-4o may generate errors or "hallucinations" [15]. To overcome this, the output of such tools requires continuous supervision and validation to ensure their reliability.
2. Data privacy: To uphold confidentiality and ethical standards, learner performance data must be protected. This may be achieved through the adoption of strict data security measures by educational establishments. However, implementing strict data security is complex, requiring closed systems, robust encryption, and governance frameworks. There are trade-offs between using real learner data, which raises confidentiality concerns, and synthetic data, which may lack realism.
3. Bias: Training data biases associated with gender, ethnicity, or socioeconomic status may influence AI outputs. As a result, initiatives to reduce these biases are essential, especially since AI tools are being used more and more to deliver and filter health care information, where transparency and trust are essential [16]. Emerging frameworks like Fairlearn and AI Fairness 360 offer tools

to audit and mitigate bias in AI systems, supporting fairer educational outcomes [17].

4. Ethics and professionalism: Although GPT-4o performs very well in simulations, professionalism or complex ethical reasoning may be more challenging for it [12]. These limitations highlight the importance of human oversight and complementary teaching methods.
5. Access: Despite its general affordability, access to GPT-4o still depends on a stable internet connection, potentially limiting its reach in underserved regions with poor connectivity [18]. Addressing this challenge requires exploring offline or low-bandwidth alternatives to ensure a broader reach.

Next Steps

Emerging empirical studies are beginning to validate GPT-4o's role in medical education. For example, a pilot by Öncü et al [2] used GPT-4o as a virtual standardized patient to simulate complex communication and crisis scenarios, demonstrating high learner engagement and feasibility in training clinical reasoning. Bicknell et al [19] evaluated GPT-4o across 4 national licensing exams, where it achieved a 90.4% accuracy rate, surpassing GPT-4 and average medical student performance. Similarly, Zhong et al [20] benchmarked GPT-4o for rare disease diagnosis using multilingual clinical data, with the model achieving the highest diagnostic accuracy among tested large language models. These studies reinforce the potential of GPT-4o to enhance assessment, simulation, and clinical education, while also underscoring the need for rigorous longitudinal evaluation frameworks.

Future research should assess how effectively GPT-4o enhances learner competencies like critical thinking, decision-making, and procedural skills. Evaluation frameworks should include multimodal fidelity measures, learner satisfaction, performance metrics, and longitudinal tracking of outcomes such as clinical preparedness. These metrics will help assess the educational impact of GPT-4o in realistic, complex learning environments. Additionally, GPT-4o could support interprofessional education by simulating collaborative scenarios between medical, nursing, and dental students. Such integration may enhance communication, teamwork, and understanding across health disciplines, particularly in complex patient care settings.

To further expand its uses, GPT's multimodal capabilities could be extended to interpret video and audio data, such as echo scans and heart sounds. To guarantee that implementation is ethical, scalable, and in line with curriculum objectives, cooperation between AI developers, educators, and legislators is equally vital. Addressing challenges such as accuracy, bias, and accessibility challenges will ensure equitable benefits for all learners. Even though GPT-4o is a significant step forward, its use in medical education should be guided by research and specific learning aims. To offer useful assistance for adoption, we propose a set of initial steps for institutions to implement GPT-4o (Textbox 1).

Textbox 1. Recommendations for institutions considering GPT-4o integration.

1. Begin with a small-scale pilot project focusing on a single competency (eg, communication).
2. Provide prompt engineering training for faculty members.
3. Develop ethical guidelines for student use.
4. Incorporate artificial intelligence literacy into medical curricula.
5. Collaborate with technology experts to conduct longitudinal assessments of performance.
6. Design simulated scenarios that support interprofessional education (eg, medical, dental, and nursing collaboration).
7. Address faculty concerns through workshops, co-designed implementation strategies, and sharing evidence from early adopters.

Conclusion

The integration of GPT-4o into medical education represents a promising shift. It not only tackles long-standing challenges by offering scalable, immersive training solutions, but its multimodal nature also results in it offering advantages over prior models in certain contexts. Consequently, both educators and learners globally stand to benefit as access to quality education becomes more equitable. However, for this integration to be effective, medical institutions must move past experimentation through pilot studies and toward embedding into formal curricula. Finally, GPT-4o must not be viewed as a quick fix but rather as a tool that demands thoughtful design, evaluation, and governance, and its integration should be guided by rigorous research, ethical considerations, and faculty collaboration.

Acknowledgments

This study was partially funded by a research grant from the University of Sharjah (grant number 2301100278).

The authors alone are responsible for the content and writing of the article. Generative artificial intelligence tools, specifically OpenAI's ChatGPT-4, were used to proofread and improve the clarity of the manuscript. All original ideas, analyses, and conclusions were solely developed by the authors.

Conflicts of Interest

None declared.

References

1. Abu Owida H, R. Hassan M, Ali AM, et al. The performance of artificial intelligence in prostate magnetic resonance imaging screening. IJECE 2024;14(2):2234. [doi: 10.11591/ijece.v14i2.pp2234-2241]
2. Öncü S, Torun F, Ülkü HH. AI-powered standardised patients: evaluating ChatGPT-4o's impact on clinical case management in intern physicians. BMC Med Educ 2025 Feb 20;25(1):278. [doi: 10.1186/s12909-025-06877-6] [Medline: 39979969]
3. Bazzari AH, Bazzari FH. Assessing the ability of GPT-4o to visually recognize medications and provide patient education. Sci Rep 2024 Nov 5;14(1):26749. [doi: 10.1038/s41598-024-78577-y] [Medline: 39501020]
4. Liu CL, Ho CT, Wu TC. Custom GPTs enhancing performance and evidence compared with GPT-3.5, GPT-4, and GPT-4o? A study on the emergency medicine specialist examination. Healthcare (Basel) 2024 Aug 30;12(17):1726. [doi: 10.3390/healthcare12171726] [Medline: 39273750]
5. Wu YC, Wu YC, Chang YC, Yu CY, Wu CL, Sung WW. Advancing medical AI: GPT-4 and GPT-4o surpass GPT-3.5 in Taiwanese medical licensing exams. PLoS ONE 2025;20(6):e0324841. [doi: 10.1371/journal.pone.0324841] [Medline: 40465748]
6. Perkins M, Pregowska A. The role of artificial intelligence in higher medical education and the ethical challenges of its implementation. AIH 2024 Oct 21;2(1):1. [doi: 10.36922/aih.3276]
7. Thomae AV, Witt CM, Barth J. Integration of ChatGPT into a course for medical students: explorative study on teaching scenarios, students' perception, and applications. JMIR Med Educ 2024 Aug 22;10:e50545. [doi: 10.2196/50545] [Medline: 39177012]
8. Duan S, Liu C, Rong T, Zhao Y, Liu B. Integrating AI in medical education: a comprehensive study of medical students' attitudes, concerns, and behavioral intentions. BMC Med Educ 2025 Apr 23;25(1):599. [doi: 10.1186/s12909-025-07177-9] [Medline: 40269824]
9. Salinas-Navarro DE, Vilalta-Perdomo E, Michel-Villarreal R, Montesinos L. Designing experiential learning activities with generative artificial intelligence tools for authentic assessment. ITSE 2024 Oct 30;21(4):708-734. [doi: 10.1108/ITSE-12-2023-0236]
10. Khumrin P. The use of clinical decision support systems for the development of medical students' diagnostic reasoning skills [Dissertation]. URL: https://minerva-access.unimelb.edu.au/handle/11343/227067 [accessed 2025-10-07]
11. Banihashem SK, Kerman NT, Noroozi O, Moon J, Drachsler H. Feedback sources in essay writing: peer-generated or AI-generated feedback? Int J Educ Technol High Educ 2024;21(1). [doi: 10.1186/s41239-024-00455-4]
12. Masters K. Artificial intelligence in medical education. Med Teach 2019 Sep;41(9):976-980. [doi: 10.1080/0142159X.2019.1595557] [Medline: 31007106]
13. Johnston JL, Kearney GP, Gormley GJ, Reid H. Into the uncanny valley: simulation versus simulacrum? Med Educ 2020 Oct;54(10):903-907. [doi: 10.1111/medu.14184] [Medline: 32314435]
14. Hamilton A. Artificial intelligence and healthcare simulation: the shifting landscape of medical education. Cureus 2024 May;16(5):e59747. [doi: 10.7759/cureus.59747] [Medline: 38840993]
15. Masters K. Medical Teacher's first ChatGPT's referencing hallucinations: lessons for editors, reviewers, and teachers. Med Teach 2023 Jul;45(7):673-675. [doi: 10.1080/0142159X.2023.2208731] [Medline: 37183932]
16. Shambour QY, Abualhaj MM, Abu-Shareha A, Hussein AH, Kharma QM. Mitigating healthcare information overload: a trust-aware multi-criteria collaborative filtering model. J Appl Data Sci 2024;5(3):1134-1146. [doi: 10.47738/jads.v5i3.297]
17. Bellamy RKE, Dey K, Hind M, et al. AI Fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. IBM J Res & Dev 2019 Jul 1;63(4/5):4. [doi: 10.1147/JRD.2019.2942287]
18. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023 Jun 1;9:e48291. [doi: 10.2196/48291] [Medline: 37261894]
19. Bicknell BT, Butler D, Whalen S, et al. ChatGPT-4 Omni performance in USMLE disciplines and clinical skills: comparative analysis. JMIR Med Educ 2024 Nov 6;10:e63430. [doi: 10.2196/63430] [Medline: 39504445]
20. Zhong W, Liu Y, Liu Y, et al. Performance of ChatGPT-4o and four open-source large language models in generating diagnoses based on China's rare disease catalog: comparative study. J Med Internet Res 2025 Jun 18;27:e69929. [doi: 10.2196/69929] [Medline: 40532199]

Abbreviations

AI: artificial intelligence
GPT-4o: GPT-4 Omni

Edited by J Gentges; submitted 18.06.25; peer-reviewed by C Inglis, M El-Kishawi; revised version received 06.08.25; accepted 10.09.25; published 15.10.25.

Please cite as:
Sabouni S, Moufti MA, Taha MH
From Hype to Implementation: Embedding GPT-4o in Medical Education
JMIR Med Educ 2025;11:e79309
URL: https://mededu.jmir.org/2025/1/e79309
doi:10.2196/79309

Viewpoint

Beyond Lectures: Reimagining Psychiatric Didactics for the Age of AI

Laurent Elkrief[1,2], MSc, MD; Alexandre Hudon[1,3,4,5], BEng, MSc, MD, PhD; Giovanni Briganti[6,7], MD, PhD; Paul Lespérance[1,2], MSc, MD

[1]Département de Psychiatrie et d'Addictologie, Faculté de Médecine, Université de Montréal, Montréal, QC, Canada

[2]Centre Hospitalier de l'Université de Montréal, Montreal, QC, Canada

[3]Department of Psychiatry, Institut universitaire en santé mentale de Montréal, Montréal, QC, Canada

[4]Department of Psychiatry, Institut Philippe Pinel de Montréal, Montreal, QC, Canada

[5]Groupe Interdisciplinaire de Recherche sur la Cognition et le Raisonnement Professionnel, Université de Montréal, Montréal, QC, Canada

[6]Service de Médecine computationnelle et neuropsychiatrie, Faculté de Médecine, Pharmacie, et Sciences Biomédicales, University of Mons, Mons, Belgium

[7]Département des Sciences Cliniques, Faculté de Médecine, University of Liège, Liège, Belgium

Corresponding Author:
Laurent Elkrief, MSc, MD
Département de Psychiatrie et d'Addictologie
Faculté de Médecine
Université de Montréal
2900, boul. Édouard-Montpetit
Montréal, QC, H3T 1J4
Canada
Phone: 1 514 343 5803
Email: laurent.elkrief@umontreal.ca

Abstract

The increasing use of generative large language models (LLMs) necessitates a fundamental reevaluation of traditional didactic lectures in medical education, particularly within psychiatry. The specialty's inherent diagnostic ambiguity, biopsychosocial complexity, and reliance on nuanced interpersonal skills demand an educational model that transcends mere information transfer, focusing instead on cultivating sophisticated clinical reasoning. This viewpoint argues for a shift from passive knowledge transmission to active, facilitated development of higher-order thinking, aligning with the Bloom taxonomy. We describe four core propositions: (1) shifting foundational knowledge acquisition to faculty-curated asynchronous artificial intelligence (AI)–assisted micromodules; (2) transforming synchronous time into "Ambiguity Seminars" for discussing nuanced cases, biopsychosocial formulation, and ethical dilemmas, leveraging faculty expertise in guiding reasoning; (3) integrating live LLM critical interaction drills to develop prompt engineering skills and critical appraisal of AI outputs; and (4) realigning assessment methods (eg, objective structured clinical examinations [OSCEs], reflective writing) to evaluate clinical reasoning and integrative skills rather than rote recall. Successful implementation requires comprehensive faculty development, explicit institutional investment, and a phased approach that addresses scalability across varying resource settings. This reimagined approach aims to cultivate clinical wisdom, equipping psychiatric trainees with adaptive reasoning frameworks essential for excellence in an AI-mediated future.

(JMIR Med Educ 2025;11:e78110) doi:10.2196/78110

KEYWORDS

large language models; medical education; didactic lecture; artificial intelligence; educational technology

Introduction

The advent of powerful, publicly accessible large language models (LLMs) like ChatGPT marks an inflection point for medical education. Specifically, these tools are driving a shift from information scarcity to abundance, which directly challenges the traditional role of the didactic lecture as the main medium of information transfer. Consequently, the widespread adoption of these tools necessitates a fundamental rethinking of this lecture model. The necessity to evolve beyond traditional

didactics is amplified in psychiatry, a specialty with inherent diagnostic ambiguity, profound biopsychosocial complexity, and a fundamental reliance on nuanced interpersonal competencies and the interpretation of subjective human experience. These defining features demand an educational model that transcends mere information, one that actively cultivates the sophisticated clinical reasoning and integrative skills essential for practice. Such a model aligns with established pedagogical frameworks like the Bloom taxonomy, aiming to engage trainees across a spectrum of cognitive processes, from foundational understanding to higher-order thinking and complex problem-solving. Consequently, we argue that psychiatric didactic time must pivot from passive knowledge transmission toward an active, facilitated development of clinical reasoning, with faculty evolving from information repositories ("sage on the stage") [1] into catalysts for critical thinking and contextualization. This proposed evolution aligns with the core tenets of competency-based medical education, which prioritizes the demonstration of integrated professional capabilities over time-based training and simple knowledge acquisition [2]. It is a direct application of the call to cultivate "adaptive expertise," the ability to flexibly and innovatively apply deep conceptual knowledge to novel problems, which stands in contrast to the routine expertise fostered by traditional didacticism [3]. This viewpoint outlines four core propositions for this necessary transformation.

Psychiatry's inherent complexity imposes a high intrinsic cognitive load, yet traditional lectures often increase extraneous load through passive delivery, hindering deep processing and schema learning [4]. Seeking greater efficiency, trainees increasingly forgo these sessions. This mismatch contributes significantly to declining attendance as trainees prioritize the flexibility of alternative resources [5]. Beyond attendance effects, meta-analyses indicate that active learning consistently improves achievement and reduces failure compared with traditional lectures [6], with flipped designs in medical education showing measurable gains when preclass work is structured and accountability is built in [7].

The Shifting Landscape: AI's Impact on Medical Education and Psychiatry's Unique Needs

The widespread availability of LLMs for educational purposes [8-10] drastically accelerates this trend by collapsing traditional knowledge asymmetries. However, alongside potential benefits, the unguided use of these tools necessitates critical artificial intelligence (AI) literacy due to inherent risks of these systems, mainly around inaccuracies and "hallucinations" [10-15] as well as the potential propagation of embedded societal biases [16-18]. Additionally, the risk of automation bias influencing clinical judgment [13], coupled with outputs often lacking the critical nuance essential for psychiatric practice [10,15,16], further underscores current AI limitations. Effectively navigating this landscape demands not only critical evaluation skills but also proficiency in prompt engineering [14,19]. Current didactic structures, largely reliant on outdated lecture formats, are ill-equipped for this complex new reality, failing to prepare trainees for an information environment increasingly mediated by AI. This challenge is not unique to the AI era; for decades, educational theorists have argued for the necessity of active learning methodologies to move beyond the passivity of the lecture format and better cultivate the complex reasoning skills required for professional practice [1]. This aligns with meta-analytic evidence from health profession education and the broader higher education literature showing flipped/active approaches outperform lecture-based formats [6,7].

Beyond general AI challenges, psychiatric education faces unique demands. Diagnosis relies on subjective interpretation and negotiated constructs, not definitive tests, with evolving models adding complexity. Effective practice requires sophisticated biopsychosocial formulation, integrating diverse data (biology, psychology, narrative, and social context), which is a reasoning skill poorly served by simple fact delivery. Current LLMs struggle with the nuance, empathy, subjectivity, and deep biopsychosocial integration vital for psychiatry [15,16,20]. Given that navigating uncertainty and ambiguity are core competencies, psychiatric education must prioritize cultivating robust clinical reasoning, metacognition, and critical thinking to develop "clinical wisdom" over mere recall [1,4].

Developing such clinical wisdom demands a pedagogical evolution. Because LLMs reduce the challenge of accessing factual knowledge, faculties' comparative advantage shifts to fostering higher-order cognitive skills such as critical thinking, contextual reasoning, and the synthesis of information, evolving their role from primary knowledge sources to catalysts for these deeper learning processes. The following propositions operationalize this shift.

Core Propositions for Reimagining Psychiatric Didactics

First, the acquisition of foundational knowledge, corresponding to initial cognitive levels in the Bloom taxonomy such as "Remembering and Understanding," shifts to faculty-curated AI micromodules. These short asynchronous resources, perhaps AI-drafted [15,19,20] but rigorously vetted for accuracy/nuance [10,14-16], free synchronous time and reduce extraneous cognitive load [21,22]. This vetting process would involve cross-referencing AI-generated content against established clinical guidelines and seminal texts, scrutinizing for embedded biases [16-18], verifying the authenticity of citations [11], and ensuring the material aligns with local practice standards and the appropriate learner level.

Second, synchronous time becomes an "Ambiguity Seminar," where psychiatric complexity is addressed directly. Faculty use nuanced vignettes to teach how to reason, framing the clinical problem and uncertainties first, rather than delivering more facts [1]. To achieve this, faculty would use techniques such as Socratic questioning to probe assumptions and guide hypothesis generation ("What evidence supports that diagnosis over others?"). They would also focus on metacognitive modeling, verbalizing their own reasoning process when faced with uncertainty ("Here is why I am prioritizing this intervention, despite these conflicting data points...") to demonstrate how

experts navigate ambiguity, thereby shifting the focus from finding a single correct answer to developing a robust and defensible reasoning process. While AI may help draft initial cases [18], instructors refine them toward situations in which diagnoses straddle categories and pharmacological guideline algorithms do not neatly apply. Learners then generate competing hypotheses with confirming and disconfirming data, craft a succinct biopsychosocial formulation, and propose a first-line plan that goes beyond the textbook, stating trade-offs and safety contingencies in the patient's context. LLMs can be used as a sounding board, but outputs are treated as claims to be tested; their breakdowns become teachable moments about limits and bias. The aim is disciplined, creative problem-solving, yielding brief original formulations and plans that learners can defend aloud.

Third, seminars integrate live LLM critical interaction drills. Trainees query an LLM with case questions, then critique the output: checking accuracy, bias, relevance, and citations. This requires prompt engineering instruction [14,19,20]. Engaging in such exercises helps trainees develop sound habits for evaluating information, improves their AI literacy, and equips them to counter automation bias [13]. Prompting students to critique LLM responses encourages them to use the AI as a sounding board to refine their own clinical judgment; this process also lessens the risk of AI-driven inaccuracies and develops crucial practical abilities.

Finally, to ensure learning objectives are met, assessment must be realigned with reasoning skills, moving beyond recall. While the Bloom taxonomy can delineate how AI might assist with foundational knowledge tasks, evaluation must focus on higher-order thinking that is inherently resistant to artificial augmentation. Specifically, objective structured clinical examinations (OSCEs) should be designed to assess not just the analysis of complex information but also the trainee's real-time interpersonal skills and their ability to adapt to unexpected information from a standardized patient. To ensure integrity, these OSCEs must be conducted in proctored environments where the use of external AI tools is prohibited. Similarly, reflective writing assignments can be made more robust by requiring trainees to integrate highly specific, personal patient interactions—details an AI could not fabricate—or by using in-class timed "reflection stems" that demand immediate synthesis of a shared experience. A mandatory oral defense of these reflections then becomes a nonnegotiable component to validate the authenticity of the reasoning and personal insights presented. These methods, by directly targeting the upper echelons of the Bloom taxonomy and evaluating skills requiring embodied clinical presence and personal experience, offer a more authentic assessment of the competencies that current AI systems struggle to replicate.

A Framework for Implementation: Addressing Practical Challenges

We proposed a pragmatic blueprint that flips factual acquisition to curated micromodules, reclaims synchronous time for faculty-facilitated ambiguity seminars, and integrates AI-critical drills, with assessments aligned to higher-order clinical reasoning. This section aims to translate that design into an implementable institutional plan while acknowledging costs and constraints.

Translating these propositions from theory into practice requires a pragmatic strategy that directly addresses the significant challenges of institutional inertia, resource allocation, and faculty development. A successful rollout is not merely a technical task but a complex exercise in change management. The most significant barrier is often faculty resistance, which may stem from the substantial workload of curriculum redesign and a perceived evaluation of traditional lecturing expertise. Consequently, this change must be framed as an elevation of the faculty role, shifting members from information transmitters to expert guides who model and cultivate complex clinical reasoning. We acknowledge that this viewpoint presents a theoretical framework and that its efficacy has yet to be established through empirical research; its primary goal is to provide a road map for such investigation.

To manage this transition, a dedicated faculty development program is essential, requiring protected (and renumerated) time for hands-on training in advanced Socratic facilitation for the ambiguity seminars, critical AI literacy for appraising model outputs, and the skills for curriculum cocreation. Furthermore, this educational model is not resource-neutral and requires explicit institutional investment. Success is contingent on access to a stable and user-friendly learning management system; privacy-compliant AI platforms; and most critically, formally protected faculty time. This work cannot be an unfunded mandate added to existing clinical and academic responsibilities. To centralize and sustain the effort, programs might consider creating a dedicated role, such as a clinical AI education lead. Recognizing that institutional capacities vary widely, this framework is designed to be scalable. High-resource programs might implement the full model, while lower-resource settings can adopt a "low-fidelity" version using freely available language models and multi-institutional consortia for open-access materials. Crucially, this scalability must extend to individual trainees to ensure equitable participation. Programs should provide accessible, screen reader–friendly materials and use privacy-compliant AI platforms, offering device support or non-LLM analytic pathways where live access is infeasible. By embedding these accessibility measures, the core principles of flipping the classroom and focusing synchronous time on facilitated reasoning can be maintained inclusively across all settings.

Central to this framework is a commitment to educational equity. The integration of AI tools risks exacerbating existing disparities related to socioeconomic status, disability, or access to technology. Therefore, programs must actively ensure equitable implementation. This includes providing institutional access, where possible, to privacy-compliant AI platforms to avoid financial barriers for trainees; offering device support; and ensuring all digital materials are fully accessible and screen reader–friendly. Furthermore, the development of "non-LLM analytic pathways" is crucial; these are alternative assignments that achieve the same core learning objectives of critical reasoning and evidence appraisal but do not require live AI

interaction, ensuring that technological barriers do not impede a trainee's educational progress.

Finally, a phased 3-year timeline can make this significant reform manageable. Year one would focus on a pilot within a single teaching block to establish the proof of concept and gather feasibility data. Year two would involve expansion and refinement, using pilot data to roll the model out to other blocks. Year three would target full integration and sustainability, making the model standard practice and shifting research toward longitudinal multisite evaluation to assess broader generalizability.

Beyond implementation, systematic and rigorous evaluation is essential. Pilot studies should assess primary outcomes (OSCEs for formulation/reasoning/appraisal, reflective writing) and secondary measures, including engagement versus historical data [5], ambiguity tolerance, and satisfaction [1]. Process evaluation and qualitative focus groups should explore reasoning, AI trust, and cognitive load [4]. Future research needs longitudinal tracking, cost-utility analysis, AI comparisons, and scalability assessments, prioritizing methodological rigor [9,14] to address gaps in psychiatric AI education research [15,16].

Conclusion

This reimagined approach aims to redefine psychiatric education for a new era defined by the widespread availability of knowledge through LLMs. Faculty should pivot from primarily dispensing facts toward cultivating clinical wisdom, defined as sound judgment under uncertainty. Accordingly, this viewpoint proposes retiring lectures focused on the transfer of facts in favor of curated micromodules, thereby reclaiming synchronous time for facilitated reasoning seminars that incorporate critical AI interaction. We hope programs pilot this model (or similar ones), focusing didactic time on core competencies like biopsychosocial formulation and ethical deliberation. Equipping trainees to interrogate machines, not just query them, requires moving beyond outdated methods. In an AI-mediated future, the cultivation of adaptive reasoning frameworks will be fundamental to clinical excellence.

Acknowledgments

AH declares funding from Fonds opérationnels d'IVADO et Fondation de l'IUSMM. The other authors have no funding declarations. This paper was written with the assistance of generative artificial intelligence (AI; gemini-2.5-pro-preview-05-06; Multimedia Appendix 1). Specifically, generative AI was used in the brainstorming portions of the project. It was also used for editing.

Conflicts of Interest

LE is a founder at OneCare Biotechnologies, a mental health biotechnology start-up. His work at OneCare is not in any way related to the present work.

Multimedia Appendix 1
Description of the use of generative artificial intelligence by the authors.
[DOCX File , 67 KB - mededu_v11i1e78110_app1.docx]

References

1. Sandrone S, Berthaud JV, Carlson C, Cios J, Dixit N, Farheen A, et al. Active learning in psychiatry education: current practices and future perspectives. Front Psychiatry 2020;11:211 [FREE Full text] [doi: 10.3389/fpsyt.2020.00211] [Medline: 32390876]
2. Frank JR, Snell LS, Cate OT, Holmboe ES, Carraccio C, Swing SR, et al. Competency-based medical education: theory to practice. Med Teach 2010;32(8):638-645. [doi: 10.3109/0142159X.2010.501190] [Medline: 20662574]
3. Hatano G, Inagaki K. Two courses of expertise. Hokkaido University Collection of Scholarly and Academic Papers. 1984 Mar. URL: https://eprints.lib.hokudai.ac.jp/dspace/bitstream/2115/25206/1/6_P27-36.pdf [accessed 2025-10-27]
4. Jordan J, Wagner J, Manthey DE, Wolff M, Santen S, Cico SJ. Optimizing lectures from a cognitive load perspective. AEM Educ Train 2020 Jul;4(3):306-312 [FREE Full text] [doi: 10.1002/aet2.10389] [Medline: 32704604]
5. Gardner G, Feldman M, Santen SA, Mui P, Biskobing D. Determinants and outcomes of in-person lecture attendance in medical school. Med Sci Educ 2022 Aug;32(4):883-890 [FREE Full text] [doi: 10.1007/s40670-022-01581-2] [Medline: 35821745]
6. Freeman S, Eddy SL, McDonough M, Smith MK, Okoroafor N, Jordt H, et al. Active learning increases student performance in science, engineering, and mathematics. Proc Natl Acad Sci U S A 2014 Jun 10;111(23):8410-8415 [FREE Full text] [doi: 10.1073/pnas.1319030111] [Medline: 24821756]
7. Chen F, Lui AM, Martinelli SM. A systematic review of the effectiveness of flipped classrooms in medical education. Med Educ 2017 Jun;51(6):585-597. [doi: 10.1111/medu.13272] [Medline: 28488303]
8. Aster A, Laupichler MC, Rockwell-Kollmann T, Masala G, Bala E, Raupach T. ChatGPT and other large language models in medical education - scoping literature review. Med Sci Educ 2025 Feb;35(1):555-567. [doi: 10.1007/s40670-024-02206-6] [Medline: 40144083]

9. Hallquist E, Gupta I, Montalbano M, Loukas M. Applications of artificial intelligence in medical education: a systematic review. Cureus 2025 Mar;17(3):e79878. [doi: 10.7759/cureus.79878] [Medline: 40034416]
10. Arif TB, Munaf U, Ul-Haque I. The future of medical education and research: is ChatGPT a blessing or blight in disguise? Med Educ Online 2023 Dec;28(1):2181052 [FREE Full text] [doi: 10.1080/10872981.2023.2181052] [Medline: 36809073]
11. Aljamaan F, Temsah M, Altamimi I, Al-Eyadhy A, Jamal A, Alhasan K, et al. Reference hallucination score for medical artificial intelligence chatbots: development and usability study. JMIR Med Inform 2024 Jul 31;12:e54345 [FREE Full text] [doi: 10.2196/54345] [Medline: 39083799]
12. Arun G, Perumal V, Urias FPJB, Ler YE, Tan BWT, Vallabhajosyula R, et al. ChatGPT versus a customized AI chatbot (Anatbuddy) for anatomy education: a comparative pilot study. Anat Sci Educ 2024 Oct;17(7):1396-1405. [doi: 10.1002/ase.2502] [Medline: 39169464]
13. Nguyen T. ChatGPT in medical education: a precursor for automation bias? JMIR Med Educ 2024 Jan 17;10:e50174 [FREE Full text] [doi: 10.2196/50174] [Medline: 38231545]
14. Kıyak YS. Beginner-level tips for medical educators: guidance on selection, prompt engineering, and the use of artificial intelligence chatbots. Med Sci Educ 2024 Dec;34(6):1571-1576. [doi: 10.1007/s40670-024-02146-1] [Medline: 39758489]
15. Prégent J, Chung V, El Adib I, Désilets M, Hudon A. Applications of artificial intelligence in psychiatry and psychology education: scoping review. JMIR Med Educ 2025 Jul 28;11:e75238 [FREE Full text] [doi: 10.2196/75238] [Medline: 40720804]
16. Lee QY, Chen M, Ong CW, Ho CSH. The role of generative artificial intelligence in psychiatric education- a scoping review. BMC Med Educ 2025 Mar 25;25(1):438 [FREE Full text] [doi: 10.1186/s12909-025-07026-9] [Medline: 40133891]
17. Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large language models propagate race-based medicine. NPJ Digit Med 2023 Oct 20;6(1):195. [doi: 10.1038/s41746-023-00939-z] [Medline: 37864012]
18. Smith A, Hachen S, Schleifer R, Bhugra D, Buadze A, Liebrenz M. Old dog, new tricks? Exploring the potential functionalities of ChatGPT in supporting educational methods in social psychiatry. Int J Soc Psychiatry 2023 Dec;69(8):1882-1889. [doi: 10.1177/00207640231178451] [Medline: 37392000]
19. Birks S, Gray J, Darling-Pomranz C. Using artificial intelligence to provide a 'flipped assessment' approach to medical education learning opportunities. Med Teach 2025 Aug;47(8):1377-1384 [FREE Full text] [doi: 10.1080/0142159X.2024.2434101] [Medline: 39616548]
20. Thomae AV, Witt CM, Barth J. Integration of ChatGPT into a course for medical students: explorative study on teaching scenarios, students' perception, and applications. JMIR Med Educ 2024 Aug 22;10:e50545 [FREE Full text] [doi: 10.2196/50545] [Medline: 39177012]
21. Hurtubise L, Hall E, Sheridan L, Han H. The flipped classroom in medical education: engaging students to build competency. J Med Educ Curric Dev 2015;2:10.4137/JMECD.S23895 [FREE Full text] [doi: 10.4137/JMECD.S23895] [Medline: 35187252]
22. Lopez S. Impact of cognitive load theory on the effectiveness of microlearning modules. Euro J Education Pedagogy 2024 Mar 15;5(2):29-35. [doi: 10.24018/ejedu.2024.5.2.799]

Abbreviations

AI: artificial intelligence
LLM: large language model
OSCE: objective structured clinical examination

Edited by T Gladman; submitted 02.06.25; peer-reviewed by M Mansoor, A Verma; comments to author 15.07.25; revised version received 18.09.25; accepted 22.10.25; published 31.10.25.

Please cite as:
Elkrief L, Hudon A, Briganti G, Lespérance P
Beyond Lectures: Reimagining Psychiatric Didactics for the Age of AI
JMIR Med Educ 2025;11:e78110
URL: https://mededu.jmir.org/2025/1/e78110
doi:10.2196/78110
PMID:

Fostering Multidisciplinary Collaboration in Artificial Intelligence and Machine Learning Education: Tutorial Based on the AI-READI Bootcamp

Taiki W Nishihara[1], BA; Fritz Gerald P Kalaw[1,2], MD; Adelle Engmann[3], MS; Aya Motoyoshi[4], MD, PhD; Paapa Mensah-Kane[5], BPharm, MPhil, PhD; Deepa Gupta[6], PhD; Victoria Patronilo[1], BA; Linda M Zangwill[1], PhD; Shahin Hallaj[1,2], MD; Amirhossein Panahi[7], MS; Garrison W Cottrell[8], MS, PhD; Bradley Voytek[6,7,9,10], PhD; Virginia R de Sa[6,7,9], MS, PhD; Sally L Baxter[1,2], MSc, MD

[1]Viterbi Family Department of Ophthalmology and Shiley Eye Institute, Hamilton Glaucoma Center, Division of Ophthalmology Informatics and Data Science, University of California, San Diego, 9415 Campus Point Drive, La Jolla, CA, United States

[10]Kavli Institute for Brain and Mind, University of California, San Diego, La Jolla, CA, United States

[2]Department of Medicine, Division of Biomedical Informatics, University of California, San Diego, La Jolla, CA, United States

[3]Stanford Deep Data Research Center, Department of Genetics, Stanford University, Palo Alto, CA, United States

[4]Department of Ophthalmology, University of Washington, Seattle, WA, United States

[5]School of Pharmacy, South University, Savannah, GA, United States

[6]Department of Cognitive Science, University of California, San Diego, La Jolla, CA, United States

[7]Halıcıoğlu Data Science Institute, University of California, San Diego, La Jolla, CA, United States

[8]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, United States

[9]Neurosciences Graduate Program, University of California, San Diego, La Jolla, CA, United States

Corresponding Author:
Sally L Baxter, MSc, MD
Viterbi Family Department of Ophthalmology and Shiley Eye Institute, Hamilton Glaucoma Center, Division of Ophthalmology Informatics and Data Science, University of California, San Diego, 9415 Campus Point Drive, La Jolla, CA, United States

Abstract

Background: The integration of artificial intelligence (AI) and machine learning (ML) into biomedical research requires a workforce fluent in both computational methods and clinical applications. Structured, interdisciplinary training opportunities remain limited, creating a gap between data scientists and clinicians. The National Institutes of Health's Bridge to Artificial Intelligence (Bridge2AI) initiative launched the Artificial Intelligence–Ready and Exploratory Atlas for Diabetes Insights (AI-READI) data generation project to address this gap. AI-READI is creating a multimodal, FAIR (findable, accessible, interoperable, and reusable) dataset—including ophthalmic imaging, physiologic measurements, wearable sensor data, and survey responses—from approximately 4000 participants with or at risk for type 2 diabetes. In parallel, AI-READI established a year-long mentored research program that begins with a 2-week immersive summer bootcamp to provide foundational AI/ML skills grounded in domain-relevant biomedical data.

Objective: To describe the design, iterative refinement, and outcomes of the AI-READI Bootcamp, and to share lessons for creating future multidisciplinary AI/ML training programs in biomedical research.

Methods: Held annually at the University of California San Diego, the bootcamp combines 80 hours of lectures, coding sessions, and small-group mentorship. Year 1 introduced Python programming, classical ML techniques (eg, logistic regression, convolutional neural networks), and data science methods, such as principal component analysis and clustering, using public datasets. In Year 2, the curriculum was refined based on structured participant feedback—reducing cohort size to increase individualized mentorship, integrating the AI-READI dataset (including retinal images and structured clinical variables), and adding modules on large language models and FAIR data principles. Participant characteristics and satisfaction were assessed through standardized pre- and postbootcamp surveys, and qualitative feedback was analyzed thematically by independent coders.

Results: Seventeen participants attended Year 1 and 7 attended Year 2, with an instructor-to-student ratio of approximately 1:2 in the latter. Across both years, postbootcamp evaluations indicated high satisfaction, with Year 2 participants reporting improved experiences due to smaller cohorts, earlier integration of the AI-READI dataset, and greater emphasis on applied learning. In Year 2, mean scores for instructor effectiveness, staff support, and overall enjoyment were perfect (5.00/5.00). Qualitative feedback emphasized the value of working with domain-relevant, multimodal datasets; the benefits of peer collaboration; and the applicability of skills to structured research projects during the subsequent internship year.

Conclusions: The AI-READI Bootcamp illustrates how feedback-driven, multidisciplinary training embedded within a longitudinal mentored research program can bridge technical and clinical expertise in biomedical AI. Core elements—diverse trainee cohorts, applied learning with biomedical datasets, and sustained mentorship—offer a replicable model for preparing health professionals for the evolving AI/ML landscape. Future iterations will incorporate additional prebootcamp onboarding modules, objective skill assessments, and long-term tracking of research engagement and productivity.

(JMIR Med Educ 2025;11:e83154) doi:10.2196/83154

KEYWORDS

artificial intelligence; machine learning; biomedical research; interdisciplinary training; data science; curriculum development; translational research; medical education

Introduction

Artificial intelligence (AI) has demonstrated transformative potential in health care, with deep learning algorithms now able to screen for diabetic retinopathy from fundus photographs and predict patient-specific glycemic fluctuations [1] at performance levels comparable to expert clinicians. However, despite these advances, a persistent gap remains between model developers and clinical end users.

Clinicians often lack formal training in AI and machine learning (ML) or data science. Only about 28% of published AI/ML model development studies include clinician involvement, and their contributions are frequently limited [2]. Likewise, in the United Kingdom, just 13.8% of trainee physicians reported feeling adequately prepared for the integration of AI into clinical practice [3]. Conversely, engineers and computer scientists are rarely trained in the clinical, regulatory, or ethical complexities of health care delivery. This persistent disconnect constrains interdisciplinary collaboration, limits translational impact, and risks generating AI systems that perform poorly in real-world clinical settings [4-6].

Recognizing these interdisciplinary and workforce gaps, the National Institutes of Health (NIH) launched the Bridge to Artificial Intelligence (Bridge2AI) initiative in 2022 to promote the creation of FAIR (findable, accessible, interoperable, and reusable) multimodal datasets while advancing coordinated skills and workforce development [7]. Among its flagship data generation projects (DGPs), the Artificial Intelligence–Ready and Exploratory Atlas for Diabetes Insights (AI-READI) is curating a comprehensive dataset integrating ophthalmic imaging, physiologic measurements, wearable sensor data, and survey responses from approximately 4000 individuals with or at risk for type 2 diabetes [8,9].

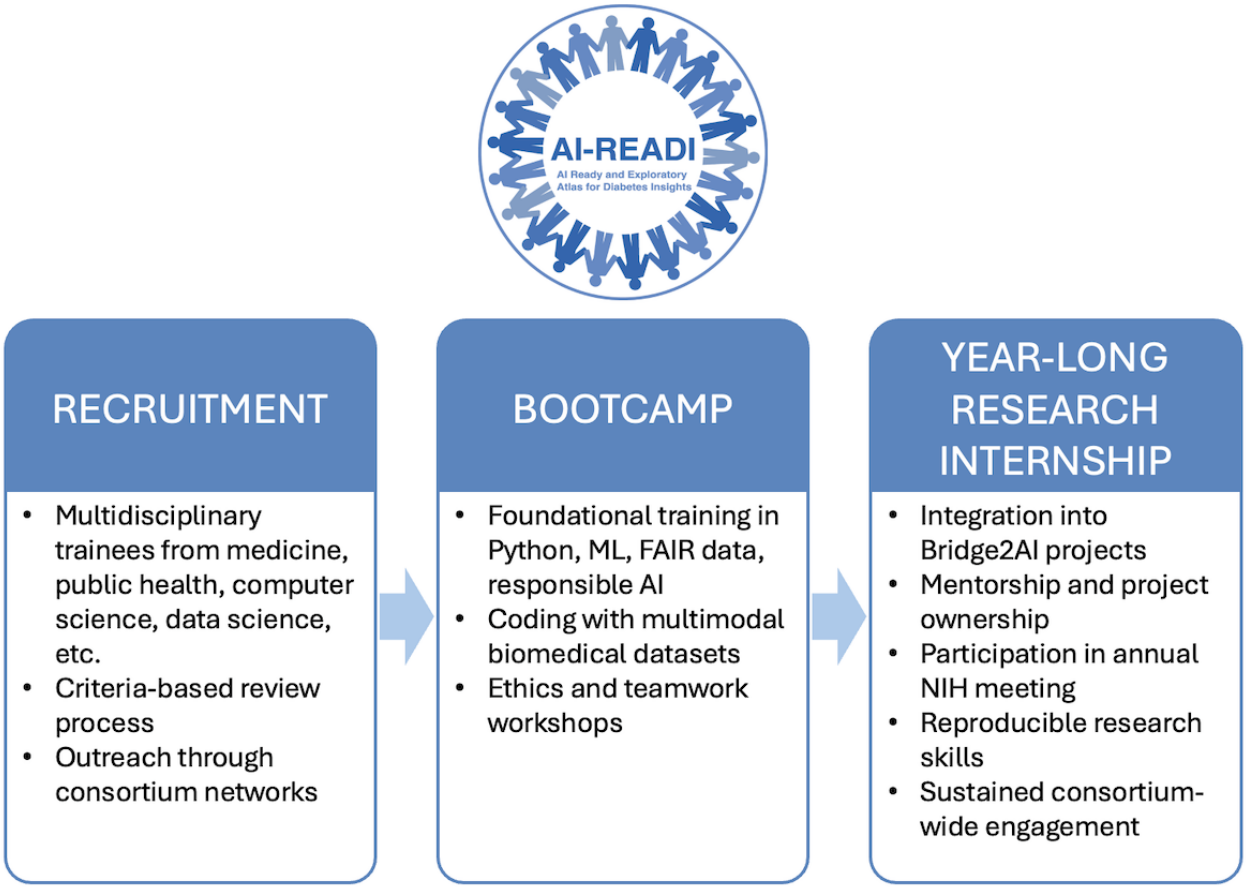
Beyond data generation, AI-READI integrates a 3-phase educational framework within the Bridge2AI ecosystem (Figure 1). This framework recruits multidisciplinary trainees, delivers an intensive 2-week AI/ML bootcamp, and transitions participants into a year-long mentored research internship. Together, these stages provide a continuum of learning that couples foundational instruction with sustained, project-based engagement, directly aligned with the consortium's goal of building a diverse and AI-ready biomedical workforce.

Although a growing number of AI training programs exist—from massive open online courses to short-term institutional electives—many rely on generic or narrowly scoped datasets, offer limited interdisciplinary interaction, or lack sustained mentorship [10-12]. In contrast, the AI-READI Bootcamp was intentionally developed for trainees from diverse disciplines to collaborate on structured, domain-relevant biomedical datasets that reflect the complexity of real-world research. Led by faculty with experience in NIH- and National Science Foundation–funded training initiatives, the curriculum integrates seminars in ML, statistics, and responsible AI with notebook-based coding laboratories anchored in the AI-READI dataset.

Each cohort (Year 1=2023; Year 2=2024) was independently developed and iteratively refined in response to structured participant feedback. This feedback-driven design aligns with broader trends in AI education emphasizing modular content, scaffolded mentorship, and interdisciplinary collaboration [11-13].

This manuscript details the design and iteration of the AI-READI Bootcamp, summarizes participant characteristics and evaluation outcomes, and distills key lessons for institutions aiming to build inclusive, practice-oriented AI training programs in health care.

Figure 1. AI-READI skills and workforce development module within the NIH Bridge2AI initiative. AI: artificial intelligence; AI-READI: Artificial Intelligence–Ready and Exploratory Atlas for Diabetes Insights; Bridge2AI: Bridge to Artificial Intelligence; FAIR: findable, accessible, interoperable, and reusable; ML: machine learning; NIH: National Institutes of Health.



Methods

Ethical Considerations

This study involved the evaluation of an educational training program and analysis of aggregated, deidentified survey and program evaluation data collected as part of routine program assessment. In accordance with institutional and US federal guidelines (45 CFR §46) [14], formal institutional review board review was not required, as the project constituted educational program evaluation with minimal risk to participants and did not involve the collection of identifiable private information. Participation in surveys and program evaluations was voluntary. Participants were informed that their responses could be used for research and dissemination purposes and that participation or nonparticipation would not affect their standing in the program. Completion of the surveys was considered to imply informed consent. All data were analyzed in a deidentified and aggregated manner. No direct identifiers were collected or retained. Data were stored on secure, access-controlled institutional systems, and only study personnel had access to the data. Privacy and confidentiality were maintained throughout the study. Participants did not receive financial compensation for survey participation beyond the educational benefits associated with participation in the AI-READI Bootcamp.

AI-READI Intern Recruitment and Bootcamp Participant Selection

Participants for the year-long AI-READI internship program were recruited from diverse academic and professional backgrounds, including computer science, engineering, medicine, public health, nursing, pharmacy, and allied health fields. Selection prioritized quantitative aptitude, scientific curiosity, and interdisciplinary interest rather than prior coding experience. Coursework in calculus, linear algebra, or statistics was preferred but not required for eligibility. Recruitment strategies included an informational brochure on the AI-READI website; dissemination through faculty web pages, journals, mailing lists, and social media; and outreach by alumni, mentors, and current trainees through word-of-mouth engagement.

Applicants completed an online application comprising educational history, research experience, a 750-word personal statement, and one faculty recommendation. Reviewers scored submissions on a 1 - 5 scale across 4 domains: academic achievement, technical skills, research experience, and strength of recommendation. Top-ranked candidates received AI-READI–funded internship positions (6 in Year 1; 5 in Year 2), while additional high-scoring applicants were invited to participate in the bootcamp as unfunded trainees (11 in Year 1; 2 in Year 2).

Bootcamp Structure and Educational Objectives

The AI-READI Bootcamp is a 2-week immersive, in-person educational program hosted annually at the University of California San Diego. It functioned as both the foundational training phase and the launchpad for the year-long mentored research internship. The curriculum emphasized collaborative, application-oriented learning tailored to participants' diverse disciplinary backgrounds and varying levels of technical preparedness.

Educational objectives were to (1) develop proficiency in core programming workflows using Python, Jupyter Notebooks, and GitHub; (2) introduce foundational principles of AI/ML, including supervised and unsupervised learning; (3) provide hands-on experience through structured coding laboratories using multimodal, domain-relevant biomedical data; (4) promote reproducible and ethical research practices; and (5) foster interdisciplinary collaboration, critical thinking, and cohort cohesion.

The curriculum intentionally targeted multiple domains of Bloom's taxonomy [15], integrating didactic instruction to build knowledge (cognitive), using applied coding to develop technical skills (psychomotor), and facilitating ethics discussions to promote responsible AI use (affective).

Participants completed approximately 80 hours of lectures, coding tutorials, and small-group mentorship sessions. Dormitory housing facilitated peer learning, collaborative debugging, and informal knowledge exchange. Instruction was led by a multidisciplinary faculty team spanning computer science, data science, medicine, public health, and ethics. Curriculum content was refined iteratively between cohorts in response to postbootcamp feedback (see Results).

Participant Characteristics and Baseline Data Collection

Before the bootcamp, all participants completed a standardized intake form capturing demographic and educational information, including age, gender, highest degree attained, primary discipline, prior experience with programming languages (eg, Python, R, SQL), and self-reported exposure to AI/ML. These responses informed real-time instructional adjustments and enabled instructors to tailor laboratory groupings, pacing, and mentorship to each cohort's skill profile.

Postbootcamp Survey and Feedback Analysis

After completing the bootcamp, participants were invited to complete an anonymous evaluation survey assessing instructional quality and overall experience.

- Quantitative items: 7 core statements rated on a 5-point Likert scale (1=strongly disagree; 5=strongly agree) evaluated lecture usefulness, facility quality, instructor effectiveness, alignment with expectations, staff support, organizational quality, and overall enjoyment.
- Qualitative items: open-ended prompts solicited feedback on the most and least valuable components, suggested improvements, and logistical factors such as scheduling, pacing, and housing.

Quantitative data were summarized using descriptive statistics. Qualitative responses underwent independent dual-coder thematic analysis; discrepancies were resolved through discussion. Resulting insights informed iterative refinements to curriculum structure, pacing, and instructional methods between program years.

Results

Participant Characteristics

A total of 17 trainees participated in Year 1 and 7 in Year 2 of the AI-READI Bootcamp. As summarized in Table 1, participants came from varied academic and professional backgrounds, including ophthalmology, public health, pharmacology, neuroscience, engineering, and computer science. Educational attainment and programming experience also varied substantially, reflecting the program's deliberate design to attract learners with strong quantitative potential regardless of prior coding experience.

Table . Participant characteristics of bootcamp cohorts.

Characteristic	Year 1 (n=17)	Year 2 (n=7)
Age (years), mean (SD)	33 (3.4)	32 (9.5)
Race, n (%)		
Asian	8 (47.1)	3 (42.9)
African American	3 (17.6)	2 (28.6)
White	4 (23.5)	1 (14.3)
Other	2 (11.8)	0 (0)
Sex, n (%)		
Male	9 (52.9)	2 (28.6)
Female	8 (47.1)	5 (71.4)
Highest degree, n (%)		
PhD or MD	15 (88.2)	4 (57.1)
MA, MS, or MPH	2 (11.8)	1 (14.3)
BA or BS	0 (0)	2 (28.6)
Funding, n (%)		
AI-READI–funded[a]	6 (35.3)	5 (71.4)
Non-AI-READI–funded	11 (64.7)	2 (28.6)
Disciplinary background, n (%)		
Ophthalmology	12 (70.6)	1 (14.3)
Public Health	1 (5.9)	1 (14.3)
Pharmacology	1 (5.9)	1 (14.3)
Neuroscience	1 (5.9)	0 (0)
Engineering	1 (5.9)	0 (0)
Biochemistry	1 (5.9)	0 (0)
Behavioral Science	0 (0)	1 (14.3)
Computer Engineering	0 (0)	1 (14.3)
Medicine	0 (0)	1 (14.3)
Molecular Biology	0 (0)	1 (14.3)
Familiarity with programming language, n (%)		
Python	3 (27.3)	4 (57.1)
R	5 (45.5)	4 (57.1)
SQL	1 (9.1)	2 (28.6)
JAVA	0 (0)	1 (14.3)
MATLAB	2 (18.2)	1 (14.3)
Julia	0 (0)	1 (14.3)
C	0 (0)	1 (14.3)
C++	3 (27.3)	1 (14.3)

[a]AI-READI: Artificial Intelligence–Ready and Exploratory Atlas for Diabetes Insights.

Survey Feedback and Satisfaction Outcomes

Postbootcamp surveys were completed by 13 of 17 (76%) participants in Year 1 and 4 of 7 (57%) in Year 2. Across both cohorts, quantitative ratings were high, with notable improvements in Year 2 (Table 2). Year 2 participants assigned mean scores of 5.00 in 3 categories—instructor effectiveness, staff support, and overall enjoyment—while all other domains, including lecture usefulness, organizational quality, and facility adequacy, averaged between 4.50 and 4.75. No item received a mean rating below 4.0.

Qualitative feedback from Year 1 highlighted several key areas for improvement. Participants valued the conceptual depth of lectures but desired greater emphasis on applied content aligned with their upcoming research projects. Some noted that the larger cohort size made coding laboratories challenging to manage and recommended smaller groups to support individualized troubleshooting. Hands-on laboratories and mentorship—both from faculty and peers—were consistently described as the most valuable program components. Participants also emphasized the benefits of shared housing and peer interaction in fostering collaboration and community.

Table . Postbootcamp evaluation: comparison of Year 1 and Year 2.

Item	Year 1 (n=13)	Year 2 (n=4)
	Average (SD; range)	Average (SD; range)
The lectures were helpful to my learning and development	4.46 (1.05; 2.00-5.00)	4.75 (0.50; 4.00-5.00)
The bootcamp facility was in an accessible location and adequate	4.46 (1.00; 2.00-5.00)	4.75 (0.50; 4.00-5.00)
The instructors helped me understand the subject matter	4.31 (1.11; 1.00-5.00)	5.00 (0.00; 5.00-5.00)
The bootcamp met my educational needs and expectations	4.31 (1.07; 1.00-5.00)	4.75 (0.50; 4.00-5.00)
I had adequate support from the program staff and faculty	4.46 (1.20; 1.00-5.00)	5.00 (0.00; 5.00-5.00)
The bootcamp was well organized	4.23 (1.24; 1.00-5.00)	4.50 (0.58; 4.00-5.00)
I enjoyed the bootcamp overall	4.46 (1.00; 2.00-5.00)	5.00 (0.00; 5.00-5.00)

The redesigned Year 2 curriculum addressed many of these concerns. Participants emphasized the benefits of earlier integration of the AI-READI dataset, closer alignment between instructional content and research projects, and strengthened peer collaboration. Several respondents noted that receiving materials and agendas in advance would have enhanced preparation for more technical sessions. Overall, Year 2 participants described the bootcamp as a strong foundation for subsequent research activities, increasing both confidence and competence.

Together, these findings validate the bootcamp's iterative design and demonstrate that refinements in Year 2 enhanced the learning experience while preserving core strengths in applied instruction, mentorship, and interdisciplinary collaboration.

Curriculum Iteration and Structure

The curriculum was iteratively refined between Year 1 and Year 2 based on participant feedback and faculty debriefings. Year 1 focused on establishing foundations in Python workflows, core ML algorithms, and introductory discussions on ethics and bias. In Year 2, the instructional team implemented several structural and pedagogical updates to enhance applied learning and mentorship. The cohort size was reduced to achieve an approximately 1:2 faculty-to-student ratio, and modules on Git/GitHub version control and environment setup were moved earlier to reinforce reproducibility. Structured, domain-relevant data from the AI-READI project—including retinal images and clinical variables—were integrated throughout the curriculum, allowing trainees to engage directly with multimodal biomedical data that mirrored the complexity of biomedical research workflows.

Additional modules were introduced in Year 2 to align with evolving learner needs and faculty expertise. Mini-seminars on FAIR data principles, AI-READI schema design, and agile project management helped participants navigate the practical aspects of large-scale dataset curation. A dedicated half-day session on large language models introduced transformer architectures and encouraged discussion of the opportunities and limitations of generative AI in health care. The Year 2 capstone project centered on fine-tuning RETFound, a retina-specific foundation model for institutional classification of retinal images, fostering critical reflection on domain generalizability and algorithmic bias.

Tables 3-5 summarize the curriculum's progression from foundational programming to applied AI/ML methods, highlighting key updates, instructional content, and the shift toward hands-on, clinically relevant learning experiences.

Table . Summary of curriculum refinements between Year 1 and Year 2 of the AI-READI[a] Bootcamp.

Dimension	Year 1 focus	Year 2 iteration and rationale
Programming foundations	Introduction to Python and Jupyter via guided exercises	Added pandas and real-world data operations to support independent analysis
Tools and environment	Introduced GitHub and IDEs[b] for version control	Moved earlier to establish reproducibility from the start
Machine learning concepts	Covered regression, SoftMax, convolutional neural networks, and backpropagation	Added large language models and expanded gradient descent laboratories
Data science techniques	Applied principal component analysis, K-means, and spectral clustering to face images	Expanded to exploratory data analysis, digital signal processing, and feature extraction
Applied learning	Face clustering and glucose modeling laboratories	Shifted to retinal image analysis using the AI-READI dataset
Ethics and fairness	Discussed racial bias in pain expression	Broadened to data pitfalls and fairness across AI[c] pipelines
Clinical integration	Minimal use of clinical data	Incorporated AI-READI clinical variables and retinal images
Student engagement	Lunch talks and informal discussions	Added structured mini-seminars and interactive sessions for peer learning

[a]AI-READI: Artificial Intelligence–Ready and Exploratory Atlas for Diabetes Insights.

[b]IDE: integrated development environment.

[c]AI: artificial intelligence.

Table . Year 1 AI-READI[a] Bootcamp curriculum.

Day	Session topics	Format
1	Bootcamp orientation; introduction to Python, Jupyter notebooks	Lecture + hands-on laboratories
2	Python, IDEs[b], Jupyter workflows	Lecture + coding practice
3	Introduction to machine learning, perceptrons, gradient descent; logistic and SoftMax regression	Lecture + laboratories
4	Backpropagation, deep learning fundamentals; representation learning	Lecture + laboratories + discussion
5	Convolutional neural networks	Lecture + laboratories
6	GitHub version control; introduction to pandas	Lecture + coding
7	Linear algebra review; regression models; regression laboratories	Lecture + regression laboratories
8	Principal component analysis, face laboratories, K-means, spectral clustering	Lecture + laboratories
9	Clustering laboratories; discussion on racial bias in data; pitfalls in data science	Lecture + laboratories + ethics discussion
10	Digital signal processing; glucose laboratories; closing reflections	Lecture + laboratories + closing

[a]AI-READI: Artificial Intelligence–Ready and Exploratory Atlas for Diabetes Insights.

[b]IDE: integrated development environment.

Table . Year 2 AI-READI[a] Bootcamp curriculum.

Day	Session topics	Format
1	Introduction to Python, Jupyter, IDEs[b]; GitHub version control	Lecture + hands-on laboratories
2	Python pandas, joining datasets, exploratory data analysis	Lecture + coding exercises
3	Correlations, health sheet overview, data visualization	Lecture + laboratories
4	Digital signal processing, feature extraction, basic image processing	Lecture + laboratories
5	Linear algebra, regression models (linear, nonlinear, ridge, lasso)	Lecture + regression laboratories
6	Principal component analysis (PCA), image alignment, introduction to clustering	Lecture + PCA laboratories
7	Clustering (K-means), pitfalls in data science	Lecture + laboratories + discussion
8	Machine learning introduction, perceptrons, logistic/SoftMax regression	Lecture + gradient descent laboratories
9	Backpropagation, deep learning, representation learning, machine learning best practices	Lecture + eigenface laboratories
10	Convolutional neural networks, introduction to large language models	Lecture + coding demos

[a]AI-READI: Artificial Intelligence–Ready and Exploratory Atlas for Diabetes Insights.

[b]IDE: integrated development environment.

Discussion

Principal Results and Learner Outcomes

Quantitative and qualitative outcomes demonstrate that the AI-READI Bootcamp effectively delivered foundational AI/ML education to multidisciplinary biomedical trainees, yielding consistently high satisfaction and confidence across 2 consecutive years. In Year 2, mean postbootcamp ratings improved across all domains—from 4.23 - 4.46 (Year 1: n=13) to 4.50-5.00 (Year 2: n=4)—with 3 categories (instructor effectiveness, staff support, and overall enjoyment) achieving perfect 5.00 scores. These findings align with Kirkpatrick's training evaluation model (Levels 1 and 2 outcomes), reflecting strong learner satisfaction and perceived knowledge gains, while qualitative feedback suggested enhanced confidence and readiness for independent research (Level 3 outcome).

Bridging Disciplinary Divides in AI/ML Education

As AI continues to transform biomedical research and clinical practice, a persistent skills gap remains between data scientists and clinicians [16-18]. Engineers may lack clinical context, whereas physicians often have limited exposure to algorithmic reasoning and data analytics—constraints that can hinder translational innovation and collaboration.

The AI-READI Bootcamp was intentionally designed to bridge this divide by uniting trainees from medicine, neuroscience, computer science, public health, and pharmacology under a shared mentorship model spanning technical and clinical faculty. This approach reflects best practices identified in recent AI curriculum review [5,11,12] and parallels pedagogical strategies in health professions education—such as interprofessional and team-based learning—that foster cross-disciplinary problem-solving and shared understanding [19].

Building Engagement Through Iterative Refinement

Iterative curriculum refinement was central to sustaining engagement and relevance. Key adjustments between Year 1 and Year 2—including smaller cohort size, earlier integration of AI-READI datasets, and increased time for small-group coding—were guided directly by participant feedback. Year 2 trainees highlighted the benefits of multimodal biomedical data, individualized mentorship, and peer collaboration as core strengths.

Anchoring abstract AI/ML concepts in domain-specific datasets proved particularly effective. By analyzing fundus photographs and structured clinical variables from the AI-READI dataset, participants investigated issues such as site-level variability and domain shift, deepening their understanding of real-world data challenges. This experiential approach aligns with educational frameworks emphasizing authentic data environments and iterative feedback [20], supporting evidence that short-format programs can achieve substantial impact when paired with applied learning and sustained mentorship.

Situating the Bootcamp in the National AI Training Ecosystem

The AI-READI Bootcamp contributes to a growing ecosystem of NIH-supported initiatives advancing the biomedical AI/ML workforce. Within the Bridge2AI program, AI-READI complements other DGPs—including VOICE, which hosts AI summer schools and hackathons on precision public health, and CHoRUS, which provides continuing medical

education–accredited clinical AI workshops on dataset curation, pair programming, and mentored laboratories [9].

Beyond Bridge2AI, the AIM-AHEAD Consortium extends these efforts through part-time fellowships, faculty development programs, and mentored research opportunities aimed at graduate students, health care professionals, and community partners [21]. Parallel innovations are emerging at the institutional level: Stanford University engages students in interdisciplinary teams applying ML to clinical challenges; the Duke Institute for Health Innovation pairs medical trainees with data scientists; and programs at the University of Florida and Carle Illinois College of Medicine integrate clinician-engineer coteaching models. Collectively, these efforts underscore the increasing national and institutional commitment to embedding AI within health professions education, though many remain short-term or elective in scope.

Distinctive Features and Broader Applicability

Although numerous national initiatives—such as AIM-AHEAD, VOICE, and CHoRUS—have expanded AI/ML education through workshops and fellowships, the AI-READI Bootcamp occupies a distinctive niche within this ecosystem. By embedding a 2-week immersive experience within a year-long mentored research internship, it integrates foundational instruction with sustained, project-based engagement. In addition to mastering core AI/ML methods and completing mentored research projects, participants contribute to consortium-wide Bridge2AI initiatives focused on data standardization, FAIR and ethical data practices, biorepository optimization, and workforce development. This multifaceted structure positions the AI-READI Bootcamp as both a pilot and a scalable framework for cultivating interdisciplinary expertise in biomedical AI.

Lessons Learned and Recommendations

Our experience designing and refining the AI-READI Bootcamp suggests several important lessons for future initiatives. Modular, scaffolded content enables learners with varying backgrounds to progress in parallel. The use of curated, domain-relevant datasets grounds abstract concepts in applied contexts, fostering deeper engagement. Participants valued the theoretical framing but reported that they learned most effectively through practical, hands-on components, suggesting future bootcamps should emphasize applied coding while keeping lectures concise and focused.

Equally important are the program's structural features. Maintaining a low faculty-to-student ratio supports real-time troubleshooting and individualized feedback. Embedding bootcamps into longitudinal research structures promotes meaningful skill transfer and project ownership. Structured and informal peer support—through shared housing, collaborative debugging, and group presentations—strengthens technical skills, enhances problem-solving, and builds lasting professional networks. These practices align with curriculum frameworks emphasizing structure, assessment, real-world alignment, and longitudinal mentorship [5,10-12].

Scalability, Replication, and Sustainability

The AI-READI Bootcamp model was intentionally designed for scalability and replication through a modular curriculum organized into discrete instructional units. In alignment with FAIR and open science principles, all lectures, laboratories, and onboarding materials are publicly available on GitHub, enabling other institutions to adapt content to their own technical and educational contexts. The program's structure, which pairs short-term immersive instruction with a longitudinal mentored research experience, offers a reproducible framework that can be integrated into diverse training pipelines, including graduate education, residency programs, and interdisciplinary research initiatives. The AI-READI model also promotes long-term sustainability through its open educational resources within the Bridge2AI consortium, embedded mentorship networks, and continued dissemination of curricular updates and trainee outcomes across the Bridge2AI community.

Limitations and Future Directions

This study is limited by the absence of standardized, performance-based assessments, which are necessary to measure knowledge retention and applied competency. Postbootcamp evaluations for both years relied on self-reported survey data, which are subject to response and recall bias and may not directly reflect objective skill acquisition. In addition, the study is limited by its single-site implementation and small sample size (Year 1: n=17; Year 2: n=7), which may affect generalizability. The Year 2 postbootcamp ratings, while high, are based on only 4 respondents.

To address these gaps, future iterations will incorporate pre- and postbootcamp knowledge assessments and coding exercises aligned with Kirkpatrick Level 2 outcomes to objectively measure learning gains. Longitudinal tracking of trainee outputs such as publications, presentations, and continued engagement in AI-related research will further assess sustained impact. Broader dissemination of the curriculum and collaboration across Bridge2AI partner DGPs may also enhance reproducibility and external validation of outcomes. We also plan to strengthen prebootcamp onboarding (Table 6) through structured preassessment materials, curated readings, and practice exercises to improve baseline preparedness and maximize in-person learning.

Through these refinements, the AI-READI Bootcamp aims to evolve from a formative, single-site pilot to a scalable, high-impact model for interdisciplinary AI/ML education in biomedicine. To promote transparency and adoption, all bootcamp materials, readings, and onboarding instructions are openly available via the AI-READI Bootcamp GitHub repository (Table 6).

Table . Prebootcamp onboarding module and recommended readings.

Category and resource	Details and access
Core online text	
Dive into Deep Learning (D2L)	Website [22]
Bootcamp GitHub repository	
AI-READI[a] Bootcamp GitHub	Web page on GitHub [23]
Readings (author, year)	
Berisha et al, 2021 [24]	Available on Bootcamp GitHub Readings page [23]
Bishop, 2006 [25]	Chapter 1; introduction to chapter 9 & section 9.1; introduction to chapter 12 & section 12.1; Appendix C
Ezer & Whitaker, 2019 [26]	Available on Bootcamp GitHub Readings page [23]
Obermeyer et al, 2019 [27]	Available on Bootcamp GitHub Readings page [23]
Rumelhart et al, 1987 [28]	Available on Bootcamp GitHub Readings page [23]
Strang, 2016 [29]	Also see YouTube lectures [30]
Wilkinson et al, 2016 [31]	Available on Bootcamp GitHub Readings page [23]
Zou & Schiebinger, 2018 [32]	Available on Bootcamp GitHub Readings page [23]
Helpful links	
Introduction to Python	Available on Bootcamp GitHub Readings page [23]
Git terminology	Available on Bootcamp GitHub Readings page [23]
Setting up Git	Available on Bootcamp GitHub Readings page [23]

[a]AI-READI: Artificial Intelligence–Ready and Exploratory Atlas for Diabetes Insights.

Acknowledgments

We thank the faculty instructors, teaching assistants, and administrative staff at University of California San Diego for their invaluable contributions to the planning and delivery of the Artificial Intelligence–Ready and Exploratory Atlas for Diabetes Insights (AI-READI) Bootcamp. We also thank the AI-READI interns for their active participation and constructive feedback, which directly informed improvements to the Year 2 curriculum.

Funding

This work was supported by the National Institutes of Health (NIH) through the Bridge2AI program (award OT2OD032644). The views expressed in this manuscript are those of the authors and do not necessarily represent the official views of the NIH. AM received additional individual support from the SHISEIKAI Scholarship Fund to Study Abroad.

Data Availability

The postbootcamp evaluation data (quantitative survey responses and deidentified qualitative feedback) are not publicly available to preserve participant confidentiality but are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: SLB (lead), TWN, FGPK, AE, AM, PM-K, DG, VP, LMZ
Data curation: TWN, AE, AM, PM-K, DG, VP
Formal analysis: TWN, AE, AM, PM-K, DG, VP
Methodology: TWN, AE, AM, PM-K, DG, VP, SH, AP, SLB, LMZ, GWC, BV, VRdS
Supervision: SLB
Writing – original draft: TWN (lead), FGPK, AE, AM, PM-K, DG
Writing – review & editing: TWN (lead), FGPK, AE, AM, PM-K, DG, VP, SH, AP, SLB, LMZ, GWC, BV, VRdS

Conflicts of Interest

None declared.

References

1. Faruqui SHA, Du Y, Meka R, et al. Development of a deep learning model for dynamic forecasting of blood glucose level for type 2 diabetes mellitus: secondary analysis of a randomized controlled trial. JMIR Mhealth Uhealth 2019 Nov 1;7(11):e14452. [doi: 10.2196/14452] [Medline: 31682586]
2. Schwartz JM, Moy AJ, Rossetti SC, Elhadad N, Cato KD. Clinician involvement in research on machine learning-based predictive clinical decision support for the hospital setting: A scoping review. J Am Med Inform Assoc 2021 Mar 1;28(3):653-663. [doi: 10.1093/jamia/ocaa296] [Medline: 33325504]
3. Scheetz J, Rothschild P, McGuinness M, et al. A survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology and radiation oncology. Sci Rep 2021 Mar 4;11(1):5193. [doi: 10.1038/s41598-021-84698-5] [Medline: 33664367]
4. Chen S, Yu J, Chamouni S, Wang Y, Li Y. Integrating machine learning and artificial intelligence in life-course epidemiology: pathways to innovative public health solutions. BMC Med 2024 Sep 2;22(1):354. [doi: 10.1186/s12916-024-03566-x] [Medline: 39218895]
5. Valikodath NG, Cole E, Ting DSW, et al. Impact of artificial intelligence on medical education in ophthalmology. Transl Vis Sci Technol 2021 Jun 1;10(7):14. [doi: 10.1167/tvst.10.7.14] [Medline: 34125146]
6. Armoundas AA, Narayan SM, Arnett DK, et al. Use of artificial intelligence in improving outcomes in heart disease: a scientific statement from the American Heart Association. Circulation 2024 Apr 2;149(14):e1028-e1050. [doi: 10.1161/CIR.0000000000001201] [Medline: 38415358]
7. Baxter SL, de Sa VR, Ferryman K. AI-READI: rethinking AI data collection, preparation and sharing in diabetes research and beyond. Nat Metab 2024 Dec;6(12):2210-2212. [doi: 10.1038/s42255-024-01165-x] [Medline: 39516364]
8. AI-READI Consortium. Flagship dataset of type 2 diabetes from the AI-READI project. FAIRhub. 2024.
9. Bridge to Artificial Intelligence (Bridge2AI). NIH Common Fund. URL: https://commonfund.nih.gov/bridge2ai [accessed 2025-08-08]
10. Charow R, Jeyakumar T, Younus S, et al. Artificial intelligence education programs for health care professionals: scoping review. JMIR Med Educ 2021 Dec 13;7(4):e31043. [doi: 10.2196/31043] [Medline: 34898458]
11. Tolentino R, Baradaran A, Gore G, Pluye P, Abbasgholizadeh-Rahimi S. Curriculum frameworks and educational programs in AI for medical students, residents, and practicing physicians: scoping review. JMIR Med Educ 2024 Jul 18;10:e54793. [doi: 10.2196/54793] [Medline: 39023999]
12. Mir MM, Mir GM, Raina NT, et al. Application of artificial intelligence in medical education: current scenario and future perspectives. J Adv Med Educ Prof 2023 Jul;11(3):133-140. [doi: 10.30476/JAMP.2023.98655.1803] [Medline: 37469385]
13. National Academies of Sciences, Engineering, and Medicine. Artificial Intelligence in Health Professions Education: Proceedings of a Workshop: The National Academies Press; 2023. [doi: 10.17226/27174]
14. eCFR :: 45 CFR Part 46 -- Protection of Human Subjects. Electronic Code of Federal Regulations (eCFR). URL: https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46 [accessed 2025-12-19]
15. Orgill BD, Nolin J. Learning taxonomies in medical simulation. In: StatPearls: StatPearls Publishing; 2025. URL: http://www.ncbi.nlm.nih.gov/books/NBK559109/ [accessed 2025-12-03]
16. Chen JH, Asch SM. Machine learning and prediction in medicine - beyond the peak of inflated expectations. N Engl J Med 2017 Jun 29;376(26):2507-2509. [doi: 10.1056/NEJMp1702071] [Medline: 28657867]
17. Wartman SA, Combs CD. Medical education must move from the information age to the age of artificial intelligence. Acad Med 2018 Aug;93(8):1107-1109. [doi: 10.1097/ACM.0000000000002044] [Medline: 29095704]
18. Mesko B. The role of artificial intelligence in precision medicine. Expert Rev Precis Med Drug Dev 2017 Sep 3;2(5):239-241. [doi: 10.1080/23808993.2017.1380516]
19. Haidet P, Kubitz K, McCormack WT. Analysis of the team-based learning literature: TBL comes of age. J Excell Coll Teach 2014;25(3-4):303-333. [Medline: 26568668]
20. Kjelvik MK, Schultheis EH. Getting messy with authentic data: exploring the potential of using data from scientific research to support student data literacy. In: Gardner S, editor. CBE Life Sci Educ 2019 Jun;18(2):es2. [doi: 10.1187/cbe.18-02-0023] [Medline: 31074698]
21. AIM-AHEAD programs. AIM-AHEAD. URL: https://www.aim-ahead.net/programs/ [accessed 2025-12-03]
22. Dive into Deep Learning. URL: https://d2l.ai/ [accessed 2025-12-03]
23. AI-READI bootcamp. GitHub. URL: https://github.com/voytek/AI-READI-Bootcamp [accessed 2025-12-03]
24. Berisha V, Krantsevich C, Hahn PR, et al. Digital medicine and the curse of dimensionality. NPJ Digit Med 2021 Oct 28;4(1):153. [doi: 10.1038/s41746-021-00521-5] [Medline: 34711924]
25. Bishop C. Pattern Recognition and Machine Learning: Springer; 2006.
26. Ezer D, Whitaker K. Data science for the scientific life cycle. Elife 2019 Mar 6;8:e43979. [doi: 10.7554/eLife.43979] [Medline: 30839275]
27. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]

28. Rumelhart H, Williams C. Learning internal representations by error propagation. In: Parallel Distributed Processing, Volume 1: Explorations in the Microstructure of Cognition: Foundations 1987:318-362 URL: https://dl.acm.org/doi/10.5555/104279.104293 [accessed 2025-12-03]
29. Strang C. Eigenvalues and eigenvectors. In: Introduction to Linear Algebra, 5th edition: Wellesley-Cambridge Press; 2016.
30. Lec 1 | MIT 18.06 linear algebra, spring 2005. MIT OpenCourseWare YouTube page. 2009 May 6. URL: https://www.youtube.com/watch?v=ZK3O402wf1c&list=PL49CF3715CB9EF31D&index=1 [accessed 2025-12-17]
31. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3:160018. [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
32. Zou J, Schiebinger L. AI can be sexist and racist - it's time to make it fair. Nature New Biol 2018 Jul;559(7714):324-326. [doi: 10.1038/d41586-018-05707-8] [Medline: 30018439]

Abbreviations

AI: artificial intelligence
AI-READI: Artificial Intelligence–Ready and Exploratory Atlas for Diabetes Insights
Bridge2AI: Bridge to Artificial Intelligence
DGP: data generation project
FAIR: findable, accessible, interoperable, and reusable
ML: machine learning
NIH: National Institutes of Health

Edited by A Hasan Sapci; submitted 28.08.25; peer-reviewed by B Niehaves, M Kozak, S Sharma; revised version received 30.10.25; accepted 09.11.25; published 29.12.25.

Please cite as:
Nishihara TW, Kalaw FGP, Engmann A, Motoyoshi A, Mensah-Kane P, Gupta D, Patronilo V, Zangwill LM, Hallaj S, Panahi A, Cottrell GW, Voytek B, de Sa VR, Baxter SL
Fostering Multidisciplinary Collaboration in Artificial Intelligence and Machine Learning Education: Tutorial Based on the AI-READI Bootcamp
JMIR Med Educ 2025;11:e83154
URL: https://mededu.jmir.org/2025/1/e83154
doi:10.2196/83154

Creation of the ECHO Idaho Podcast: Tutorial and Pilot Assessment

Ryan Wiet*, PhD; Madeline P Casanova*, PhD; Jonathan D Moore*, PhD; Sarah M Deming*, PhD; Russell T Baker Jr*, PhD, DAT

WWAMI Medical Education Program, Idaho Office of Underserved and Rural Medical Research, University of Idaho, 875 Perimeter Drive, Moscow, ID, United States

*all authors contributed equally

Corresponding Author:
Russell T Baker Jr, PhD, DAT
WWAMI Medical Education Program, Idaho Office of Underserved and Rural Medical Research, University of Idaho, 875 Perimeter Drive, Moscow, ID, United States

Abstract

Background: Project ECHO (Extension for Community Health Outcomes) is an innovative program that uses videoconferencing technology to connect health care providers with experts. The model has been successful in reaching health care providers in rural and underserved areas and positively impacting clinical practice. ECHO Idaho, a replication partner, has developed programming that has increased knowledge and confidence of health care professionals throughout the state of Idaho, United States. Although the ECHO model has a demonstrated ability to recruit, educate, and train health care providers, barriers to attending Project ECHO continuing education (CE) programs remain. The asynchronous nature of podcasts could be used as an innovative medium to help address barriers to CE access that health care professionals face. The ECHO Idaho "Something for the Pain" podcast was developed to increase CE accessibility to rural and frontier providers, while upscaling their knowledge of and competence to treat and assess substance use disorders, pain, and behavioral health conditions.

Objective: This paper describes the creation and preliminary assessment of the ECHO Idaho "Something for the Pain" podcast.

Methods: Podcast episodes consisted of interviews with individuals as well as didactic lectures. Audio from these recordings were edited for content and length and then professionally reviewed by subject matter experts (eg, featured episode speakers). Target audiences consisted of health care providers and community members interested in behavioral health and substance use disorders. Metrics on podcast listeners were assessed using SoundCloud's RSS feed, continuing education survey completion, and iECHO.

Results: The ECHO Idaho "Something for the Pain" podcast's inaugural season comprised 14 episodes with 626 minutes of CE material. The podcast series received a total of 2441 listens from individuals in 14 different cities across Idaho, and 63 health care providers listened and claimed CE credits. The largest professional group was social workers (n=22; 35%).

Conclusions: We provide preliminary evidence that podcasts can be used to provide health care providers with opportunities to access CE material. Health care providers listened to and claimed CE credits from the ECHO Idaho "Something for the Pain" podcast. Project ECHO programs should consider creating podcasts as an additional platform for disseminating ECHO material.

(JMIR Med Educ 2025;11:e55313) doi:10.2196/55313

KEYWORDS

Project ECHO; ECHO Idaho; medical education; medical training; medication teaching; medical knowledge; rural health care; rural medicine; underserved population; underserved people; substance use; substance use disorder; SUD; drug abuse; drug use; alcoholism; addiction; pain; behavioral health; podcast; webinar

Introduction

Project ECHO (Extension for Community Health Outcomes), founded by the University of New Mexico, provides accessible education and specialty training to rural and underserved professionals through an all-teach, all-learn model that uses interactive video conferencing (ie, Zoom technology) [1]. In 2018, through a collaboration between the University of Idaho WWAMI (Washington, Wyoming, Alaska, Montana, Idaho) Regional Medical Education Program and the Idaho Department of Health and Welfare Statewide Healthcare Innovation Plan, ECHO Idaho became the only Project ECHO replication partner in the state of Idaho [2]. Since 2018, ECHO Idaho has provided rural and underserved communities with educational opportunities and cultivated a community of health care professional learners across the state. Researchers assessing the Project ECHO model have found that attendees increase their

knowledge, clinical skills, and confidence, thus increasing the quality of care for patients in rural and underserved communities [1-10]. Health care professionals participating in the ECHO Idaho series have similarly reported increases in both their knowledge and in their confidence to provide specialty care for their patients [2,11].

Despite the myriad documented benefits of the Project ECHO model, challenges relevant to maximizing attendance and increasing overall reach of the program (eg, recruiting participants from rural areas) still exist [2-4,12]. For example, scheduling conflicts and time constraints have been reported as barriers to attending ECHO sessions [5,10,13]. Although some hubs have modified the ECHO model (eg, changed session lengths), no conclusive changes have been recommended and minimal research has been conducted to assess how modifications to the model could positively impact and enhance provider training [10,13]. Other models of continuing education (CE) (eg, e-learning) have been reported to be effective in changing knowledge, self-efficacy, and skills [14]. However, the effect of such models on professional and clinical practice, as well as patient outcomes, remains largely unknown.

Thus, exploring and researching novel or innovative platforms to disseminate information and increase clinical skills for health care providers, while subsequently decreasing potential barriers to attendance, are necessary. One proposed approach is the use of a podcast platform to deliver Project ECHO materials. A podcast may be an effective, interactive, and cost-efficient way to deliver Project ECHO programming while addressing some barriers reported by participants. For example, the asynchronous nature of podcasts would allow providers the flexibility to engage with the material on their own time, and to pause and resume listening when needed. Research has reported that podcasts can be an impactful model to train and enhance professional development [15-17]. Additionally, due to the COVID-19 pandemic dramatically changing the landscape of health care and education [18], the use of podcasts to deliver medical education curricula has increased [19-22]. Researchers have found podcasts increase participant knowledge and self-efficacy while also indirectly impacting professional and clinical practice [19-22]; however, the full scope of the impact of educational podcasts remains unknown [16].

To test the efficacy of the podcast model for delivering medical education, ECHO Idaho developed the "Something for the Pain" podcast, with support from the Idaho WWAMI Medical Education Program and in partnership with the Valley County Opioid Response Project [23]. The podcast was meant to increase CE accessibility for rural and frontier providers, while upscaling their knowledge of and competence to treat and assess behavioral health conditions. The purpose of this paper was to present the development and implementation process of the ECHO Idaho "Something for the Pain" podcast's inaugural season, as well as to discuss the preliminary findings regarding the reach of the podcast, lessons learned, and proposed future uses of this innovative platform.

Methods

Program Development and Implementation

Program Description

Due to the increased need for more specialty training related to behavioral health, opioids, pain, and substance use disorder (SUD), in 2021, the ECHO Idaho "Something for the Pain" podcast was developed as an innovative approach to disseminate ECHO Idaho materials. The inaugural season contained 14 episodes that presented best-practices and resources for behavioral health, opioid use disorder, and SUD prevention, treatment, and recovery specific to Idaho. The podcast was free to access, and eligible providers were able to obtain no-cost CE credits for listening to an episode.

Episode Development

The ECHO model learning framework typically includes two sections: (1) didactic lectures and (2) a case-based presentation and discussion. To mirror the framework of the ECHO model, the 14 podcast episodes were primarily broken down into two parts: (1) interviews with individuals and (2) didactic presentations. The interviews were with health care professionals and community members who had first-hand experience and knowledge of evidence-based practices and resources pertinent to the Idaho context. The didactic presentations were taken from previously recorded ECHO Idaho sessions across 4 series (ie, Behavioral Health in Primary Care; Opioids, Pain and Substance Use Disorders; Counseling Techniques for Substance Use Disorders; and Viral Hepatitis and Liver Care). Audio from these recordings were edited for content and length and then professionally reviewed by subject matter experts. The subject matter experts also assisted with generating CE-eligible episode assessment questions that could be used to gauge audience engagement.

Audience Recruitment

The primary intended audience for the podcast was health care providers and community members with an interest in SUD and behavioral health. Initial recruitment occurred by using ECHO Idaho's network of prior ECHO Idaho session attendees, which, at the time of the first episode's release in May 2021, consisted of approximately 3000 diverse Idaho health care professionals (eg, those who held an MD/DO, PA, NP, MCSW, or LCPC). Members of the ECHO Idaho staff invited participants from the pre-existing ECHO network to listen to the podcast as an additional means of earning CE credit at their convenience. Additionally, ECHO staff developed a targeted marketing campaign that involved personal and bulk emails, newsletter announcements, weekly announcements, paid advertisements, social media posts, print bulk postcard mailings, as well as advertising on other podcast programs. The podcast was housed on SoundCloud, which increased its accessibility through user search engines and algorithms.

Podcast Data Metrics

Listener engagement was tracked using SoundCloud's RSS feed. The feed provided click metrics (ie, how many listens occurred) for each podcast episode consumed within a specified

timeframe. Eeds, an electronic CE management system, was used to track CE credits claimed following episode and assessment completion. Additionally, to gain insight into listener experience, participants were required to answer a user experience question in Eeds. Lastly, individuals interested in listening to the podcast could register on the ECHO Idaho website. Registration information collected included demographic questions like profession, credentials, primary practice location, sex, and age. iECHO, a web-based proprietary program and management software database, was used to manage participant demographic data (ECHO Institute, University of New Mexico) entered on the registration form.

Data Analysis

Data were exported from Eeds and iECHO and a descriptive statistical analysis was performed using SPSS, version 28 (IBM Corp).

Ethical Considerations

The project was certified as exempt (protocol #23‐150) by the Institutional Review Board at the University of Idaho. Data were deidentified for the purposes of analysis, and informed consent was obtained from all participants prior to their involvement in the study.

Results

RSS Feed

The first season of the ECHO Idaho “Something for the Pain” podcast included 14 episodes (released May 2021 to June 2022); 13 episodes were related to perinatal SUD, and 1 bonus episode gave a brief history of Project ECHO and Vandal Theory (Table 1). The average length of an episode was 45 (SD 11.9) minutes and the average number of listens per episode was 188 (SD 46.5) (Table 1). The season provided a total of 626 minutes of CE material available for perpetual access. The podcast was released on the streaming platforms SoundCloud, Apple Podcasts, Google Podcasts, Spotify, Sticher, and iHeartRadio. As of April 19, 2023, the initial season of the podcast had garnered over 2000 listeners from various parts of the United States, with most of the listeners based in Idaho.

Table . Information on each episode including release date, title, length, and total listens of each episode.

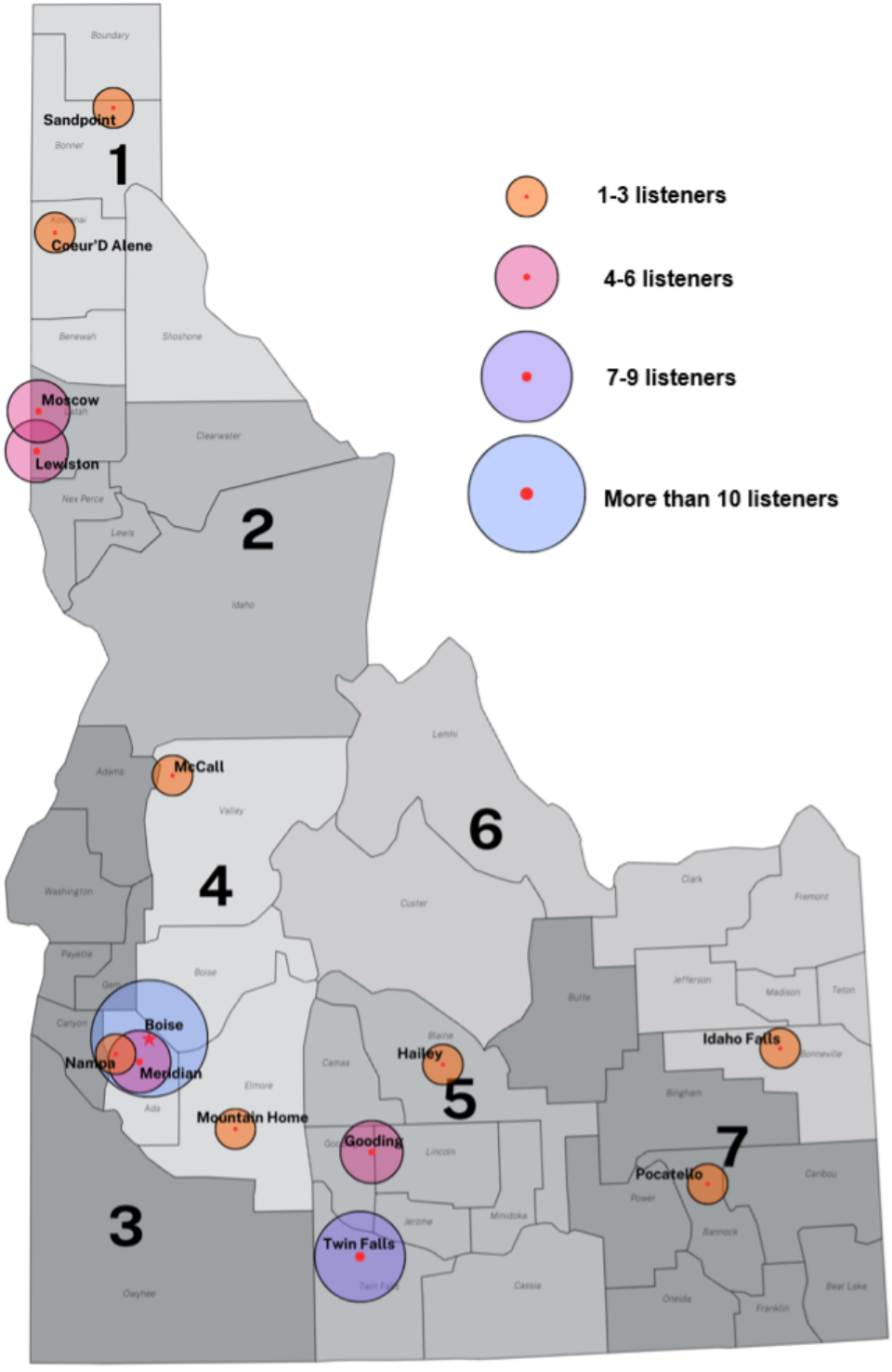
Release date	Episode	Length (minutes)	Listens
5/11/2021	Episode 1: Framework for Addiction as Disease (feat. Craig Lodis, PhD)	31	286
5/17/2021	Episode 2: State of Substance Use in Idaho (feat. Amy Jeppesen, LCSW, ACADC)	47	212
6/3/2021	Bonus Episode: Project ECHO Origin Story & The Vandal Theory (feat. Sanjeev Arora, MD)	37	N/A[a]
7/1/2021	Episode 3: De-escalation Techniques and the Valley County Court Services' Diversion Program (feat. Abby Abbondondalo and Skip Clapp)	45	159
7/11/2021	Episode 4: Harm Reduction and Valley County's Opioid Response Project (feat. Brenda Hoyt, NP, Courtney Boyce and Shelly Hitt)	47	165
7/23/2021	Episode 5: Motivational Interviewing and Donnelly's The Change Clinic (feat. Deb Thomas and Barbara Norton)	60	191
8/9/2021	Episode 6: LaDessa Foster Talks Levels of Care in Substance Use Disorder Treatment (feat. LaDessa Foster)	32	164
8/23/2021	Episode 7: Monica Forbes Talks SMART Recovery, Stigma and Reentering Society Post-Incarceration (feat. Monica Forbes)	35	178
9/1/2021	Episode 8: Marjorie, Wilson Talks Idaho's Syringe Service Programs (feat. Marjorie, Wilson and Ian Trosoyer, DNP)	51	185
9/13/2021	Episode 9: LaDessa Foster Talks Managing Clinical Services for Patients and Providers (feat. LaDessa Foster)	30	274
3/2/2022	Episode 10: Lindsay Brown Talks Peer Recovery Supports (feat. Lindsay Brown)	55	168
3/25/2022	Episode 11: Talking Telehealth in SUD Care | McCall Mobile Medicine (feat. Drew Holliday, MSW)	45	188
5/4/2022	Episode 12: Deborah Seltzer Talks Coding and Billing for Substance Use Disorders (feat. Deborah Seltzer)	51	122
6/13/2022	Episode 13: SUD Treatment for Justice-Involved Patients | Day One Program (feat. Radha Sadacharan, MD, MPH and Rebecca Lee)	60	149

[a]Not available.

CE Credit

A total of 63 unique health care professionals representing 14 distinct cities and spanning all 7 public health districts within the state of Idaho, as well as out-of-state listeners, claimed CE credit for season 1 of the podcast (Figure 1). The podcast drew a diverse array of professionals from various disciplines, with social workers being the most widely represented group among those who claimed CE credits (Multimedia Appendix 1). Three questions were generated for each of the 13 episodes eligible for CE credits. The average percentage of correct answers for all episodes was 83% (SD 11.6%) (Multimedia Appendix 2), thus providing preliminary evidence related to audience engagement and potential impact on learning outcomes.

Figure 1. Health care professionals claiming continuing education credits from the ECHO Idaho "Something for the Pain" podcast within the state of Idaho. Values 1-7 represent Idaho's 7 public health districts (map adapted from MapChart).



Discussion

Principal Findings

ECHO Idaho's "Something for the Pain" podcast included 14 episodes providing best practices and resources for behavioral health, opioid use disorder, and SUD prevention, treatment, and recovery specific to Idaho. The podcast included interviews with health care professionals and community members who had first-hand experience and knowledge of evidence-based practices, as well as didactic presentations taken from previously recorded ECHO Idaho sessions. Audience recruitment primarily focused on health care providers and community members interested in the behavioral health and SUD field. Metrics such as click metrics, CE credits, and user experience questions were used to track listener engagement. Data collected from Eeds and iECHO systems were analyzed using SPSS, providing a descriptive statistical analysis.

The ECHO Idaho podcast effectively engaged a diverse group of health care professionals throughout Idaho, demonstrating the utility of podcasts as a versatile tool for professional development and outreach. This highlights the possibility of leveraging podcasts to not only provide CE but also to serve as a means of attracting more health care professionals to Project ECHO sessions. The widespread accessibility of podcasts suggests their potential as a potent recruitment tool for future Project ECHO initiatives.

This report presents a preliminary assessment of an innovative platform to engage with health care professionals, as well as community members, by providing insightful information and professional development material via the medium of podcasts.

Limitations

Although the initial season of the podcast engaged with diverse health care providers across the state, the podcast's impact on clinical practice remains unknown. Previous literature has suggested internet-based continuing medical education can increase knowledge, change physician behavior, and indirectly impact clinical practice [16,24]. Therefore, future research should collect feedback from nonlisteners and previous listeners to understand their preferences for this innovative platform as well as assess the direct impact of podcast-based CE on listeners' knowledge and professional and clinical practice through quantitative and qualitative analyses.

Additionally, although several data analytics methods were used, some challenges remain. The RSS data records clicks (ie, listens) on each episode, but there is no way to track how long someone listened or if they completed the full episode. Therefore, interpretation of those data points should be made with caution. Similarly, due to data constraints, it is difficult to assess the benefits of the ECHO podcast compared to the more traditional ECHO program model. As a result, we were unable to fully evaluate the podcast's impact, particularly in terms of reach, impact on learning outcomes, and its influence on clinical practice. Future research should focus on assessing these variables, as they are crucial for understanding the effectiveness of this type of CE material. Lastly, the effectiveness of our recruitment strategies remains undetermined. Future research should prioritize evaluating these strategies to optimize the return on investment and enhance the overall impact of the podcast.

Lessons Learned

Overall, the goal of the podcast was to provide clinicians across the state with an easily accessible CE program. We learned that podcasts could be used to engage Idaho health care providers and provide access to CE opportunities. Although we provide evidence of successfully reaching listeners, a robust marketing campaign was not used. Furthermore, the topics were selected by a team of subject matter experts without input from listeners or previous ECHO Idaho attendees. In the future, assessing additional indicators (eg, listener preferences related to episode length, topic selection, impact on clinical practice, perceived barriers to listening to podcasts) could ensure that new podcast series and episodes meet the needs of current and potential listeners.

Conclusions

The ECHO Idaho "Something for the Pain" podcast was developed as a new and innovative way of providing CE for health care providers. We provide evidence that this approach was successful in its efforts to engage a sizable audience of listeners who claimed CE credits for participation. The podcast had listeners from diverse health care professions representing cities from across Idaho and the United States. Future research efforts should include the collection of additional information such as listener preferences, knowledge change, professional and clinical impact, and patient outcomes to guide the implementation of Project ECHO information into an effective CE podcast format.

Acknowledgments

ECHO (Extension for Community Health Outcomes) Idaho's "Something for the Pain" (episodes 1-13) podcast was created by ECHO Idaho and the University of Idaho, supported by the WWAMI Medical Education Program, in partnership with the Valley County Opioid Response Project. ECHO Idaho's "Something for the Pain" (episodes 1-13) podcast was made possible by a grant from the US Department of Health and Human Services Health Resources and Services Administration (HRSA; grant GA1RH39585). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of Central District Health or HRSA. The authors would like to thank Samuel Steffen for his work with podcast development and implementation.

Authors' Contributions

RW was involved in the concept and design, data analysis, data interpretation, manuscript writing, and approval of the final manuscript. MPC was involved in the concept and design, data interpretation, manuscript writing, and approval of the final manuscript. JDM, SMD, and RTB were involved in the concept and design, manuscript writing, and approval of the final manuscript. All authors agreed to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Unique health care professionals claiming continuing education credit from the ECHO Idaho “Something for the Pain” podcast.
[DOCX File, 22 KB - mededu_v11i1e55313_app1.docx]

Multimedia Appendix 2
Percentage of correct answers for questions of continuing education credit–eligible episodes.
[DOCX File, 22 KB - mededu_v11i1e55313_app2.docx]

References

1. Komaromy M, Duhigg D, Metcalf A, et al. Project ECHO (Extension for Community Healthcare Outcomes): a new model for educating primary care providers about treatment of substance use disorders. Subst Abus 2016;37(1):20-24. [doi: 10.1080/08897077.2015.1129388] [Medline: 26848803]
2. Baker RT, Casanova MP, Whitlock JN, Smith LH, Seegmiller JG. Expanding access to health care: evaluating project Extension for Community Health Care Outcomes (ECHO) Idaho’s tele-education behavioral health program. Journal of Rural Mental Health 2020;44(4):205-216. [doi: 10.1037/rmh0000157]
3. Katzman JG, Qualls CR, Satterfield WA, et al. Army and Navy ECHO pain telementoring improves clinician opioid prescribing for military patients: an observational cohort study. J Gen Intern Med 2019 Mar;34(3):387-395. [doi: 10.1007/s11606-018-4710-5] [Medline: 30382471]
4. Shimasaki S, Bishop E, Guthrie M, Thomas JFF. Strengthening the health workforce through the ECHO stages of participation: participants’ perspectives on key facilitators and barriers. J Med Educ Curric Dev 2019;6:2382120518820922. [doi: 10.1177/2382120518820922] [Medline: 30729170]
5. Salvador J, Bhatt S, Fowler R, et al. Engagement with Project ECHO to increase medication-assisted treatment in rural primary care. Psychiatr Serv 2019 Dec 1;70(12):1157-1160. [doi: 10.1176/appi.ps.201900142] [Medline: 31434561]
6. Dubin RE, Flannery J, Taenzer P, et al. ECHO Ontario chronic pain & opioid stewardship: providing access and building capacity for primary care providers in underserviced, rural, and remote communities. Stud Health Technol Inform 2015;209:15-22. [doi: 10.3233/978-1-61499-505-0-15] [Medline: 25980700]
7. Katzman JG, Comerci G Jr, Boyle JF, et al. Innovative telementoring for pain management: project ECHO pain. J Contin Educ Health Prof 2014;34(1):68-75. [doi: 10.1002/chp.21210] [Medline: 24648365]
8. Mazurek MO, Brown R, Curran A, Sohl K. ECHO Autism. Clin Pediatr (Phila) 2017 Mar;56(3):247-256. [doi: 10.1177/0009922816648288] [Medline: 27169714]
9. Arora S, Kalishman S, Thornton K, et al. Expanding access to hepatitis C virus treatment--Extension for Community Healthcare Outcomes (ECHO) project: disruptive innovation in specialty care. Hepatology 2010 Sep;52(3):1124-1133. [doi: 10.1002/hep.23802] [Medline: 20607688]
10. Agley J, Delong J, Janota A, Carson A, Roberts J, Maupome G. Reflections on project ECHO: qualitative findings from five different ECHO programs. Med Educ Online 2021 Dec;26(1):1936435. [doi: 10.1080/10872981.2021.1936435] [Medline: 34076567]
11. Casanova MP, Nelson MC, Blades KC, Smith LH, Seegmiller JG, Baker RT. Evaluation of an opioid and addiction treatment tele-education program for healthcare providers in a rural and frontier state. J Opioid Manag 2022;18(4):297-308. [doi: 10.5055/jom.2022.0725] [Medline: 36052928]
12. Furlan AD, Zhao J, Voth J, et al. Evaluation of an innovative tele-education intervention in chronic pain management for primary care clinicians practicing in underserved areas. J Telemed Telecare 2019 Sep;25(8):484-492. [doi: 10.1177/1357633X18782090] [Medline: 29991316]
13. Shea CM, Gertner AK, Green SL. Barriers and perceived usefulness of an ECHO intervention for office-based buprenorphine treatment for opioid use disorder in North Carolina: a qualitative study. Subst Abus 2021;42(1):54-64. [doi: 10.1080/08897077.2019.1694617] [Medline: 31809679]
14. Rouleau G, Gagnon MP, Côté J, et al. Effects of e-learning in a continuing education context on nursing care: systematic review of systematic qualitative, quantitative, and mixed-studies reviews. J Med Internet Res 2019 Oct 2;21(10):e15118. [doi: 10.2196/15118] [Medline: 31579016]

15. Block J, Lerwick P. 1064: educational preferences among residents in the ICU. Crit Care Med 2019;47(1):509. [doi: 10.1097/01.ccm.0000551809.99859.12]
16. Malecki SL, Quinn KL, Zilbert N, et al. Understanding the use and perceived impact of a medical podcast: qualitative study. JMIR Med Educ 2019 Sep 19;5(2):e12901. [doi: 10.2196/12901] [Medline: 31538949]
17. Purdy E, Thoma B, Bednarczyk J, Migneault D, Sherbino J. The use of free online educational resources by Canadian emergency medicine residents and program directors. CJEM 2015 Mar;17(2):101-106. [doi: 10.1017/cem.2014.73] [Medline: 25927253]
18. Grafton-Clarke C, Uraiby H, Gordon M, et al. Pivot to online learning for adapting or continuing workplace-based clinical learning in medical education following the COVID-19 pandemic: a BEME systematic review: BEME Guide No. 70. Med Teach 2022 Mar;44(3):227-243. [doi: 10.1080/0142159X.2021.1992372] [Medline: 34689692]
19. Chartier LB, Hansen S, Lim D, et al. P023: Code Resus - using a quality improvement approach to improve health care provider response during resuscitations. CJEM 2016 May;18(S1):S86-S86. [doi: 10.1017/cem.2016.199]
20. McCarthy J, Porada K. Serving up Peds Soup: podcast-based paediatric resident education. Med Educ 2020 May;54(5):456-457. [doi: 10.1111/medu.14113] [Medline: 32185820]
21. Reid Burks A, Nicklas D, Owens J, Lockspeiser TM, Soranno D. Urinary tract infections: pediatric primary care curriculum podcast. MedEdPORTAL 2016 Aug 5;12:10434. [doi: 10.15766/mep_2374-8265.10434] [Medline: 31008213]
22. Roth J, Chang A, Ricci B, Hall M, Mehta N. Why not a podcast? Assessing narrative audio and written curricula in obstetrical neurology. J Grad Med Educ 2020 Feb;12(1):86-91. [doi: 10.4300/JGME-D-19-00505.1] [Medline: 32089798]
23. ECHO Idaho. Something for the Pain. SoundCloud. URL: https://soundcloud.com/user-658492948 [accessed 2025-03-05]
24. Fordis M, King JE, Ballantyne CM, et al. Comparison of the instructional efficacy of internet-based CME with live interactive CME workshops: a randomized controlled trial. JAMA 2005 Sep 7;294(9):1043-1051. [doi: 10.1001/jama.294.9.1043] [Medline: 16145024]

Abbreviations

CE: continuing education
ECHO: Extension for Community Health Outcomes
SUD: substance use disorder
WWAMI: Washington, Wyoming, Alaska, Montana, Idaho

Edited by B Lesselroth; submitted 08.12.23; peer-reviewed by J Woods, LE Tomedi, Y Lunsky; revised version received 15.08.24; accepted 25.02.25; published 21.03.25.

Please cite as:
Wiet R, Casanova MP, Moore JD, Deming SM, Baker Jr RT
Creation of the ECHO Idaho Podcast: Tutorial and Pilot Assessment
JMIR Med Educ 2025;11:e55313
URL: https://mededu.jmir.org/2025/1/e55313
doi:10.2196/55313

Cardiac Implantable Electronic Device Educational Application for Cardiac Anesthesiology Trainees: Tutorial on App Development

Ahmed Zaky[1*], MD, MPH, MBA, MSHQS, CMQ; Aisha Waheed[2*], MD, MS; Brittany Hatter[1*], MD, MS, MEd; Srilakshmi Malempati[3*], PhD; Sai Hemanth Maremalla[3*], MSc; Ragib Hasan[3*], PhD; Yuliang Zheng[3*], PhD; Scott Snyder[4*], PhD

[1]Department of Anesthesiology and Perioperative Medicine, University of Alabama at Birmingham, 619 South 19th Street, JT 804, Birmingham, AL, United States

[2]Department of Internal Medicine, Princeton Baptist Health System, Birmingham, AL, United States

[3]Department of Computer Sciences, University of Alabama at Birmingham, Birmingham, AL, United States

[4]Department of Education, University of Alabama at Birmingham, Birmingham, AL, United States

[*]all authors contributed equally

Corresponding Author:
Ahmed Zaky, MD, MPH, MBA, MSHQS, CMQ
Department of Anesthesiology and Perioperative Medicine, University of Alabama at Birmingham, 619 South 19th Street, JT 804, Birmingham, AL, United States

Abstract

Despite the exposure of cardiothoracic anesthesiology trainees to patients with cardiac implantable electronic devices (CIEDs), there is a paucity of formal curricula on this subject. Major impediments to educating cardiothoracic anesthesiology trainees on CIEDs include busy clinical schedules, short staffing, inconsistent trainees' exposure to CIEDs, multiplicity of vendors, and a "millennial" mentality of the new generation of learners. As a result, cardiothoracic anesthesiology trainees graduating from their residency and fellowship programs may lack the competency to manage patients with CIEDs. Herein, we report our systematic approach to designing, validating, mapping, evaluating, and delivering a CIED curriculum on the first mobile app of its kind on this subject. Development of the CIED curriculum proceeded through the Kern 6-step approach of problem identification, determining and prioritizing content, writing goals and objectives, selecting instructional strategies, implementation of the material, and evaluation and applications of lessons learned. This was followed by the delivery of the curriculum in the form of a user-study app and administrator-type app with functionalities in the assessment of the learners' gains, experience, and satisfaction as well as the administrator's capability to update the educational content based on the feedback of the learners and the emerging technology. As such, the CIED app allows asynchronous learning at the pace of the learners and allows, through a multiplicity of educational materials, the ability to digest this complex and understudied subject. We report on the pilot phase of the project. We benefit from the experience of a multidisciplinary team of anesthesiologists, computer scientists, and educators in accomplishing this project.

(JMIR Med Educ 2025;11:e60087) doi:10.2196/60087

KEYWORDS

pacemakers; curriculum; mobile; education; rhythm devices; anesthesiology trainees; technology; medical education; cardiac; cardiac implantable; electronic device; application; app; anesthesiology; cardiothoracic anesthesiology; trainee; assessment

Introduction

Cardiac implantable electronic devices (CIEDs) are electronic devices placed in patients experiencing permanent life-threatening arrhythmias. They include pacemakers, implantable cardioverter-defibrillators (ICD), and cardiac resynchronization therapies (CRT). The number of CIEDs implanted is increasing rapidly worldwide [1]. Due to increased longevity and technological advancement, the number of patients with CIEDs undergoing surgical procedures is steadily increasing [2]. With a growing sophistication of the technology and the potential for preventable adverse events related to CIED dysfunctionality in the perioperative period [3], there is a need for anesthesiology trainees to gain the knowledge and skills necessary to manage CIEDs perioperatively [4].

Despite an increased exposure of anesthesiology trainees to patients with CIEDs, and the availability of anesthesiology-led CIED services [5], there have not been formal curricula to educate trainees on this complex subject. Furthermore, busy clinical schedules, staff shortages, relatively short duration of the cardiothoracic rotation, the unpredicted and inconsistent exposure of trainees to patients with CIEDs, and the "millennial" mentality of the current generation of learners [6], all have challenged synchronous teaching and learning modalities. As

a consequence, currently graduating anesthesiology trainees may lack the competency needed to manage patients with CIEDs in their future careers.

Mobile learning (m-learning) is a learning mode that uses a mobile device or tablet. It allows self-directed asynchronous learning at the pace of the learner and hence frees the learner of time and physical constraints. M-learning was shown to be associated with higher learning gains compared with conventional learning in a multitude of medical specialty education [7-10]. At the time of writing this study, there are only 4 programs in the country where CIEDs are perioperatively managed by anesthesiologists with inconsistencies in the educational methodologies delivered to the trainees on this subject.

To remedy the current gap, we describe in this tutorial the systematic process of development and delivery of a high-quality CIED educational curriculum through a mobile app platform. This manuscript should be of interest to anesthesiology educators seeking to incorporate new competencies and rubrics related to CIED into residency training programs.

Methods

Development of High-Quality CIED Educational Curriculum

Using the Kern 6-step approach to curriculum development [11], we created goals, learning objectives, instructional content, and pre and postassessments to be delivered via m-learning, and user feedback regarding CIEDs. The Kern framework for curriculum development has been applied for training across multiple medical specialties [12-14]. The authors expanded the 6 steps by including an expert review of the instructional objectives.

Due to the perioperative nature of CIEDs, the curriculum focuses on knowledge and skills within the Accredited Council of Graduate Medical Education (ACGME) domain of Patient Care.

The basic process of curriculum development in medical education has been outlined [12-14] as consisting of the following steps:

1. Problem identification
2. Determining and prioritizing content (focusing on the needs assessment).
3. Writing goals and objectives.
4. Selecting instructional strategies.
5. Implementation of the curriculum.

Evaluation and application of lessons learned.

Problem Identification

Several intersecting challenges motivated the development of a curriculum for CIED management. As indicated earlier, technological advances and increased life expectancy have resulted in a steady increase in the number of CIEDs implanted annually [2]. Similarly, there has been an increase in the number of patients with CIEDs undergoing surgical procedures. Due to the rates of postoperative CIED-related complications, CIED-related deaths, and hospital readmissions related to CIEDs, there is a specific need for training in the knowledge and skills needed to manage CIEDs in the perioperative period. Cardiothoracic anesthesiology trainees are extensively exposed to patients with CIEDs during their training. Training ranges from a month-long rotation in the cardiac operating rooms and cardiac surgical intensive care units as an elective or core rotation during residency to a 12-month dedicated accredited clinical fellowship. Despite evidence that the initiation of perioperative anesthesiology-based CIED service is associated with better patient outcomes [5,15-17], there are no standardized published training curricula regarding CIEDs for cardiothoracic anesthesiology trainees. While these conditions highlight the general need for a curriculum focused on CIED interrogation for cardiothoracic anesthesiology trainees and fellows, a specific problem prompted the efforts of this project. A problem involving case delays of procedures on patients with CIEDs due to the show of device representatives and cardiology fellows and the occurrence of multiple preventable patient safety events due to perioperative CIED malfunctions triggered our institution to seek the initiation of an anesthesiologists-led CIED service. A dual-trained cardiothoracic and critical care anesthesiologist with expertise and training in CIED management (AZ) championed this quality and safety initiative. The implementation of an anesthesiologist-led CIED service created an in-parallel need to train anesthesiology trainees with a similar scope of practice to sustain the service. This need was based on an average number of 400 patients with CIED undergoing cardiac and noncardiac procedures at our institution per year. We reasoned that the rising use and exposure to CIED, the complexities of perioperative management of these devices, and the clear gap resulting from the absence of a standard methodology for educating cardiothoracic anesthesiology trainees all justify the design and delivery of a high-quality CIED curriculum for these trainees.

Determining and Prioritizing Content (Needs Assessment)

We conducted a needs assessment to assess the current status of CIED education to cardiothoracic anesthesiology trainees in our institution and nationwide and determine potential areas for improvement. Our needs assessment focused on 3 aspects. First, scanning the literature and anesthesiology program websites on published educational material on CIEDs for trainees. Second, we conducted a survey via Qualtrics to cardiothoracic anesthesiology trainees at our institution and at the Ochsner Clinic. The survey included questions about the level of training, the form of CIED didactics, the fallbacks of the current teaching, and what the trainees would like to see in a CIED curriculum. The third aspect of the needs assessment was in the form of in-person informal interviews with residents and cardiothoracic anesthesia trainees at our institution. Interviewees were asked open-ended questions about the current teaching of CIEDs at our institution and areas for improvement. A total of 6 interviews were conducted, 10 - 15 minutes each, with 2 trainees at a time. The interviews took place during breaks from working hours in the anesthesia lounge. The PI took notes and observed the reactions of the interviewees during the interviews. The PI then presented the results of the needs assessment to the Cardiothoracic Fellowship Program Director and the Residency

Program Director at our institution. The results of the needs assessment are summarized in Textbox 1.

Textbox 1.

The results of our needs assessment are summarized as follows:

- At the time of writing, there were 4 institutions that deliver formal cardiac implantable electronic device (CIED) didactics to cardiothoracic anesthesia trainees and fellows.
- Didactics are in the form of website videos (2 institutions) and scheduled lectures (3 institutions).
- Didactic materials were in the form of textbooks, review articles, and on-site workshops at national societal meetings on a yearly basis.
- Most of the didactics are directed to the simpler programming of temporary epicardial pacers that are placed intraoperatively. As a consequence, there is no formal interrogation of these devices [18].
- Trainees at our institution had not received instructional material about CIEDs before the interviews and survey dates.
- Lecture-based training was the predominant instructional approach trainees experienced and was also the least desirable.
- Approximately one-third of trainees rated app-based instruction as their first or second choice of instructional method (hands-on was the most desirable approach).
- Almost three-fourths of respondents indicated that they would prefer a self-paced e-learning system for learning about CIEDs.

The collective results of our needs assessment led us to reason a need to test the design, development, and delivery of a formal comprehensive CIED curriculum to our cardiothoracic anesthesia trainees that adapts to their time constraints.

Conceptual Framework

Conceptual frameworks represent ways of thinking about a problem and responding to the complexity of an educational phenomenon. Adding a conceptual framework to best practice approaches such as the Kern 6 step-approach for curriculum design helps construct goals and objectives. Furthermore, this integration helps select the educational intervention to achieve the educational goals and objectives and determine the evaluative methodology of the educational intervention, which will eventually determine the outcomes to be assessed and the evaluative strategy to determine the success of these outcomes [19]. Overall, both best practice and conceptual framework contribute to the rigor of educational scholarship.

Around 2 theories contributed to the design and development of the CIED curriculum, based on the needs assessment. The self-regulated framework, attributed to Zimmerman and Schunk, entails that learners plan, monitor, and evaluate their own learning to achieve their goals [20]. The self-regulated theory is applicable to the busy schedule of cardiothoracic anesthesiology trainees that interferes with synchronous learning, and importantly, to the maturity level of trainees at this stage of their career. The self-regulated theory thus calls for the design of an asynchronous learning method in the form of m-learning. The self-regulated framework was used to formulate the research question: whether an app-delivered CIED curriculum is a suitable delivery method for cardiothoracic anesthesiology trainees.

The Kolb experiential learning cycle that describes the 4-stage learning cycle in the form of concrete experience, reflective observation, abstract conceptualization, and active experimentation is another framework that influenced the design of the practical modules of the curriculum [21]. The framework guided the formulating of the research question: whether video recording of the interrogation and programming procedures of CIEDs by vendor will allow trainees to gain the skills needed to perform these procedures through Kolb stage of reflective observation?

Establishing Goals and Objectives

Based on a review of textbooks and consultation with colleagues, the first author developed an initial draft of seventeen major goals and associated instructional objectives (5 - 15 objectives for each goal). Overall, curricular goals focused on knowledge, psychomotor skills, and attitudes of the learners and were tailored in a stepwise fashion from acquiring foundational knowledge to applying this knowledge to acquiring the skills in managing CIEDs. For example, the curriculum progresses from understanding and recognizing electrocardiographic patterns of pertinent tachyarrhythmias and bradyarrhythmias to understanding the basics of CIED indications, structure, and function of CIEDs, to recognizing pacer rhythms and understanding the 4 basic parameters used in interrogating and programming CIEDs on different vendors. The final modules of the curriculum are in the form of recorded videos of the performance of device interrogation and programming on the 4 contemporary device vendors. Assessment of the psychomotor skills of the trainees will be in the form of a video recording of the trainee conducting the 4 steps of device programming and interrogations in the form of (1) identification of the vendor, (2) identification of the device, (3) assessment of pacing dependency, and (4) performance of battery and leads functionality in the form of battery longevity, leads impedances, and sensing amplitude and capture threshold. An assessment rubric was created to grade the performer's skill (Multimedia Appendix 1).

Following the development of 17 goals and objectives associated with such goals, we sought to evaluate the proposed goals and objectives by surveying a group of 17 CIED subject matter experts (SME) on whether (1) the objectives are critical or important to CA learners, (2) the instructional objectives fit the learning goals, and (3) whether the objectives as written were clear. After the initial development of the goals and objectives,

the process to evaluate the objectives was similar to methods used by a previous medical curriculum validation process used in obstetrics in Canada [22].

The curriculum validation process involved the identification of a national panel of experts in curricula relating to CIEDs. The definition of expertise of an expert was based on the qualifications: board certification in Cardiology, Electrophysiology, and equivalent certification in CIEDs with greater than 5 years of experience, and an expertise in medical curricular design.

Each panelist received an online survey (Qualtrics) which presented the 17 major learning goals of the curriculum and the instructional objectives for each goal (5 - 15 objectives per goal). Respondents were asked to rate the importance of each instructional objective (not important, important, and essential), the fit of the objective to the learning goal (does not fit and aligns), and the clarity of the objective (unclear or ambiguous and clear).

Ethical Considerations

The study was approved by the University of Alabama Institutional Review Board (IRB00012550) and adheres to the applicable CONSORT (Consolidated Standards of Reporting Trials) guidelines. The requirement for patient consent was waived.

Results

Response to Validation of Goals and Objectives

A total of 17 SMEs responded to the Qualtrics survey. The results of the validation survey are in the process of publication elsewhere. In brief, all objectives were rated as important or essential by 80% or more of the 17 raters, several goals had fewer than 70% of objectives rated as essential by 73% or more of the experts, with 6 goals having no objective that was rated as “essential” by more than 73% of experts. A goal of understanding the unique features of individual CIED vendors received the lowest rating given that this subject was not pertinent to cardiothoracic anesthesiologists.

In response to SMEs’ response, we have kept objectives that were either essential or important to greater than 75% of the raters and removed the objectives that were considered important by less than 75% of raters. Furthermore, goals whose objectives were considered important by less than 75% of the raters were removed.

The modified curriculum was then named “Anesthesiology Prospective-basic Curriculum” (Textbox 2).

Textbox 2. Basic cardiac implantable electronic device (CIED) curriculum: anesthesiology perspective

- Module 1: Basic cardiac electrophysiology
  - 1A. Basics of surface electrocardiogram
  - 1B. Heart blocks
  - 1C. Pertinent tachyarrhythmias
- Module 2: Indications of pacemakers, implantable cardioverter defibrillators (ICDs), and cardiac resynchronization therapies (CRTs)
  - 2A. Permanent pacemaker (PPM) indications
  - 2B. ICD indications
  - 2C. CRT indications
- Module 3: Anatomy of pacemakers
  - 3A. Generators
  - 3B. Leads
- Module 4: Pacemaker modes and codes
  - 4A. Generic pacemaker codes and modes
  - 4B. Ventricle paced, Ventricle sensed, Inhibited response (VVI) pacing as an example of single chamber pacing
  - 4C. Dual chamber paced, Dual chamber sensed, Dual response (DDD) pacing as an example of dual chamber pacing
- Module 5: Anatomy of ICDs
- Module 6: Operation of ICDs
  - 6A. ICD recognition of arrhythmias
  - 6B. ICD treatment of arrhythmias
- Module 7: Guidelines for perioperative management of CIEDs
- Module 8: Magnet behavior
- Module 9: Programming and interrogation of CIEDs
- Module 10: University of Alabama (UAB) Anesthesia Clinical Protocol of Perioperative CIED Management

Selection of Instructional Modality and Strategies

One focus of this curriculum development project was the commitment to delivering instructional content and resources, assessment, and program evaluation (eg, usage, acceptability, and satisfaction) of the curriculum via an updatable, accessible, asynchronous electronic platform. An m-learning platform for delivering the curriculum was considered essential for accommodating the constraints experienced by cardiothoracic anesthesiology trainees. Delivering the curriculum via modular delivery of content provides a flexible, self-paced, competency-focused approach to instruction and assessment.

In order to develop the m-learning platform within a structure that would be manageable to learners, the retained goals and objectives were distributed among 11 modules and submodules. The use of modules is a common structure for m-learning platforms [23,24]. Curricular mapping between goals, objectives, the instructional resources associated with each objective, and the assessment for each objective was created for all goals and objectives within the curriculum (Table 1).

Table . Instructional map for module 1A, 1B, and 1C

Module	Goals	Objective	Instruction	Assessment
1A	• Understand basics of heart rhythm formation and propagation • Establish a foundation for recognizing paced rhythms	• Differentiate between pacemaker and myocardial cells • Differentiate the action potential for both pacemaker and myocardial cells • Understand the ionic basis responsible for different phases of action potentials • Recognize and define refractory periods on an action potential	• Text • Differentiate between pacemaker cells and myocardial cells	Written quiz with visuals
1B	• Understand basics of electrocardiogram depiction • Understand the various types of heart block	• Determine heart rates on EKG by at least 2 methods • Identify the axis on surface EKG • Identify different types and levels of heart block on surface EKG	• Visuals • Text	Written quiz with visuals

1C	• Recognize the different types of supraventricular tachycardias
	• Understand the differences between ventricular and supraventricular tachycardias
		• Describe EKG patterns of SVT
		  • Determine electrocardiogram pattern of atrial fibrillation
		  • Describe electrocardiogram patterns of AVRNTa
		  • Describe electrocardiogram patterns of atrial tachycardias
		  • Describe electrocardiogram pattern of atrial flutter
		• Recognize EKG pattern of VT
		  • Monomorphic
		  • Polymorphic
		  • VF
		  • Accelerated idioventricular rhythm
		• Understand common SA node abnormalities
		• Describe wide complex SVT
		• Differentiate wide complex SVT from VT
			• Visuals
			• Text
				Written quiz with visuals

Assessment materials were developed in the form of pre and postinstructional quizzes. Postinstructional quizzes were slightly different in order and content from prequizzes. The rationale for this difference is to assess trainees' incremental learning gains after reading the modular instructional material and to assess whether the postquiz responses were due to knowledge acquisition versus memorization of prequiz grading of wrong answers. A minimum passing score of 80% was used as the criterion for passing any quiz. Quizzes consisted of both multiple-choice questions (MCQ) and open-ended questions. The former were created to assess knowledge, comprehension, and simple applications, while the latter were created to assess

a deeper level of learning and ability to explain the reasoning behind the responses [25]. Grading of open-ended questions was performed by the instructor, while grading for MCQs was performed automatically. Quizzes could be taken indefinitely until passing scores are achieved. Assessment for modules that covered skills was in the form of virtual actual device interrogations. This was in the form of performing 4 basic steps: identification of CIED vendor, determination of the type of CIED, determination of pacer dependency, and determination of device functionality in terms of battery longevity, leads' function in terms of sensing and capture thresholds, and leads' impedances. Programming entailed adjusting the device settings to suit the site of the surgical procedure and the use of electrosurgical interference, applying a magnet when indicated, and disabling tachycardia therapy in patients with implantable cardioverter defibrillators. Device restoration included restoring of basic settings of the device after the completion of the procedure and enabling tachycardia therapies for ICDs. The interrogations, programming, and restoration were performed under the supervision of the PI on the protocol.

The CIED curriculum underwent an instructional and content validity assessment as a final evaluation step. The instructional validity is aimed at engaging a cohort of SEM to judge whether the instructional content and resources within the app provide a valid representation of each instructional objective. The content validity assessment tested whether the pre-post quizzes were adequately informed by the instructional content and resources linked to the objective. The rationale for engaging SMEs on instructional and content validity is to consolidate the validation of the alignment of goals to objectives. Engagement was in the form of the Qualtrics survey sent to the same SMEs who were provided with the assessment and instructional and assessment material via email. Both the instructional and content validity assessment provided validation of the curricular map in the form of alignment of goals to objectives and to instructional material to assessment. This validation process is conceptually distinct from the assessment of learners' satisfaction of the curricular content, which is in the form of a short survey embedded at the end of each module. The validation process in this form is consistent with the Accreditation Council of Graduate Medical Education (ACGME) [26-28] (Table 2).

Table . Alignment of curricular contents with Accreditation Council of Graduate Medical Education core competencies.

ACGME[a] core competency	Corresponding curricular material to achieve competency
Patient care: "To demonstrate compassionate, appropriate, and effective care for patients, including the ability to perform comprehensive history and physical examinations, formulate diagnostic plans, and coordinate patient care."	History and physical examination of the patient to determine pertinent information about the CIED[b] (indications, type, vendor, previous interrogation reports).
Medical Knowledge: "To possess a strong foundation of biomedical knowledge and the ability to apply it to patient care, staying current with evolving medical knowledge."	Instructional material in the form of module narrative, constantly updated narratives in response to new guidelines and latest technology (leadless pacemakers and subcutaneous ICDs)[c], library of updated articles
Practice-based learning: "to investigate and evaluate their own care, appraise scientific evidence, and continuously improve their practice through self-evaluation and lifelong learning."	Unlimited attempts at solving the quizzes, interactive platform between the user study app and the admin app to allow interactive feedback from instructor to learner on performance
Interpersonal and communication skills: "to effectively communicate with patients, families, and other healthcare professionals, building strong relationships and fostering trust."	Adequate documentation of the interrogation and programming procedure, communication with surgical and EP services on device malfunction issues in need of intervention.
Professionalism: "To demonstrate ethical conduct, respect for colleagues, and a commitment to patient well-being, upholding the highest standards of professional behavior."	Competency adequately emphasized throughout the instructional material and skill assessment of the curriculum.
Systems-based practice: "To understand and navigate healthcare systems, working effectively within teams and advocating for patients' needs within the context of the healthcare system."	Detailed instruction on navigating EMRs[d], checklists on consulting other services and contacting device reps for vendor-specific issues.

[a]ACGME: Accreditation Council of Graduate Medical Education.

[b]CIED: cardiac implantable electronic devices.

[c]ICD: implantable cardioverters defibrillators.

[d]EMR: electronic medical records.

Delivery of the Curriculum

Pilot Phase

We developed a de novo cloud-based system for the project. The CIED app (and the accompanying admin grading app) has 2 components: the mobile and web front-end, and the cloud-based back-end (Figure 1).

Figure 1. Screenshots of the cardiac implantable electronic device app.

Front-end

The CIED app and the admin grading app are developed using Ionic (OutSystems), a cross-platform mobile app framework. This allows consistent functionality and user experience across iOS and Android devices. It also allows deploying a web app, which can be accessed from mobile phones or tablets as well as desktop web browsers.

Back-end

The back-end of this system is built using the Google Cloud. More specifically, we host the back-end computing, database, and analytics services based on the Google Firebase engine. We used the Google Cloud Firestore database to store the course modules, quiz questions, and per-user data such as grades. We used Firebase's authentication engine to register and manage users.

Analytics

For detailed user analytics, we used Google Firebase Analytics engine. It provided us with details on how the users interacted with the system and viewed different modules and videos. The per-user data on grades and quiz attempts provides us with information on the quiz attempts and performance. We collected pre and postlesson quizzes to compare performance before and after each lesson.

The evaluation of the educational material will occur in three phases. First, performance analytics will be performed, which will quantify the learning gain after compared to before taking the quizzes. Second, web analytics will determine the time taken by trainees in studying and answering the quizzes. Through this evaluation, the level of complexity of the material will be evaluated. Third, usability analytics will assess the learners' satisfaction with the curricular content and app usability.

A total of 2 distinct applications were created; the User Study App, empowering cardiothoracic anesthesiology trainees to enhance their knowledge, and the Admin Grading App, streamlining quiz management and assessment. The successful integration of these apps creates a cohesive ecosystem that fosters growth, engagement, and continuous improvement within the anesthesiology community. Both apps are compatible with Android and iOS platforms.

User Study App

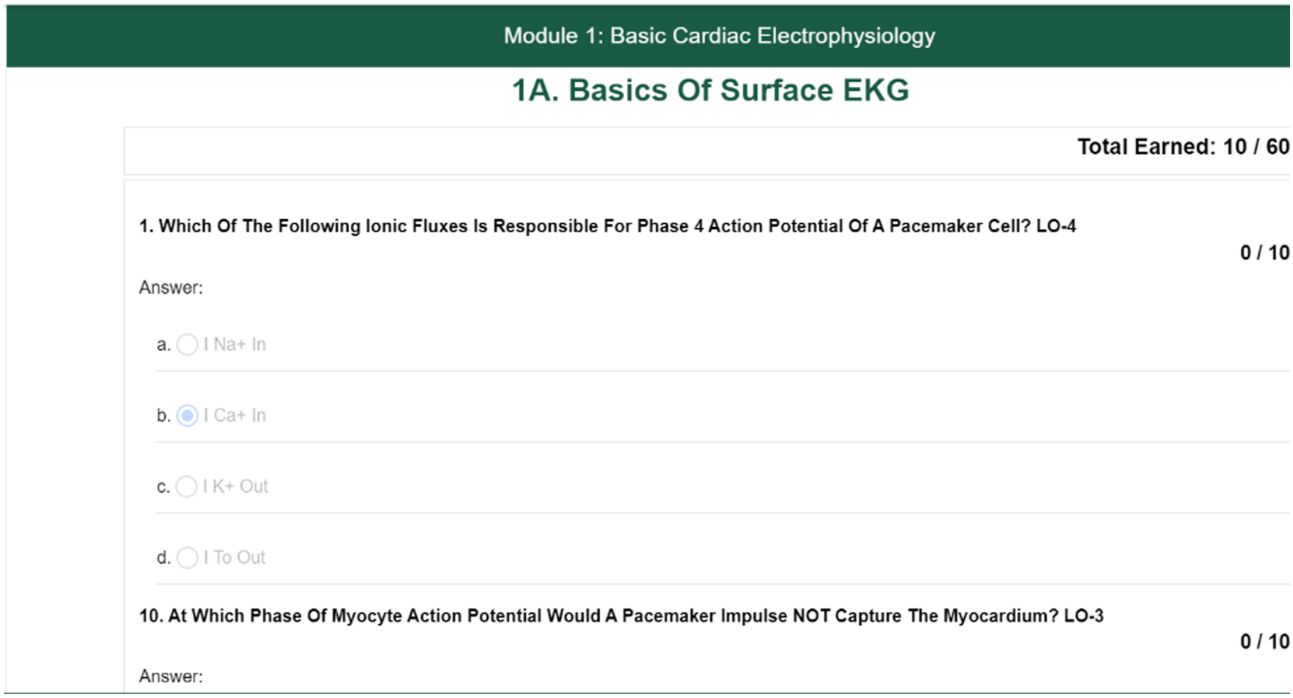
This app is a knowledge hub catering to anesthesiology practitioners' learning needs. It offers a collection of study materials, interactive resources, and self-assessment tools to foster professional growth. For usability testing, we are using the AttrakDiff test (https://www.attrakdiff.de/index-en.html). AttrakDiff is a standardized test designed to evaluate the user's experience. It is widely used to assess the attractiveness of a product in terms of usability and appearance.

The User Study App is integrated with the prequiz condition (Figure 2).

Prequiz Requirement

Users must complete a prequiz assessment before gaining access to study materials.

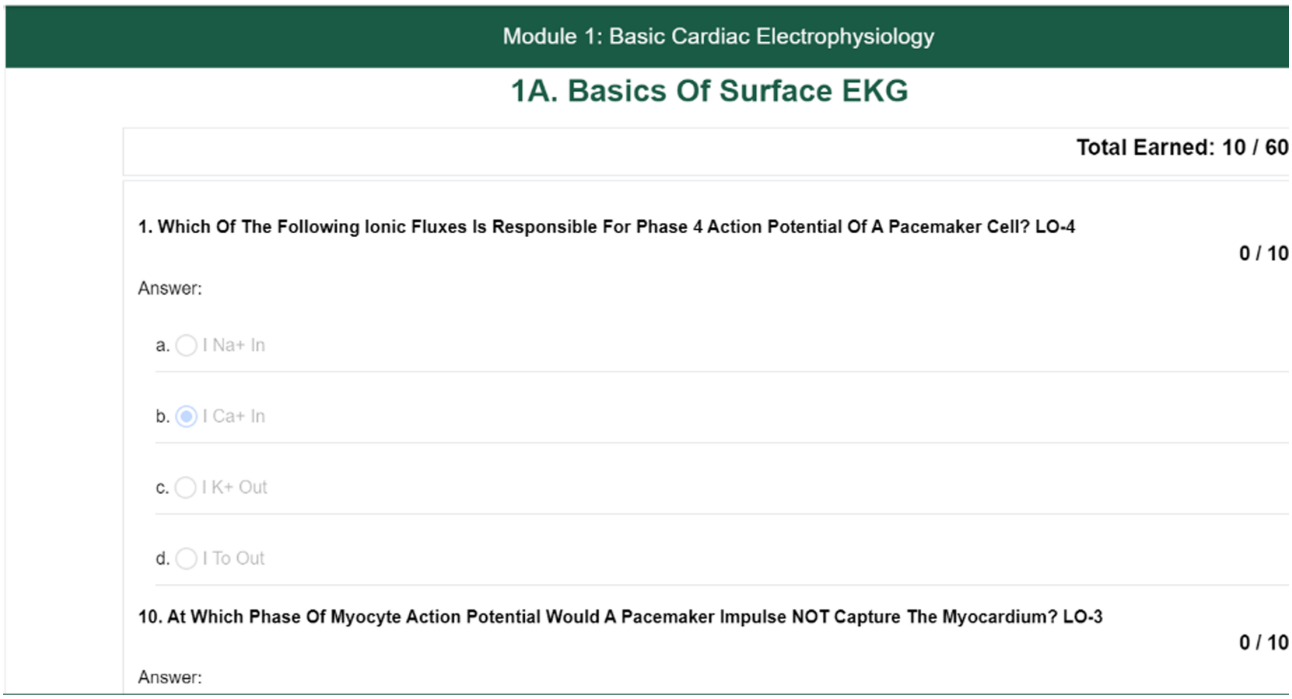
Figure 2. Screenshot of a modular instruction material-user app.



Study Materials

This includes a library of articles, multimedia content, and case studies tailored to various anesthesiology topics and expertise levels (Figure 3).

Figure 3. Screenshot of a quiz on a modular instruction material.



Interactive Quizzes

Quizzes were designed to complement the study materials and reinforce learning objectives.

Progress Tracking

Personalized user dashboards displaying prequiz scores and study progress, aiding in self-assessment and goal setting.

Learners' Satisfaction Surveys

Surveys on the learning experience of trainees were embedded at the end of each module. The surveys assess learners' views on the clarity of the educational material, alignment of the educational objectives to goals and to the instructional and assessment material, what areas need further instruction and what technical issues arose during the navigating the module (Figure 4).

Figure 4. Screenshot of the Qualtrics survey embedded at the end of each instructional module.

Admin Grading App

The Admin Grading App equips administrators with powerful tools to manage quiz content and evaluate user performance. Key features are given below.

Admin Dashboard

This is a secure login portal providing access to quiz management, user profiles, and analytics.

Quiz Creation

This is a user-friendly interface allowing administrators to create and customize quizzes, tailoring them to specific learning objectives.

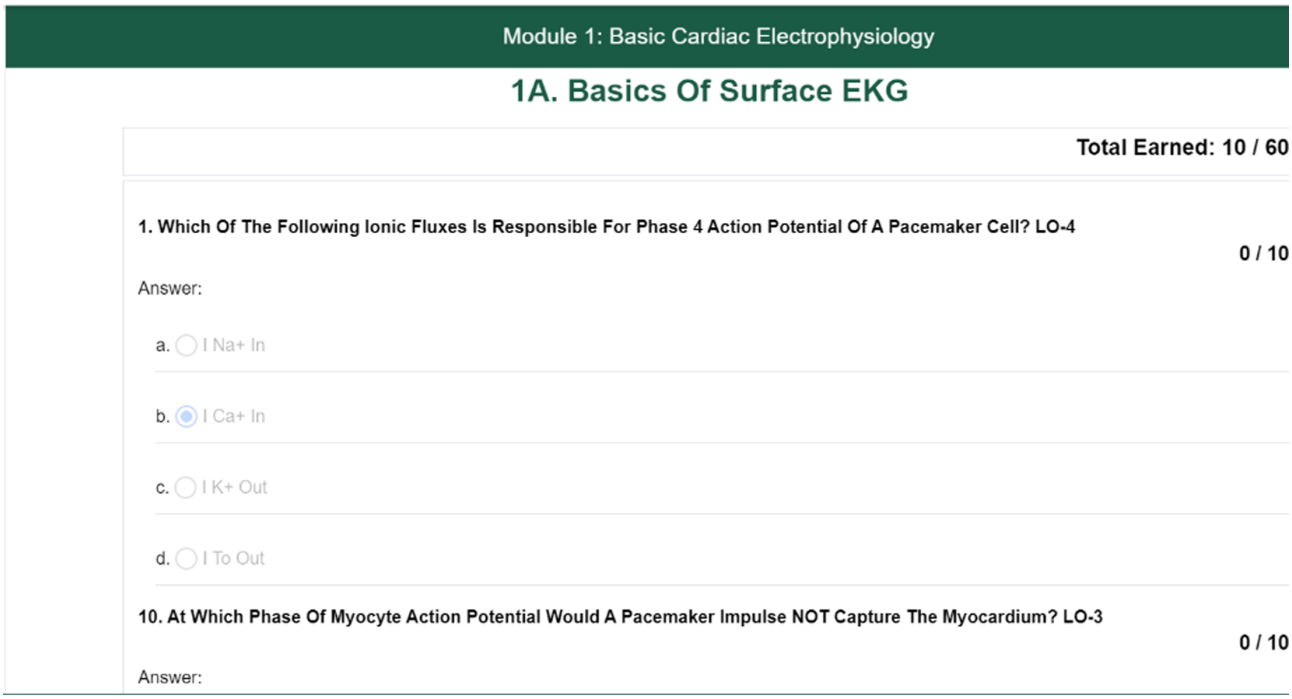
User Management

This includes comprehensive user profiles and progress reports, assisting administrators in understanding individual learning trajectories.

Grading Interface

This was a grading system enabling assessment of quiz responses (Figure 5).

Figure 5. A screenshot of grading of a quiz-administrative app.



Analytics and Reporting

Data visualization tools providing insights into overall user performance, popular study topics, and areas needing further emphasis.

Integration and Security Measures: Fostering a Cohesive Ecosystem

The integration of the User Study and Admin Grading Apps is facilitated through shared databases, ensuring data synchronization and consistency. Security measures, including strong encryption and authentication protocols, safeguard sensitive user information and preserve the platform's integrity.

Discussion

Principal Findings

This manuscript describes the process of designing, developing, and app delivering a comprehensive CIED curriculum for cardiothoracic anesthesiology trainees. This project has engaged an interdisciplinary team of cardiologists, electrophysiologists, technologists, computer scientists, and medical education. With the delivery of the curriculum, the investigators are in the process of collecting data on the platform implementation; user perceptions of usability, perceived effectiveness in promoting learning, and satisfaction with all components of the app; psychometrics of the learning assessments, and analysis of learning within each module.

While there are no previous studies on this subject to compare our work to, our mobile app goes in line with other studies that used mobile apps for educating anesthesiology trainees. Marty et al [29] designed a mobile app to facilitate programmatic assessment of anesthesiology trainees and tested its performance in five teaching institutions. The app provided insightful information about feedback completion times and documentation of learning goals. Monroe et al [30] implemented a mobile app for anesthesiology trainees rotating through a pediatric rotation in an academic institution. Similarly, Herbstreit et al [31] created a mobile platform formed of several department-specific apps and qualitatively demonstrated high satisfaction rates with the mobile learning content and pace. Compared with the above-mentioned studies, we have followed formal steps of creating a curriculum, consulted nationwide SMEs, and developed a curricular map that aligns instructional material

and assessment to goals and objectives. Distinctively, our admin app uses web, performance, and usability analytics, which are either not covered or partially covered in other studies.

The project represents a collaborative effort between multiple teams to enhance trainees' educational experience. As such, it extends the boundaries of cardiac anesthesiology, adding the new dimension of electrophysiology. Furthermore, this project represents a remarkable link between quality improvement, patient safety, and medical education. The effective implementation of our project is in line with similar improved outcomes with enhanced multidisciplinary team buildings and dynamics [32-34]. On the other hand, the program encountered some challenges before its successful implementation. First, the diversity in the learners' training levels, including second-year rotating residents and cardiothoracic fellows, posed a challenge in the selection of instructional material. Second, the process of curricular instructional and assessment validation was more of a laborious process for SMEs due to the lengthiness of the material. Third, the choice of the instructional material was more of a challenge for basic learners given the density of the subject. This has led us to a multiple iterative process to adjust the curriculum to conform to the needs of a basic anesthesiology learner.

It is important to recognize some of the limitations of our project. At the time of writing of this manuscript, the project remains descriptive. We are currently collecting data on the learning gain and usability of the app. Therefore, the process of curricular refinement remains an iterative process. There is relatively a few number of SMEs recruited for the validation process of the curriculum. At the time of this recruitment, the curriculum remains in the initial stages and is subject to more iterations when more feedback is received from trainees. At such a point, more SMEs will be recruited on further refinements of the curriculum. Ideally, validation of curricular mapping should occur in a single step. In our case, given the lengthiness of the content, the process of curricular validation proceeded through validation of alignment of learning goals and objectives, followed by the instructional content and assessment validation. Furthermore, our SMEs were not confined to those with backgrounds in anesthesiology. While this may be regarded as a limitation, the diversity of SMEs' backgrounds allows more robustness of the curriculum.

Conclusions

The CIED app is a novel application-delivered curriculum on a complex and understudied subject in anesthesiology training. The curriculum adheres to conventional steps of curricular design. The m-delivery of the curriculum and the functionality of the app in the evaluation of the learners' performance and satisfaction responds well to the challenges that face trainees in achieving synchronous learning and engages them in the learning and evaluation processes. As we continue to enhance and expand the CIED app, we envision its instrumental role in elevating the quality of anesthesiology practice worldwide, ultimately benefiting patient outcomes and health care excellence.

Acknowledgments

This work was supported by the Foundation of Anesthesia for Education and

Research REG-02-15-2021 (AZ). Funding mechanism played no role in any of the design, data synthesis, collection, analysis, writing, or choice of submission of this work.

Authors' Contributions

AZ: conception of the manuscript, writing of the curriculum, performing the need assessment, writing the initial draft, edited the final draft,
AW: conception of the app, edited and reviewed the final version of the manuscript.
BH: evaluated the curriculum, edited the final draft of the manuscript.
SM:conception, design and development of the app, editing and review of the final draft of the manuscript
SHM: conception, design and development of the app, editing and review of the final draft of the manuscript
RH: conception, design and development of the app, editing and review of the final draft of the manuscript
YZ: conception, design and development of the app, editing and review of the final draft of the manuscript.
All authors agree on the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Rubric for assessment of psychomotor skills during device interrogation.
[DOCX File, 12 KB - mededu_v11i1e60087_app1.docx]

References

1. Zhang S, Gaiser S, Kolominsky-Rabas PL. Cardiac implant registries 2006-2016: a systematic review and summary of global experiences. BMJ Open 2018 Apr 12;8(4):e019039. [doi: 10.1136/bmjopen-2017-019039] [Medline: 29654008]

2. Cardiac implantable electronic device management [corrected]. Anesthesiology 2020 Feb;132(2):225-252. [doi: 10.1097/ALN.0000000000002821] [Medline: 31939838]
3. Silva KD, Albertini CDM, Crevelari ES, et al. Complications after surgical procedures in patients with cardiac implantable electronic devices: results of a prospective registry. Arq Bras Cardiol 2016 Sep;107(3):245-256. [doi: 10.5935/abc.20160129] [Medline: 27579544]
4. Thorpe RL, Rohant N, Cryer M, et al. Inappropriately firing defibrillator: a simulation case for emergency medicine residents. MedEdPORTAL 2019 Feb 27;15:10808. [doi: 10.15766/mep_2374-8265.10808] [Medline: 30931387]
5. Zaky A, Melvin RL, Benz D, et al. Economic evaluation of anesthesiology-led cardiac implantable electronic device service. Healthcare (Basel) 2023 Jun 27;11(13):13. [doi: 10.3390/healthcare11131864] [Medline: 37444698]
6. Chaudhuri JD. Stimulating intrinsic motivation in millennial students: a new aeneration, a new approach. Anat Sci Educ 2020 Mar;13(2):250-271. [doi: 10.1002/ase.1884] [Medline: 31021529]
7. Dunleavy G, Nikolaou CK, Nifakos S, et al. Mobile digital education for health professions: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019 Feb 12;21(2):e12937. [doi: 10.2196/12937] [Medline: 30747711]
8. Chandran VP, Balakrishnan A, Rashid M, et al. Mobile applications in medical education: a systematic review and meta-analysis. PLoS ONE 2022;17(3):e0265927. [doi: 10.1371/journal.pone.0265927] [Medline: 35324994]
9. Chase TJG, Julius A, Chandan JS, et al. Mobile learning in medicine: an evaluation of attitudes and behaviours of medical students. BMC Med Educ 2018 Jun 27;18(1):152. [doi: 10.1186/s12909-018-1264-5] [Medline: 29945579]
10. Rodríguez-Ríos A, Espinoza-Téllez G, Martínez-Ezquerro JD, et al. Information and communication technology, mobile devices, and medical education. J Med Syst 2020 Mar 16;44(4):90. [doi: 10.1007/s10916-020-01559-w] [Medline: 32173765]
11. Hughes MT, Kern DE, Thomas PA. Curriculum Development for Medical Education: A Six-Step Approach: Johns Hopkins University Press; 2009:33-58 URL: https://scholar.google.com/scholar?hl=en&as_sdt=0%2C1&q=Hughes+MT%2C+Kern+DE%2C+Thomas+PA.+Curriculum+Development+for+Medical+Education%3A+A+Six-Step+Approach.+Johns+Hopkins+University+Press%3B+2009%3A33-58.+ISBN%3A+9780801893660&btnG= [accessed 2025-07-17]
12. Sweet LR, Palazzi DL. Application of Kern's Six-step approach to curriculum development by global health residents. Educ Health (Abingdon) 2015;28(2):138-141. [doi: 10.4103/1357-6283.170124] [Medline: 26609014]
13. Thomas PA, et al. Curriculum Development for Medical Education: A Six-Step Approach: JHU press; 2022. URL: https://www.press.jhu.edu/books/title/12470/curriculum-development-medical-education?srsltid=AfmBOoouBZS5m93nw4fR45Fq1r-3tX1v6M7_L0noxPxS084Lr6G-l-ii [accessed 2025-07-17] [Medline: 25899812]
14. Khamis NN, Satava RM, Alnassar SA, et al. A stepwise model for simulation-based curriculum development for clinical skills, a modification of the six-step approach. Surg Endosc 2016 Jan;30(1):279-287. [doi: 10.1007/s00464-015-4206-x] [Medline: 25899812]
15. Rooke GA, Bowdle TA. Perioperative management of pacemakers and implantable cardioverter defibrillators: it's not just about the magnet. Anesth Analg 2013 Aug;117(2):292-294. [doi: 10.1213/ANE.0b013e31829799f3] [Medline: 23881371]
16. Rooke GA, Lombaard SA, Van Norman GA, et al. Initial experience of an anesthesiology-based service for perioperative management of pacemakers and implantable cardioverter defibrillators. Anesthesiology 2015 Nov;123(5):1024-1032. [doi: 10.1097/ALN.0000000000000838] [Medline: 26352380]
17. Zaky A, Beck A, Rooke GA. Hemodynamically significant heart block after carotid artery stenting in a patient with atrial demand pacer-echocardiography-guided rescue pacing. J Cardiothorac Vasc Anesth 2020 Jan;34(1):187-191. [doi: 10.1053/j.jvca.2019.08.027] [Medline: 31526556]
18. Crowe ME, Hayes CT, Hassan ZU. Using software-based simulation for resident physician training in the management of temporary pacemakers. Simul Healthc 2013 Apr;8(2):109-113. [doi: 10.1097/SIH.0b013e31826ec3e1] [Medline: 23086515]
19. Zackoff MW, Real FJ, Abramson EL, et al. Enhancing educational scholarship through conceptual frameworks: a challenge and roadmap for medical educators. Acad Pediatr 2019 Mar;19(2):135-141. [doi: 10.1016/j.acap.2018.08.003] [Medline: 30138745]
20. Schunk DH, Zimmerman BJ. Self-Regulated Learning: From Teaching to Self-Reflective Practice: Guilford Press; 1998. URL: https://psycnet.apa.org/record/1998-07519-000 [accessed 2025-07-17]
21. Kolb DA. Experiential Learning: Experience as the Source of Learning and Development: FT press; 2014. URL: https://scholar.google.com/scholar?hl=en&as_sdt=0%2C1&q=Experiential+Learning%3A+Experience+as+the+Source+of+Learning+and+Development&btnG= [accessed 2025-07-17]
22. Cumyn A, Harris IB. A comprehensive process of content validation of curriculum consensus guidelines for a medical specialty. Med Teach 2012;34(8):e566-e572. [doi: 10.3109/0142159X.2012.668623] [Medline: 22489987]
23. Gagnon J, Gagnon MP, Buteau RA, et al. Adaptation and evaluation of online self-learning modules to teach critical appraisal and evidence-based practice in nursing: an international collaboration. Comput Inform Nurs 2015 Jul;33(7):285-294. [doi: 10.1097/CIN.0000000000000156] [Medline: 25978538]

24. Foster MJ, Shurtz S, Pepper C. Evaluation of best practices in the design of online evidence-based practice instructional modules. J Med Libr Assoc 2014 Jan;102(1):31-40. [doi: 10.3163/1536-5050.102.1.007] [Medline: 24415917]
25. Yang BW, Razo J, Persky AM. Using testing as a learning tool. Am J Pharm Educ 2019 Nov;83(9):7324. [doi: 10.5688/ajpe7324] [Medline: 31871352]
26. Swing SR. The accreditation council for graduate medical education's outcome project and its effects on graduate medical education in anesthesia. Adv Anesth 2005 Jan;23:15-39. [doi: 10.1016/j.aan.2005.06.002]
27. Hawkins RE, Lipner RS, Ham HP, et al. American board of medical specialties maintenance of certification: theory and evidence regarding the current framework. J Contin Educ Health Prof 2013;33 Suppl 1(Suppl. 1):S7-19. [doi: 10.1002/chp.21201] [Medline: 24347156]
28. Epstein RM, Hundert EM. Defining and assessing professional competence. JAMA 2002 Jan 9;287(2):226-235. [doi: 10.1001/jama.287.2.226] [Medline: 11779266]
29. Marty AP, Braun J, Schick C, et al. A mobile application to facilitate implementation of programmatic assessment in anaesthesia training. Br J Anaesth 2022 Jun;128(6):990-996. [doi: 10.1016/j.bja.2022.02.038] [Medline: 35410792]
30. Monroe KS, Evans MA, Mukkamala SG, et al. Moving anesthesiology educational resources to the point of care: experience with a pediatric anesthesia mobile app. Korean J Anesthesiol 2018 Jun;71(3):192-200. [doi: 10.4097/kja.d.18.00014] [Medline: 29739184]
31. Herbstreit S, Herbstreit F, Diehl A, et al. A novel mobile platform enhances motivation and satisfaction of academic teachers. J Eur CME 2021;10(1):2014100. [doi: 10.1080/21614083.2021.2014100] [Medline: 34925966]
32. Van Der Vegt GS, Bunderson JS. Learning and performance in multidisciplinary teams: the importance of collective team identification. AMJ 2005 Jun;48(3):532-547. [doi: 10.5465/amj.2005.17407918]
33. Hilton ML, Cooke NJ. Enhancing the Effectiveness of Team Science 2015. URL: https://nap.nationalacademies.org/catalog/19007/enhancing-the-effectiveness-of-team-science [accessed 2025-07-17] [Medline: 26247083]
34. Jackson SE. The consequences of diversity in multidisciplinary work teams. In: Handbook of Work Group Psychology 1996:53-75 URL: https://scholar.google.com/scholar?hl=en&as_sdt=0%2C1&q=The+consequences+of+diversity+in+multidisciplinary+work+teams.+In%3A+Handbook+of+Work+Group+Psychology&btnG= [accessed 2025-07-17]

Abbreviations

ACGME: Accredited Council of Graduate Medical Education
AVRNT: AV nodal reentrant tachycardia
CIED: cardiac implantable electronic devices
CRT: cardiac resynchronization therapy
ICD: implantable cardioverters defibrillators
m-learning: mobile learning
SME: subject matter experts
SVT: supraventricular tachycardia
VT: ventricular tachycardia

Edited by B Lesselroth; submitted 04.05.24; peer-reviewed by P Bertini, Z Dimassi; revised version received 17.02.25; accepted 19.04.25; published 29.07.25.

Please cite as:
Zaky A, Waheed A, Hatter B, Malempati S, Maremalla SH, Hasan R, Zheng Y, Snyder S
Cardiac Implantable Electronic Device Educational Application for Cardiac Anesthesiology Trainees: Tutorial on App Development
JMIR Med Educ 2025;11:e60087
URL: https://mededu.jmir.org/2025/1/e60087
doi:10.2196/60087

Enhancing Access to Neuraxial Ultrasound Phantoms for Medical Education of Pediatric Anesthesia Trainees: Tutorial

Leah Webb[1*], MD; Melissa Masaracchia[2*], MD; Kim Strupp[1*], MD

[1]Division of Pediatric Anesthesiology, Department of Anesthesiology, Children's Hospital Colorado, University of Colorado, Denver, CO, United States

[2]Northwell, Division of Pediatric Anesthesiology, Department of Anesthesiology, Cohen Children's Medical Center, Zucker School of Medicine at Hofstra/Northwell, New Hyde Park, NY, United States

[*]all authors contributed equally

Corresponding Author:
Leah Webb, MD
Division of Pediatric Anesthesiology, Department of Anesthesiology, Children's Hospital Colorado, University of Colorado, Denver, CO, United States

Abstract

Opportunities to learn ultrasound-guided/assisted (USGA) neuraxial techniques for pediatric patients are limited, given the inherent high stakes and small margin of error in this population. Simulation is especially valuable in pediatrics because it enhances competency and efficiency, without added risk, when learning new skills, specifically those seen with ultrasound-guided regional anesthetic techniques. However, access to simulation opportunities involving the use of phantom models in medical education is limited due to excessive costs. We describe a process for producing ultrasound phantoms by using synthetic ballistic gelatin; these ultrasound phantoms can be used for simulation and are affordable, reproducible, and indefinitely shelf stable. The ultrasound images produced by these phantoms are comparable to those obtained from a real pediatric patient, including the sacral anatomy necessary for caudal epidural blocks, as validated by practicing pediatric anesthesiologists. Phantom models offer a more cost-effective alternative to commercially prepared phantoms, thereby expanding access to realistic simulations for neuraxial ultrasound in pediatric medical education, without the prohibitively high expense.

(JMIR Med Educ 2025;11:e63682) doi:10.2196/63682

KEYWORDS

anesthesiology; pediatric; ultrasound; education; neuraxial ultrasound; medical education; pediatric anesthesia trainees; anesthesia; trainees; ultrasound-guided; neuraxial techniques; pediatric patients; efficiency

Introduction

Opportunities to learn ultrasound-guided/assisted (USGA) neuraxial techniques for pediatric patients are limited, given the inherent high stakes and small margin of error in this population. Simulation is a valuable and effective method for learners—whether used by trainees or experienced clinicians—to enhance their competency, efficiency, and confidence in performing regional anesthetic and neuraxial techniques [1-4]. Ultrasound enhances safety, decreases complications, and improves the efficacy and accuracy of neuraxial blockade in pediatric patients from preterm to adolescence [5-12]. The utility of ultrasound is even more apparent in syndromic children with unusual anatomy, patients who comprise a large subset of the pediatric population that presents for surgery at a young age [13]. Honing pediatric patient–related ultrasound skills in a simulation setting is an ideal scenario for learning without risk. Unfortunately, educational curricula and teaching models lag behind recent advancements in simulation.

Despite efforts to create affordable and reproducible ultrasound phantoms, many lack a realistic appearance, and most are not indefinitely, if at all, shelf stable or portable because they are made of water, agar, gelatin, or other substances or are derived porcine models [3,4,14,15]. The cost of manufactured models that offer all these features can be prohibitively expensive, amounting to several thousand dollars, and these models may generate an inferior simulation experience [14]. Limited access to high-fidelity ultrasound phantoms significantly restricts opportunities for learners to take advantage of low-stakes simulation training and necessitates practicing on live patients, including infants and children, to learn valuable skills—a method with varying degrees of success and much higher stakes. There is a dearth of literature describing spine phantoms that are made with the necessary anatomy to teach pediatric trainees how to use ultrasound to approach the caudal epidural space. We describe a method for creating a realistic, affordable, reproducible, and shelf-stable spine phantom model that allows for the demonstration of key ultrasound images of the spine and caudal anatomy that are required to perform USGA neuraxial techniques on pediatric patients. Furthermore, the synthetic ballistic gelatin used to produce the phantom model can be reclaimed and reused to make “fresh” models for an indefinite

period of time, allowing for multiple practice sessions without additional costs.

Methods

Overview

We present a tutorial describing the construction of an ultrasound phantom of the spine, based on similar previous descriptions [16]. Notably however, our model includes both lumbar anatomy and sacral anatomy, which are lacking in previously published iterations but are essential for learning pediatric-specific neuraxial sonoanatomy. Additionally, we completely submerged our spine model in ballistics gel, creating stable, flat surfaces surrounding the spine to facilitate scanning the model in multiple orientations, which simulates the use of ultrasound for prone, lateral, and sitting positions. Further, an anonymous survey was sent to 10 practicing attending pediatric anesthesiologists to evaluate the similarity between the ultrasound images generated from the phantom and images from a real patient. Three ultrasound views were evaluated for likeness and accuracy on a 5-point Likert scale.

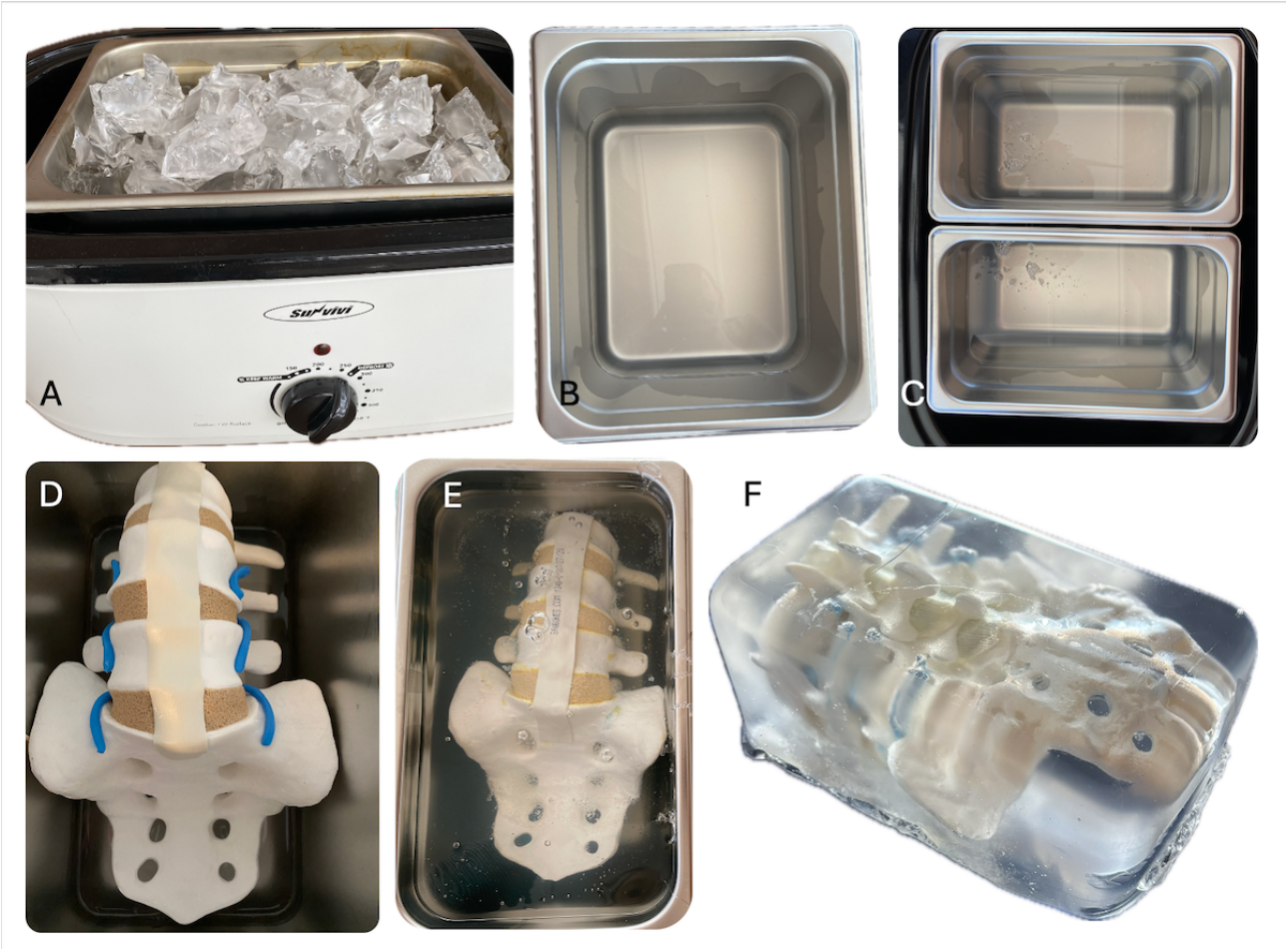
How to Create a Phantom Model

The following stepwise process can be used to create a spine phantom:

- Step 1: Preheat a portable oven to 250-270 °F (121-132 °C). A portable oven is preferable, as it can be used outside to prevent inhalation of the unpleasant smell from melting gel. It is critical to review the manufacturer's guidelines; ensure that the ballistics gel is always melted in well-ventilated areas; and ensure that caution is used to avoid overheating the gel, as it could light on fire.
- Step 2: Cut or tear ballistics gel into smaller pieces for melting.
- Step 3: Place the gel into an oven-safe pan (either a mold pan or an extra container), with the goal of melting the gel to create a 1- to 3-inch gel layer at the bottom of the pan.
  - This layer mimics the soft tissue covering the spinous processes. Add more gel to create a thicker layer, if desiring to create a model with greater depth to the epidural space.
  - Generally, it is preferable to melt the gel in an extra oven-safe container and pour it into a mold pan for each subsequent step; however, for the first layer, the gel can preferentially be melted directly in the mold pan.
- Step 4: Melt the gel in the oven until all bubbles are gone. This step takes about 2 to 4 hours, depending upon the amount of gel melted. It is critical to minimize bubbles in this layer, as this is the surface that will be scanned with ultrasound.
- Step 5: Allow the bubble-free layer to cool significantly (approximately 30-60 min). Then, place the spine model into the pan, with spinous processes facing down toward the bottom of pan and touching the gel, and press it very gently into the gel. Hold or secure the spine model in place until the gel sets and the model is not moving in the pan (10-20 min).
  - During the first cooling period, the gel should be cool enough to touch and be starting to firm up, with some resistance to pressure from a fingertip, but it should be soft enough to envelop the tips of the spine model's spinous processes.
  - If free-pouring gel into a mold pan (rather than melting gel directly in it, as is preferable), aim to pour toward one side/corner of the pan to minimize air bubbles. A ladle can be used to pour gel into the mold pan, again aiming to pour into one side/corner of the pan (rather than fanning) to minimize bubbles.
  - If many large bubbles are present, consider placing the mold pan back into the oven and cooking further until bubbles are gone (a few small bubbles are generally not problematic).
  - Hot gel will soften the spine model and result in curved spinous processes. Therefore, press only hard enough on the model to make slight contact between the tips of spinous processes and the gel; this contact secures the model to hold it in place in future steps.
- Step 6: Allow gel in mold pan (containing 1- to 3-inch gel layer) and spine model to cool completely.
- Step 7: Once cool, pour more melted gel (see step 3 for melting instructions) into the mold pan until the spine model is completely covered.
  - Pour quickly and to one corner or side at the coccyx-end of the model.
  - Bubbles in this step are not as concerning because this surface will be placed facing down and will not be scanned.
  - Bubbles will continue to rise to the top for several minutes as the gel cools; large bubbles can be popped/opened to create a flatter surface on this side of the phantom, though it is not necessary to do so.
- Step 8: Allow the mold pan (which should now contain the gel-covered spine model) to cool completely, preferably overnight, until the gel is solid.
- Step 9: Remove phantom from pan, using firm but gentle traction on the gel.
  - It may help to run an offset spatula (or another flat, thin tool, such as a butter knife) along the edges of the phantom and pan to help separate the phantom from the mold pan.
  - Once loosened, it can be helpful to stand the pan upright on the short side and slide fingers between the gel and pan as deep as possible to fully free the top side of the gel. Then, firmly push down, while continuously pulling out, on the gel until it releases from the pan.
- Step 10: Store phantom at room temperature, with spinous process side up. To clean, use water and a lint-free towel.

Figure 1 shows correlated pictures of the stepwise process and final phantom model.

Figure 1. Stages of phantom production, with A to C showing steps 1 to 4, D showing steps 5 and 6, E showing steps 7 and 8, and F showing step 9. (A): Cut gel in an extra container placed inside a portable oven set to 270 °F (132 °C). (B): Melted, bubble-free gel in the extra container. (C): Two mold pans. (D): Spine model, with anterior side up, placed in cooled, bubble-free layer. (E): Spine model submerged completely in gel and cooled. (F): Completed spine phantom that has cooled completely and has been removed from the mold pan.



Materials

Multiple options exist for the materials that are used to create a phantom model for practicing neuraxial ultrasound skills. Table 1 outlines those used by the authors, along with purchase sites and prices. Each phantom is composed of a spine model embedded in ballistics gel. Additional necessary items are reusable for multiple production cycles. Supplies include an oven that can sustain 250-270 °F (121-132 °C; US $119 for a portable oven), an oven-safe mold (US $9), and an extra oven-safe container (US $11). Optional items include a ladle or another heatproof tool for scooping melted gel. The ladle can be useful for more precision in transferring the gel into the pan. It does potentially create more bubbles than pouring directly; however, bubbles are mitigated by placing the pan back into the oven. The ladle is also useful for ensuring that the anterior side of the model is completely covered with gel and that any bubbles remaining on the anterior side do not interfere with ultrasound scanning. Furthermore, because the spine model is completely submerged in ballistics gel, an offset spatula may be helpful for loosening and releasing the phantom from its mold.

Table . List of materials, where to purchase, costs, and notes on pertinent information.

Item and description		Purchase site	Cost	Notes
Oven (portable)				
	"Sunvivi 22-Quart Roaster Oven"	Amazon.com (ASIN[a]: B07K25WBZ4)	US $119	Any oven that can sustain 250 - 270 °F (121-132 °C).
Ballistics gel				
	"10% FBI Gel Block"	Clearballistics.com (SKU[b]: 852844007000)	US $76 + shipping	Makes ≥4 phantoms.
Spine model				
	"Spine, Lumbar Vertebrae with Nerve Roots and Ligamenta Flava, L3-Sacrum, Solid Foam"	Sawbones.com (SKU: 1340-1)	US $161 + shipping	Preferred.
	"Medical Human Lumbar Spine Demonstration Model Anatomical Model Lumbar Vertebrae Sacrum & Coccyx, with Herniation Disc,for Science Classroom Study Display Teaching Medical Model 15 Inch Hight"	Amazon.com (ASIN: B074JCS4SC)	US $34	Less expensive. Requires removal of some vertebrae to fit recommended oven-safe mold. Alternative option is 3D printed model.
Oven-safe mold				
	"1/4 size 6" Deep Steam Table Pan"	Webstaurantstore.com (item number: 4070469	US $9	Any oven-safe receptacle that is similar in size to spine model. Can be purchased from local restaurant supply store.
Extra oven-safe container				
	"1/2 Size 6" Deep Steam Table Pan"	Webstaurantstore.com (item number: 4070269)	US $11	Used to melt bigger volume of gel.
Gel dye				
	"Tone dye"	Humimic Medical [17]	US $35	Used to opacify gel; comes in a variety of skin tone colors.

[a]Amazon Standard Identification Number.

[b]Stock keeping unit.

Cost

The cost for our preferred spine model is approximately US $161, but a more cost-effective version with fewer vertebrae can be purchased for US $34. The more expensive model is preferred due to the ease of placement in the mold, image quality on ultrasound scans, and representation of more neuraxial structures (spinal nerves and ligamentum flavum). Newer 3D printing technology allows for the printing of customizable and cost-effective pediatric spine models that could alternatively be used in our phantom. Ballistics gel priced at US $76 allows for the production of 4 or more phantoms. The additional items previously mentioned can amount to a cost between US $139 and US $150.

Results

There are 6 views that are critical to performing USGA neuraxial procedures; each is easily obtained from the ultrasound phantom:

- Parasagittal views (Figure 2): transverse process ("trident" sign), articular process ("camel hump" sign), and oblique interlaminar ("horse head" or "sawtooth" sign) views
- Transverse midline views (Figure 3): spinous process, interspinous process/interlaminar ("bat" or "flying bat" sign), and sacral cornua ("frog" or "frog eye" sign) views

Figure 2. Parasagittal images from phantom (**A, B, and C**), with probe placement relevant to bony anatomy and ultrasound indicator oriented cephalad, and patient (**D**), with ultrasound indicator oriented caudad. (A): Parasagittal TP view ("trident" sign). (B): Parasagittal AP view ("camel hump" sign; dashed blue line shows "camel hump" outline). (C): Parasagittal oblique interlaminar view ("horse head" or "sawtooth" sign; blue line shows "horse head" outline) in phantom. (D): Parasagittal oblique interlaminar view in patient. AC: anterior complex (interface of anterior dura and vertebral body); AP: articular process; ESM: erector spinae muscle; L: lamina; LF: ligamentum flavum; PC: posterior complex (interface of ligamentum flavum and posterior dura); TP: transverse process.

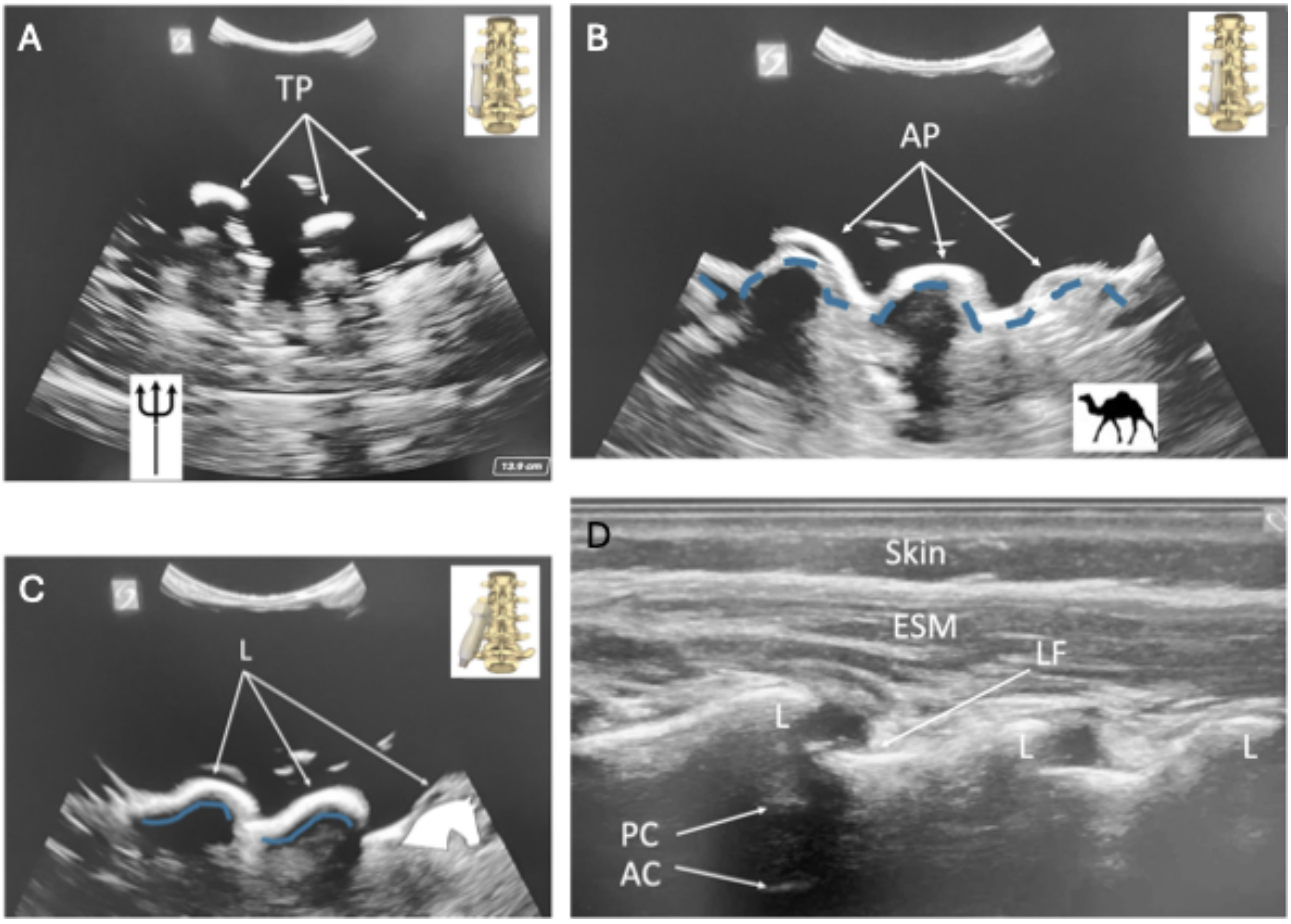
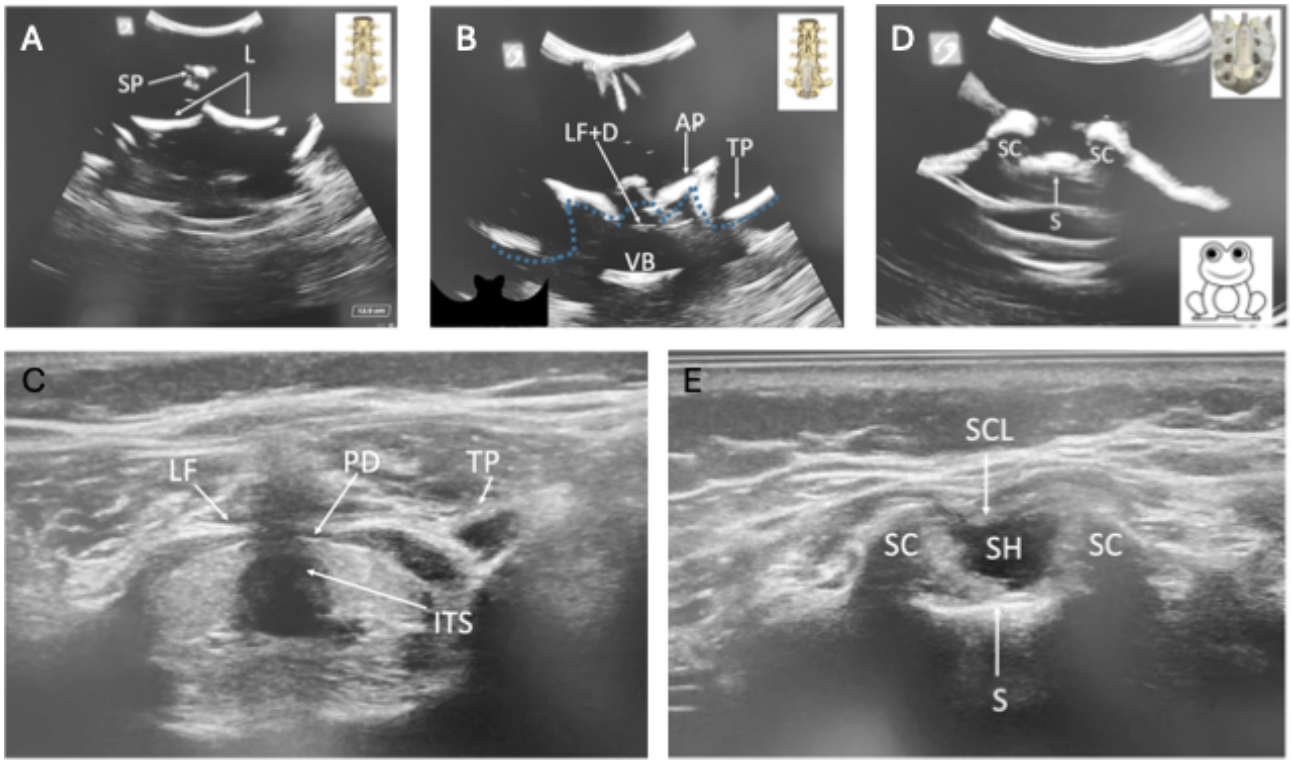


Figure 3. Transverse images from phantom (**A, B, and D**), with probe placement relevant to bony anatomy, and patient (**C and E**); ultrasound probe indicator is oriented left in all images. (A): Transverse midline SP view from phantom. (B): Transverse interspinous (interlaminar) view ("bat" or "bat wing" sign; dotted line shows "bat wing" outline) from phantom. (C) Transverse interspinous (interlaminar) view from patient. (D): Transverse SC view ("frog" or "frog eye" sign) from phantom. (E): Transverse SC view from patient. AP: articular process; ITS: intrathecal space; L: lamina; LF: ligamentum flavum; LF+D: ligamentum flavum+dura (ie, posterior complex); PD: posterior dura; S: sacrum; SC: sacral cornua; SCL: sacrococcygeal ligament; SH: sacral hiatus; SP: spinous process; TP: transverse process; VB: vertebral body.



Each of these views is demonstrated in Figures 2 and 3, with an additional image of the ultrasound probe placement in relation to the bony landmarks of the lumbosacral spine and patient sonoanatomy, where available. Several spine images from a 6-month-old male infant (written consent obtained from parent) are shown (Figures 2D, 3C, and 3E) beside the phantom images for comparison. Because the spine model lacks certain elements, not all structures appear on the phantom scan, and some cannot be obtained.

Phantom images were evaluated for likeness and accuracy via comparison to actual patient images by 10 practicing attending anesthesiologists. Each image was graded on a 5-point Likert scale for how similar it appeared to the actual patient image. All 10 respondents agreed or strongly agreed that the transverse sacral cornua view ("frog" sign) and parasagittal oblique interlaminar view ("horse head" sign) were similar to those of real patients. Of the 10 respondents, 8 agreed or strongly agreed that the transverse interspinous view ("bat wing" sign) was similar between the phantom and real patient images, 1 respondent was neutral, and 1 respondent somewhat disagreed.

Discussion

Unlike previous phantoms described by Morrow et al [16], Mashari et al [14], and others, our spine phantom generates ultrasound images and views that closely replicate the sonoanatomy of a pediatric patient (Figures 2D, 3C, and 3E). A key advancement in our design is the incorporation of the sacrum and sacral hiatus—critical structures needed for visualizing the caudal space, which is a technique that is often used in accessing the neuraxis in pediatric patients. Furthermore, by fully submersing the spine model in ballistics gel, our phantom offers superior stability during scanning and allows for repositioning to simulate sitting, lateral, and prone patient orientations. The enhanced design ensures a more realistic training experience, thereby helping practitioners develop the precise skills necessary for pediatric neuraxial techniques.

Practicing pediatric anesthesiologists overall found our phantom's ultrasound images comparable to ultrasound images of real pediatric anatomy, particularly for the transverse sacral cornua ("frog" sign) and parasagittal oblique interlaminar ("horse head" sign) views. However, while responses for the transverse interspinous (interlaminar) view ("bat wing" sign) were generally positive, some noted minor discrepancies between the phantom images and those of real patients. Given that the phantom lacked several ligaments and the spinal canal seen in real patients, this feedback provides an opportunity for improvement in future phantom models, which could be addressed by the techniques described by Morrow et al [16] (spinal canal) and Mashari et al [14] (ligaments).

Ultrasound has been used to identify anatomical landmarks for epidural or spinal neuraxial procedures and to identify placement of catheters that are inserted in the caudal space and threaded to the lumbar or thoracic space in pediatric patients [5,6,9-12]. The creation of ultrasound phantoms, as described in this paper, can increase access to ultrasound simulation and enhance opportunities for learning critical procedural skills in a

low-stakes environment [1-4,14,16]. To meet these needs, we created a phantom that is indefinitely shelf stable, reproducible, and cost-effective (approximately US $92 to US $219 per phantom, including the materials listed plus the reusable materials). By modifying the previous technique described by Morrow et al [16], our phantom was specifically designed to image the sacral cornua and to easily scan in the prone or lateral positions, which are essential features for training anesthesia clinicians in pediatric neuraxial sonoanatomy.

The use of spine phantoms was previously limited by their costs; however, budget-friendly spine phantoms created with readily available materials produce a realistic feel when palpating for anatomic landmarks [14] and generate many of the views required to perform neuraxial USGA procedures [14,16]. These phantoms also replicate sonoanatomy with high fidelity, as demonstrated by Mashari et al [14], who actually found that their low-cost model resulted in superior fidelity for ultrasound imaging when compared to an expensive, commercially available task trainer.

There are some limitations to the phantom described herein, of which many can be attributed to the absence of more complex anatomical structures. Although some views and sonoanatomy cannot be identified without these structures, easy solutions are available if needed. For example, our model has a fused sacrum, as is common in adults; therefore, scanning of the sacrum in the sagittal plane—a useful technique for performing in-plane USGA caudal epidural blocks in infants and children—is futile. This problem can be relieved by obtaining an anatomically correct spine model that is reflective of infants or young children, either through purchase or through 3D printing [14,18]. Models of pediatric spines with both normal anatomy and abnormal anatomy could be made via 3D printing, enhancing the pediatric-specific simulation experience; however, access can be limited and may be costly when considering the initial monetary investment in a 3D printer.

The phantom described also lacks contents of the spinal canal, rendering it inadequate for simulating access to the intrathecal space for spinal blockade. Inserting fluid-filled tubing into the empty spinal canal (a technique described by Morrow et al [16]) prior to pouring melted gel on the model could provide a potential solution. However, while this added feature can present itself as another useful learning tool, we found that needling the phantom degrades the image quality over time and should be considered when deciding whether to include a spinal canal in future models. Other potential options for making a more complete model include adding a ligamentum flavum by using silicone paste [14]. This technique may be useful for creating a sacrococcygeal ligament, which is an important landmark when performing USGA caudal blocks while using the transverse sacral cornua view (“frog” sign).

Of further note, we chose to use clear ballistics gel for our phantoms, since it allows for the direct visualization of spine model structures, which can be very helpful for early learners but is not realistic or comparable to scanning live patients. Products used to opacify the gel can be purchased on the internet (Table 1) if a more realistic option is desired.

Future directions for the use of our spine phantom model center on teaching critical skills and assessing knowledge of and comfort with high-stakes procedures in novice trainees. Further evaluation of our phantom should focus on the effectiveness of the phantom as a teaching tool.

By constructing a reproducible, affordable, and shelf-stable spine phantom that can be scanned to generate images and sonoanatomy of the infant and child neuraxis, trainees can be provided with a low-stakes environment in which they can learn how to perform high-stakes regional anesthesia blocks. By addressing the limitations of previous models, our phantom provides an affordable, high-fidelity tool that enhances access to realistic neuraxial ultrasound training for pediatric trainees.

Acknowledgments

Funding for the materials used to create the phantoms was provided by the University of Colorado Department of Anesthesiology. The funder was not involved in the study design; in the collection, analysis, and interpretation of the data; in the writing of the report; or in the decision to submit the paper for publication.

Conflicts of Interest

None declared.

References

1. Moore DL, Ding L, Sadhasivam S. Novel real-time feedback and integrated simulation model for teaching and evaluating ultrasound-guided regional anesthesia skills in pediatric anesthesia trainees. Paediatr Anaesth 2012 Sep;22(9):847-853. [doi: 10.1111/j.1460-9592.2012.03888.x] [Medline: 22612411]
2. Chen XX, Trivedi V, AlSaflan AA, et al. Ultrasound-guided regional anesthesia simulation training: a systematic review. Reg Anesth Pain Med 2017;42(6):741-750. [doi: 10.1097/AAP.0000000000000639] [Medline: 28759501]
3. Karmakar MK. The “Water-based Spine Phantom”? - A small step towards learning the basics of spinal sonography. Br J Anaesth 2009 Mar 12;103(eLetters Supplement). [doi: 10.1093/bja/el_4114]
4. Li JW, Karmakar MK, Li X, Kwok WH, Ngan Kee WD. Gelatin-agar lumbosacral spine phantom: a simple model for learning the basic skills required to perform real-time sonographically guided central neuraxial blocks. J Ultrasound Med 2011 Feb;30(2):263-272. [doi: 10.7863/jum.2011.30.2.263] [Medline: 21266566]

5. Boretsky KR, Camelo C, Waisel DB, et al. Confirmation of success rate of landmark-based caudal blockade in children using ultrasound: a prospective analysis. Paediatr Anaesth 2020 Jun;30(6):671-675. [doi: 10.1111/pan.13865] [Medline: 32267040]
6. Cristiani F, Henderson R, Lauber C, Boretsky K. Success of bedside ultrasound to identify puncture site for spinal anesthesia in neonates and infants. Reg Anesth Pain Med 2019 Jul 3:rapm-2019-100672. [doi: 10.1136/rapm-2019-100672] [Medline: 31273065]
7. Sidiropoulou T, Christodoulaki K, Siristatidis C. Pre-procedural lumbar neuraxial ultrasound-a systematic review of randomized controlled trials and meta-analysis. Healthcare (Basel) 2021 Apr 17;9(4):479. [doi: 10.3390/healthcare9040479] [Medline: 33920621]
8. Rubin K, Sullivan D, Sadhasivam S. Are peripheral and neuraxial blocks with ultrasound guidance more effective and safe in children? Paediatr Anaesth 2009 Feb;19(2):92-96. [doi: 10.1111/j.1460-9592.2008.02918.x] [Medline: 19207895]
9. Kil HK, Cho JE, Kim WO, Koo BN, Han SW, Kim JY. Prepuncture ultrasound-measured distance: an accurate reflection of epidural depth in infants and small children. Reg Anesth Pain Med 2007;32(2):102-106. [doi: 10.1016/j.rapm.2006.10.005] [Medline: 17350519]
10. Kil HK. Caudal and epidural blocks in infants and small children: historical perspective and ultrasound-guided approaches. Korean J Anesthesiol 2018 Dec;71(6):430-439. [doi: 10.4097/kja.d.18.00109] [Medline: 30086609]
11. Tsui BCH, Suresh S. Ultrasound imaging for regional anesthesia in infants, children, and adolescents: a review of current literature and its application in the practice of neuraxial blocks. Anesthesiology 2010 Mar;112(3):719-728. [doi: 10.1097/ALN.0b013e3181c5e03a] [Medline: 20179511]
12. Ponde VC, Bedekar VV, Desai AP, Puranik KA. Does ultrasound guidance add accuracy to continuous caudal-epidural catheter placements in neonates and infants? Paediatr Anaesth 2017 Oct;27(10):1010-1014. [doi: 10.1111/pan.13212] [Medline: 28795472]
13. Park SK, Bae J, Yoo S, et al. Ultrasound-assisted versus landmark-guided spinal anesthesia in patients with abnormal spinal anatomy: a randomized controlled trial. Anesth Analg 2020 Mar;130(3):787-795. [doi: 10.1213/ANE.0000000000004600] [Medline: 31880632]
14. Mashari A, Montealegre-Gallegos M, Jeganathan J, et al. Low-cost three-dimensional printed phantom for neuraxial anesthesia training: development and comparison to a commercial model. PLoS One 2018 Jun 18;13(6):e0191664. [doi: 10.1371/journal.pone.0191664] [Medline: 29912877]
15. Cole JH, Fishback JE, Hughey SB. Cadaveric porcine spines as a model for the human epidural space. Comp Med 2019 Aug 1;69(4):308-310. [doi: 10.30802/AALAS-CM-18-000133] [Medline: 31340882]
16. Morrow DS, Cupp JA, Broder JS. Versatile, reusable, and inexpensive ultrasound phantom procedural trainers. J Ultrasound Med 2016 Apr;35(4):831-841. [doi: 10.7863/ultra.15.04085] [Medline: 26969595]
17. Dye. Humimic Medical. URL: https://humimic.com/product-category/molds-and-accesories/dye/ [accessed 2025-04-25]
18. Marsh-Armstrong BP, Ryan JF, Mariano DJ, Suresh PJ, Supat B. Building affordable, durable, medium-fidelity ballistic gel phantoms for ultrasound-guided nerve block training. J Vis Exp 2024 Feb 9(204). [doi: 10.3791/66194] [Medline: 38407270]

Abbreviations

USGA: ultrasound-guided/assisted

Edited by B Lesselroth; submitted 26.06.24; peer-reviewed by JG Bailey, S Narayanasamy; revised version received 28.02.25; accepted 19.03.25; published 12.05.25.

Please cite as:
Webb L, Masaracchia M, Strupp K
Enhancing Access to Neuraxial Ultrasound Phantoms for Medical Education of Pediatric Anesthesia Trainees: Tutorial
JMIR Med Educ 2025;11:e63682
URL: https://mededu.jmir.org/2025/1/e63682
doi:10.2196/63682

Designing Personalized Multimodal Mnemonics With AI: A Medical Student's Implementation Tutorial

Noor Elabd[1*]; Zafirah Muhammad Rahman[1*]; Salma Ibrahim Abu Alinnin[1]; Samiyah Jahan[1]; Luciana Aparecida Campos[2], PhD; Ovidiu Constantin Baltatu[1,2], MD, PhD

[1]College of Medicine, Alfaisal University, PO Box 50927, Riyadh, Saudi Arabia

[2]Center of Innovation, Technology, and Education, Anhembi Morumbi University, Sao Jose dos Campos, Brazil

[*]these authors contributed equally

Corresponding Author:
Ovidiu Constantin Baltatu, MD, PhD
College of Medicine, Alfaisal University, PO Box 50927, Riyadh, Saudi Arabia

Abstract

Background: Medical education can be challenging for students as they must manage vast amounts of complex information. Traditional mnemonic resources often follow a standardized approach, which may not accommodate diverse learning styles.

Objective: This tutorial presents a student-developed approach to creating personalized multimodal mnemonics (PMMs) using artifical intelligence tools.

Methods: This tutorial demonstrates a structured implementation process using ChatGPT (GPT-4 model) for text mnemonic generation and DALL-E 3 for visual mnemonic creation. We detail the prompt engineering framework, including zero-shot, few-shot, and chain-of-thought prompting techniques. The process involves (1) template development, (2) refinement, (3) personalization, (4) mnemonic specification, and (5) quality control. The implementation time typically ranges from 2 to 5 minutes per concept, with 1 to 3 iterations needed for optimal results.

Results: Through systematic testing across 6 medical concepts, the implementation process achieved an initial success rate of 85%, improving to 95% after refinement. Key challenges included maintaining medical accuracy (addressed through specific terminology in prompts), ensuring visual clarity (improved through anatomical detail specifications), and achieving integration of text and visuals (resolved through structured review protocols). This tutorial provides practical templates, troubleshooting strategies, and quality control measures to address common implementation challenges.

Conclusions: This tutorial offers medical students a practical framework for creating personalized learning tools using artificial intelligence. By following the detailed prompt engineering process and quality control measures, students can efficiently generate customized mnemonics while avoiding common pitfalls. The approach emphasizes human oversight and iterative refinement to ensure medical accuracy and educational value. The elimination of the need for developing separate databases of mnemonics streamlines the learning process.

(JMIR Med Educ 2025;11:e67926) doi:10.2196/67926

KEYWORDS

medical education; personalized learning; prompt engineering; multimodal learning; memory techniques; dual-coding theory; student-centered approach; student-centered; large language model; natural language processing; NLP; machine learning; AI; ChatGPT; medical student; digital literacy; health care professional

Introduction

Problem Statement

Medical education presents students with the challenge of managing vast amounts of complex information. Mnemonics, memory techniques using associations and patterns, have demonstrated efficacy in improving the encoding and retrieval of medical knowledge [1]. These aids enhance learning and recall by transforming information into more memorable formats through elaborative encoding, retrieval cues, and imagery [2]. However, traditional standardized approaches often fail to accommodate diverse learning preferences, necessitating flexible applications that cater to individual needs.

Theoretical Framework

Paivio's dual-coding theory provides the theoretical foundation for this tutorial, supporting the integration of multimodal tools in education. By encoding both verbal and visual information through separate but interconnected pathways, students' understanding of academic vocabulary can be enhanced [3]. This theory underpins the potential effectiveness of multimodal mnemonics in medical education, particularly when combined

with personalization. Research indicates that personalized mnemonic techniques yield superior recall performance compared to standard strategies, with students using self-generated mnemonics demonstrating better performance on recall tasks [4,5].

Current State of Artificial Intelligence in Medical Education

Recent advances in generative artificial intelligence (AI) technologies have created new opportunities for personalized learning aid creation [6]. AI tools such as ChatGPT and DALL-E have demonstrated proficiency in generating creative and personalized content [7]. ChatGPT, a large language model, uses natural language processing to understand context and generate human-like text responses [8]. It can create diverse textual outputs, including various types of mnemonics. DALL-E, on the other hand, is an AI model designed to generate images from textual descriptions [9]. However, most AI applications in medical education currently focus on data analysis and pattern recognition rather than creative content generation for learning support [10].

Tutorial Aims and Target Audience

This tutorial aims to provide medical students with practical guidance for creating personalized multimodal mnemonics (PMMs) using AI tools. Through a systematic approach, we detail the step-by-step process of generating text and visual mnemonics using ChatGPT and DALL-E, incorporating templates and examples for effective prompt engineering. The tutorial shares strategies for personalizing mnemonics based on individual learning preferences while addressing common challenges and their solutions in AI-assisted mnemonic creation.

The primary audience includes medical students seeking to enhance their learning through personalized AI-assisted mnemonics. Secondary audiences include medical educators interested in implementing these tools in their teaching practice and students in other health care fields who can adapt these methods to their specific needs. By following this tutorial, readers will learn to create personalized learning aids that combine text and visual elements, potentially improving their ability to retain and recall complex medical information. The approach emphasizes practical implementation while maintaining academic rigor, making it accessible to both novice and experienced users of AI tools in educational settings.

Methods

PMMs: Tool Selection and Rationale

This tutorial uses ChatGPT and DALL-E 3 as the primary AI tools for creating PMMs. ChatGPT (GPT-4 model) was selected for text mnemonic generation due to its advanced natural language processing capabilities and ability to generate diverse textual outputs [11]. DALL-E 3 was chosen for visual mnemonic creation based on its proficiency in generating detailed, concept-relevant images from textual descriptions [12]. These tools were selected for their complementary strengths in producing textual and visual content, respectively, allowing for the creation of comprehensive, multimodal mnemonics. Both tools are accessible through OpenAI's platform, with ChatGPT available through both free and paid subscriptions, and DALL-E 3 using a credit-based system.

Configuration Settings and Access

For optimal results in medical mnemonic generation, we recommend using ChatGPT's GPT-4 model with a temperature setting of 0.7, which provides an effective balance between creativity and accuracy in medical content generation. For DALL-E 3, the high-quality setting ensures maximum detail and clarity in visual representations. While both tools offer free tiers, a professional subscription is recommended for consistent access to the latest model versions and enhanced capabilities.

Prompt Engineering Framework and Quality Assessment

The implementation of PMMs requires a systematic approach to prompt engineering and quality assessment. The process begins with zero-shot learning, where we provide clear instructions without examples, allowing the AI to generate mnemonics based purely on prompt structure. When initial results require refinement, we implement few-shot learning by providing 1 to 2 successful examples to guide the AI. For complex medical concepts, we use chain-of-thought prompting to break down the mnemonic creation process into logical steps.

Zero-shot prompting (providing direct instructions without examples) allows the AI to generate outputs based on its pretrained knowledge [13]. For example, a simple prompt like "Create a memorable mnemonic for the Krebs cycle intermediates" tests the AI's baseline capabilities without additional guidance.

Few-shot prompting (including 1 to 2 successful examples before the target prompt) helps guide the AI by demonstrating desired outputs [13,14]. For example, showing a successful biochemical pathway mnemonic before requesting one for the Krebs cycle improves output quality by providing clear examples of the expected format and style.

Chain-of-thought prompting breaks complex tasks into logical steps, improving accuracy through structured reasoning [15]. For example, "First, list intermediates. Then, identify key features. Finally, create a mnemonic." This systematic approach helps ensure comprehensive and accurate outputs, particularly for complex medical concepts.

The implementation followed a five-stage iterative process: (1) template development, in which adaptable prompt templates for text and visual mnemonics are created; (2) refinement, which optimizes prompts through testing of various structures and keywords; (3) personalization, which integrates learning preferences and personal associations and adds options for imagery, humor, and clinical relevance [16]; (4) mnemonic specification, in which prompts for various mnemonic types (acronyms, phrases, rhymes) are created, with corresponding visual representations; and (5), quality control, which is based on peer-to-peer review discussions.

The mnemonics generated through this process were evaluated in peer-to-peer discussions among student authors, facilitated and guided by mentors. This iterative feedback process, integral to quality control, not only ensured medical accuracy and

educational value by leveraging student insights and expert guidance, but also trained students to critically assess the quality, accuracy, and effectiveness of AI-generated content. Identified inaccuracies or areas for improvement directly informed subsequent prompt adjustments.

The medical concepts used in this study were carefully selected from ongoing medical courses, ensuring immediate relevance to current learning needs. This selection process focused on identifying complex topics that students found challenging to memorize while ensuring diverse representation of medical subjects and validation against standard medical resources.

For text mnemonic generation, we developed a basic template structure that incorporates medical accuracy, memorability, and personalization elements: "Create a memorable [mnemonic type] for [medical concept]. Focus on [key aspects]. Make it [characteristics: funny/clinical/etc]. Include [specific elements] that relate to [learning context]."

For visual mnemonic creation, the template emphasizes clarity and medical accuracy: "Generate a [style] image depicting [mnemonic content] for [medical concept]. Emphasize [key visual elements]. Ensure medical accuracy and clarity." These templates serve as starting points and can be customized based on individual learning preferences and specific medical concepts.

Common Challenges and Solutions

Through our implementation process, we identified several common challenges. Inaccurate medical terminology can be addressed by including specific medical terms in prompts. Unclear visual representations are improved by specifying anatomical or clinical details. When mnemonics become overly complex, requesting step-by-step breakdowns helps maintain clarity and usability. These solutions emerged from practical experience and continue to evolve as we refine the process.

Ethical Considerations

This educational methodology development did not require formal ethical review as it did not involve human subjects research, collected no personal data, and used only publicly available AI tools as part of regular educational activities.

Results

The implementation of PMMs using AI tools demonstrated both potential and limitations across a range of medical concepts. Through systematic testing and refinement, we identified key performance metrics, quality assessment outcomes, and practical implementation challenges.

Generation Performance

The PMM generation process exhibited consistent performance characteristics. Text mnemonic generation via ChatGPT consistently required 2 to 3 minutes per concept. Generating corresponding visual mnemonics with DALL-E 3 required 3 to 5 minutes per concept. Reaching a satisfactory mnemonic typically involved 1 to 3 iterative attempts. The initial success rate, defined as achieving acceptable output on the first attempt, was 85%. After applying quality control and refinement procedures, this success rate increased to 95%. These data suggest that generating PMMs is feasible within a reasonable timeframe, particularly when incorporating iterative refinement.

Quality Assessment and Examples

Table 1 presents 6 examples of AI-generated PMMs, illustrating the range of mnemonic types generated and highlighting observed successes and areas for improvement. The table includes the target medical concept, the text mnemonic generated by ChatGPT, the visual prompt provided to DALL-E 3, the resulting visual mnemonic, and specific limitations encountered.

Through systematic documentation of the implementation process, we observed that achieving satisfactory results typically required 1 to 3 iterations per concept. Initial attempts often needed refinement in areas such as the precision of medical terminology, visual clarity, or text-image alignment. The examples in Table 1 showcase both successful implementations and instances where refinement was necessary, providing practical insights for users of this tutorial.

The development of a mnemonic for bacterial cell wall antibiotics serves as a representative example. The initial AI-generated mnemonic incorrectly used the term "harbipenems" instead of "carbapenems." This error, corrected in subsequent iterations, demonstrates the need for human validation, specifically by the medical student or a medical professional. Other examples in Table 1 illustrate the variety of generated mnemonic types, including acronyms, rhymes, and short phrases, paired with their visual counterparts. While generally effective, some visual mnemonics presented challenges in terms of clarity, anatomical accuracy, and alignment with the corresponding text mnemonic.

Table . Artifical intelligence (AI)–generated personalized multimodal mnemonics.

Medical concept	Prompts used and results	Text mnemonic	Visual and text mnemonic	Observations and limitations
Antibiotics targeting cell wall	First prompt: "Make a mnemonic for antibiotics that target the cell wall, you can group the antibiotics by their mechanism and make it memorable." Second prompt: "Using the mnemonic, make a memorable photo." Result: photo was not memorable. Third prompt: "Retry this time make a funny mnemonic." Result: the cat image.	Mnemonic: "very big purple cats bake happy eggs." • V: vancomycin • B: beta-lactams • P: penicillins • C: cephalosporins • B: bacitracin • H: harbapenems		Creative and memorable representation, but incorrect text
Types of leukocytes	Prompt: "Develop a mnemonic to remember the different types of leukocytes. Also, create a funny and cartoonish visual that represents this mnemonic in a memorable way. The illustration should effectively depict the mnemonic."	Mnemonic (for leukocytes): "never let monkeys eat bananas." • N: neutrophils • L: lymphocytes • M: monocytes • E: eosinophils • B: basophils		Engaging visualization, but incorrect text
Life cycle of malaria	Prompt: "Create a mnemonic to help remember the malaria life cycle, and design a humorous, cartoon-style visual to represent it in a memorable way. The illustration should clearly depict the mnemonic for easy recall."	Mnemonic (each word stands for a stage of the malaria parasite's life cycle): "salty mice bite liver, chilling red blood cells." • S: sporozoites (transmitted via mosquito bite) • M: migrate to the liver • B: burst out of the liver as merozoites • L: liver stages • C: chills and fevers (as the parasites enter red blood cells) • Red blood cells: parasites invade and replicate in red blood cells		The AI was occasionally unable to fully adhere to detailed prompts in image generation. Additionally, the text included in the image was incorrect.
Renin-angiotensin-aldosterone system (RAAS)	First prompt used: "Generate a funny mnemonic for Renin-Angiotensin-Aldosterone System (RAAS) but make it simple and easy to remember." Second prompt: "Create simple and memorable image of this mnemonic 'Keen Elephants Always Raise Blood.' Use less texts, clear and correct spelling. Place the mnemonic as a title."	Mnemonic: "keen elephants always raise blood." • K: kidneys release renin when blood pressure or sodium levels are low. • E: enzyme (renin) converts angiotensinogen to angiotensin I. • A: ACE (angiotensin-converting enzyme) converts angiotensin I to angiotensin II. • R: raises blood pressure by vasoconstriction (angiotensin II constricts blood vessels). • B: blood volume increases as aldosterone causes sodium and water retention.		This mnemonic emphasizes the kidney's keen response to blood pressure regulation. Text is not clear.

Medical concept	Prompts used and results	Text mnemonic	Visual and text mnemonic	Observations and limitations
Non-selective COX inhibitors	Prompt used: Make a funny memorable picture mnemonic for me to remember the *Non-Selective COX Inhibitors*. They include: Salicylates: Aspirin; Pyrazalone derivatives; Propionic acid derivatives: Ibuprofen, Naproxen; Indole Derivatives: Indomethacin; Anthranilic acid derivative; Aryl acetic acid derivative: Diclofenac; Oxicam derivative; Pyrrole derivatives."	Mnemonic: "a perfect brew is delightfully aromatic, outstandingly pleasant." • A: aspirin • P: pyrazalone derivatives • B: ibuprofen • I: indole derivatives • D: diclofenac • A: anthranilic acid derivative • O: oxicam derivative • P: pyrrole derivatives		The coffee-themed elements like cups, beans, and mugs may be too abstract and unrelated to the actual topic. The picture lacks clarity and is not memorable.
Naeglaria fowleri	First prompt used: "Make an image to remember naeglaria fowleri, brain eating heat loving free living with a mneumonic for its major facts." Second prompt: "Make it funny: >" Response: "I couldn't generate the image because the content did not align with the guidelines for image creation. If you'd like, we can adjust the concept or explore a different approach. Let me know how you'd like to proceed!" Prompt: "Try again" Response: "The guidelines for image creation restrict certain content related to how medical or potentially dangerous organisms are portrayed, especially if humor or exaggerated characteristics could be seen as trivializing the topic. Creating a funny depiction of a 'brain-eating' amoeba might fall under those restrictions."	—[a]		The AI refused to generate the content, highlighting challenges in discerning educational needs from potentially inappropriate content, emphasizing the need for human oversight and refinement.

[a]Not applicable.

Generic Templates for Prompt Engineering

For text mnemonic generation (with ChatGPT), the following generic prompt template was developed: "Create a memorable sentence mnemonic for [medical concept]. The mnemonic should be [characteristic 1: funny/simple/easy to remember] and [characteristic 2: relevant to clinical practice/focused on key steps/highlight main components]. Each word or part of the sentence should represent a key aspect of the concept. If possible, incorporate [optional element: wordplay/alliteration/vivid imagery]. Make it relatable to [personal preference: a specific scenario/everyday objects/animals]."

For creation of the visual mnemonic (with DALL-E 3), the following generic prompt template was used: "Generate a [style: cartoon/funny/medical illustration] depicting the mnemonic '[text mnemonic]' for [medical concept]. The image should be [characteristic 1: visually engaging/humorous/clear] and [characteristic 2: memorable/related to the mnemonic words]. Incorporate [specific visual elements: anthropomorphized objects/exaggerated features/relevant symbols]. Ensure any text is minimal, clear, and correctly spelled. Place the mnemonic sentence as a title."

These templates were iteratively refined based on the quality and relevance of the AI-generated outputs. Examples of specific prompts based on these templates are in Table 1. These templates and examples provide a framework for creating diverse and engaging sentence mnemonics while allowing for customization based on the specific medical concept and desired learning outcomes.

Implementation Challenges and Refinement Strategies

Three key challenges emerged during implementation, leading to the development of targeted refinement strategies. The first

challenge was medical accuracy. Maintaining medical accuracy necessitated continuous verification against established medical resources. Initial outputs occasionally exhibited terminology errors or incomplete conceptual coverage. These issues were addressed by incorporating specific medical terminology in the prompts and implementing a systematic review process involving medical experts.

Second, achieving consistent visual clarity and anatomical accuracy in the AI-generated images presented challenges. Some images lacked clarity or contained inconsistencies between textual and visual elements. We improved visual quality through prompt refinement, including more precise anatomical descriptions and requests for simplified representations of complex concepts.

The third challenge was in content integration. Ensuring seamless integration between the text mnemonic and visual representation required careful prompt design and quality control. A structured review process was implemented to verify that both components effectively reinforced the target medical concept and functioned synergistically to enhance learning.

These findings offer practical observations for educators considering the use of AI-assisted mnemonic generation. While the PMM approach holds promise for personalized learning, our results underscore the essential role of human oversight, domain expertise, and iterative refinement in ensuring accuracy, clarity, and educational value.

Discussion

Principal Findings and Implications

This tutorial demonstrates a practical approach to generating PMMs using readily available AI tools. Our findings highlight the feasibility of creating customized mnemonics within a reasonable timeframe (2-5 minutes per concept, with 1-3 iterative attempts). The combination of text and visual elements aligns with dual-coding theory [3,17], potentially enhancing learning and recall. However, challenges related to medical accuracy, visual clarity, and content integration underscore the crucial role of human oversight and domain expertise. The “harbipenems” error, for example, emphasizes the need for medical professionals to validate AI-generated content. These findings suggest that AI-assisted PMM generation can be a valuable tool for personalized learning, but careful attention to quality control and prompt refinement is essential.

Comparison to the Literature

This tutorial’s approach aligns with the growing interest in applying AI for personalized learning in medical education. While much of the current research focuses on AI for tasks like data analysis [10,18], this tutorial explores the relatively novel application of AI for generating personalized learning content. Our emphasis on multimodal learning resonates with the principles of dual-coding theory [3], which suggests that combining visual and textual representations can enhance learning and memory. Furthermore, the challenges we encountered regarding accuracy and clarity in AI-generated content echo broader concerns in the literature about the need for human oversight in AI-driven educational applications [8,19].

Strengths and Limitations

This tutorial provides a practical, step-by-step guide for generating PMMs using AI, offering readily adaptable prompt templates and illustrative examples. The student-centered perspective offers valuable insights into the practical challenges and potential benefits of this approach.

This tutorial has several limitations. First, the AI models used may exhibit biases, potentially limiting the diversity and novelty of generated PMMs. Second, inaccuracies in visual representations, such as misspellings or mismatches with the text mnemonic, require careful review and correction. Third, current AI models may refuse to generate content for sensitive medical topics, necessitating alternative strategies or manual content creation. Finally, the lack of a formal evaluation with medical students limits the generalizability of our findings and prevents definitive conclusions about the effectiveness of PMMs on learning outcomes.

Future Directions

Future research should investigate the effectiveness of PMMs on learning outcomes through controlled studies comparing PMMs to traditional learning methods. Such studies should use objective measures of learning, such as recall accuracy, learning efficiency, and student satisfaction. Further research should also explore the long-term impact of PMMs on knowledge retention and application. The scalability and adaptability of the PMM approach across diverse medical subjects and educational settings warrant investigation. Additionally, future work should address the ethical considerations surrounding AI-generated educational content, including data privacy, bias, and overreliance on technology [8]. Developing guidelines for the ethical and effective use of AI in mnemonic creation and medical education more broadly will be crucial as this field evolves [19].

Conclusion

This tutorial presents a practical approach to generating PMMs for medical education using the AI tools ChatGPT and DALL-E 3. This approach emphasizes AI as a tool to enhance, rather than replace, traditional learning methods. Originating from medical students seeking to improve their own learning, this tutorial describes a step-by-step process involving prompt engineering, iterative refinement, and quality assessment, illustrated with examples for 6 medical concepts. The personalized nature of the mnemonics, coupled with the multimodal approach, demonstrates potential for enhancing student engagement and facilitating the retention of complex medical concepts. We also highlight key challenges related to medical accuracy, visual clarity, and content integration, underscoring the importance of human oversight and domain expertise in refining AI-generated content. This student-led exploration offers practical guidance and a valuable starting point for educators and students alike interested in leveraging AI for personalized learning in medical education.

Conflicts of Interest

None declared.

References

1. Rosi-Schumacher M, DeGiovanni JC. Using the lessons of learning science to improve medical education in otolaryngology. Ear Nose Throat J 2022 Nov;101(9_suppl):16S-19S. [doi: 10.1177/01455613231160509] [Medline: 36825609]
2. Mocko M, Lesser LM, Wagler AE, Francis WS. Assessing effectiveness of mnemonics for tertiary students in a hybrid introductory statistics course. J Stat Educ 2017 Jan 2;25(1):2-11. [doi: 10.1080/10691898.2017.1294879]
3. Soper K, Geske JA, Bronner L, Godfrey M. Improving student understanding of academic assessment vocabulary words using visual cues: a collaborative effort. J Community Engagem Scholarsh 2022 Jul 29;15(1):1. [doi: 10.54656/jces.v15i1.52] [Medline: 36081415]
4. Dirette DP. A comparison of self-generated versus taught internal strategies for working memory. NeuroRehabilitation 2015;36(2):187-194. [doi: 10.3233/NRE-151206] [Medline: 25882198]
5. Wheeler RL, Gabbert F. Using self-generated cues to facilitate recall: a narrative review. Front Psychol 2017;8:1830. [doi: 10.3389/fpsyg.2017.01830] [Medline: 29163254]
6. Altintas L, Sahiner M. Transforming medical education: the impact of innovations in technology and medical devices. Expert Rev Med Devices 2024 Sep;21(9):797-809. [doi: 10.1080/17434440.2024.2400153] [Medline: 39235206]
7. Waikel RL, Othman AA, Patel T, et al. Generative methods for pediatric genetics education. medRxiv. 2023 Aug 2. [doi: 10.1101/2023.08.01.23293506] [Medline: 37790417]
8. Leng L. Challenge, integration, and change: ChatGPT and future anatomical education. Med Educ Online 2024 Dec 31;29(1):2304973. [doi: 10.1080/10872981.2024.2304973] [Medline: 38217884]
9. Adams LC, Busch F, Truhn D, Makowski MR, Aerts H, Bressem KK. What does DALL-E 2 know about radiology? J Med Internet Res 2023 Mar 16;25:e43110. [doi: 10.2196/43110] [Medline: 36927634]
10. Tolentino R, Baradaran A, Gore G, Pluye P, Abbasgholizadeh-Rahimi S. Curriculum frameworks and educational programs in AI for medical students, residents, and practicing physicians: scoping review. JMIR Med Educ 2024 Jul 18;10:e54793. [doi: 10.2196/54793] [Medline: 39023999]
11. Shoham OB, Rappoport N. MedConceptsQA: Open source medical concepts QA benchmark. Comput Biol Med 2024 Nov;182:109089. [doi: 10.1016/j.compbiomed.2024.109089] [Medline: 39276611]
12. Temsah MH, Alhuzaimi AN, Almansour M, et al. Art or artifact: evaluating the accuracy, appeal, and educational value of AI-generated imagery in DALL·E 3 for illustrating congenital heart diseases. J Med Syst 2024 May 23;48(1):54. [doi: 10.1007/s10916-024-02072-0] [Medline: 38780839]
13. Sivarajkumar S, Kelley M, Samolyk-Mazzanti A, Visweswaran S, Wang Y. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: algorithm development and validation study. JMIR Med Inform 2024 Apr 8;12:e55318. [doi: 10.2196/55318] [Medline: 38587879]
14. Drori I, Zhang S, Shuttleworth R, et al. A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level. Proc Natl Acad Sci U S A 2022 Aug 9;119(32):e2123433119. [doi: 10.1073/pnas.2123433119] [Medline: 35917350]
15. Khademi S, Palmer C, Dimaguila GL, Javed M, Buttery J. Exploring large language models for detecting online vaccine reactions. Stud Health Technol Inform 2024 Sep 24;318:30-35. [doi: 10.3233/SHTI240887] [Medline: 39320177]
16. Tuttle JJ, Moshirfar M, Garcia J, Altaf AW, Omidvarnia S, Hoopes PC. Learning the Randleman criteria in refractive surgery: utilizing ChatGPT-3.5 versus internet search engine. Cureus 2024 Jul;16(7):e64768. [doi: 10.7759/cureus.64768] [Medline: 39156271]
17. Clark JM, Paivio A. Dual coding theory and education. Educ Psychol Rev 1991 Sep;3(3):149-210. [doi: 10.1007/BF01320076]
18. Tozsin A, Ucmak H, Soyturk S, et al. The role of artificial intelligence in medical education: a systematic review. Surg Innov 2024 Aug;31(4):415-423. [doi: 10.1177/15533506241248239] [Medline: 38632898]
19. Franco D'Souza R, Mathew M, Mishra V, Surapaneni KM. Twelve tips for addressing ethical concerns in the implementation of artificial intelligence in medical education. Med Educ Online 2024 Dec 31;29(1):2330250. [doi: 10.1080/10872981.2024.2330250] [Medline: 38566608]

Abbreviations

AI: artificial intelligence
PMM: personalized multimodal mnemonic

Edited by B Lesselroth; submitted 24.10.24; peer-reviewed by M Alshiekh, R Sajja; revised version received 20.03.25; accepted 06.04.25; published 08.05.25.

Please cite as:
Elabd N, Rahman ZM, Abu Alinnin SI, Jahan S, Campos LA, Baltatu OC
Designing Personalized Multimodal Mnemonics With AI: A Medical Student's Implementation Tutorial
JMIR Med Educ 2025;11:e67926
URL: https://mededu.jmir.org/2025/1/e67926
doi:10.2196/67926

Faculty Retreats in Academic Medicine: Tutorial

Rachel Skains[*], MD, MSPH; Julie Brown[*], MD; Erin F Shufflebarger[*], MD, MSPH; Justine McGiboney[*], MD; Sherell Hicks[*], MD; Laine McDonald[*], MD, MPH; Katherine B Griesmer[*], MD, MPH; Christine Shaw[*], MD; Emily Grass[*], MD, MEHP, MSc; Marie-Carmelle Elie[*], MD; Lauren A Walter[*], MD, MSPH

Department of Emergency Medicine, Heersink School of Medicine, University of Alabama at Birmingham, 519 19th St S, Birmingham, AL, United States

[*]all authors contributed equally

Corresponding Author:
Lauren A Walter, MD, MSPH
Department of Emergency Medicine, Heersink School of Medicine, University of Alabama at Birmingham, 519 19th St S, Birmingham, AL, United States

Abstract

Faculty development is a cornerstone of academic medicine, supporting personal growth, professional advancement, and departmental effectiveness across all stages of a faculty member's career. Among the tools available, faculty retreats have increasingly emerged as a high-impact strategy to foster collaboration, advance strategic planning, and address individual and collective goals in a structured, reflective setting. While retreats are widely used in other sectors, practical guidance tailored to the academic medicine context remains limited. This tutorial offers a comprehensive, step-by-step framework for planning and implementing faculty retreats within academic departments. Key elements of effective retreat design are outlined, including (1) conducting a preretreat needs assessment to align goals with faculty priorities, (2) selecting an appropriate format (eg, in-person or hybrid), (3) fostering psychological safety to enhance participation, and (4) using facilitation techniques that promote inclusive dialogue and actionable outcomes. The tutorial also emphasizes logistical considerations, such as agenda design, timing, and participant engagement strategies, alongside mechanisms to ensure follow-up and accountability after the retreat. In addition to highlighting common barriers, such as resource limitations, scheduling constraints, and engagement disparities, the tutorial provides practical solutions drawn from real-world examples in academic medicine. By integrating thoughtful planning, evidence-informed facilitation, and postretreat follow-through, faculty retreats can serve as transformative experiences that support both individual development and departmental cohesion. This resource aims to fill a gap in the literature by equipping leaders in academic medicine with a structured approach to designing, executing, and sustaining the benefits of faculty retreats.

(JMIR Med Educ 2025;11:e71622) doi:10.2196/71622

KEYWORDS

faculty development; faculty retreat; academic medicine; organizational leadership; medical education

Introduction and Background

As faculty development in academic medicine has evolved, it now emphasizes career-long growth and attention to both individual and departmental dynamics [1,2]. Faculty retreats have emerged as a valuable tool to enhance performance and institutional effectiveness by bringing faculty together outside routine settings for collaboration, strategic planning, and professional development. Retreats also offer a structured space to address challenges, build consensus, resolve conflicts, and foster collegiality (Textbox 1; [3-12]).

Textbox 1. Benefits of faculty development retreats.

Skill enhancement

Retreats provide dedicated time for faculty to engage in specific professional development activities [4,5].

Improved collaboration and networking

Retreats foster interdisciplinary collaboration and networking among faculty, leading to new research partnerships, coteaching opportunities, and the sharing of best practices [5,6].

Reflection and constructive feedback

Faculty retreats can offer time for reflection on teaching, research, and administrative responsibilities. Structured feedback from peers during retreats helps faculty identify areas for improvement and develop actionable plans to enhance their performance [7].

Burnout prevention and well-being

Retreats provide a break from the daily pressures of academic and clinical life, allowing faculty to focus on self-care and stress management. This can lead to improved well-being, reduced burnout, and increased job satisfaction, which positively affect performance [8].

Strengthened culture and community

There is a strong correlation between team building and company or department culture. The mutual collaboration that happens over the course of team-building exercises fosters a sense of community. This, in turn, helps department culture to develop and evolve smoothly and in a way that is true to the departmental vision [9].

Innovation and development

Faculty retreats often serve as spaces for brainstorming and developing innovative teaching strategies, curricular changes, new research, or quality improvement ideas [10].

While retreats have long been common in the business world, their broader use in academic medicine is more recent [13]. However, practical guidance for planning and implementing faculty retreats remains limited and fragmented. This tutorial offers a comprehensive overview of the purpose, value, and implementation of faculty retreats within academic departments, highlighting current trends and best practices.

Solution

Faculty Retreat Implementation Strategies

Preretreat Preparation: Needs Assessment

Conducting a needs assessment prior to the faculty retreat ensures that the most pressing issues are addressed, the goals of the intended group are aligned, and the participants are engaged effectively [14]. It is important to communicate the objectives of the needs assessment to the faculty prior to response solicitation by clarifying the purpose. It should be explained clearly that the data are intended to be used to direct faculty development objectives, including retreat planning. The needs assessment should gather information to achieve these goals. Further, the scope of the needs assessment should be carefully considered, whether it will focus on individual faculty members, a whole department, or the institution as a whole. This helps target the assessment appropriately.

It is important to ensure all faculty members who will be the intended participants of the retreat have an opportunity to contribute to the needs assessment. This increases buy-in and ensures the retreat meets their expectations. If the retreat involves interdisciplinary work or collaborations with external partners, gather input from relevant stakeholders to understand their expectations and needs as well. Finally, the needs assessment participants may constitute a larger group than those subsequently engaged in a future retreat, as the data may be used more broadly for additional faculty development efforts.

Consider the use of multiple methods for needs assessment data collection [15]. A combination of qualitative and quantitative methods provides a more robust picture of faculty needs by gathering diverse perspectives. This can include surveys and questionnaires (eg, Google Forms, SurveyMonkey (SurveyMonkey Inc), Qualtrics (Qualtrics International Inc), etc, for easy distribution and analysis). Use a mix of open-ended and multiple-choice questions to cover various areas (Textbox 2). Finally, ensure that survey responses are anonymous or use anonymous identifiers to provide distinction between responders, if needed.

Textbox 2. Example of preretreat needs assessment questions.

- What are the main challenges you face in your teaching or research?
- What are your goals for the upcoming academic year?
- What topics would you like to see covered at the retreat (eg, leadership development, team building, teaching strategies, and research collaboration)?
- How do you feel about the current team dynamics within the department/faculty?
- How can the retreat best support your professional development?
- Ranked Priorities: Ask faculty to rank the importance of potential retreat topics, such as:
- Well-being and self-care/self-compassion
- Mentorship and sponsorship
- Negotiating effectively
- Effective communication
- Strategic planning
- Team building
- Increasing scholarly activity (eg, research productivity)
- Leadership and management
- Promotion and tenure

Needs assessment data collection may also involve interviews or focus groups [16]. One-on-one in-depth interviews with leadership, faculty leaders, and administration can obtain deeper insights into institutional or departmental strategic priorities. This ensures that the retreat is aligned with larger institutional and departmental goals. Ask about specific challenges within departments, faculty development gaps, and potential solutions. Use these interviews to uncover nuanced issues that might not emerge in a survey. Organize small focus groups to encourage open discussion among faculty members. This is particularly useful for gathering collaborative input on broader issues, like curriculum design or faculty culture. Consider having a neutral facilitator to ensure open, honest dialogue.

In addition to solicited participant input, also consider reviewing objective institutional or departmental data. This can include performance data, such as performance metrics, teaching evaluations, faculty publication records, and other key indicators. This data can help identify key areas where faculty need support, such as improving student engagement or increasing research output. Review any existing data from faculty climate surveys or teaching evaluations to identify patterns or concerns that need to be addressed during the retreat. Finally, if available, analyze feedback from previous retreats or faculty development initiatives to identify ongoing or unresolved issues.

Once the needs assessment is obtained, identify key themes and prioritize needs [17]. Organize the feedback into categories, such as personal development needs, leadership skills, work-life balance, mentoring, etc. Before the retreat, share a summary of the needs assessment findings with faculty. This transparency builds trust and helps participants understand how their input has shaped the retreat agenda. Once developed, provide a draft of the retreat agenda, highlighting how it addresses the needs and priorities gathered from the assessment.

Finally, consider the needs assessment timeline. It should be conducted well enough in advance of the anticipated retreat to provide ample time for analysis and planning but not so far ahead of time that the data obtained and reviewed are no longer timely or pertinent. By conducting a thorough needs assessment, the faculty retreat can be tailored to the specific needs of the group, which will increase engagement, relevance, and the overall success of the retreat.

Preretreat Preparation: Setting Retreat Goals and Objectives

A comprehensive consideration of needs assessment data (see above) should identify and prioritize specific objectives. By addressing common areas of interest and import, clear and meaningful goals and objectives can be set to guide the retreat's agenda and outcomes [18]. In addition, early engagement of key participants, including faculty leaders and key stakeholders, will ensure that the retreat's objectives align with the broader mission and vision of the institution or department. Consider that retreat funders may impart influence on a retreat's scope and attempt to align an instructed scope with identified needs assessment priorities to improve the effectiveness of the retreat.

Finally, create SMART (Specific, Measurable, Achievable, Relevant, and Time-bound) goals to provide a clear roadmap for retreat planning and postretreat follow-up [19]. These goals help ensure that retreat objectives are not only aspirational but also operational. For example, instead of a vague aim, such as "foster collaboration," a SMART goal would be: "Improve cross-departmental collaboration by initiating at least three joint research proposals within the next academic year." Using SMART goals helps organizers and participants align around shared expectations and provides a framework for evaluating success. Specific goals target a defined area for improvement; Measurable goals allow for tracking progress; Achievable goals ensure feasibility given available resources; relevant goals align

with broader institutional priorities; and Time-bound goals set a clear deadline. During planning, facilitators should work with stakeholders to codevelop 2-3 SMART goals that reflect the most pressing needs and strategic aims of the department or institution. These goals can then guide session design, inform postretreat evaluation metrics, and promote accountability by linking retreat outcomes to tangible follow-up actions.

Prioritization of objectives will be required if the needs assessment identified several themes or areas of interest. Differentiate between objectives that can be addressed during the retreat versus those that require ongoing effort. Retreat goals can include long-term objectives, but given time constraints, the retreat may only be a start or a contributing component to that objective. Aligning the retreat content with long-term objectives can be effective from a strategic planning perspective. Finally, focus on a manageable number of objectives to avoid overwhelming participants; consider what is feasible in the retreat time allotted.

Preretreat Preparation: Resource Allocation

Determining resource allocation for a faculty retreat involves careful planning to ensure that funds, time, and human resources are used efficiently to meet the retreat's objectives [20]. One of the most important first steps is to set the budget, considering which elements are critical (resources essential to the success of the retreat) versus optional. Review the available budget from the department, institution, or external grants. Be sure to account for all funding sources and restrictions (Textbox 3).

Textbox 3. Potential retreat costs for budget consideration.

Venue costs

Space rental for meetings, workshops, and team activities

Meals and refreshments

Catering for meals, snacks, and coffee breaks

Accommodation (if overnight)

Lodging for multiday retreats

Travel expenses

Transportation costs (eg, shuttles, mileage reimbursement, train, or flight tickets)

Facilitator or speaker fees

Costs for external facilitators, speakers, or consultants

Workshop materials

Supplies for activities (eg, printed materials, flipcharts, and projectors)

Recreational activities

Funding for team-building exercises, social activities, or outings

Miscellaneous

Any additional costs (eg, technology support, insurance, and contingency funds for unexpected expenses)

Plan for contingencies

Buffer for unexpected costs. Set aside a portion (typically 5%-10%) of the overall budget as a contingency fund for unforeseen expenses, such as last-minute speaker fees, additional supplies, or travel delays

In addition to financial costs, consider human resources allocation, including the creation of a retreat organizing committee. Identify staff or faculty members who will be responsible for planning, logistics, and facilitating the retreat. Ensure these roles are clearly defined, including responsibilities for agenda setting, coordinating logistics with the venue and vendors, and managing participant communication. If the retreat requires significant logistical support, allocate part of the budget to administrative staff, event planners, or student assistants to assist with setup, note-taking, or technical support during the event.

Preretreat Preparation: Selecting Participants and Faculty

Selecting participants for a faculty retreat involves careful consideration to ensure that the right mix of individuals is involved, based on the retreat's purpose and goals (see above) [21]. Based on retreat objectives, consider prioritizing a balanced representation. This can involve fostering diversity in experience by engaging faculty across the career spectrum, from early-career to senior members, to enrich perspectives and strike a balance between tradition and innovation. This can also include cultural diversity to ensure background and social diversity to promote inclusive discussions and a wide range of viewpoints. Finally, tailor role-specific involvement to the retreat's focus by including teaching and education faculty for curriculum development, research-active faculty and grant managers for research goals, and both faculty and professional staff for career growth discussions.

When appropriate, encourage voluntary participation. Identify motivated participants by soliciting a call for interest. Send an open call to faculty and allow those with an interest in the retreat's objectives to self-nominate. However, a more selective

invitation process may be required in some cases to ensure that the right people are present, especially for leadership-focused retreats.

Keep the group size manageable, typically between 15 and 25 participants, depending on the format. On some occasions, an even smaller cohort may be preferred, no more than a dozen, to ensure all voices are heard and universal participation is feasible. Smaller groups are more conducive to intimate, productive discussions, while larger groups can handle broader, more diverse topics. For larger retreats, plan for breakout sessions that enable smaller, focused discussions among participants. If relevant, consider inviting external facilitators or industry experts who can offer outside perspectives and guide discussions more effectively. Finally, account for logistical and financial constraints incurred by the number of retreat participants, including travel, budget constraints, and scheduling availability.

Preretreat Preparation: Choosing Location and Setting

Choosing a location for a faculty retreat involves several key considerations to ensure that the setting supports the retreat's goals and fosters collaboration, relaxation, and productivity [22]. For example, if the focus is academic, such as strategic planning, curriculum development, or research collaboration, a location with quiet meeting spaces and minimal distractions might be best. If the focus is team building or bonding, consider locations that offer outdoor activities or team-building opportunities, such as nature resorts or retreat centers. Proximity and accessibility should be considered. Consider a central location within a reasonable distance easily accessible by car, train, or public transport, especially if it is a 1- or 2-day event or if participants are coming from multiple campuses. If it is remote, ensure there are clear directions and transport options with ample parking for faculty members that are driving.

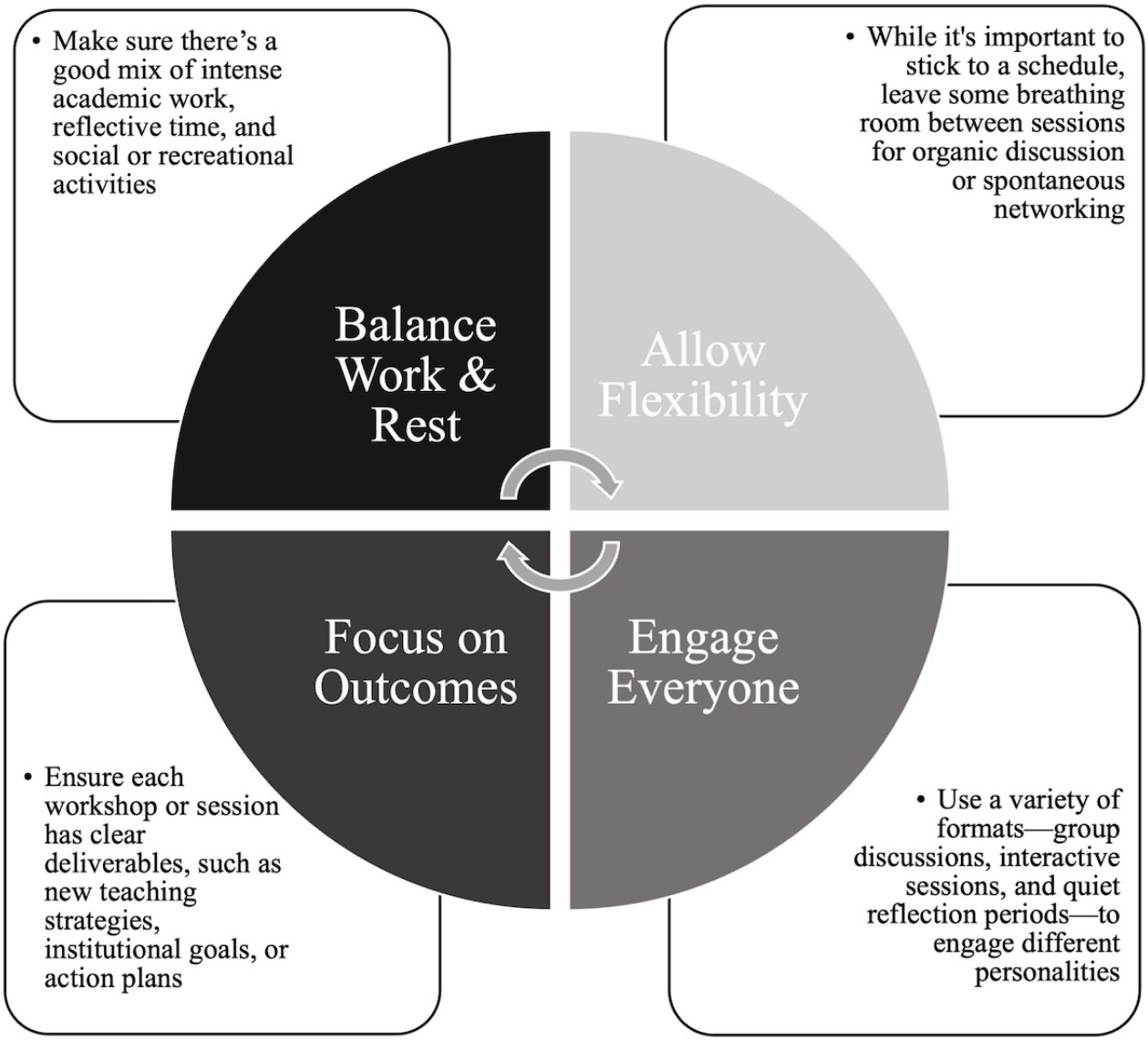
Regarding the actual meeting space and facilities, again consider the size of the group. The venue should have meeting rooms that accommodate the entire group comfortably. Look for spaces with flexible seating arrangements, good acoustics, and appropriate lighting. Also, consider technology needs and ensure that the facility can support presentations or workshops with the necessary audiovisual equipment (eg, projectors, screens, microphones, and Wi-Fi). If smaller group discussions or breakout sessions are planned, look for venues with multiple meeting rooms or areas conducive to group work. If it is a multiday retreat requiring an overnight stay, ensure the location offers comfortable amenities and accommodations for faculty. Confirm dining options available at the facility, including considerations for dietary restrictions. If incorporating relaxation or leisure time, look for venues that offer recreational amenities, such as hiking trails, swimming pools, or wellness facilities. Finally, consider the cost; a venue must fit the budget without compromising on essential needs. Consider that some universities or institutions have partnerships with certain venues or offer discounts for academic groups.

When choosing a location, group dynamics, inclusivity, and environment should also be considered. Consider accessibility for faculty members with mobility issues, and ensure that spaces accommodate various needs (eg, lactation room). If you have a diverse group, make sure the venue respects cultural differences, including dietary needs, religious observances, and inclusive spaces. Make sure the location is available for your preferred dates by planning early as popular retreat centers and venues often book up months in advance. Finally, ambience and environment are important elements to also consider for enhancing faculty engagement and collaboration. Consider seasonal considerations, as outdoor activities might be less enjoyable during rainy or cold seasons, and ensure the venue is prepared for inclement weather. Many faculty retreats benefit from being in scenic or natural settings (eg, mountains, lakesides, or parks) as they promote relaxation and creativity. Some institutions may prefer locations with cultural or academic significance, such as university-affiliated retreat centers, historical sites, or museums.

Preretreat Preparation: Structuring the Agenda and Activities

Before creating the agenda, establish the key objectives (see above). The agenda should be designed explicitly to meet these goals and reflected in any workshops, seminars, or team-building exercises (Figure 1; [23,24]). Allowing the identified key objectives to guide and define retreat content creates a clear path for a more meaningful faculty experience. While noncomprehensive, a list of potential retreat content has been provided in Textbox 4 along with an example retreat agenda (one and a half-day; Table 1).

Figure 1. Key agenda design tips.



Textbox 4. Agenda content ideas.

Professional development workshops

- Effective mentorship across career stages: Techniques for becoming or finding a mentor; case-based discussions.
- Navigating promotion and tenure: Department-specific guidance on academic advancement pathways.
- Time management for academic clinicians: Focused strategies on balancing clinical, teaching, and research responsibilities.
- Grant writing and funding strategies: Tips for National Institutes of Health/Health Resources and Services Administration (NIH/HRSA) submissions, developing specific aims, or finding pilot funds.
- Leadership skills for faculty: Training on conflict resolution, delegation, and leading teams.
- Work-life integration and wellness in academic medicine: Interactive session focused on reducing burnout and enhancing resilience.

Culture-focused sessions

- Implicit bias in clinical and academic settings: Facilitated discussion with activities for self-reflection and behavioral change.
- Building an inclusive department culture: Group exercises around shared values, microaggressions, and allyship.
- Health equity in research and education: Incorporating equity principles in curriculum, research, or patient care.

Strategic planning and innovation

- SWOT (strengths, weaknesses, opportunities, and threats) analysis breakout groups: Analyze departmental strengths, weaknesses, opportunities, and threats.
- Visioning and goal-setting workshops: Define shared goals for the next 1 - 3 years using techniques like appreciative inquiry.
- Innovation challenge or hackathon: Small groups generate and pitch new ideas (eg, curriculum and QI; quality improvement projects).
- Artificial intelligence (AI) in academic medicine (opportunities and risks): A future-facing panel or demo of tools for teaching, research, or clinical decision-making.

Team-building activities

- “Mission Possible” challenge: Groups solve department-relevant scenarios collaboratively (eg, resource allocation or onboarding plans).
- Human bingo or speed networking: Fast-paced way to connect across faculty and discover shared interests.
- Story circles or narrative medicine exercise: Faculty share brief, structured stories on meaningful clinical or academic experiences.
- Escape room or puzzle challenge (in person or virtual): Promotes collaboration and informal interaction.
- Personality type-based communication exercise (eg, MBTI; Myers-Briggs Type Indicator, StrengthsFinder, and DISC; Dominance, Influence, Steadiness, and Conscientiousness) to explore working styles and improve collaboration.

Webinars or prerecorded content

- “Academic Medicine in 2030 and Beyond” (invited guest speaker): Shown at the retreat or asynchronously, followed by discussion groups.
- NIH/funding agency updates: Presented by the internal grants office or invited program officers.
- Faculty wellness microlearning modules: Paired with breakout reflections or personal wellness planning.

Reflection and closure activities

- Commitment cards: Participants write one commitment they will take forward, which is collected and mailed to them months later.
- Gallery walk: Groups post ideas/goals on flip charts around the room and rotate to add or comment.
- Shared success board: Space to acknowledge team wins and individual accomplishments over the past year.

Table . Example retreat agenda (one and a half days).

Day, theme, and time	Session	Description
Day 1: faculty retreat (full day); theme: Advancing Together: Strategy, Scholarship, and Community		
8-8:30 AM	Arrival and Breakfast	• Light breakfast and coffee • Informal networking
8:30-9 AM	Welcome and Overview	• Opening remarks by the Chair or Vice Chair • Goals for the retreat
9-10:15 AM	Strategic Visioning Workshop	• Facilitated SWOT[a] analysis and goal-setting exercise in breakout groups
10:15-10:30 AM	Break	—[b]
10:30 AM-12 PM	Faculty Development Breakouts	• Choice of concurrent sessions • Grant writing tips and resources • Promotion pathways and CV[c] building • Burnout prevention and work-life integration
12-1:00 PM	Lunch (working or social)	• Optional discussion tables (eg, “Mentorship in Medicine” and “DEI[d] in Action”)
1-2:30 PM	Team Challenge: Mission Possible	• Small-group scenario activity focused on real departmental challenges (eg, onboarding, DEI, and teaching loads)
2:30-2:45 PM	Break	—
2:45-4 PM	Psychological Safety and Inclusive Culture Workshop	• Facilitated session with case examples and strategies for fostering trust and belonging
4-4:30 PM	Gallery Walk: Big Ideas Board	• Faculty rotate among flipcharts capturing ideas from earlier sessions and add input
4:30-5 PM	Day 1 Wrap-Up and Preview	• Recap of themes, reflection exercise, and preview of day 2
5-6:30 PM	Optional Social Hour	• Informal reception or outdoor gathering (may include dinner or group activity)
Day 2: Faculty retreat (half day); theme: From Ideas to Action		
8-8:30 AM	Breakfast and Networking	• Casual start with time for reflection
8:30-9:45 AM	Lightning Rounds: Faculty Innovations	• 5-minute presentations by faculty on projects, teaching methods, or QI[e] work
9:45-10:30 AM	Small Group Action Planning	• Based on day 1 priorities (eg, mentorship, wellness, and diversity), groups draft next steps
10:30-10:45 AM	Break	—
10:45-11:30 AM	Department Town Hall	• Leadership addresses submitted questions, strategic updates, and resource planning
11:30 AM-12 PM	Closing Reflections and Commitment Cards	• Faculty write one goal or takeaway • Cards mailed to them in 3-6 months
12 PM	Adjourn	• Optional lunch or grab-and-go meal

[a]SWOT: Strengths, Weaknesses, Opportunities, Threats.

[b]Not applicable.

[c]CV: curriculum vitae.

[d]DEI: diversity, equity, and inclusion.

[e]QI: quality improvement.

Conducting the Retreat

When it comes to conducting the retreat, several elements can ensure that things run smoothly. Primarily, identify a retreat coordinator, typically someone on the planning committee involved in the retreat planning process and familiar with all the retreat elements and plans. These individuals, a faculty or staff member, will be responsible for monitoring the retreat agenda, timekeeping, and serving as a day-of contact for any outside speakers, moderators, or vendors. In addition, these individuals should have the capacity or resources to troubleshoot any potential hiccups, technology-related or otherwise. Depending on the size and scope of the retreat, delegation of specific roles to additional individuals may be required. Clear responsibilities, set ahead of time, allow for smooth session delivery and a more effective retreat.

Another element to consider while conducting the retreat is ongoing assessment of active engagement. Even with careful preretreat needs assessment analysis and planning, engagement can wax and wane during any retreat. As mentioned previously, incorporating a mixture of interactive sessions, small-group discussions, and hands-on workshops to maintain energy and involvement can make engagement drop-off less likely. Depending on retreat length, ample break time and access to refreshments can also assist with energy levels.

Finally, for sessions that might be more interpersonal, it can be important to recognize and acknowledge participant vulnerability and to normalize all responses and emotions. Creating a psychologically safe environment—where individuals feel comfortable expressing themselves without fear of embarrassment, rejection, or retribution—is foundational to fostering trust and vulnerability in retreat settings [25]. Psychological safety has been shown to improve team learning, engagement, and innovation in both health care and academic environments [26,27]. Faculty are more likely to share honest perspectives, disclose challenges, and collaboratively problem-solve when they perceive the environment as respectful, nonjudgmental, and supportive [28-33].

To cultivate psychological safety, facilitators should cocreate ground rules with participants, such as “assume good intent,” “confidentiality is expected,” and “all voices matter.” Setting norms early helps frame the retreat as a shared, inclusive space [34]. Facilitators can further normalize openness by modeling vulnerability themselves, sharing personal experiences or acknowledging areas of uncertainty, which signals that it is safe to take interpersonal risks.

Evidence-based facilitation techniques also enhance psychological safety. These include structured turn-taking (eg, round-robin formats), the use of anonymous input tools (eg, polling apps or sticky-note activities), and small group breakouts that lower the stakes for participation [35]. Active facilitation, such as naming group dynamics, gently redirecting dominant voices, and explicitly inviting quieter participants to share, further supports inclusion. Attending to emotional cues, validating contributions, and pacing sessions to allow reflection all contribute to a respectful tone and a sense of collective care.

By embedding these practices into the retreat design, organizers can create the conditions for open dialogue, team cohesion, and shared commitment to change. Further, anticipating common retreat challenges and potential solutions in advance can prove invaluable for both planners and participants (Textbox 5).

Textbox 5. Common retreat challenges and solutions.

Budget constraints

- Consider solicitation of either institutional or departmental support [28]. This funding source typically requires early engagement of high-level leadership and stakeholders along with well-defined anticipated outcomes. In the absence of a leadership-led retreat, this strategy may take initiative on the part of other faculty members.
- Self-funding or donation funding can be considered. Creation of a "Faculty Well-Being Fund" or "Development Fund" with voluntary donations can be a strategy option. While potentially effective, this approach could result in participation bias, with self-selected faculty or donors more likely to engage [29].
- Focus on local resources and simple activities, which do not require extensive budgets. Similarly, consider if a virtual retreat might be a sufficient option.
- Finally, consider holding the retreat at work during the workweek. While this limits the impact of an external environment, it might allow any existing budget to be nonetheless maximized [30].

Schedule conflicts

- Tagged, protected time can be considered. This is an often-used strategy for trainees, typically resident, retreats. This might be more challenging for academicians with 24-hour call or service schedules.
- Careful, extreme advanced planning and faculty notification, to avoid busy times of the year, whether seasonal or department-specific busier times, may help with turnout and engagement [33].
- Limiting the length of a retreat, generally recommended not to extend beyond 2 days, can also limit schedule conflicts [30].

Participant engagement

- When budget and logistics allow, consider separating the retreat site from work, as this may prevent work-related distractions and allow the group to stay on point [32].
- If the retreat is held at a destination or vacation location and families attend as well, consider limiting attendance at the actual retreat to faculty only, again to discourage distractions and promote focus. Family engagement may be appropriate at associated retreat social events at the organizers' discretion.
- If occurring outside of usual work hours or expectations, consider whether provision of additional compensation may be possible. A notable limitation to consider here includes budget and introduction of participant bias.
- If the agenda allows, incorporation of social events is heavily encouraged. Scheduled social events can make it feel like an actual "retreat" from work. These events will also foster faculty camaraderie [30].
- Make sure there's interest. Consider a preretreat needs assessment to gauge. Careful assessment of responses can avoid the subsequent expenditures of unnecessary efforts.

Hybrid model considerations

- To avoid participant inequity, adopt a "remote-first" mindset when planning content and facilitation strategies. Ensure all materials are digital and accessible in real time. Design activities that are inclusive for all participants, not just those in the room.
- Designate a cofacilitator or team member to advocate for and monitor remote participant inclusion.
- Intentionally structure breakout groups to mix in-person and virtual participants or keep them separate but equally resourced.
- Create informal spaces and scheduled time for social interaction across modalities.
- Include shared virtual coffee breaks, games, or team-building sessions where all participants engage in a similar experience, ideally via a single platform.
- Document key discussions and decisions in real time on shared platforms.

Agenda Pitfalls

- Avoid overpacking an agenda. If time is limited, concentrate on a singular objective directed by the needs assessment prioritization [33].
- Steer clear of a dull agenda. Avoid didactic-like structure [30].
- Postretreat neglect.
- Designate a "retreat secretary" to keep and distribute minutes. Establish an action item list, with deadlines, for identified postretreat interventions [33].
- Conduct a postretreat assessment; solicit feedback.

Postretreat Considerations

The conclusion of the retreat marks not the end, but the continuation of the organizers' and leaders' responsibilities. To be most effective, a good retreat involves solicitation and consideration of participant feedback. During the planning process, retreat objectives and outcomes should have been defined to outline how success will be measured. Tailor subsequent postretreat feedback instruments so that reviews can be considered in the context of the stated retreat goals and objectives. For instance, if team building is a goal, consider postretreat surveys to gauge improved collaboration. Feedback instruments can be survey-based and/or involve focus groups, like the process used for preretreat needs assessment data gathering. Consider both quantitative and qualitative feedback content. In addition, organizers should capitalize on the groundwork laid or the momentum gained by the retreat and seek future opportunities to weave positive retreat outcomes into continuous improvement and ongoing development (Multimedia Appendix 1).

To ensure that momentum continues beyond the retreat, organizers should establish structured follow-up mechanisms that translate insights into action. Begin by forming small working groups around priority areas identified during the retreat. These groups should be assigned clear deliverables and timelines and report progress during faculty meetings or through a shared digital platform. Designating retreat "champions" or coleads within each group can help maintain accountability and enthusiasm. Within 1 month post retreat, schedule a follow-up meeting to review early progress, reinforce shared goals, and recalibrate as needed. In addition, consider incorporating retreat themes into existing faculty development programs or launching targeted initiatives—such as peer mentoring circles, writing groups, or leadership development cohorts—aligned with retreat goals. Finally, build in evaluation checkpoints at 3, 6, and 12 months to assess progress, solicit feedback, and celebrate milestones. These actions foster sustained engagement and embed retreat outcomes into the academic fabric of the department.

Timeline Considerations

Although mentioned above, it deserves reiterating, as a final note in the planning process, that implementation of any retreat takes significant time, from organizing and conducting preretreat needs assessment through to evaluating and considering postretreat feedback and impact. When possible, at the onset, retreat planners should attempt to develop a "retreat timeline" (Textbox 6) that considers each aspect of the retreat development and implementation process. This timeline, which can doubly serve as a retreat checklist, should be retreat-specific and will be highly dependent on the anticipated breadth and expanse of the retreat. Dynamic timeline adjustments may need to be made pending results of preretreat assessments or other variables, including desired venue availability, participant work schedule, etc.

Textbox 6. Example retreat timeline.

6 months (or more) in advance

- Announcement of Faculty Retreat Intentions
- Open Invitation to Self-Nominate Organizing Committee

4-6 months before

- Preretreat needs assessment and analysis

3-4 months before

- Agenda creation
- Save-the-date creation to participants
- Speaker/moderator invitation
- Location and accommodation reservations

1-2 months before

- Agenda confirmation
- Formal faculty invitation and RSVP (répondez s'il vous plaît [French])

1 week after

- Immediate postretreat feedback solicitation and consideration

1-3 months after

- Feedback follow-up
- Action items follow-up

Case Studies and Examples

Reviewing examples of prior faculty retreats in academic medicine can be particularly valuable and can demonstrate both areas of success and lessons learned. In 2022, Lee et al [36] published about their experience with the Abdominal Radiology Division at Harvard Medical School, Brigham and Women's Hospital. Held in 2021, this retreat aimed to specifically address division stressors and challenges experienced during the COVID-19 pandemic. Twenty-eight faculty participated in a 2 and a half-hour virtual retreat via Zoom (Zoom Communications). A preretreat survey was used to establish discussion topics as well as determine current faculty satisfaction. A postretreat survey assessed retreat effectiveness and faculty satisfaction regarding stated objectives. With regard to limitations, it was suggested that the size of the larger group (N=28) may have caused hesitancy in discussion responders, and the authors also noted that the retreat was relatively short in duration. However, the authors ultimately concluded that the retreat was successful with regards to several outcomes, including promoting group camaraderie, provisioning ideas to improve the work environment (eg, adjusting shift times or coverage), and identifying specific faculty academic needs (eg, mentorship and sponsorship).

Birx et al [37] published their experience with a retreat held for nursing faculty at Radford University. Specifically designed within a positive psychology framework, the stated retreat goal was to build faculty relationships to create a cohesive team and to accomplish institutional goals. Held at a nature conservatory owned by the university and involving 29 nursing faculty, the daylong team-building event used challenge course activities, described as "an experiential adventure program…involving mental, physical, and emotional risk taking." The authors used a mixed method evaluation, both pre- and post retreat, considering qualitative and quantitative data. They also incorporated an extended postevaluation, which was completed at the end of the semester to consider potential longer-term effects of the retreat. The authors concluded that the retreat offered an immediate positive effect on their targeted outcomes; however, these were not necessarily sustained throughout the semester, suggesting that a one-time intervention may not be sufficient to impart continued impact. This experience highlights the importance of building on the momentum achieved at a retreat and using it as a foundation for ongoing development.

While previously unpublished, the authors of this tutorial also recently planned and implemented a faculty retreat. Of note, a paucity of available free "how to" content to guide the authors was the impetus for this tutorial paper as well as a specific scholarly objective of the held retreat. The inaugural EmpowerED Faculty Retreat for women in the University of Alabama at Birmingham (UAB) Department of Emergency Medicine was held September 4 - 6, 2024, at a resort and outdoor leisure destination. Designed to foster leadership development, professional growth, wellness, and team building among faculty, the retreat featured didactic sessions, small group discussions, and wellness activities in a reflective, off-campus setting. A postretreat survey, completed by 11 of the 12 participants, revealed high levels of satisfaction and strong support for continuing this initiative. All respondents agreed that the retreat met or exceeded their expectations, citing a well-balanced agenda that integrated learning, networking, and self-care. The workshops—Dare to Dialogue: Navigating Difficult Conversations and Microskills for Professional Success—were particularly well received. Nearly 3 quarters of attendees reported applying retreat content within 1 month, including strategies for communication, time management, and relationship-building. Participants praised the retreat location and logistics, noting that the peaceful setting supported creativity and deeper connection. However, several suggested that centralized lodging would have enhanced informal interaction, particularly in the mornings and between sessions. Feedback also highlighted a desire for more small group engagement, 1:1 mentoring opportunities, and dedicated time to discuss department-specific topics. Looking ahead, faculty expressed strong interest in future retreats, identifying topics, such as negotiating, attaining promotion, and mentorship as high priorities. The most favored strategies for sustaining the work of the retreat included annual offsite gatherings, quarterly check-ins, and peer mentoring. Overall, the EmpowerED retreat was viewed as a transformative and affirming experience, supporting both individual development and a stronger sense of community within the department.

Future Directions and Emerging Trends

As models of faculty development, including retreat incorporation, continue to expand and evolve, several recent trends have been noted. These trends align with the current state of medicine, reflecting key developments, such as the emphasis on clinician well-being and integrating the growth of advanced technologies in both clinical practice and academic settings.

The COVID-19 pandemic was associated with a record-high physician burnout rate. While the American Medical Association (AMA) reports that these numbers have improved more recently, percentage rates remain astonishingly high, over 50% for some subspecialties [38]. There has been a noted emphasis on integrating wellness, mindfulness, resiliency training, and structured time for reflection into today's faculty and trainee retreats [32,39].

Considering advanced technology, faculty development can both use and integrate content specific to emerging technology, particularly artificial intelligence (AI), data analytics, and virtual platforms, which offer new opportunities to enhance the planning, delivery, and impact of faculty retreats. AI-driven tools may be used to streamline various aspects of retreat implementation. For example, natural language processing tools can analyze preretreat survey responses to identify common themes and tailor agendas to faculty needs. AI scheduling assistants can coordinate availability across large faculty groups, reducing the administrative burden on organizers. During the retreat itself, AI-powered collaboration platforms (eg, Miro with AI clustering features or Microsoft Copilot) can help synthesize live input from participants, identify emerging priorities, and generate summaries in real time. In the realm of personalized learning, AI-enabled learning management systems can recommend postretreat faculty development resources aligned

with individual career goals, teaching portfolios, or scholarly interests. These platforms may also track progress and provide nudges for continued engagement, thereby extending the retreat's value over time.

Virtual and hybrid retreat models, initially born out of necessity, now offer sustainable and scalable solutions for faculty development—particularly in geographically dispersed departments. With the integration of AI-enhanced virtual facilitators, breakout room optimization, and sentiment analysis, retreat organizers can gain immediate feedback and adjust facilitation strategies accordingly. When appropriate, a hybrid approach to a retreat (combining in-person and virtual participation) might allow for greater flexibility, particularly when inviting outside speakers. Virtual or hybrid models may also have the added benefit of decreasing overall budget. When considering the incorporation of virtual components, it is important to ensure that the audiovisual technology is high-functioning, and a moderator (in person, if hybrid) should facilitate virtual speakers [40]. As these technologies continue to mature, their thoughtful application will be critical to ensuring that innovation complements, rather than replaces, the relational and reflective core of faculty retreats. Further, a careful consideration of outcome comparisons in virtual versus hybrid versus in-person retreats should be a consideration for future analysis.

Conclusion

In conclusion, professional development for academic medicine faculty is a goal-directed, continuous effort pursued over one's entire career span and is vital to improve and hone skills. Faculty retreats, an increasingly popular form of professional development, can lead to improved faculty performance, effectiveness, and wellness. Initial retreat planning should include a needs assessment from the targeted audience, the results of which should be used to define retreat objectives. Then, SMART goals should be set to guide the retreat's agenda and outcomes. In order to meet retreat objectives, other important considerations include resource allocation (including personnel coordination and oversight) and participant selection, as well as location and setting choice. Following the event, solicited postretreat feedback can translate positive retreat outcomes into ongoing faculty improvement and development. Faculty retreats can represent a diverse, expansive, and valuable development tool for academic medicine. Careful planning consideration and subsequent implementation can help ensure their success.

Acknowledgments

We wish to thank Professor Resa E Lewiss for conceptualizing this work. Their intellectual contribution and guidance in shaping the research direction were invaluable to this project.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Example postretreat assessment.
[DOCX File, 35 KB - mededu_v11i1e71622_app1.docx]

References

1. Brown GM, Lang E, Patel K, et al. A national faculty development needs assessment in emergency medicine. CJEM 2016 May;18(3):161-182. [doi: 10.1017/cem.2015.77] [Medline: 26350557]
2. Pandit K, Coates WC, Diercks D, Gupta S, Siegelman J. Faculty development for academic emergency physicians: a focus group analysis. Cureus 2022 Aug;14(8):e27596. [doi: 10.7759/cureus.27596] [Medline: 36059367]
3. Steinert Y, Mann K, Anderson B, et al. A systematic review of faculty development initiatives designed to enhance teaching effectiveness: a 10-year update: BEME Guide No. 40. Med Teach 2016 Aug;38(8):769-786. [doi: 10.1080/0142159X.2016.1181851] [Medline: 27420193]
4. Cable CT, Boyer D, Colbert CY, Boyer EW. The writing retreat: a high-yield clinical faculty development opportunity in academic writing. J Grad Med Educ 2013 Jun;5(2):299-302. [doi: 10.4300/JGME-D-12-00159.1] [Medline: 24404277]
5. Buffington ALH, Lange C, Bakker C, et al. The collaborative scholarship intensive: a research-intensive course to improve faculty scholarship. Fam Med 2021 May;53(5):355-358. [doi: 10.22454/FamMed.2021.534614] [Medline: 34019681]
6. Bligh J. Faculty development. Med Educ 2005 Feb;39(2):120-121. [doi: 10.1111/j.1365-2929.2004.02098.x] [Medline: 15679676]
7. Hargreaves A, Fullan M. Mentoring in the new millennium. Theory Pract 2000 Feb;39(1):50-56 [FREE Full text] [doi: 10.1207/s15430421tip3901_8]
8. Shanafelt T, Goh J, Sinsky C. The business case for investing in physician well-being. JAMA Intern Med 2017 Dec 1;177(12):1826-1832. [doi: 10.1001/jamainternmed.2017.4340] [Medline: 28973070]
9. Culture-driven team building: building high-performing teams. Penn Online Learning. URL: https://platform.onlinelearning.upenn.edu/offering/culture-driven-team-building-building-high-performing-teams-a0Q2E00000JmMNWUA3 [accessed 2024-10-31]

10. Thibault GE, Neill JM, Lowenstein DH. The Academy at Harvard Medical School: nurturing teaching and stimulating innovation. Acad Med 2003 Jul;78(7):673-681. [doi: 10.1097/00001888-200307000-00005] [Medline: 12857684]
11. Organizing a retreat. Community Tool Box. URL: https://ctb.ku.edu/en/table-of-contents/structure/training-and-technical-assistance/retreats/main [accessed 2024-10-31]
12. Advancing by retreating. SERC. URL: https://serc.carleton.edu/departments/heads_chairs/retreat.html#:~:text=A%20faculty%20retreat%20can%20be,and%20facilitate%20discussion%2C%20and%20camaraderie [accessed 2024-10-21]
13. Viera AJ, Kramer R. Management and Leadership Skills for Medical Faculty: A Practical Handbook: Springer; 2016. [doi: 10.1007/978-3-319-27781-3]
14. Retreat planning guide. : University of Cincinnati; 2024 URL: https://www.uc.edu/content/dam/refresh/studentaffairs-62/student-activities-leadership-development/student-org/RetreatPlanningGuide_SALD.pdf [accessed 2024-10-31]
15. Siow Y, Scot M, Darabi H, Mashayek F. A faculty retreat model featuring collaborative and active learning. Presented at: 2019 ASEE IL-IN Section Conference URL: https://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1060&context=aseeil-insectionconference [accessed 2025-10-24]
16. Centers for Disease Control. Assessing training needs: conducting needs analysis. URL: https://www.cdc.gov/training-development/php/about/assess-training-needs-conducting-needs-analysis.html [accessed 2024-10-31]
17. Seven steps for conducting a successful needs assessment. NICHQ. URL: https://nichq.org/blog/seven-steps-conducting-successful-needs-assessment [accessed 2024-10-31]
18. Moser K. A step-by-step guide: how to plan a company retreat. Go Gather. URL: https://gogather.com/blog/guide-to-planning-a-successful-company-retreat [accessed 2024-10-31]
19. Bjerke MB, Renger R. Being SMART about writing SMART objectives. Eval Program Plann 2017 Apr;61:125-127. [doi: 10.1016/j.evalprogplan.2016.12.009] [Medline: 28056403]
20. Wingo L. 10 smart ways to maximize your company retreat budget. CO— by US Chamber of Commerce. URL: https://www.uschamber.com/co/run/human-resources/company-retreat-budget-planning#:~:text=If%20you%20don’t%20know,Consider%20team%20size [accessed 2024-10-31]
21. Planning for education retreats: the ultimate guide. Hobnob. URL: https://hobnob.app/planning-for-educational-retreats-the-ultimate-guide [accessed 2024-10-31]
22. Planning the perfect faculty retreat: steps to foster collaboration and innovation. Edith Macy Center. URL: https://www.edithmacy.com/blog/planning-a-faculty-retreat [accessed 2024-10-31]
23. Smorynski HW. Using academic retreats to enhance academic affairs performance. Academic Briefing. URL: https://www.academicbriefing.com/human-resources/faculty-development/using-academic-retreats/#:~:text=Finally%2C%20retreats%20are%20a%20time,of%20the%20university’s%20academic%20affairs [accessed 2024-10-31]
24. Patel S, O’Brien BC, Dulay M, Earnest G, Shunk RL. Team retreats for interprofessional trainees and clinic staff: Accelerating the development of high-functioning teams. MedEdPORTAL 2018 Dec 21;14:10786. [doi: 10.15766/mep_2374-8265.10786] [Medline: 30800986]
25. A guide to building psychological safety on your team. Harvard Buisiness Review. URL: https://hbr.org/2022/12/a-guide-to-building-psychological-safety-on-your-team [accessed 2024-10-31]
26. Edmondson A. Psychological safety and learning behavior in work teams. Adm Sci Q 1999 Jun;44(2):350-383. [doi: 10.2307/2666999]
27. Newman A, Donohue R, Eva N. Psychological safety: a systematic review of the literature. Hum Resour Manag Rev 2017 Sep;27(3):521-535. [doi: 10.1016/j.hrmr.2017.01.001]
28. O’Shea J, Dannenfelser M, White M, Osborne A, Moran TP, Lall MD. A resident retreat with emergency medicine specific mindfulness training significantly reduces burnout and perceived stress. Journal of Wellness 2022;4(1):3. [doi: 10.55504/2578-9333.1114]
29. Bailey AK, Sawyer AT, Tao H, et al. Evaluating the feasibility and impact of a well-being retreat for physicians and advanced practice providers. Journal of Wellness 2023;5(1):9. [doi: 10.55504/2578-9333.1186]
30. Ponomarenko J, Garrido R, Guigó R. Ten simple rules on how to organize a scientific retreat. PLOS Comput Biol 2017 Feb;13(2):e1005344. [doi: 10.1371/journal.pcbi.1005344] [Medline: 28151954]
31. The ultimate guide to corporate retreat planning for success and team bonding. Geneva Point Center. URL: https://www.genevapoint.org/the-ultimate-guide-to-corporate-retreat-planning-for-success-and-team-bonding [accessed 2024-10-31]
32. Egan DJ, He C, Leslie Q, Clark MA, Lewiss RE. The emergency medicine resident retreat: creating and sustaining a transformative and reflective experience. Cureus 2022 Aug;14(8):e27601. [doi: 10.7759/cureus.27601] [Medline: 36059321]
33. Hills L. Team building organizing a practice retreat. J Med Pract Manage 2020;19(2):97-99 [FREE Full text]
34. Edmondson AC, Lei Z. Psychological safety: the history, renaissance, and future of an interpersonal construct. Annu Rev Organ Psychol Organ Behav 2014 Mar 21;1(1):23-43 [FREE Full text] [doi: 10.1146/annurev-orgpsych-031413-091305]
35. Frazier ML, Fainshmidt S, Klinger RL, Pezeshkan A, Vracheva V. Psychological safety: a meta‐analytic review and extension. Pers Psychol 2017 Feb;70(1):113-165 [FREE Full text] [doi: 10.1111/peps.12183]

36. Lee LK, Krajewski KM, Suarez-Weiss KE, Silverman SG, Shinagare AB. Learning from experience: confronting challenges and adapting to change in a large academic abdominal radiology practice: insights from a faculty retreat. Curr Probl Diagn Radiol 2022;51(6):818-822. [doi: 10.1067/j.cpradiol.2022.06.002] [Medline: 35842346]
37. Birx E, Lasala KB, Wagstaff M. Evaluation of a team-building retreat to promote nursing faculty cohesion and job satisfaction. J Prof Nurs 2011;27(3):174-178. [doi: 10.1016/j.profnurs.2010.10.007] [Medline: 21596358]
38. Burnout falls, but still hits these 6 physician specialties most. American Medical Association. URL: https://www.ama-assn.org/practice-management/physician-health/burnout-falls-still-hits-these-6-physician-specialties-most [accessed 2023-10-31]
39. Cornelius A, Cornelius BG, Edens MA. Increasing resident wellness through a novel retreat curriculum. Cureus 2017 Jul 28;9(7):e1524. [doi: 10.7759/cureus.1524] [Medline: 28966896]
40. Lewis AR, Choong GM, Cathcart-Rake E, et al. Preparing hematology/oncology fellows for success: implementing an annual career development and research retreat. J Canc Educ 2024 Feb;39(1):58-64. [doi: 10.1007/s13187-023-02375-9]

Abbreviations

AI: artificial intelligence
AMA: American Medical Association
SMART: Specific, Measurable, Achievable, Relevant, and Time-bound
UAB: University of Alabama at Birmingham

Edited by T Gladman; submitted 22.01.25; peer-reviewed by AS Lee, JCC Chen; revised version received 02.08.25; accepted 28.08.25; published 03.11.25.

Please cite as:
Skains R, Brown J, Shufflebarger EF, McGiboney J, Hicks S, McDonald L, Griesmer KB, Shaw C, Grass E, Elie MC, Walter LA
Faculty Retreats in Academic Medicine: Tutorial
JMIR Med Educ 2025;11:e71622
URL: https://mededu.jmir.org/2025/1/e71622
doi:10.2196/71622

Tutorial

Twelve Practical Tips for Integrating AI Into Medical Education: Tutorial to Support Educators Across Teaching, Research, Administration, and Ethical Domains

Alireza Jalali[1], MD; Kadidja Harbi Houssein[1], BSc; Salomon Fotsing[1], MD, MA Ed

Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada

Corresponding Author:
Alireza Jalali, MD
Faculty of Medicine
University of Ottawa
451 Smyth Rd
Ottawa, ON, K1H 6H8
Canada
Phone: 1 613 562 5800
Email: ajalali@uottawa.ca

Abstract

Artificial intelligence (AI) is rapidly reshaping medical education, offering new opportunities to personalize learning, enhance research, and streamline administration. The aim of this study is to provide 12 practical, evidence-informed tips by drawing on current literature and real-world examples to guide the integration of AI into medical education, supporting educators across teaching, research, administration, and ethical domains. Key strategies include using adaptive learning platforms to tailor educational content, using AI tools to provide timely feedback, and incorporating AI-generated clinical scenarios in case-based learning. The importance of fostering AI literacy among students is emphasized, as well as utilizing AI-powered tools for efficient literature reviews, data analysis, and manuscript preparation. Administrative applications such as automating routine tasks, supporting strategic planning through data analysis, and enhancing faculty development with AI-driven platforms are also discussed. Ethical considerations are highlighted, with a focus on ensuring transparency, fairness, and accountability in all AI applications. By following these 12 tips, medical educators can leverage the benefits of AI to improve educational outcomes, increase efficiency, and prepare future clinicians for a technology-driven health care environment.

(JMIR Med Educ 2025;11:e81297) doi:10.2196/81297

KEYWORDS

artificial intelligence; ethics; educational innovation; AI literacy; medical education; educational technology

Introduction

Artificial intelligence (AI) is rapidly transforming medical education. It presents new opportunities for personalized learning, enhanced feedback, and streamlined administrative processes. Its impact is clear across multiple areas in medical schools, including medical student teaching, research, administrative governance, leadership paradigms, and ethical considerations.

In teaching, AI facilitates adaptive educational tools, intelligent clinical simulators, and automated assessment platforms. These technologies allow greater personalization of learning, helping students acquire clinical competencies in realistic simulated environments. Illustrative examples include intelligent tutoring systems and virtual patient simulators, which permit students to engage in risk-free clinical decision-making. Nevertheless, the integration of these tools into established curricula and the requisite faculty training present notable pedagogical challenges [1,2].

In the research field, AI serves as a catalytic agent in research by expediting the analysis of extensive biomedical datasets, accelerating the discovery of novel therapeutic interventions, and advancing personalized medicine. University researchers employ machine learning algorithms to discern intricate correlations within genomic, clinical, and imaging data. AI also automates labor-intensive tasks such as literature reviews and cohort selection, thereby allocating more time for critical analysis and experimental design. These advancements are substantiated by extensive work from institutions like Inserm (French National Institute of Health and Medical Research) and in research pertaining to health data integration [3,4].

On the administrative and leadership front, AI is redefining governance practices within medical education institutions. Its applications include optimizing human resource management, scheduling academic timetables, forecasting training requirements, and supporting strategic decision-making. Academic leaders are thus compelled to cultivate enhanced digital literacy to effectively oversee the ethical and efficient implementation of these technologies. The Quebec AI in Health Master Plan provides a pertinent example, outlining a strategic roadmap for AI integration across health and educational entities [5,6].

Finally, ethical considerations assume a central role in the integration of AI within medical education. Key issues encompass algorithmic transparency, the safeguarding of personal data for students and simulated patients, and the mitigation of potential biases within evaluation or diagnostic systems. These concerns underscore the critical need for reinforced ethical training for aspiring physicians, ensuring their responsible utilization of AI tools. Recent scholarship has identified over 70 ethical challenges associated with AI in education—many directly relevant to medical training [7,8]. Despite the substantial opportunities and expansive potential afforded by AI across these 4 sectors of medical education, a significant impediment persists: a considerable number of medical educators lack adequate training and resources for effective AI integration into their pedagogical practices. This deficiency is characterized by low digital literacy, a limited understanding of algorithmic principles, and a hesitancy to adopt tools perceived as complex or ethically ambiguous [7]. Furthermore, empirical studies indicate a pervasive sense of disempowerment among educators, attributable to the absence of specialized training and clear guidelines [2]. This paper will therefore delineate 12 practical recommendations designed to facilitate the integration of AI for the advancement of student instruction, research endeavors, administrative and leadership functions, and ethical considerations within medical education. Each recommendation appears to be informed by contemporary literature and best practices.

Teaching

Tip 1: Use AI to Personalize Learning Paths

Adaptive learning platforms powered by AI can tailor educational content to individual student needs, addressing unique learning requirements and competency rates. Sharma et al [9] highlight that adaptive learning is the “final piece of technology-enhanced learning” in medical education, demonstrating significant improvements in understanding and engagement. Commercial implementations such as Elsevier’s Cerego and McGraw Hill Education’s partnership exemplify the practical benefits of AI-driven personalization. Kellman’s [10] research on adaptive learning further supports this, showing educational improvements through repeated knowledge delivery and mastery criteria, with clinical trials in dermatology histopathology at University of California, Los Angeles, reporting significant pretest and posttest score improvements (P<.001).

Tip 2: Enhance Feedback With AI Tools

Building on this personalization, AI can also strengthen another essential component of learning: feedback. AI tools provide timely, formative feedback during clinical simulations, fostering clinical judgment and reflective practice. Howell’s [11] study on AI-enhanced debriefing demonstrates that AI can generate real-time, personalized feedback and Socratic questions, leading to improved learner outcomes and knowledge retention. Practical applications such as Western Technical College’s AI chatbot Jennifer West loaded with open education resources and health care simulation standards illustrate how AI can support consistent, high-quality debrief sessions. Facilitators benefit by focusing on student interaction, while students report enhanced self-reflection and peer feedback.

Tip 3: Integrate AI Into Case-Based Learning

Expanding on the role of AI in enhancing feedback, the next tip explores how AI can be integrated into case-based learning to further enrich clinical education. AI-generated clinical scenarios and virtual patients enrich problem-based learning by creating interactive, adaptive environments. The Diagnostic Reasoning and Its Development Group, Inc Group’s (2025) [12] research shows that AI-driven virtual patients can simulate real emotions, symptoms, and conditions, responding dynamically to student actions. This approach enhances clinical decision-making and interview skills in a safe, controlled setting. AI-powered tools prompt critical self-reflection such as asking, “What was your rationale for the administration of metoprolol?” or “How could holding the metoprolol change the patient’s outcome?”

Tip 4: Train Students in AI Literacy

As AI becomes increasingly embedded in clinical education, it is equally important to ensure that students are equipped to use these tools responsibly. The next tip focuses on developing AI literacy to prepare learners for ethical and effective engagement with emerging technologies. Preparing students to critically evaluate and use AI tools is essential. The American Medical Association (2023) [13] emphasizes the importance of AI literacy in medical education, advocating for curricula that teach students to assess AI applications ethically and effectively. Although specific studies on AI literacy training are limited, the broader literature recognizes its critical role in preparing future clinicians for technology-driven health care environments.

Research

Tip 5: Accelerate Literature Reviews With AI

AI-powered summarization and search tools streamline systematic reviews by processing vast amounts of literature efficiently. Sharma et al [9] note that technology-enhanced learning has risen due to easy internet access, supporting the logical extension of AI tools for research. AI-powered platforms such as Perplexity.ai can quickly retrieve, summarize, and synthesize academic papers, accelerating the research process and helping summarize and synthesize relevant academic sources.

Tip 6: Employ AI in Data Analysis

Beyond literature reviews, AI contributes to research through powerful tools for data analysis. Machine learning algorithms can analyze complex educational datasets, predict learner performance, and optimize content delivery [14]. The success of adaptive learning platforms in analyzing educational data patterns suggests similar benefits for research datasets in medical education [15].

Tip 7: Collaborate With AI for Writing and Publishing

After supporting data analysis, AI can also help researchers with writing and publishing by assisting in drafting, editing, and formatting manuscripts. AI tools assist in drafting, editing, and formatting manuscripts, improving academic writing efficiency. Akbar's [16] mixed methods study on doctoral students' use of AI tools provides foundational evidence for AI's role in academic research and writing processes. Researchers can use AI to brainstorm ideas, draft initial versions, and improve formatting but should ensure their final manuscript reflects their own critical thinking and contributions.

Administration and Leadership

Tip 8: Automate Routine Administrative Tasks

AI can automate scheduling, email management, and documentation, freeing faculty and staff for higher-value activities [17]. Although specific research on administrative automation in medical education is limited, the broader technology-enhanced learning literature supports its potential for efficiency gains [18].

Tip 9: Use AI for Strategic Planning

Alongside simplifying daily administrative tasks, AI can guide strategic planning. By analyzing institutional data, it supports better decisions for curriculum design and resource use. AI can analyze institutional data to inform curriculum development and resource allocation. Bixler and Ceballos's [19] conceptual model explores how AI supports educational leadership and decision-making, demonstrating its utility for strategic planning in medical education.

Tip 10: Enhance Faculty Development With AI

In addition to guiding strategic planning, AI can also play a key role in supporting educators themselves. Through adaptive tools, it provides personalized learning and feedback to strengthen faculty development. AI-driven platforms can provide personalized learning and feedback for faculty, mirroring the benefits seen in student learning. The principles of adaptive learning apply equally to faculty professional development, enabling customized experiences and ongoing support [20].

Ethical Considerations

Tip 11: Uphold Ethical Use of AI

Transparency, fairness, and accountability must guide all AI applications in medical education. Weidener and Fischer's [21] scoping review of AI ethics teaching in medical education highlights the importance of ethical instruction and practice. The American Medical Association (2023) [13] further reinforces the need for ethical standards, advocating for the responsible integration of AI. Additionally, educators should consider nuanced challenges such as the impact of AI-generated assessments or content creation on student autonomy and academic integrity, ensuring that AI supports rather than replaces critical thinking and professional development.

Future

Tip 12: Stay Informed and Adaptive

Maintaining ethical standards goes hand in hand with staying up-to-date on AI's rapid developments. By continuously learning, educators can adapt responsibly and uphold best practices in medical education. The rapid evolution of AI in medical education necessitates continuous learning and adaptation. Sharma et al [9] emphasize that ongoing attention to emerging developments is essential, as the field is likely to change significantly over time.

Practical Tips for Integrating AI Into Medical Education

Table 1 outlines 12 actionable strategies (Multimedia Appendix 1) for effectively integrating AI into medical education, with suggested tools and expected outcomes.

Table 1. Twelve practical tips for integrating AI[a] into medical education.

Tip title	Description	Example tools/platforms	Key outcomes
Tip 1: Use AI to personalize learning paths	Adaptive platforms tailor content to individual student needs, improving engagement and mastery	• Cerego • Smart • Sparrow • McGraw Hill ALEKS	• Improved learner engagement • Personalized progression
Tip 2: Enhance feedback with AI tools	AI delivers immediate, personalized feedback in clinical simulations, promoting reflective learning	• Jennifer West (AI chatbot) • SimConverse • FeedbackFruits	• Accelerated skill development • Deeper reflection
Tip 3: Integrate AI into case-based learning	AI generates dynamic clinical scenarios and virtual standardized patients for realistic decision-making practice	• Diagnostic Reasoning and Its Development Group, Inc Clinician • Body Interact • Open-Source Clinical Application Resource	• Improved diagnostic reasoning • Clinical readiness
Tip 4: Train students in AI literacy	Teaching AI literacy prepares students to evaluate, interpret, and use AI tools ethically and effectively	• Artificial Intelligence for Health (AI4HealthEd) • American Medical Association AI Curriculum • Google • Teachable Machine	• Greater digital competency • Ethical awareness
Tip 5: Use AI to accelerate literature reviews	AI simplifies literature searches and synthesis, saving time and broadening evidence coverage	• Perplexity.ai • Elicit • ResearchRabbit	• Faster research preparation • Improved evidence integration
Tip 6: Use AI for data analysis	AI analyzes complex datasets to predict performance and optimize learning strategies	• RapidMiner • Orange • IBM SPSS Modeler	• Data-driven decision-making • Targeted interventions
Tip 7: Collaborate with AI for writing and publishing	AI assists in drafting, editing, and formatting manuscripts while ensuring responsible authorship and academic integrity	• Grammarly • ChatGPT • Writefull	• Enhanced writing quality • Efficient publishing workflow
Tip 8: Automate routine administrative tasks	AI manages scheduling, email correspondence, and documentation to free educators' time for strategic work	• Explainable Artificial Intelligence (x.ai) • Clara • Google Duplex	• Increased faculty productivity • Streamlined operations
Tip 9: Use AI for strategic planning	AI analyzes institutional data to inform curriculum design, forecasting, and resource allocation	• Tableau • IBM Watson • Power Business Intelligence	• Informed planning • Optimized resource management
Tip 10: Enhance faculty development with AI	AI platforms deliver adaptive learning and feedback for continuous professional growth	• LinkedIn Learning • Coursera AI Tracks • Education Application (EdApp)	• Ongoing educator improvement • Personalized learning
Tip 11: Uphold ethical use of AI	Implement transparent, fair, and accountable AI practices in all academic and clinical settings	• Ethics guidelines (American Medical Association, United Nations Educational, Scientific and Cultural Organization) • Explainable AI tools	• Trustworthy AI adoption • Regulatory compliance
Tip 12: Stay informed and adaptive	Continuous learning helps educators keep pace with AI advancements and emerging best practices	• AI newsletters • PubMed alerts • arXiv.org	• Sustained innovation • Up-to-date knowledge

[a]AI: artificial intelligence.

Limitations

Although all 12 tips are grounded in published literature, they have not yet undergone external evaluation or been tested in educational practice. The paper synthesizes existing evidence but does not include longitudinal outcome data. Future studies could assess the practical impact and effectiveness of these recommendations in real-world settings.

Conclusion

The integration of AI into medical education is supported by robust evidence, particularly in teaching applications, with emerging support in research, administrative, and ethical domains. AI enhances personalized learning, improves feedback quality, enriches case-based instruction, and supports institutional decision-making. However, maintaining human-centered approaches, ensuring ethical implementation, and continuously updating knowledge are critical. Future research should focus on long-term outcomes, comparative effectiveness, and expanded investigation of AI applications in administrative and research functions.

Acknowledgments

The authors acknowledge that generative AI tools were used only for language and idea refinement (not idea generation) and for table formatting during manuscript preparation, under full human supervision. Tools used were Copilot Pro and ChatGPT. The responsibility for all content rests entirely with the authors. AI tools are not listed as authors and bear no responsibility for the manuscript.

Authors' Contributions

Conceptualization: AJ
Methodology: AJ, SF
Writing – original draft: AJ, SF, KHH
Writing – review & editing: AJ, SF, KHH

Conflicts of Interest

None declared.

Multimedia Appendix 1
A visual poster of the 12 actionable strategies for effectively integrating artificial intelligence into medical education.
[PDF File (Adobe PDF File), 601 KB - mededu_v11i1e81297_app1.pdf]

References

1. Khoyratty B. L'intelligence artificielle dans l'enseignement supérieur universitaire et le développement de compétences du 21ème siècle [doctoral thesis project]. Université de Limoges. 2023. URL: https://theses.fr/s379508 [accessed 2025-12-03]
2. Chevalier L, Garcia F. L'intelligence artificielle générative dans l'enseignement supérieur, une course perdue d'avance? 2024 May Presented at: 29e Conférence de l'Association Information et Management; May 27-29, 2024; Montpellier – La Grande-Motte, France URL: https://shs.hal.science/halshs-04631049v1/document
3. Cuvex-Combaz B. Intégration de l'intelligence artificielle et des données de santé pour une médecine de plus en plus personnalisée. Université Grenoble Alpes. 2021. URL: https://dumas.ccsd.cnrs.fr/dumas-03414167/document [accessed 2025-12-03]
4. Intelligence artificielle et santé, La science pour la santé. Inserm. URL: https://www.inserm.fr/dossier/intelligence-artificielle-et-sante/ [accessed 2025-12-03]
5. Stratégie d'intégration de l'intelligence artificielle dans l'administration publique 2021-2026. Gouvernement du Québec. 2021. URL: https://www.quebec.ca/gouvernement/politiques-orientations/strategie-integration-ia-administration-publique-2021-2026 [accessed 2025-12-03]
6. IA et leadership: comment l'intelligence artificielle transforme la prise de décision et les compétences des dirigeants. Valtus. 2024 Nov 7. URL: https://www.wip-valtus.com/blog/2024/11/07/ia-quels-impacts-sur-le-leadership/ [accessed 2025-07-15]
7. Collin S, Lepage A, Nebel L. Enjeux éthiques et critiques de l'intelligence artificielle en éducation: une revue systématique de la littérature. Can J Learn Technol 2023:1-29 [FREE Full text] [doi: 10.21432/cjlt28448]
8. Corfmat M, Martineau JT, Régis C. High-reward, high-risk technologies? An ethical and legal account of AI development in healthcare. BMC Med Ethics 2025 Jan 15;26(1):4 [FREE Full text] [doi: 10.1186/s12910-024-01158-1] [Medline: 39815254]
9. Sharma N, Doherty I, Dong C. Adaptive learning in medical education: the final piece of technology enhanced learning? Ulster Med J 2017 Sep;86(3):198-200 [FREE Full text] [Medline: 29581634]
10. Kellman PJ. Adaptive and perceptual learning technologies in medical education and training. Military Medicine 2013 Oct;178(10S):98-106. [doi: 10.7205/milmed-d-13-00218]

11. Howell J. AI enhanced debriefing for meaningful learning in clinical simulation. Healthy Simulation. 2025 Mar 2. URL: https://www.healthysimulation.com/healthcare-simulation-ai-debriefing/ [accessed 2025-07-15]
12. AI in PBL: transforming learning with virtual patients, dynamic case adjustments, and automated decision analysis. DxR Development Group. 2025. URL: https://dxrgroup.com/ai-in-pbl-transforming-learning-with-virtual-patients-dynamic-case-adjustments-and-automated-decision-analysis/ [accessed 2025-07-15]
13. Advancing AI in medical education through ethics, evidence and equity. American Medical Association. 2023 Oct 21. URL: https://www.ama-assn.org/practice-management/digital-health/advancing-ai-medical-education-through-ethics-evidence-and [accessed 2025-12-03]
14. Chen L, Chen P, Lin Z. Artificial intelligence in education: a review. 2020 Presented at: 2020 IEEE International Conference on Artificial Intelligence and Education (ICAIE); June 26, 2020; Tianjin, China URL: https://ieeexplore.ieee.org/document/9069875 [doi: 10.1109/access.2020.2988510]
15. Baker R, Inventado P. Educational Data Mining and Learning Analytics. New York: Springer; 2014.
16. Akbar MN. Use of artificial intelligence tools by doctoral students: a mixed-methods explanatory-sequential investigation. J Furth High Educ 2025 Jun 05;49(7):995-1013. [doi: 10.1080/0309877x.2025.2515135]
17. Shaw K, Henning MA, Webster CS. Artificial intelligence in medical education: a scoping review of the evidence for efficacy and future directions. Med Sci Educ 2025 Jun;35(3):1803-1816. [doi: 10.1007/s40670-025-02373-0] [Medline: 40625971]
18. Holmes W, Bialik M, Fadel C. Artificial Intelligence in Education: Promises and Implications for Teaching and Learning. Boston, MA: Center for Curriculum Redesign; 2019.
19. Bixler BA, Ceballos M. Principals leading AI in schools for instructional leadership: a conceptual model for principal AI use. Leadership and Policy in Schools 2025;24(1):137-154 [FREE Full text]
20. Zawacki-Richter O, Marín VI, Bond M, Gouverneur F. Systematic review of research on artificial intelligence applications in higher education – where are the educators? Int J Educ Technol High Educ 2019 Oct 28;16(1):39. [doi: 10.1186/s41239-019-0171-0]
21. Weidener L, Fischer M. Teaching AI ethics in medical education: a scoping review of current literature and practices. Perspect Med Educ 2023;12(1):399-410 [FREE Full text] [doi: 10.5334/pme.954] [Medline: 37868075]

Abbreviations

AI: artificial intelligence

Edited by A Stone; submitted 25.07.25; peer-reviewed by DP Rai, S Ito, N Bievre, T Yusuff; comments to author 09.09.25; revised version received 24.11.25; accepted 25.11.25; published 12.12.25.

Please cite as:
Jalali A, Harbi Houssein K, Fotsing S
Twelve Practical Tips for Integrating AI Into Medical Education: Tutorial to Support Educators Across Teaching, Research, Administration, and Ethical Domains
JMIR Med Educ 2025;11:e81297
URL: https://mededu.jmir.org/2025/1/e81297
doi:10.2196/81297
PMID:

Exploring the Implementation of Multiple Telementoring ECHO Programs From an Institutional and Organizational Perspective: Qualitative Study

M Gabrielle Pagé[1,2], PhD; Élise Develay[2], MSc; Annie Talbot[2], MD; Rania Khemiri[2], MSc; Claire Wartelle-Bladou[2], MD

[1]Department of Anesthesiology and Pain Medicine, Faculty of Medicine, Université de Montréal, 850 St-Denis, Office S03-910, Montreal, QC, Canada
[2]Research Center, Centre Hospitalier de l'Université de Montréal, Montreal, QC, Canada

Corresponding Author:
M Gabrielle Pagé, PhD
Department of Anesthesiology and Pain Medicine, Faculty of Medicine, Université de Montréal, 850 St-Denis, Office S03-910, Montreal, QC, Canada

Abstract

Background: Project Extension for Community Healthcare Outcomes (ECHO) is an innovative model to increase capacity to treat patients in their community. Despite a growing body of evidence supporting its effectiveness, little is known about the implementation processes of multiple ECHO programs within an institution from the perspective of executives and institutional leaders.

Objective: The study objective was to explore from an institutional and organizational standpoint the systemic characteristics that influence the implementation of Project ECHO programs, their growth within an ecosystem, and their sustainability.

Methods: Focus groups and individual interviews were carried out with executives and leaders from an institution that implemented 3 Project ECHO programs, and verbatim were analyzed based on organizational readiness and implementation tools for Project ECHO.

Results: This study highlighted the rarely reported perspectives of executives and institutional partners, shedding light on the organizational components that are essential to the deployment and sustainability of Project ECHO. Results reflect the intricate balance between institutional resources and its broader mission within a provincial, public health care system. In terms of acceptability, the fit between the projects and the institution's values of innovation, contribution to the broader community, and improving patient trajectory was central from the organizational leaders' standpoint. The structure of the projects and their rapid growth within the institution confirmed the adequacy with the institution. The projects benefited from temporary funds initially, and the lack of performance indicators that were easily measurable and the lack of recognition for invested time from clinicians were barriers to moving toward sustainability. Organizational characteristics, including a decentralized management structure and ministerial support for innovative educational practices, increased the perceived feasibility of implementing and maintaining these programs.

Conclusions: This qualitative study of institution leaders and directors highlighted the challenges and facilitators to the deployment of an innovative continuous education model aimed at building capacity in the community for the management of various health conditions. Despite limitations, such as temporary initial funding, challenges in collecting performance indicators, most valued, and rigidity of the projects' structure, results also show many characteristics (innovative model, alignment with the institution's mission, and simplicity of its deployment) that helped move these projects toward sustainability within the institution. Results offer learning experiences that will be relevant to other settings evolving within a similar public health care system, wanting to implement this model.

(JMIR Med Educ 2025;11:e75844) doi:10.2196/75844

KEYWORDS

Extension for Community Healthcare Outcomes; ECHO; qualitative; implementation; hepatitis C; chronic pain; concomitant disorders; sustainability; barriers; facilitators

Introduction

Wait times to access specialized care in Canada are 242% longer than they were in 1993, now at about 12.6 weeks [1]. Rates of referrals to specialized care are influenced by many factors, ranging from practitioners' degree of comfort and patient-doctor relationship to the occurrence of a global health crisis impacting referral pathways and delays, as demonstrated with the COVID-19 pandemic [2-5]. Given the rapidly evolving knowledge across the various fields of medicine and considering the limited capacity of specialized centers to absorb all of the patient referrals, it is essential to increase the capacity and comfort levels of primary care providers to manage increasingly complex cases.

Several innovative continuous education models have emerged, among which one is Project Extension for Community Healthcare Outcomes (ECHO) [6-8]. This continuous telementoring model aims to empower community-based health care professionals to provide care to individuals with specific conditions within their community. The model relies on videoconferencing meetings between hub experts and spokes (participants from the community) where there is a case discussion and didactic presentation. Four principles guide fidelity of the model when applied to new settings and new conditions: using technology to leverage scarce resources, using best practices to reduce disparity, using a case-based learning approach to better understand the complexity of patient presentation, and monitoring outcomes [9]. The benefits for health care providers of participating in such telementoring offerings in terms of knowledge, confidence, self-efficacy, and care delivery are well-established [6,10-15]. The model has been replicated and deployed in 202 countries (2813 programs) to address common and complex diseases [16]. Yet, there are known barriers to implementation and participation, including conflicting priorities and competing demands, issues of inclusiveness, time constraints, technology requirements, and a low degree of endorsement from the clinic leadership [17-23].

Beyond the characteristics of the intervention, from an organizational perspective, many inner and outer setting parameters, such as structural characteristics, population needs, local context, sources of funding, and external incentives, can influence the successful implementation of Project ECHO [24]. Little is known about systemic factors that must be considered when planning and implementing such a program and monitoring its growth within the institution, and how it might benefit more broadly the local health care ecosystem. Most implementation studies explored the postimplementation perspectives of health care professionals, whether participants (spokes) or hub experts, with little attention to date given to the perspectives of institutional leaders and managers.

In addition, many institutions harbor several ECHO programs within a single center [16]. However, data on the implementation processes of multiple programs within a single site are scarce. Agley and colleagues [25] explored the implementation of 5 programs within 1 site and found that Project ECHO is associated with change in practice and knowledge and reduced isolation of practitioners, but that rigorous studies of such a model are warranted.

The Consolidated Framework for Implementation Research (CFIR), created by Damschroder, comprising a set of 39 constructs spread across 5 domains, successfully guided the implementation of several projects across various health domains [26]. An adapted CFIR was developed by an ECHO implementation team to create ECHO-specific tools to help organizations assess their readiness and capacity to support an ECHO project and provide a checklist to support a successful implementation process [24].

The study objective was to explore, from an institutional and organizational standpoint, the systemic characteristics that influence the implementation of Project ECHO, their growth within an ecosystem, and their sustainability. Results will help guide other teams to successfully identify optimal implementation conditions for multiple programs within the same institution.

Methods

Study Setting

This study is part of a larger project aiming at evaluating the implementation of 3 Project ECHO programs (Hepatitis C, chronic pain, and concurrent mental health and substance use disorder) within a tertiary care, university-affiliated hospital center [27]. The pilot Project ECHO Centre hospitalier de l'Université de Montréal (CHUM) Hepatitis C program was launched in April 2017 to enable health care professionals to increase the rate of Hepatitis C screening, assessment, and treatment in Quebec remote areas. Shortly thereafter, the Project ECHO CHUM *Douleur Chronique* was launched in September 2017. Its goal is to foster knowledge regarding best practices in chronic pain management and enable collaboration between health care professionals of different degrees of specialization. Finally, a third program, Project ECHO CHUM Troubles Concomitants (concurrent mental health and substance use disorder), began in September 2018 to help health care professionals from different backgrounds gain knowledge in assessment, treatment, and referral of patients with concomitant mental health and substance use disorders.

Study Design

This is a qualitative study comprised of focus groups and individual interviews conducted between November 2020 and April 2021 with stakeholders, directors, health care managers, project coordinators, and medical program leads within the hospital or the university network, and Project ECHO medical leads.

Ethical Considerations

The study was approved by the research ethics committee of the CHUM (19.281). Written informed consent was obtained from all participants, and they could opt out from the study at any time. Data were deidentified prior to analyses. Participants did not receive a compensation for taking part in this study. The description of the study methods followed the Consolidated

Criteria for Reporting Qualitative Research (COREQ) guidelines [28].

Participants

Potential participants were targeted by the research team based on their role within and outside of the institution, to cover all relevant directions that were involved with the project, including (1) medical leads, coordinators, and managers involved directly or indirectly with the implementation and delivery of Project ECHO; (2) directors and health care managers of various hospital services and programs involved with Project ECHO (communication services, professional services, etc); and (3) stakeholders within the larger university network (Réseau universitaire intégré de santé et de services sociaux de l'Université de Montréal). A total of 30 potential participants were thus approached. Of the total, 13 participants were excluded because they refused to participate (n=5) or did not respond to our invitation (n=8). The 17 study participants included are presented in Table 1, which illustrates the roles and responsibilities of each participant within Project ECHO. Project ECHO spokes and experts were interviewed separately, and these data have been presented elsewhere [27]. Descriptions of participants' characteristics are presented in Table 1.

Table . Characteristics of study participants.

Participants' characteristics	Frequency, n (%)
Role	
ECHO[a] medical leads	3 (18)
Project ECHO coordinators	4 (23)
Health care managers	2 (12)
Directors (teaching, academics, communication, and operations) and stakeholders	8 (47)
Gender	
Women	13 (77)
Men	4 (23)
Age (y)	
25 - 34	1 (6)
35 - 44	5 (29)
45 - 54	3 (18)
55 - 64	7 (41)
65+	1 (6)
Number of years involved in Project ECHO	
1	3 (18)
2	2 (12)
3	12 (70)
Employer	
Centre hospitalier de l'Université de Montréal	13 (77)
Réseau universitaire intégré de santé et de services sociaux de l'Université de Montréal (Integrated University Network for Health and Social Services at Université de Montréal)	4 (23)

[a]ECHO: Extension for Community Healthcare Outcomes.

Recruitment

Potential participants were identified by the study investigators and those involved in the deployment of each of the 3 studied Project ECHO programs. Potential participants received an email inviting them to participate in a focus group. They were asked to express their interest by replying to the email, after which a research assistant explained the study procedure to them and scheduled the focus group. Those interested but unable to take part in a focus group because of scheduling conflicts were invited to participate in an individual interview. They were informed of the study's objectives, namely to understand the barriers and facilitators to the deployment of ECHO programs and their sustainability, but they were not informed of the interviewers' reasons or interests in the research topic.

Procedure

Before the scheduled interview, participants received a link through email to complete an online consent form and a brief sociodemographic questionnaire. In total, 3 semistructured focus groups (2 with medical leads, project coordinators, and health care managers and 1 with directors and stakeholders) and 4

semistructured individual interviews were conducted online using the secured videoconferencing platform Zoom (Zoom Communications) by 1 of 2 women research team members (MGP [PhD] or ÉD [MSc]) trained and experienced (5+ y) in qualitative methods and were audio-recorded. Both interviewers had been working at the institution where the study was conducted for more than 5 years and knew many of the study participants in the context of other research projects or collaborations. Semistructured interview guides were built with open-ended questions aimed to explore topics such as barriers and facilitators to program implementation, interest toward this program, program fit within the larger institution and provincial health care ecosystem, and sustainability (Multimedia Appendix 1). Audio recordings were transcribed verbatim (transcripts were not returned to participants for comments due to time constraints). Field notes were taken during the interviews and used to supplement the data analysis. Focus groups lasted between 55 and 59 minutes (median 58 min) and varied in size between 3 and 6 participants, while individual interviews lasted between 18 and 28 minutes (median 24 min). Interviews were carried out in French, and the selected quotes were translated into English using a forward-backward translation process.

Data Analysis

A qualitative content analysis with both deductive and inductive coding was used [29,30]. A deductive approach was used as a theoretical lens through which the data were interpreted [31], with initial domains mapping onto the Organizational Readiness and Implementation Tools for Project ECHO [24] (an adaptation of the CFIR [26]). These domains represented an initial starting point to understand the data, but other domains could be added to the framework as needed. More specifically, 1 analyst (MGP) read the transcripts many times and identified units of meaning that mapped into the participants' perceptions of implementation barriers and facilitators of the projects within the institution. A coding sheet identifying the dimensions of the CFIR framework that fit the project's context (acceptability, appropriateness, cost and trialability, and feasibility) was collaboratively identified with team members (ÉD and CW-B). Within each of these dimensions, inductive coding was used to identify subcategories and units of meaning. This step was performed by GP with frequent meetings with other analysts (ÉD and CW-B). When disagreements arose regarding data interpretation, we examined whether they stemmed from differences in backgrounds and sought to integrate these viewpoints to enrich the analysis. Memos were written throughout the analysis process. NVivo software (Lumivero) [32] was used to conduct the analyses. Participants' feedback on the study results was not solicited.

Reflexivity Statement

Members of the research team were involved in the development of ECHO programs at the institution (CW-B) or in the coordination of the research project (AT, RK, ÉD, and MGP). They come from the disciplines of medicine (AT and CW-B), social science (ÉD), environmental and occupational health (RK), and psychology (MGP). Team members directly involved in the execution of the research project and analysis (MGP and ÉD) engaged in reflexivity throughout the project, but most intensely during the elaboration of the study material (ie, development of interview guide), throughout data collection, and during the analysis, with the intent of reflecting on how our perspectives and experiences within the institution but also with regards to the research topic, influenced the orientation of the project, our understanding of the data, and how it shaped the findings. Reflexivity involved individual reflections through memos and note-taking as well as collaborative meetings.

Results

Overview

We generated themes based on available data pertaining to Project ECHO's acceptability, appropriateness, cost and trialability, and feasibility (refer to Textbox 1 for a summary of the themes).

Textbox 1. Summary of domains and themes.

Acceptability

- Knowledge transfer as a core value in line with the institution's academic and public health mission
- Innovation as an institutional pillar
- Institutional benefits—optimizing the patient flow in a 2-way stream

Appropriateness

- Versatility and structure of the ECHO (Extension for Community Healthcare Outcomes) model
- Meeting a growing need for education on rapidly evolving complex health conditions
- Rapid growth of ECHO programs

Cost and trialability

- Institutional sources of funding and human resources that benefit the outside community
- Lack of an easily measurable performance indicator
- Lack of recognition for time and resources invested

Feasibility

- Decentralized management structure
- Ministerial and academic characteristics of the outer setting bolstering the institution's innovative educational practices

Acceptability

The perceived acceptability of the Project ECHO programs was influenced by cultural aspects specific to the institution and its role within the larger health care network. The structure of this innovative program also contributed to increasing its acceptability.

Knowledge Transfer as a Core Value in Line With the Institution's Academic and Public Health Mission

The institutional support for Project ECHO stemmed in part from the program's fit with the institution's academic mission. Being a university health network institution, its mandate beyond patient care involves knowledge transfer and supporting the wider health care community. From that perspective, Project ECHO uses an innovative continuous education model that enables expertise within the university health network and academic workforce to be shared with the broader medical community, which facilitates the execution of the institution's academic mission as a center of expertise.

> *I'd say that we have a great deal of specific expertise at the [institution], and our challenge is to share this expertise with partners who could benefit from it. Projects ECHO are in line with this and at the [institution], we try to generate knowledge, and from that knowledge we generated, how can we play our role as a member of a university health network within our network and how can we maybe support teams within this network. It builds pride in what they do, what they achieve, so I would say that there are not super demanding activities, not overly costly, but so worth it on multiple fronts. You cannot not support these initiatives and encourage them.* [P4, Director]

Innovation as an Institutional Pillar

The Project ECHO distinguishes itself from other more conventional learning modalities, as it is based on ongoing real clinical cases and offers adaptability to have participants' questions answered by a panel of clinical experts, using personalized solutions in an "all teach, all learn" approach. This leads to rich exchanges, making it possible for participants to have access to the experts' reflections and decision-making processes that led to the proposed treatment path. Those components were identified as central to the success of such a learning model. Comments from managers and stakeholders underlined that one of the institutions' characteristics that facilitated the deployment of those programs was the crucial importance given to innovative practices. This implied an agility to recognize innovative approaches and to maximize existing resources to bring them to life.

> *However, to be an innovative institution, its also to see things differently, you know to have other ways of doing things. And so it was that, another way of looking at this and to say well, we have given a lot of training, of information, but telementoring, we had never done it. So, the openness that this brings…* [P1, Director]

The emphasis put on innovative practices not only facilitated the implementation of the programs, but it also improved the likelihood of successful sustainability. Justifying the financial expenses and required resources for running Project ECHO was facilitated by a good comprehension of the project's fit within the larger institutional academic mission and its core value of innovative practices.

Institutional Benefits-Optimizing the Patient Flow in a 2-Way Stream

Stakeholders described perceiving important institutional benefits from the Project ECHO programs in place. The projects were in line with the institution's vision of patients receiving the right treatment at the right place at the right time. By helping health care providers within the broader health care system to care for their patients with complex health conditions, the institution contributed to facilitating optimal patient flow. This ensured that specialized services were used for the most complex cases, while less complex patients could access the right care in their community.

> *Well, I think that it's the main strength because otherwise our institution has 772 beds, not all of them are open, we cannot admit all of the patients so we must develop knowledge and expertise about specific clienteles and after that, help our network partners to care for these clienteles. With Projects ECHO, it helps do that. So, I would say that the strength of such programs is that it reinforces education, it reinforces partnership, it helps reinforcing multidisciplinary work and also the actors from different sectors.* [P4, Director]

One way that Project ECHO facilitated this was through connecting and empowering first-line health care providers to treat more complex cases.

> *We make sure that patients who are physically coming to the CHUM really need to be here, that they are the ones that fit with our mission. Because each time we accept a patient that could have been treated elsewhere, well we worsen access for another patient who probably must be seen at the CHUM. So I see a clear impact in terms of access, the famous "the right patient at the right place," there really is an impact at that level. So for the CHUM, to allow the institution to recenter itself on its tertiary and quaternary mission, well it's always a good thing that there are activities targeting primary and secondary care, we are after all a teaching institution.* [P2, Director]

In addition, the experiences of clinical experts within the program benefited not only participants but the experts themselves. Through the constructive discussions taking place on complex cases, experts gained a better understanding of constraints other health care professionals are confronted with in their practice. These were perceived as factors enhancing the acceptability of the program by directors, managers, and stakeholders.

> *Well those who do it, the team in fact, it helps them rethink, and then hearing, listening to what people working in primary and secondary care have to say, well it helps them too to discover other aspects that perhaps they might not have focused on previously. And I think that it brings the team closer because they to do that together. So even if these are people who work together, well in this case we can be a microbiologist who comes on board and the pharmacists is there… and that, that brings a lot.* [P1, Director]

Appropriateness

Appropriateness of Project ECHO fluctuated based on the structure of its delivery and on its fit with the outer setting.

Versatility and Structure of the ECHO Model

Project ECHO was perceived as attractive by managerial instances because of its operationalized structure that demonstrated effectiveness across various chronic health conditions and settings. This prestigious program's history and the standardization of its delivery facilitated buy-in from various instances as little investment was needed to deploy the programs.

> *Also, it's a proven program, so it's not just us, it's elsewhere in the world… in Canada, in the United States, it's well implemented. So, it's had already proven its worth before we imported it here. That's an important aspect.*
>
> *[FG7, Director]*

On the other hand, the rigidity around the delivery of Project ECHO turned out in some instances to be an obstacle to its spread within the institution. This was either because it was perceived as less suitable for some pathologies since associated with higher costs than other telementoring alternatives, or because it lacked flexibility to use other technological platforms approved by ministerial instances.

> *I don't know if the ministerial position about the use of TEAMS, with for example teleconsultations, if that can be an obstacle. Well yes and no, maybe not. But if we would arrive at the conclusion that the two technologies do exactly the same thing, but with ECHO we must work with Zoom… I'll give you an example. We are involved in the First Nations dossier. If we would realize that using Zoom for Project ECHO® is technologically not feasible in this setting… well I think it would be a strong enough justification to use another model. I mean, the first priority, is to meet the needs. Then, if we can answer those needs through a standardized approach, it accelerates things, its more efficient in terms of deployment, maintenance, putting resources in common, etc. But it must meet the needs, otherwise we must find something else that will.* [P2, Director]

Meeting a Growing Need for Education on Rapidly Evolving Complex Health Conditions

Successful implementation of the program depended in part on the adequacy between what was being offered and the need for such services within the broader medical community. In recent years, the increased complexity observed in various patient populations, the rapid evolution of therapeutic approaches, and the lack of educational resources targeting multiple concurrent diseases appeared to positively impact enrollment of participants in the program.

> *For us, we have a large number of participants, but I think it answered a need, that people were a bit lost,*

> *people were helpless to know what to do with this clientele that is complex. So, I think that when they saw this, they jumped on it because there were no other services that provided this… It's really difficult to be able to solve the problems this clientele has. So, to know that in one place, at the same time, many professionals from different disciplines. I think it was very tempting to have enrollees.* [FG6, Coordinator]

By palliating gaps in the health care system and facilitating patient care within the network, this program was also perceived as indirectly benefiting the institution. This was perceived as one of the main advantages of Project ECHO from stakeholders' perspectives.

> *It is part of our mandate, so the admission to the institution, the return back to the network, all of that is part of the big network umbrella. So, it's a bit for that reason that I was interested in Projects ECHO, because for us, obviously, we aim to optimize healthcare trajectory of patients who must come to our center or return to the network… So it is in this context that Projects ECHO are really interesting for our mission, because we have realized, us, while working with many clinical teams, that very often, a medical care is offered in our institution to compensate for a discomfort from certain partners in the network. And also, there are other ways of responding to this than having patients come here. So that's how Projects ECHO fall in part within our institutional mission.* [P3, Director]

Project ECHO was viewed as a valuable addition to the institutional service offerings because it had the potential to reach individuals beyond the actual participants and clinical experts involved in these sessions. In fact, the knowledge sharing that occurred between participants and their peers, as well as with other tertiary care centers, was significant. Several health care managers and stakeholders noted the snowball effects associated with the visibility of Project ECHO and the improvement of clinical practices within the network.

> *So, to have a platform where when we do an intervention, we reach well 50, 70 other professionals, but in fact we maybe reach much more, because when the social worker from the team of first psychotic episodes in a clinic, often this person will talk to others during team meetings of 20 other professionals, and all of that. So, we reach a very large proportion of people.* [FG5, Medical lead]

Rapid Growth of ECHO Programs

There has been a rapid growth within all 3 Project ECHOs over time. All 3 programs were implemented by individual medical leaders within an 18-month period with a rate of participation rapidly increasing over time. This rapid growth could be accommodated by the institution from a technology support standpoint. This expansion raised both awareness and interest within the institution and also within the broader medical community and other provincial health care networks.

> *We always say that the CHUM is there for all individuals in Quebec, so ECHO is one way to help people from outside… The idea is to know, if you tell me tomorrow morning "we will go from 3 to 10 [programs]," well then we would need to assess the technical needs, how many times per week [technicians] need to be there, what this represents so that we can do this work together, so that it is a collaboration that works well. Because the idea is to support but we must plan well.* [P3, Director]

Cost and Trialability

Alongside the elements that impact acceptability and appropriateness of the programs, the absence of a long-term source of funding and the lack of easily measurable performance indicators remained important barriers to further program expansion and long-term sustainability of the existing programs.

Institutional Sources of Funding and Human Resources That Benefit the Outside Community

Basic costs of these programs involve access to telediffusion material such as webcams, screens, and computers, along with an available technology staff to troubleshoot any problems that can arise during live sessions. Considering the material and conference room being shared by the 3 programs, investments were optimized over the years.

Unlike other programs elsewhere in the country or in other countries, there was minimal funding from public sources to set up these 3 projects within the institution. The in-kind services offered by the institution, such as access to technology material and services, were made possible because of the fit between Project ECHO and the institutional mission. However, the absence of a permanent program coordinator position and of a dedicated paid medical lead posed a threat to the sustainability of the projects.

> *In the United States, they often finance it, but in Toronto as well, they have a pretty big team compared to what we had. In terms of researchers, there are two doctors who are paid to do only ECHO with one salary. So, there is a really big paradigm between what is being asked here. And the risk is that it gets off track, and then stops, or the quality decreases, meaning that I can make recommendations in 15 minutes, but it won't be of the same quality as if I were taking longer to write them. So, there is that organizational, financial aspect.* [FG5, Medical lead]

Despite the recognized adequation between Project ECHO and the institutional mission, the lack of a permanent source of funding was identified as a constant threat to their survival.

> *And let me tell you, I consider that it is not going fast enough, we should have more [programs], but we always hit the same problem when we say "I will need a budget to do that.*" [P1, Director]

This instability might be due in part to the dual mission of the project, which is at the crossroads between educational and health care mandates. This overlap can generate confusion and a lack of accountability among funding bodies, who may assume

that the program falls under the responsibility and the funding of other authorities.

> *You know, its complex Project ECHO. There are multiple aspects. But at some level, it's clear that after the meeting, at least from what I understood, there is an intention to help a patient from the start, which motivates, there is a discussion of real cases that will influence medical acts eventually. But in addition, in parallel, there is knowledge that people acquire more and more to become independent. So, there are these two aspects. One is clinical and one integrates knowledge transfer. And that, this aspect, is really novel.* [FG7, Director]

Lack of an Easily Measurable Performance Indicator

Project ECHO programs are innovative educational models that benefit health care professionals within the community and patients. Currently, there are no easily accessible measurement metrics or dedicated resources to collect performance outcomes within the network. This poses a barrier to the long-term sustainability of the program. For example, it is difficult to document the exact number of patients that benefited from recommendations made within the ECHO sessions, or to estimate the number of patients who were not referred to specialized centers because of the knowledge and expertise gained by their primary health care provider. Given the very low budget allocated to Project ECHO, it became difficult for individual projects to collect empirical data to demonstrate their effectiveness. Because the data are limited, providing a strong rationale to maintain and expand the ECHO model within the institution was challenging. Such an issue is being recognized and must be addressed.

> *We are going to have to think about how we will measure outputs from this process… and what is reasonable to invest considering the population gains in the end, but also for the knowledge development of our professionals. And… we cannot forget our professionals, our researchers, our experts, who learn a lot from people in the different regions, and not to remain disconnected from reality, it is important that communication goes in both directions. So, this has a worth. In my opinion, we probably do not measure it enough, and we probably don't have enough indicators from a managerial standpoint to advertise the program to the level of the real impact it has for the population and for clinicians.* [FG7, Manager]

Lack of Recognition for Time and Resources Invested

Despite the enthusiasm reported above, the time required to set up Project ECHO programs and maintain them reduced its acceptability. Such programs require leadership and skills to offer a high-quality experience. The tasks are often complex, including within and outside of the institution, to finalize the program logistics, create a cohesive hub team, recruit participants from the community, and carry out direct and indirect tasks related to the program. There was often a lack of recognition for the efforts invested. Beyond the actual time spent in an ECHO session, there was significant legwork required behind the scenes, which was not recognized financially.

> *But Projects ECHO raise a lot of small questions, that are not evident from the start. And people think we put on a show you know, that it's fun to do that, … but there is the visible side, but also the invisible side of ECHO. You know, there is the show, but after that, well you have to take care of it after. So it's not that simple, when there is an investment from so many people. I think there are a lot of people who work for nothing in this program. We always work more than what we get paid for I think. And, that's a reality for people who innovate, I think not necessarily the responsibility, but the impact that it has on each person. So, that's what innovation looks like in [region], we never receive anything for the efforts that we put in.* [FG5, Coordinator]

Feasibility

Feasibility was enhanced by the managerial structure in place and also by the ministerial and academic characteristics of the outer setting that supported the institution's innovative educational practices.

Decentralized Management Structure

The institution's managerial structure is to assign a manager for each group of clientele, favoring, as a result, proximity management. This was identified as an organizational characteristic of the inner setting that facilitated the deployment of the Project ECHO model but also made it more difficult.

On the one hand, the decentralized management structure allowed each service to provide authorization for their staff to free up clinical and medico-administrative time to participate in the program, without needing approval from higher-level managers.

> *Yes, the decentralized management structure in the patient populations I think is favorable because it is medical co-management and medico-administrative and proximity management so near sectors of care.* [P4, Director]

On the other hand, however, decentralization increased the reliance on medical leaders to ensure the deployment and maintenance of programs over time. Project ECHO relied on the leadership of specific clinicians within each of the programs to set them up, mobilize their clinical teams, and obtain support from managers and stakeholders to access resources and disseminate the program within and outside of the institution. While there was an overwhelming institutional buy-in, the concrete steps to make this happen relied solely on those medical leaders.

For many participants, the existing structure, which provided a high level of autonomy within each clinical service, was a barrier to adequate communication within different levels of management about Project ECHO being put in place. Each of the 3 programs was developed through individual leaders' initiatives, and as such, they were less visible within the hospital's ecosystem. This led to suboptimal sharing of resources

that could have accelerated the implementation of each program, had an institutional vision been in place earlier. For example, not all of the managers interviewed were aware of the different programs operating within the institution more than 3 years after their launch.

> *I don't know to what extent the Projects ECHOs are known from the larger CHUM community. Me I know them because they operate in sectors within my responsibility so we discuss them regularly. Do we hear about them more broadly, is everyone aware that we have initiatives such as Projects ECHO in other sectors? I have the impression that it's less known and can be a barrier to further deployment of these programs.* [P4, Director]

Ministerial and Academic Characteristics of the Outer Setting Bolstering the Institution's Innovative Educational Practices

The institution was the first in the province to implement Project ECHO programs, so that their success depended on the support received at the ministerial and the integrated university network for health and social services (Réseau universitaire intégré de santé et de services sociaux de l'Université de Montréal [RUISSS]) levels. The RUISSS mission is to federate the university and its affiliated health and social services institutions to facilitate collaborations and set up special projects in line with its ministerial education and health mission. Project ECHO, being at the intersection of teaching and clinical practices, was a perfect fit with this mission. While the process took a few months to target the appropriate type and level of resources needed to implement the first program and required initial funding from pharmaceutical sponsors, the subsequent Project ECHO programs deployed benefited from the already established corridor.

> *So there has really been a year and a half of processes, finding a key person at the RUISSS level has been key, because that person believed in the project from the start and helped me and the manager working there, who knew that person well, to meet the dean of pharmacy, make those connections for us…* [FG5, Medical lead]

This collaboration between the RUISSS and the institution, early in the implementation phase, helped increase visibility of the projects within the medical community, which facilitated recruitment of participants and also increased awareness at the ministerial level of this type of program. The notoriety of Project ECHO worldwide, and the fact that the institution was the first university academic center to offer Project ECHO in French, highlights and contributes to the institution's and the province's visibility in innovative practices. Such progress could potentially lead to additional deployment of projects within the institution and within the province and play a role at the international level with the deployment of French-speaking Project ECHO programs.

> *Let's not forget that if we look at the strategic planning from the ministry for 2019-2023, there is an item about access to specialized care. But this type of ECHO, it's exactly about making a specialty accessible in regions. So it could fall under that mandate…. Because it has been a while since we went to the ministry about that. But let's not forget also that there is some pride. These are the first francophone ECHOs. We can rely on that. We can be francophone leaders for ECHOs.* [FG7, Director]

Discussion

Overview

This study highlighted the rarely reported perspectives of executives and institutional partners, shedding light on the organizational components that are essential to the implementation and sustainability of Project ECHO. Results reflect the intricate balance between institutional resources and its broader mission within a provincial, public health care system. In terms of acceptability, the fit between the projects and the institution's values of innovation, contribution to the broader community, and improving patient trajectory was central from the organizational leaders' standpoint. The structure of the projects and their rapid growth within the institution confirmed an adequate fit with the organization. Initially, the projects benefited from temporary funds, but the absence of easily measurable performance indicators and a lack of recognition for clinicians' invested time posed barriers to progress toward sustainability. Organizational characteristics, including a decentralized management structure and ministerial support for innovative educational practices, increased the perceived feasibility of implementing and maintaining these programs.

These results are in coherence with Mintzberg's *Professional Bureaucracy* structural configuration of the institution under study [33,34], which is characterized by a strong focus on professionals placed as the operating core within the organization, standardization of skills to coordinate the different mechanisms, and the use of both vertical and horizontal decentralization so that professionals can benefit from a high degree of autonomy. Goals of such a professional bureaucracy structure are typically to innovate and provide high-quality health care services within a large, complex yet stable environment [33,34]. These key characteristics of the organization's structure facilitate but also hinder the deployment and sustainability of initiatives like Project ECHO.

Principal Results

The 3 initial Project ECHO programs launched at the institution were initiated by local experts who aimed to meet various needs, including building a stronger primary care network to reduce referrals to their specialized clinics and contribute to knowledge diffusion. In this context, the organization had little involvement in the initial setup of these programs, but its endorsement was essential to the sustainability and growth of the ECHO model within the institution. This is worrisome considering that a recent survey showed that across more than 1000 Project ECHO programs in 68 different countries, funding was temporary at the start [35]. The public health mission of this tertiary and quaternary care center and the potential disruptive nature and value creation associated with these projects were central to

obtaining this endorsement, which enabled the launch of 3 additional Project ECHO programs since the completion of this study.

Considering Mintzberg's organizational structures [33,34] can help understand how the inner setting (structural characteristics, culture, and mission alignment) and outer setting (financing, local conditions, and external pressure) characteristics highlighted in the CFIR framework can guide the development of an implementation strategy for Project ECHO. In this case, for example, professionals' autonomy has been central to setting up the first operational year for each of these 3 projects, with little direct involvement from higher-level governance. This might not be possible in all organizations, especially if the organizational structure is less focused on professional autonomy, in which case the initial implementation strategy might need to include more rapid institutional leads.

Coherent with the professional bureaucracy organization structure, the executives also highly valued innovation in all of its forms. Project ECHO represented new ways within the province to train clinicians that showed promising results for both the recipients (primary care providers) and the deliverers (hub experts) and the institution. The benefits perceived by the participants of the 3 projects have been previously reported [27]. These bidirectional benefits were highly valued and helped the organization set up an infrastructure that would be able to maintain some growth of the innovation. These findings are similar to those reported in Australia, where executives highlighted key elements of Project ECHO that facilitated its deployment, including the alignment with the institution's strategic priorities, innovative potential, and potential for integrated care [36].

The Project ECHO programs are highly standardized, including how sessions are carried out and the telehealth services needed for the sessions (Zoom). This had significant advantages as it minimized the program development required to launch a program, keeping it simple to generate the didactic content specific to each project. Conversely, some of these standardized procedures were incompatible with governmental commendations for the platforms to use for telehealth, which did not include Zoom. At the same time, it was possible for the 3 projects to share the room and licenses to deliver the sessions.

Even after the pandemic, Canada remains largely behind Commonwealth countries in terms of scheduling appointments online (22% in 2019 and 38% in 2022 in Canada compared with 56% in 2019 and 57% in 2022 in the Commonwealth countries), but the use of technology has been improving, particularly for the use of electronic medical records which increased by 20% from 2019 to 2022 [37]. It is likely that the initial technological barriers discussed in this study during the implementation of the ECHO programs before the pandemic have widely decreased by now.

The complexity of the health care ecosystem in the province made it difficult to collect performance indicators that are directly associated with health care performance. For example, the institution executives would like to understand how many patients are not referred to the institution because they received care from their provider locally. However, not all participants in the program would refer to this institution, as they might be living in a geographical region that is under the authority of a different tertiary care center. This lack of a direct performance indicator represents an important barrier to the sustainability of these programs. This difficulty in assessing the impact of the model on a population level is well-recognized [38]. A recent scoping review of patient and community health outcomes associated with Project ECHO showed that out of the 15 studies included, only 1 provided data on outcomes changed at the community level [38].

Finally, results highlighted some possible barriers to sustainability, namely, regarding the availability of human resources, such as dedicated time and funding for hub members to participate in sessions and write recommendations, as well as funding to cover operational costs. A 2023 review of financial structures of North American Project ECHO programs highlights that while many programs initially start up using institutional funds, those are rarely renewed in time, and finding permanent sources of funding becomes crucial [39]. There is a diversity of successful strategies that have been put in place, including aligning the program with state, provincial, or federal health priorities or finding external partners (eg, practice-based networks and translational institutes) [39].

Limitations

This study provides the unique perspectives of organizational executives and leads as their institution implemented 3 different Project ECHO programs. Results are relevant to those institutions operating within public health care systems and want to implement and maintain these programs. Notwithstanding, this study has limitations. Results were collected in a single institution, and it would be important to validate these findings in other provinces and countries with a similar health care structure. Interviews took place after the implementation of the 3 projects, and thus, there might be a recall bias around questions pertaining to the early phases of implementation.

Conclusions

This qualitative study, focusing on institution leaders and directors, highlighted the challenges and facilitators to the deployment of ECHO projects aiming at building capacity in the community for the management of various health conditions. Despite temporary initial funding, challenges in collecting performance indicators most valued, and perceived rigidity of the projects' structure, our results show many characteristics (innovative model, alignment with the institution's mission, and simplicity of its deployment) that helped move these projects toward sustainability within the institution. These learning experiences will be relevant to other institutions evolving within a similar public health care system, wishing to implement this model.

This study focused on the identification of factors at various levels that can influence the readiness for implementation but also the sustainability of an innovation and used the CFIR as a determinant framework for this purpose [40]. The CFIR is less adapted, however, to the collection of quantitative measures of implementation success [41]. It might be helpful for future

studies to combine this model with other models, such as Proctor's taxonomy of outcomes, to also document implementation outcomes.

Acknowledgments

The project was funded by MEDTEQ under the second call for projects of le Fonds de soutien à l'innovation en santé et en services sociaux (FSISSS). MGP is a Junior 2 research scholar from the Fonds de recherche du Québec–santé.

Data Availability

Data can be made available upon reasonable request to the corresponding author, pending approval from the research ethics board of the Centre hospitalier de l'Université de Montréal.

Authors' Contributions

Conceptualization: CW-B, AT
Funding acquisition: CW-B, AT
Investigation: ÉD, MGP
Methodology: RK, ÉD, MGP
Project administration: RK, ÉD, MGP
Resources: RK
Writing - review & editing: CW-B, AT, RK
Writing - original draft: ÉD, MGP

Conflicts of Interest

None declared.

Multimedia Appendix 1
Semistructured interview guide.
[DOCX File, 22 KB - mededu_v11i1e75844_app1.docx]

References

1. Moir M, Barua B. Waiting Your Turn: Wait Times for Health Care in Canada: Fraser Institute; 2022:86.
2. Sperling S, Andretta CDL, Basso J, et al. Telehealth for supporting referrals to specialized care during COVID-19. Telemed J E Health 2022 Apr;28(4):544-550. [doi: 10.1089/tmj.2021.0208] [Medline: 34314637]
3. Burton C, Bajpai R, Mason KJ, et al. The impact of the COVID-19 pandemic on referrals to musculoskeletal services from primary care and subsequent incidence of inflammatory rheumatic musculoskeletal disease: an observational study. Rheumatol Adv Pract 2023;7(2):rkad044. [doi: 10.1093/rap/rkad044] [Medline: 37251663]
4. Soltani SA, Fallah M, Marvi A, et al. Performance trend of the family physician referral system before and during the COVID-19 pandemic: a study in northern Iran. BMC Public Health 2024 Aug 7;24(1):2142. [doi: 10.1186/s12889-024-19648-7] [Medline: 39112993]
5. Tzartzas K, Oberhauser PN, Marion-Veyron R, Bourquin C, Senn N, Stiefel F. General practitioners referring patients to specialists in tertiary healthcare: a qualitative study. BMC Fam Pract 2019 Dec 1;20(1):165. [doi: 10.1186/s12875-019-1053-1] [Medline: 31787078]
6. Zhou C, Crawford A, Serhal E, Kurdyak P, Sockalingam S. The impact of Project ECHO on participant and patient outcomes: a systematic review. Acad Med 2016 Oct;91(10):1439-1461. [doi: 10.1097/ACM.0000000000001328] [Medline: 27489018]
7. Socolovsky C, Masi C, Hamlish T, et al. Evaluating the role of key learning theories in ECHO: a telehealth educational program for primary care providers. Prog Community Health Partnersh 2013;7(4):361-368. [doi: 10.1353/cpr.2013.0043] [Medline: 24375176]
8. Katzman JG, Galloway K, Olivas C, et al. Expanding health care access through education: dissemination and implementation of the ECHO model. Mil Med 2016 Mar;181(3):227-235. [doi: 10.7205/MILMED-D-15-00044] [Medline: 26926747]
9. Katzman JG, Comerci G Jr, Boyle JF, et al. Innovative telementoring for pain management: project ECHO pain. J Contin Educ Health Prof 2014;34(1):68-75. [doi: 10.1002/chp.21210] [Medline: 24648365]
10. McBain RK, Sousa JL, Rose AJ, et al. Impact of Project ECHO models of medical tele-education: a systematic review. J Gen Intern Med 2019 Dec;34(12):2842-2857. [doi: 10.1007/s11606-019-05291-1] [Medline: 31485970]
11. Anderson D, Zlateva I, Davis B, et al. Improving pain care with Project ECHO in community health centers. Pain Med 2017 Oct 1;18(10):1882-1889. [doi: 10.1093/pm/pnx187] [Medline: 29044409]

12. Mazurek MO, Parker RA, Chan J, Kuhlthau K, Sohl K, ECHO Autism Collaborative. Effectiveness of the extension for community health outcomes model as applied to primary care for autism: a partial stepped-wedge randomized clinical trial. JAMA Pediatr 2020 May 1;174(5):e196306. [doi: 10.1001/jamapediatrics.2019.6306] [Medline: 32150229]
13. Mehrotra K, Chand P, Bandawar M, et al. Effectiveness of NIMHANS ECHO blended tele-mentoring model on integrated mental health and addiction for counsellors in rural and underserved districts of Chhattisgarh, India. Asian J Psychiatr 2018 Aug;36:123-127. [doi: 10.1016/j.ajp.2018.07.010] [Medline: 30086513]
14. Nhung LH, Kien VD, Lan NP, Cuong PV, Thanh PQ, Dien TM. Feasibility, acceptability, and sustainability of Project ECHO to expand capacity for pediatricians in Vietnam. BMC Health Serv Res 2021 Dec 9;21(1):1317. [doi: 10.1186/s12913-021-07311-5] [Medline: 34886871]
15. Ball S, Stryczek K, Stevenson L, et al. A qualitative evaluation of the pain management VA-ECHO program using the RE-AIM framework: the participant's perspective. Front Public Health 2020;8:169. [doi: 10.3389/fpubh.2020.00169] [Medline: 32500053]
16. Arora S. Project ECHO 2024 annual report. : University of New Mexico; 2024.
17. Walters SM, Li WP, Saifi R, et al. Barriers and facilitators to implementing Project ECHO in Malaysia during the COVID-19 pandemic. J Int Assoc Provid AIDS Care 2022;21:23259582221128512. [doi: 10.1177/23259582221128512] [Medline: 36177542]
18. Becevic M, Smith E, Golzy M, et al. Melanoma extension for community healthcare outcomes: a feasibility study of melanoma screening implementation in primary care settings. Cureus 2021 May 29;13(5):e15322. [doi: 10.7759/cureus.15322] [Medline: 34221770]
19. Shea CM, Gertner AK, Green SL. Barriers and perceived usefulness of an ECHO intervention for office-based buprenorphine treatment for opioid use disorder in North Carolina: a qualitative study. Subst Abus 2021;42(1):54-64. [doi: 10.1080/08897077.2019.1694617] [Medline: 31809679]
20. Salvador J, Bhatt S, Fowler R, et al. Engagement with Project ECHO to increase medication-assisted treatment in rural primary care. Psychiatr Serv 2019 Dec 1;70(12):1157-1160. [doi: 10.1176/appi.ps.201900142] [Medline: 31434561]
21. de la Garza Iga FJ, Mejía Alvarez M, Cockroft JD, et al. Using the Project ECHO™ model to teach mental health topics in rural Guatemala: an implementation science-guided evaluation. Int J Soc Psychiatry 2023 Dec;69(8):2031-2041. [doi: 10.1177/00207640231188038] [Medline: 37477264]
22. Mubanga B, Fwoloshi S, Lwatula L, et al. Effects of the ECHO tele-mentoring program on HIV/TB service delivery in health facilities in Zambia: a mixed-methods, retrospective program evaluation. Hum Resour Health 2023 Mar 20;21(1):24. [doi: 10.1186/s12960-023-00806-8] [Medline: 36941682]
23. Stevenson L, Ball S, Haverhals LM, Aron DC, Lowery J. Evaluation of a national telemedicine initiative in the Veterans Health Administration: factors associated with successful implementation. J Telemed Telecare 2018 Apr;24(3):168-178. [doi: 10.1177/1357633X16677676] [Medline: 27909208]
24. Serhal E, Arena A, Sockalingam S, Mohri L, Crawford A. Adapting the Consolidated Framework for Implementation Research to create organizational readiness and implementation tools for Project ECHO. J Contin Educ Health Prof 2018;38(2):145-151. [doi: 10.1097/CEH.0000000000000195] [Medline: 29505486]
25. Agley J, Delong J, Janota A, Carson A, Roberts J, Maupome G. Reflections on Project ECHO: qualitative findings from five different ECHO programs. Med Educ Online 2021 Dec;26(1):1936435. [doi: 10.1080/10872981.2021.1936435] [Medline: 34076567]
26. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci 2009 Aug 7;4:50. [doi: 10.1186/1748-5908-4-50] [Medline: 19664226]
27. Develay É, Wartelle-Bladou C, Talbot A, et al. Implementation of Project ECHO in a university health network: contrasting and comparing experiences across health conditions through a qualitative approach in a Canadian tertiary care centre. BMJ Open 2024 Sep 17;14(9):e082947. [doi: 10.1136/bmjopen-2023-082947] [Medline: 39289013]
28. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
29. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008 Apr;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
30. Elo S, Kääriäinen M, Kanste O, Pölkki T, Utriainen K, Kyngäs H. Qualitative content analysis: a focus on trustworthiness. SAGE Open 2014. [doi: 10.1177/2158244014522633]
31. Crabtree B, Miller W. A template approach to text analysis: developing and using codebooks. In: Doing Qualitative Research: Sage; 1999:163-177.
32. QSR International. : NVivo12; 2017.
33. Lunenburg FC. Organizational structure: Mintzberg's framework. Int J Schol Acad Intellect Divers 2012;14(1):1-8 [FREE Full text]
34. Mintzberg H. Mintzberg on Management Inside Our Strange World of Organizations: Free Press/Collier Macmillian; 1989. [doi: 10.1007/978-1-349-20317-8_23]

35. Moss P, Hartley N, Russell T. Project ECHO®: a global cross-sectional examination of implementation success. BMC Health Serv Res 2024 May 3;24(1):583. [doi: 10.1186/s12913-024-10920-5] [Medline: 38702685]
36. Moss P, Hartley N, Ziviani J, Newcomb D, Russell T. Executive decision-making: piloting Project ECHO® to integrate care in Queensland. Int J Integr Care 2020 Dec 4;20(4):23. [doi: 10.5334/ijic.5512] [Medline: 33335464]
37. Canadian Institute for Health Information, The Expansion of Virtual Care in Canada: New Data and Information: CIHR; 2023.
38. Osei-Twum JA, Wiles B, Killackey T, Mahood Q, Lalloo C, Stinson JN. Impact of Project ECHO on patient and community health outcomes: a scoping review. Acad Med 2022 Sep 1;97(9):1393-1402. [doi: 10.1097/ACM.0000000000004749] [Medline: 35612913]
39. Larson R, Day SK, Dodsworth-Rugani K, et al. PROJECT ECHO implementation: guidance from the field: frequently asked questions. : Diffusion Associates; 2023 URL: https://www.researchgate.net/publication/369550690 [accessed 2025-11-03]
40. Nilsen P. Making sense of implementation theories, models and frameworks. Implement Sci 2015 Apr 21;10:53. [doi: 10.1186/s13012-015-0242-0] [Medline: 25895742]
41. GBD 2021 Low Back Pain Collaborators. Global, regional, and national burden of low back pain, 1990-2020, its attributable risk factors, and projections to 2050: a systematic analysis of the Global Burden of Disease Study 2021. Lancet Rheumatol 2023 Jun;5(6):e316-e329. [doi: 10.1016/S2665-9913(23)00098-X] [Medline: 37273833]

Abbreviations

CFIR: Consolidated Framework for Implementation Research
CHUM: Centre hospitalier de l'Université de Montréal
ECHO: Extension for Community Healthcare Outcomes
RUISSS: Réseau universitaire intégré de santé et de services sociaux de l'Université de Montréal

Edited by D Chartash; submitted 11.04.25; peer-reviewed by B Mishra, H Sadeghsalehi; revised version received 23.09.25; accepted 13.10.25; published 13.11.25.

Please cite as:
Pagé MG, Develay É, Talbot A, Khemiri R, Wartelle-Bladou C
Exploring the Implementation of Multiple Telementoring ECHO Programs From an Institutional and Organizational Perspective: Qualitative Study
JMIR Med Educ 2025;11:e75844
URL: https://mededu.jmir.org/2025/1/e75844
doi:10.2196/75844

Investigating Learning Effects Through the Implementation of Teledermatology Consultations Among General Practitioners in Germany: Mixed Methods Process Evaluation

Andreas Polanc[1], MScPH; Inka Roesel[1,2], MSc; Elke Feil[1]; Peter Martus[2], Prof Dr rer nat; Stefanie Joos[1], Prof Dr med; Roland Koch[1], Dr med

[1]Institute for General Practice and Interprofessional Care, University Hospital Tübingen, Osianderstraße 5, Tübingen, Germany
[2]Institute for Clinical Epidemiology and Applied Biometry, University Hospital Tübingen, Tübingen, Germany

Corresponding Author:
Andreas Polanc, MScPH
Institute for General Practice and Interprofessional Care, University Hospital Tübingen, Osianderstraße 5, Tübingen, Germany

Abstract

Background: The increasing prevalence of dermatological diseases will pose a growing challenge to the health care system and, in particular, to general practitioners (GPs) as the first point of contact for these patients. In many countries, primary care physicians are supported by teledermatology services.

Objective: The aim of this study was to detect learning effects and gains among GPs through teledermatology consultations (TCs) in daily practice.

Methods: As part of a mixed methods study embedded in a cluster-randomized controlled trial (TeleDerm), a full survey and semiguided face-to-face interviews were conducted among GPs of participating intervention practices using the telemedicine approach. A TC assessment tool (TC-AT) was developed to evaluate the quality of clinical data and images of TCs conducted during the run-in and intervention phases, with a score ranging from 0 (lowest quality) to 10 (highest quality). Mixed methods analysis triangulated qualitative content analysis, survey data with a growth curve model calculated from TC-AT data, comparing subjective experiences of GPs with objective process data.

Results: A total of 487 TCs of 33 practices were analyzed. Questionnaires from n=46 GPs (practice-level response rate: 69.9%) were included in the quantitative analysis. Two-thirds of the GPs (n=31; 67.4%) in the written survey rated the TCs as helpful for differential diagnosis and treatment management. Improved self-reported confidence in diagnosing skin diseases due to the timely clinical feedback from dermatologists was reported by more than half of the responding GPs (n=25; 54.3%). In the interviews (n=13), teleconsultations were mainly seen as a learning opportunity by the GPs. Regarding the quality of TCs, a mean TC-AT score of 7.4 (SD 1.7, range 0 - 10) was observed. In the growth curve model, a simple linear time trend provided the best fit to the TC-AT score trajectory across the observed study period. A significant time * TC-AT start score interaction was found (F_{452}=30.66, P<.001). While regardless of the initial TC-AT score, repeated TCs lead to process quality improvements over time, post hoc probing of the TC-AT start score as a moderator of the learning effect over time revealed the highest improvements among GP practices with a lower initial TC-AT score (−1 SD: standardized slope=0.59, P<.001; mean: standardized slope=0.38, P<.001; +1 SD: standardized slope=0.18, P<.001).

Conclusions: TCs have been shown to be an effective method of education for GPs in terms of “learning on the job” in daily practice. The telemedicine approach seems to be an easily implementable and effective tool to support continuing medical education in the field of dermatology. Strategies could be developed to train GPs and medical students in the use of TC to adequately prepare them for the increasing technological demands of their future profession in primary care.

Trial Registration: German Clinical Trials Register DRKS00012944; https://drks.de/search/en/trial/DRKS00012944/details

(JMIR Med Educ 2025;11:e65915) doi:10.2196/65915

KEYWORDS

telemedicine; teledermatology; teleconsultation; experiential learning; competence gain; general practice; primary care; competence; learning effects; consultations; mixed methods; general practitioners; Germany

Introduction

Background

Demographic and environmental factors, climate-related health risks, and lifestyle changes will contribute to a significant increase in skin diseases in the near future, placing a particular burden on health care systems [1-7]. General practitioners (GPs) as the first point of contact for patients with skin conditions play a crucial role in the early detection, accurate diagnosis, and effective management of skin diseases [8]. The care of patients with skin lesions is becoming increasingly difficult due to the growing shortage of specialists, especially in underserved areas.

Teledermatology consultations (TCs) offer an opportunity to alleviate the escalating demand for dermatology services [9-11]. International studies have shown that the telemedicine approach is as effective as traditional medical referrals in reducing costs, enhancing care coordination, and improving patients' quality of life [12-21].

In addition to the aforementioned benefits of TC for patient care, international studies have reported an increase in confidence and learning effects among GPs through TC with their dermatology colleagues [8,15,18,22-28]. Learning effects were achieved, for example, by refreshing old knowledge, sharing new therapeutic concepts, or exchanging difficult cases that were considered essential for their daily work [15,29]. It has been shown that the percentage of correctly prediagnosed cases by GPs increased through TC [8].

Studies have demonstrated teledermatology (TD) to be an effective educational method and an opportunity for dermatologists' and medical students' professional development [30-34]. Lee et al [10] recommended TD as an educational tool for health care providers to interact with for diagnostic support and guidance, especially if they lack specialized dermatology training. In a systematic scoping review of educational programs on melanoma diagnosis for GPs, only 16 of 31 educational interventions included training in dermoscopic diagnosis. Hereof, only one RCT study included TD feedback [35]. To date, little is known about the competence gains, learning effects, and skills acquired by GPs through TC in general practice.

Among these competencies, communication between specialists and primary care is of paramount importance [36]. In a cross-sectional survey by Scaioli et al [37] of more than 7100 GPs in 34 OECD (Organisation for Economic Co-operation and Development) countries on GP-specialist communication in the referral process, the extent and intensity of communication between GPs and medical specialists in a country where a gatekeeping system exists is associated with the organization of the primary health care system. Especially in societies without a mandatory gatekeeping system, as in Germany, there is a higher risk of poor communication between different levels of care [36,37]. In addition, the high workload, the shortage of doctors, and the associated time pressure on doctors may also have a negative impact on communication behavior. According to OECD statistics, Germany had an average of 9.8 doctor-patient contacts per capita in 2019, compared with an OECD average of 6.8 doctor contacts per capita [38].

Despite the need for a more direct exchange between specialists and GPs, written communication still remains the most common form of interdisciplinary exchange between specialists and primary care [36]. Rübsam et al [] [39] identified a need for optimization of interdisciplinary exchange with dermatologists among German GPs, mainly because written feedback from dermatologists was often missing or incomplete, which led to poor communication with dermatology colleagues and also made it very difficult for GPs to pursue or even effectively implement appropriate diagnoses and therapies and deprived them of the opportunity to receive further training based on written feedback 39.

The use of health information technology and a closer collaboration between GPs and specialists have been shown to improve communication [36]. For example, teledermatology has proven to be a useful tool to support direct communication between dermatologists and GPs [23].

While communication via telemedicine was already an integral part of patient care in many European countries, in Germany, liability, professional and data protection regulations, a heterogeneous organization of regional health care due to federalism, and a rather conservative attitude of users have made it difficult to introduce health technology on a large scale [40-43]. It was not until the amendment of the Medical Practitioners' Code (MBO-Ä) in 2017 and the new regulation of exclusive telemedical treatment in the Professional Code of the Medical Association of Baden-Württemberg in June 2020 that the way was paved for the implementation of telemedicine services beyond pilot projects in Baden-Württemberg, about 20 years later than in other European countries [42,44].

Apart from the regulatory framework, it has been proven that adequate reimbursement for teledermatology services plays a critical role in provider participation and the sustainability of such programs [45]. With the amendment of the reimbursement regulation in the statutory insurance on October 1, 2020, a TC request as well as the evaluation of the TC by a medical specialist from then on can be reimbursed [46]. The latter includes not only the assessment of the medical problem, but also a written report to the GP who requested the TC.

Given the persistent deficiencies in interdisciplinary communication, the integration of asynchronous store-and-forward (SaF) TC approach seems to be a promising way to close this communication gap, while at the same time providing German GPs with an on-the-job learning opportunity.

Research Aim and Questions

There is little evidence that TCs lead to learning effects in "real-world" GP settings. The aim of our substudy within the randomized controlled TeleDerm trial was thus to identify learning effects in GPs, defined as subjective reporting of competency gains in combination with objective improvement in process quality. The accompanying mixed methods process evaluation addressed the following research questions: (1) to what extent can the SaF approach be considered an effective training tool for GPs in daily practice in the sense of "learning

on the job", (2) what is the impact of case-based interdisciplinary communication through direct feedback using the SaF approach on the competence and confidence of GPs in the diagnosis of patients with skin complaints and in the assessment of dermatologic diseases, and (3) does the frequency of TC requests in combination with direct feedback from dermatologists support an improvement in the process quality improvement in TCs of GPs over time, indicating competency gains? Based on the results, strategies for the use of TC in GP training and medical education, and for the implementation in primary health care can be derived. The results from this study may benefit health care researchers and persons involved in the continuous professional development of GPs, especially in training scenarios in the context of day-to-day clinical practice.

Methods

Setting and Study Design

TeleDerm was designed as a 2-arm, cluster-randomized confirmatory trial with an accompanying mixed methods process evaluation. It was conducted from 2017 to 2020 in 4 rural and semirural intervention counties in the German federal state of Baden-Württemberg [42,47]. Inclusion criterion for all GP intervention practices (IPs) and patients with skin complaints aged 18 years and older was their participation within the GP-centered health care program of the General Local Health Insurance Fund Baden-Württemberg (AOK-BW). All patients who gave written consent to participate in the TeleDerm trial underwent TC, while patients who refused TC were treated as usual. Based on standardized case documentation, the TC system within our study provided asynchronous (SaF) recording, transmission, and reporting of the patient history (eg, medical history, differential diagnosis, medications, therapy, information on complaints, duration, and changes in shape, color, or size of the lesion). In addition, images of the affected skin area were taken by the primary care physician using a digital camera. The assessment of the TC was to be completed within 48 hours. Via a secure web interface, the dermatologists provided direct, timely, and comprehensive case-based feedback including *ICD-10* (*International Statistical Classification of Diseases and Related Health Problems, 10th Revision*) code and management advice in free text to the requesting GP. In case of further questions from either the GP or the dermatologist, a one-time "question loop" was set up. This meant that when a TC request was made, either the GP or the dermatologist had a one-off opportunity to communicate directly with their medical colleague via a secure web interface to discuss and clarify any outstanding issues relating to the specific TC request. Multimedia Appendix 1 shows the detailed SaF process steps for TC requests in the TeleDerm study. Further information on the study design can be found in the peer-reviewed TeleDerm study protocol [47]. A detailed overview of the mixed methods flow chart within our study is depicted in Figure 1.

Figure 1. Mixed methods study flow chart. GP: general practitioner; TC: teledermatology consultations; TC-AT: teledermatology consultation assessment tool.

Survey

The paper-based questionnaire was developed based on literature and informal interdisciplinary exchange during the preparatory phase of the study and covers the categories of learning effects and competence gain with 9 questions. Additional free text entries were permitted in the questionnaire. In February 2019, the questionnaire was piloted (n=5) with representatives of the target groups using the think-aloud method, and between May and June 2019, the survey was conducted in the participating GP practices (n=46 IP). During its development, the questionnaire was presented several times in an interprofessional plenary session at the Quantitative Methods Research Workshop at the University Hospital of Tübingen and discussed methodologically. Descriptive statistics were calculated by AP using SPSS Statistics version 28.0.0.0 (IBM Corp.). Categorical variables were expressed as frequencies and percentages. A chi-square test (χ^2) was performed to compare user behavior (frequent users vs infrequent users) and learning effect. Free text responses were categorized thematically, using the categories and subcategories developed in 2.3 as a template. Quotes from the free text fields of the survey in the article are marked with the abbreviation "QuesTC." Missing values are excluded from further analysis.

Interviews

Semistructured face-to-face interviews with GPs explored both the impact on interprofessional collaboration and communication between GPs and dermatologists, as well as the competence gain of GPs through feedback from dermatologists. Before piloting the interview guideline with representatives of the target group, the guideline was presented and methodologically discussed in an interprofessional plenary session at the Qualitative Methods Research Workshop at the University Hospital of Tübingen. The interview guideline was piloted by AP in February 2019 (n=5) using the think-aloud method [48]. The participants were asked to give their interpretation of the questions and to explain their answers. According to their feedback, no changes to the interview guideline were necessary.

A stratified selection of interviewees was made according to age (young, middle-aged, and older), gender (male, female), county of origin, and TC use behavior (nonresponder, occasional user, and frequent user). An initial sample size of n=20 interviews with GPs was determined using Malterud 5-dimensional approach of "information power". The dimensions comprise study aim (exploratory), sample specificity (high), established theory (experiential learning [49]), quality of dialogue (expected: mediocre, which means low information density in the transcripts), and analysis strategy [50]. Should the estimated sample size prove insufficient during the analysis, we reserve the right to recruit additional participants. Within the mixed methods evaluation, semistructured face-to-face interviews (AP and EF) were conducted starting March 2019, after informed consent to participate was obtained. Digital recordings of the interviews were transcribed verbatim.

Analysis started with an initial reading of the transcripts in April 2019 to get to know the material and assess the quality of dialogue. Based on this initial reading, we determined in May 2015 that no further information would be obtained through additional interviews due to a high degree of repetition in the interview transcripts and a better dialogue quality than expected. At that point, recruitment to additional interviews was stopped at 13 interviews, and we began with the qualitative analysis.

For qualitative analysis, we assumed a constructivist paradigm that individuals make sense of their subjective reality through narration [51]. Health care professionals learn continuously based on concrete experience [49,52-55]. Qualitative analysis was performed with software support using MaxQDA (VERBI Software GmbH; version 20.1.0) according to Mayring and Fenzl [56]. First, an initial coding frame was built deductively based on the questions of the interview guideline. Both authors (AP and EF) independently applied the coding frame to the same 3 transcripts to assess consistency and reach a shared understanding of how the codes were to be applied. According to the principles of qualitative content analysis, the coding frame was organized into main and subcategories and inductively extended by systematically paraphrasing and generalizing text segments, followed by the formulation of new categories based on recurring content and consensus between coders. Under the supervision of RK, the researchers (AP and EF) iteratively compared new data against existing codes or categories. This ongoing process allowed us to refine and validate the coding frame, ensuring it reflects the evolving understanding of the data. The refined coding frame was then used by AP and EF to code the remaining transcripts, adding and refining categories whenever new information had to be integrated. The coding frame and initial results were also presented and discussed in an interprofessional plenary session at the Qualitative Methods Research Workshop at the University Hospital of Tübingen, in an initial stage as well as the final frame with its detailed categories and subcategories. Quotes from the interviews in the article are marked with the abbreviation "Int."

Development and Reliability of a TC Assessment Tool

Dermatologists need comprehensive clinical data and high-quality images for effective, efficient TC. Especially image quality is crucial for ensuring the same quality of care as a face-to-face consultation [57-59]. A TC assessment tool (TC-AT) was developed to provide a retrospective evaluation of the quality of the TC process. In addition to assessing the quality of the uploaded images, the tool was designed to assess whether the dermatologist was provided with all of the patient information necessary for diagnosis, a detailed description of the dermatologic changes, and the diagnostic question. The results of the AC-TC assessments should show the extent to which the participating GPs were able to perform the relevant process steps in the same way throughout the study and to "deliver" the information relevant to the dermatologist's diagnosis with regard to improvement over time.

Referring to the training concept developed for the participating GPs in the TeleDerm study and the recommendations of the TC guidelines of the German Dermatological Society (DDG) and the Federal Association of German Dermatologists (BVDD) [60], 2 of the authors (AP and RK) developed an assessment tool focusing on the process quality of clinical data and images for SaF TC. Based on the evaluation of randomly selected TCs by 2 raters (AP and RK), the TC-AT was developed and

evaluated within a 2-step process. To ensure the uniformity of the raters' evaluation and to guarantee the simplicity of the criteria catalog, it was streamlined after the first evaluation and the criteria were formulated more specifically (Step 1). Both the pairwise Cohen kappa statistic for single item TC assessment scores (κ=0.852; 95% CI 0.776 - 0.928) as well as the intraclass correlation coefficient (ICC) for overall item TC assessment scores (ICC=0.871; 95% CI 0.697 - 0.945) based on a final assessment of randomly selected TCs (n=24) showed a very good test-retest reliability (Step 2).

Whereas the interrater reliability calculation for assessment of randomly selected TCs (n=24) for single scores was carried out by means of the pairwise Cohen kappa statistic, the reliability concerning the overall TC assessment scores was analyzed by using ICC estimates and their 95% CIs based on a mean-rating (k=2), absolute-agreement, 2-way random-effects model, both times using SPSS Statistics (IBM Corp.; version 28.0.0.0).

The adjusted and finalized qualitative evaluation criteria covered the following aspects: patient history, description of skin area, dynamics of changes, therapy, and medical assessment, as well as the quality of images regarding exposure, contrast, and acuity (Table 1). Each item with respect to clinical data was scored with either 0 (quality criterion not fulfilled) or 1 point (fulfilled). To appropriately weight the importance of the image quality in the evaluation process, image quality was scored from a minimum of 0 to a maximum of 5 points. Item scores were summed up, resulting in an overall TC-AT score range from 0 (poorest quality) to 10 points (highest quality) per TC.

Table . Teledermatology consultation assessment tool (TC-AT).

Item number	Item name	Criteria	Points
A. Clinical data			
A.1	Patient history[a]	Symptoms and duration of complaints	1
A.2.	Skin area[b]	Description of localization, size, and color	1
A.3.	Dynamic[c]	Changes in size, shape, and color of the affected skin area	1
A.4.	Therapy	Information on medical treatment so far	1
A.5.	Medical assessment	Formulation of a presumptive diagnosis or question for the dermatologist	1
B. Images			
B.1.	Quality of images[d]	Exposure, contrast, and acuity	5
Overall TC assessment score		Maximum points	10

[a]Patient history: at least 1 out of 2 listed criteria is to be specified

[b]Skin area: at least 2 out of 3 listed criteria are to be specified, otherwise the item is scored 0 points

[c]Dynamic: at least 1 out of 3 listed criteria is to be specified

[d]Quality of images: 3 out of 3 criteria: 5 points (all fulfilled); 2 out of 3 criteria: 4 points (exposure, contrast); 1 of 3 criteria: 3 points (exposure); 0 of 3 criteria: 0 points (nothing fulfilled, not usable)

Statistical Analysis of TC-AT Score Time Trend

To explore the evolution of TC quality over time, a mixed model was fitted to the TC-AT data. As the total number of TCs per GP practice differed and time points of TCs were unevenly spaced, time was introduced as a continuous variable. In the model building process, linear and quadratic time terms were introduced as fixed effects. Furthermore, the interaction between the starting TC-AT score at baseline and time was assessed. In addition, we investigated whether introducing the overall number of TCs per GP practice as a covariate added significantly to the improvement in model fit. A random intercept and slope for GP practices were added to account for cluster effects. Due to convergence problems, a simple diagonal covariance structure for random effects was chosen. For the residuals, an autoregressive error structure AR(1) was assumed. Model selection was achieved by forward selection: terms were only added when the model significantly improved, as tested with a log-likelihood ratio test. In addition, the Akaike information criterion and Bayesian information criterion were evaluated, deeming models with lower fit criteria as more appropriate.

The final mixed model included time (linear) and a time*baseline TC-AT start score interaction as fixed effects and a random intercept for GP practices. Simple slopes analysis as post hoc probing was conducted to determine the nature of the interaction. We furthermore applied the Johnson-Neyman technique to identify the region of significance. Model assumptions were checked visually and found adequate. A *P* value <.05 (2-sided) was considered statistically significant and adjusted for multiple testing with Bonferroni correction where necessary. All data analyses were implemented in R (version 4.1.3; R Core Team) and R Studio (version 2022.02.1; Posit Software, PBC). Linear mixed models were fitted using the *lme4* [61] and *nlme* [62] packages.

To assess whether the frequency of TC also impacted the subjective impression of competency and confidence gain, the GPs' user behavior was categorized into frequent users (≥11 TC) or low users (≤10 TC) on the practice level. The cutoff value was based on the median number of teleconsultations per practice. A chi-square test was used to compare the GPs' user behavior and their subjective learning effect. No expected cell frequencies were below 5.

Ethical Considerations

The TeleDerm study was approved by the Ethics Review Board of the Eberhard Karls University of Tübingen (Ref. No. 395/2017BO1). Participation in the study was voluntary. Participants were informed of the study's purpose and their right to withdraw and prohibit the use of their data at any time. They gave written consent to participate in the study, and to collect and to use their data for the study's purposes.

Results

Characteristics of Participants

A total of n=49 GPs participated in the full survey. One GP was excluded from further analysis of the survey because this IP did not conduct any TC. In addition, 2 other GPs were excluded from further analysis because they did not respond to the questions on learning effects and competence gain. Thus, questionnaires from n=46 GPs (response rate of 69.6% on practice level) were included in the further quantitative evaluation.

Table 2 summarizes the main characteristics of the study population at the level of GP practices and health care providers in the full survey of GP intervention practices and in the semistructured interviews with GPs.

Table . Practice and provider-level characteristics of the study population.

Characteristics	Survey	Semistructured interviews
GPs in total, n (% on GP-practice level)	46 (69.6)	13 (28.3)
Male, n (%)	33 (71.7)	8 (61.5)
Age (years), mean (SD, range)	56.87 (7.963, 37 - 75)	59.0 (8.617, 40 - 75)
Job experience (years), mean (SD, range)	22.09[a] (9.273, 1 - 38)	Not recorded

[a]n=1 missing

Educational Value of TCs

According to the survey, 9 out of 10 GPs stated that they learn more about skin conditions through TC than by regular referrals, of which 20 GPs (43.4%) strongly agreed with this item (Figure 2, item C5).

In the interviews, several GPs elaborated that in contrast to normal referrals, TC offers a direct, case-based exchange between specialist and GP:

> *In the past, patients were sent to the dermatologist, who examined them and wrote a letter. Today, this feedback from dermatologists is almost non-existent. And that's a pity because in the meantime we no longer know whether the tentative diagnosis was correct or not, what the dermatologist is doing. So, this learning effect, which we used to have automatically through the report of the specialist, has largely disappeared in the meantime. And this would now be an opportunity to reconnect with this learning effect that we used to have automatically.* [(IntA3-29)]

The GPs emphasized the benefit of a timely, direct reporting from the dermatologist for their learning effect, as opposed to regular referrals

> *For me, the learning effect is good, simply also because of the quick feedback* [(IntA12-64)]

and their therapy recommendations for the further treatment of the GP patients

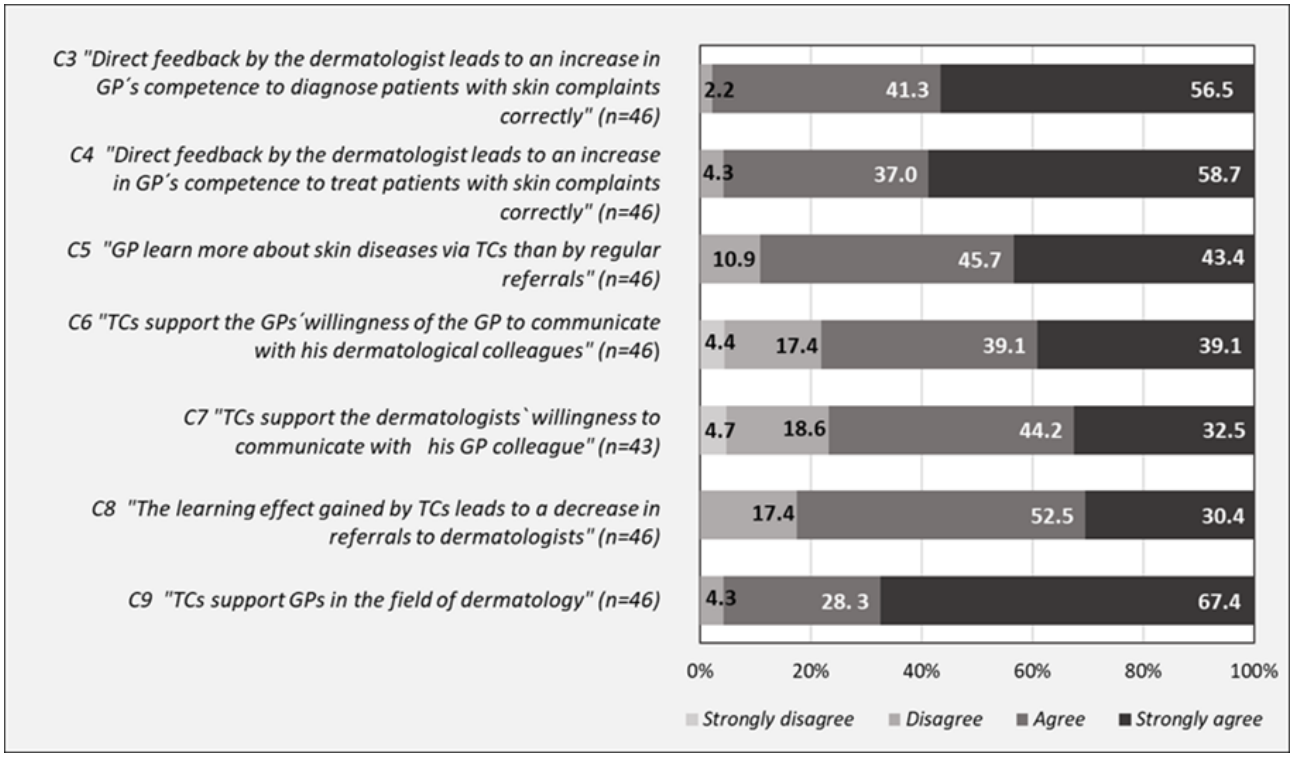
> *We started the therapy on the recommendation of the dermatologists, and it was quite successful.* [(IntA5-24)]

In addition, interviewees indicated that personal feedback helps physicians to validate their own findings and thus actively provides an educational impetus for the primary care physicians:

> *Sometimes you are confirmed, and sometimes new aspects are added, which you nevertheless include and perhaps treat.* [(IntA7-38)]

Despite the many positive effects of direct and personal interaction between GPs and their dermatology colleagues, some respondents were skeptical about the use of telemedicine for communication between the 2 groups. According to the survey results, more than 1 in 5 physicians disagreed with the statement that TCs promote GPs' willingness to communicate with dermatologists, while 39.1% strongly agreed with this point (item C6). GPs were even more skeptical about the willingness of dermatologists to communicate with GPs via TC. Only one-third (32.5%) of GPs strongly agreed that telemedicine increased dermatologists' willingness to communicate with GP colleagues, while 23.3% disagreed, including 4.7% who strongly disagreed (Figure 2, item C7).

Figure 2. Learning effect (Survey).



When asked about the support of GPs by TCs in the field of dermatology, the survey revealed an agreement among the participating physicians about the benefits for their daily practice. Using a 4-point Likert scale, two-thirds of the respondents (67.4%) strongly agreed that TCs support them, while only a small number of respondents (4.3%) strongly disagreed with their colleagues in this regard (Figure 2, item C9).

Increase in Confidence of Diagnostic and Management Skills in GPs

The GPs were asked about their feeling of learning and competence gain during the TeleDerm project. On a 3-point Likert scale, 54.3% of the GPs reported an improvement in their ability to correctly diagnose patients with skin diseases (item C1). In total, 47.8% of the respondents stated that their competence in performing and assessing dermoscopy in patients with skin diseases had increased during the TeleDerm study (item C2), 77.3% of them were male GPs. One in 2 GPs had not noticed any change in either of these areas.

Interview participants reported that direct interaction between GPs and dermatologists has a positive impact on physicians' confidence and acquisition of dermatologic knowledge regarding the treatment and management of common skin diseases and lesions.

> *You have a tentative diagnosis and now, through the TC, you have the possibility to get confirmation. You gain confidence and that is of course an enormous learning effect.* [(IntA3-29)]

In the survey, almost all GPs agreed that the direct feedback from the dermatologists leads to an increase in GPs' competence to correctly diagnose (Figure 2, item C3) and treat (Figure 2, item C4) patients with skin complaints.

Physicians also see qualitative potential in the learning effect in terms of increased treatment safety, strengthening of their role, and simplification of processes in the care system:

> *[The learning effect] of course also offers a safety factor that can be gained for oneself as a therapist and makes many steps simply unnecessary* [(IntA10-34)]

This statement was further substantiated by a GP, especially regarding a possible relief of the dermatologists as well as an extension of the approach to other areas of work in family medicine:

> *TeleDerm-approach is a good introduction to further telemedicine applications in the classic family doctor/specialist division of labor. There is thus a possibility of avoiding the clogging of the specialist's schedule with unnecessary consultations* [(QuesTC43-04)]

While GPs were overall very positive about the previous items presented, the respondents seemed to be more cautious about the statement that the learning effect gained by TC will lead to a decrease in referrals of patients with skin complaints to dermatologists. Only 30.4% of GPs strongly agreed with this statement, while almost 1 in 5 GPs (17.4%) disagreed with their colleagues (Figure 2, item C8).

Learning Curve Through Iterative Processes

The number of TC requests during our study had a positive effect on GPs' subjective competence gain in frequent users. Among GPs who frequently used the service, 64.3% (n=18) reported an increase. The findings of this study, as evidenced by the results of the survey, demonstrated a significance between user behavior and competence in performing and assessing

dermoscopy in patients with skin disease, as indicated by the following statistical analysis: χ^2 (1)=7.769, *P*=.005, ϕ=0.411.

Interview participants also reflected on the association between the frequency of TC and GPs' professional expertise in the field of diagnostic and therapy of patients with skin diseases:

> *I would have liked it better if we could have presented many more patients, because the knowledge gain for each of us would have been correspondingly higher.* [(IntA4-27)]

Next to repetitive TCs, the use of a dermascope was associated with learning from the GP's perspective:

> *By [.] working with the dermascope, you look at it even more closely [.] and when you get the feedback, you've already learned something from it - in case of a repetition, it's easier to recognize what it was*
>
> *(IntA1-47)*

TC Data Flow

Of the 568 TC datasets, 71 TCs had to be excluded due to GP dropouts, duplicates, TC sham datasets, or non-compliance with inclusion criteria.

In addition, 5 nonresponding practices did not conduct any TCs. So far, there is no evidence of when a learning effect can be expected from repeated TC requests. However, since curve fitting requires at least 3 points to derive a trend, only IPs with more than 2 TCs were included in the further analysis of the learning effect. Therefore, n=8 practices with less than 3 TCs per practice during the run-in and intervention phases were excluded, for a total of n=10 TCs. Finally, a total of 33 practices with 487 TCs were included in the subsequent statistical analysis to examine the learning effect over time.

Learning Effects

The median number of TCs on the practice level was 10 (IQR 6, range 3 - 75), and the mean assessment TC-AT score evaluated by both raters reached a value of 7.44 (SD 1.68; median 8, IQR 2, range 0 - 10). It has to be pointed out that 3 "*power-user*" IPs were responsible for more than one-third (n=174; 35.72%) of all TC requests during the run-in and the intervention phase. The last TC was conducted 426 days after baseline (ie, after the first TC of the respective IP).

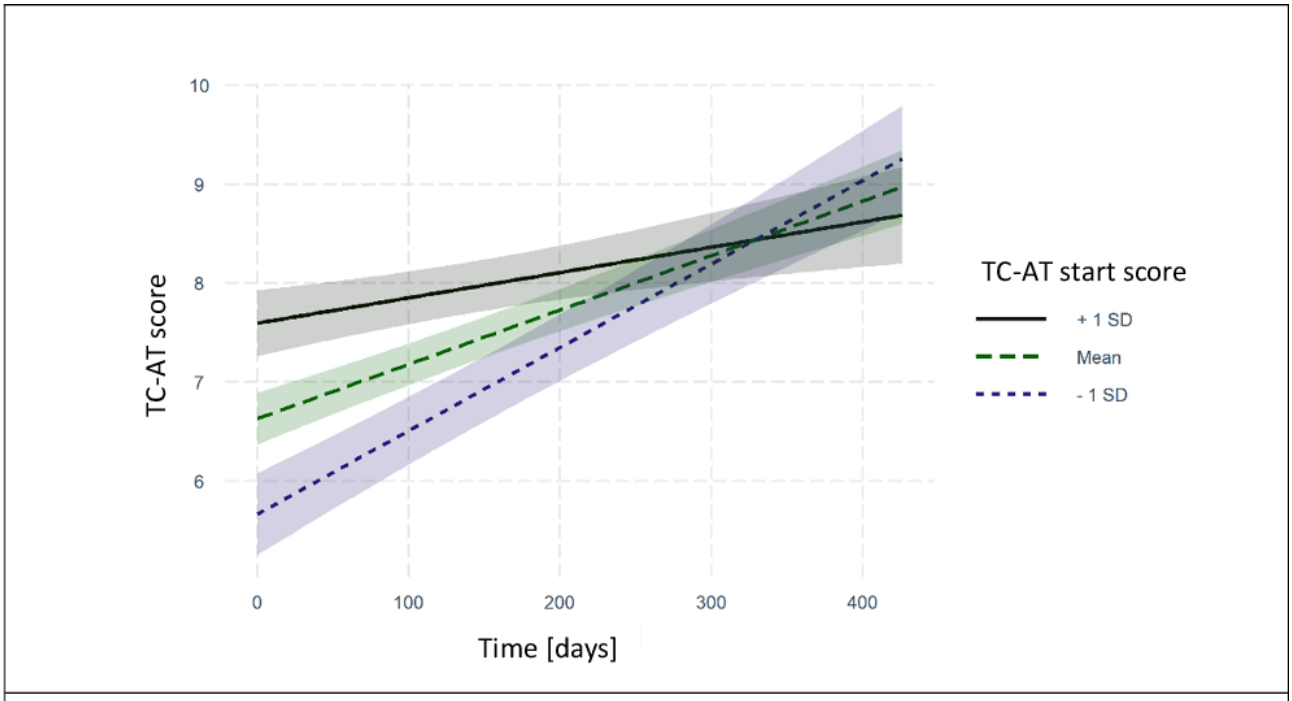
Results of the final linear mixed model are given in Table 3. A statistically significant time * TC-AT start score interaction was found (F_{452}=30.663, *P*<.001), indicating that progression of TC quality over time was influenced by the initial TC-AT score. In Figure 3, post hoc probing of the interaction is displayed by simple slopes at the mean (=average TC start score; dashed line) and 1 SD below (=low start score; dotted line) and above the mean TC-AT score level (=high start score; solid line). At all 3 moderator levels, there was a statistically significant positive effect of time on the improvement in TC quality (*P*<.001). The learning process over time was strongest for GPs with a lower-than-average TC-AT start score (standardized slope=0.590) and less pronounced for GP practices starting with a higher-than-average TC-AT start score (standardized slope=0.179). The Johnson-Neyman plot (Multimedia Appendix 2) revealed that the interaction was statistically significant below 249 days.

Table . Linear mixed model results for the evolvement of the teledermatology consultation assessment tool score over time.

	TC-AT score			
Fixed effects	B[a]	β[b]	95% CI	*P* value
(Intercept)	3.649	.03	−0.09 to 0.15	<.001
Time	0.014	.38	0.31 - 0.46	<.001
TC-AT start score	0.460	.30	0.17 - 0.43	<.001
time * TC-AT start score	−0.001	−.21	−0.28 to −0.13	<.001
Random effects				
σ^2[c]	1.90	—[d]	—	—
$\tau 00$[e]	0.16	—	—	—
ICC[f]	0.08	—	—	—
$N_{practices}$	33	—	—	—
Observations	487	—	—	—
Marginal R^2	0.252	—	—	—
Conditional R^2	0.309	—	—	—
AIC[g]	1730.031	—	—	—

[a]B: unstandardized beta coefficient.
[b]β: standardized beta.
[c]σ^2: Within-group (residual) variance.
[d]Not applicable.
[e]τ00: Between-group variance.
[f]ICC: intraclass correlation coefficient.
[g]AIC: Akaike information criterion.

Figure 3. Post hoc probing of the moderation effect of the teledermatology consultation assessment tool start score on the learning effect over time. TC-AT: teledermatology consultation assessment tool.



Moderator level				
TC-AT start score	Slope	SE	Standardized Slope	*P*
+1 SD	0.00256	0.000745	0.179	<.001
Mean	0.00549	0.000566	0.385	<.001
-1 SD	0.00843	0.000805	0.590	<.001

Discussion

Principal Results

Our findings show subjective improvements in GPs' competency paired with observable improvement in TC process quality over time assessed through the TC-AT score. The mixed methods analysis revealed that 2 main aspects contribute to these learning effects: interdisciplinary communication and learning on the job. Both aspects will be discussed in the context of the current research situation in the following section.

Interdisciplinary Communication Increases GP Expertise

According to our study, timely feedback and direct exchange with specialists on diagnosis, therapy, and treatment options seem to promote learning by validating the GP's own findings. Thus, it may increase the confidence of physicians in treating patients with skin diseases in daily practice.

As with TeleDerm, German GPs in the study by Rübsam et al [39] criticized the communication with the dermatologists and the lack of written feedback, which meant that the learning effects that GPs had previously gained from the specialists' reports were lost, making it difficult for GPs to understand and effectively implement an appropriate diagnosis or treatment. Apart from a few critical voices, the majority of our GPs were of the opinion that the TC approach was a supporting tool and a very good opportunity to increase the willingness to communicate and thus to improve the interdisciplinary exchange on both sides in the future. Also, based on the findings of van den Akker et al [23], TC is viewed as a useful tool for interpersonal communication between GPs and dermatologists.

The standardization of TC web templates to ensure the collection of necessary information not only promotes their structured implementation but also provides guidelines for structured interdisciplinary communication between GPs and dermatologists, thus increasing diagnostic and management concordance [25]. This process is particularly supported by the provision of good diagnostic images. According to the results of our main article on the TeleDerm study, 79.8% of the specialists stated that the image quality provided by the GPs during the intervention phase was sufficient for the diagnosis, underlining the importance of this approach [42].

By supporting this interdisciplinary exchange with specialists, the telemedicine approach proves to be a helpful training tool for physicians that can be used in daily practice and at the same time offers the opportunity to close the lamented communication gap between German GPs and dermatologists. Especially the recent changes in the framework conditions in Germany (changes in data protection and the medical code of conduct and the possibility of billing for teleconsultations) may contribute to a sustainable change in communication behavior and stimulate direct exchange not only between GPs and dermatologists but also with other disciplines. Although practitioners frequently perceive standardized forms as a source of double work, this study indicates that structured data

processing through such tools may enhance the quality of care by organizing relevant information more effectively.

On-the-Job Learning in Daily Practice

Practices with initially low TC-AT scores showed the steepest improvement in the quality of TC requests over time, indicating a stronger learning effect in inexperienced practices. The improvement in the process quality of TC requests over time can be explained as a result of this on-the-job learning among GPs [42]. Mohan et al [24] also demonstrated a statistically significant improvement in dermatological knowledge among frequent users of SaF TC among GPs. An effect between the time factor and the frequency of use and repetition in everyday practice is to be expected, especially if the TC requests are also used repeatedly by both sides for refresher and training purposes, for example by discussing differential diagnoses, exchanging alternative treatment approaches, or also in the case of reasons for differences of opinion [15,26]. Frequent use appears to be the most effective way for GPs to build competence with the tools. Educational formats should therefore aim to support the initial adoption phase and help overcome early barriers to use.

Based on the Kolb experiential learning cycle, our results show that frequent use of TC and access to an expert opinion on skin complaints help GPs to reflect on their daily practice. International research supports that TC is a time-saving and effective training tool for GPs and supports continuous learning [15,18,33,34]. The interdisciplinary exchange and communication between GPs and dermatologists could promote the continuous and experiential learning process on many medically relevant aspects—therapy, diagnostics, lesion description, and image quality. This low-level implementation facilitates sustainable and continuous medical education for generalists. The integration of TC into the daily clinical practice of GPs appears to have a positive impact on both perceived and self-reported gains in dermatological knowledge, as well as confidence in diagnostic skills and care management. Furthermore, the feedback and self-assessment by GPs participating in the study indicated that the adoption of TC seems to support improvements in GPs' professional competence, particularly in the assessment of patients with dermatologic complaints. Our results thus follow the findings of international studies that direct and case-based exchange with specialists provides an educational impetus for GPs by refreshing previously acquired knowledge, but also by sharing new therapeutic concepts, especially when used to discuss differential diagnoses, alternative treatment strategies, or even reasons for disagreement [15,18,25,26]. GP's reflective practice is supported by iterative, direct feedback by specialists.

In Germany, telemedicine was opened about 2 decades later than in many other European countries. In order to cope with the increasing digitalization of health care, today's medical students must be prepared for the complex challenges of their future professional life [63]. For this reason, the necessary practice-oriented approaches to digitalization and the teaching of interdisciplinary and intersectoral skills should be included in the compulsory curriculum of medical faculties as a useful combination and should be taught there at an early stage [63,64]. By introducing TC early in medical education, the next generation of practitioners will have already navigated the initial challenges, allowing them to focus more effectively on "learning on the job."

Strengths and Limitations

The TeleDerm study is one of the first studies to investigate the effects of "learning on the job" among primary care physicians in the context of dermatology teleconsultations. It should be emphasized that the mixed methods process evaluation is based on 3 different qualitative and quantitative data sources. While the process data collected at the practice level from teleconsultation requests provided a large amount of objectively evaluable data from the "real world" to answer the research question, the full survey allowed a representative assessment of the GPs. The interviews also supported an in-depth qualitative investigation of individual experiences and self-assessment of learning effects and perceived competence gains in relation to dermatological expertise of the target group under investigation. The triangulation of the different data sources thus allowed for a comprehensive analysis and a multifaceted consideration of the research questions from the perspective of the GPs at both group and individual levels.

The operationalization of learning effects as subjective competency gains in combination with objective improvements in process quality is limited. A randomized controlled trial design with the main focus on these learning effects with a stronger emphasis on clinical aspects (eg, by issuing a test of dermatologic skills) would have been preferable but was effectively impossible to implement into the main study.

The TC-AT was developed and tested to evaluate the TCs conducted. All TCs included in the evaluation were assessed and objectively evaluated on the basis of these quality criteria. As a limiting factor in the process evaluation, it should be noted that the evaluation matrix on which the qualitative assessment was based was developed post hoc by the authors. Another limiting factor is that only the formal process indicators collected during the study period were evaluated. A professional evaluation of the medical plausibility, such as a professional evaluation of the diagnosis, therapy, or even the treatment success of the teleconsultations, was not provided for in the study design and was therefore not carried out. Nevertheless, the matrix tool allowed a continuous recording of process and data quality over time, and we were able to describe central aspects of GP's education through TC based on real-world data.

Conclusions

The telemedicine approach proves to be an effective training tool for enhancing the diagnostic and therapeutic skills of GPs. The positive learning effects are enabled by direct and timely interprofessional exchange. With the increasing digitalization of our health care system, telemedicine has the potential to enrich medical education as well as training of GPs or specialists. Furthermore, incorporating telemedicine into medical education will ensure that future health care professionals are equipped to meet the evolving needs of modern health care delivery.

Acknowledgments

TeleDerm was funded by the Innovation Fund at the Federal Joint Committee (Innovationsausschuss beim Gemeinsamen Bundesausschuss (G-BA)), Berlin, Germany (grant number 01NVF16012). We thank all the GPs and dermatologists involved in the study for their support and cooperation. We acknowledge the support from the Open Access Publication Fund of the University of Tübingen.

Disclaimer

The sponsors had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability

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

Authors' Contributions

AP led the manuscript drafting (Introduction, Methods, Results, and Discussion) and coordinated the drafting process. RK and IR contributed to drafting the Methods, Results, and Discussion sections. Statistical analysis and data visualization were performed by AP and IR. Process data analysis was conducted by AP, RK, and IR. Quantitative and qualitative data analysis were performed by AP, RK, and EF. Project coordination was overseen by AP. Statistical planning and supervision were provided by PM. SJ was responsible for project planning, study design, funding acquisition, and overall project supervision. All authors reviewed and revised the final manuscript under the supervision of AP. The final version of the manuscript was compiled by AP.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Main process steps for teledermatology requests in the TeleDerm study.
[DOCX File, 21 KB - mededu_v11i1e65915_app1.docx]

Multimedia Appendix 2
Johnson-Neyman plot indicating the size and significance of the slope of the teledermatology consultation assessment tool score across time. Shaded areas indicate 95% CIs.
[PNG File, 12 KB - mededu_v11i1e65915_app2.png]

References

1. Maybaum T. Dermatologie: Zahl der Hautkrebsfälle drastisch gestiegen [Article in German]. Dtsch Arztebl Int 2019;116:1808 [FREE Full text]
2. Böhm K, Mardorf S, Nöthen M, et al. Gesundheit Und Krankheit Im Alter [Book in German]: Robert Koch-Institut; 2009:323. [doi: 10.25646/3145]
3. Sawada Y, Nakamura M. Daily lifestyle and cutaneous malignancies. Int J Mol Sci 2021 May 14;22(10):34069297. [doi: 10.3390/ijms22105227] [Medline: 34069297]
4. Holman DM, Ragan KR, Julian AK, Perna FM. The context of sunburn among U.S. adults: common activities and sun protection behaviors. Am J Prev Med 2021 May;60(5):e213-e220. [doi: 10.1016/j.amepre.2020.12.011] [Medline: 33589300]
5. Leiter U, Garbe C. Epidemiology of melanoma and nonmelanoma skin cancer--the role of sunlight. Adv Exp Med Biol 2008;624(89-103):89-103. [doi: 10.1007/978-0-387-77574-6_8] [Medline: 18348450]
6. Leiter U, Eigentler T, Garbe C. Epidemiology of skin cancer. Adv Exp Med Biol 2014;810(120-40):120-140. [doi: 10.1007/978-1-4939-0437-2_7] [Medline: 25207363]
7. Augustin J, Stephan B, Augustin M. Klimawandelbedingte Veränderungen in Der UV-Exposition: Herausforderungen Fu¨r Die Prävention UV-Bedingter Hauterkrankungen [Book in German]: Klima Und Gesundheit. Medizinisch Wissenschaftliche Verlagsgesellschaft; 2021:119-131. [doi: 10.32745/9783954666270-9]
8. Tensen E, van Sinderen F, Witkamp L, Jaspers MWM, Peute LWP. The value of teledermoscopy to the expertise of general practitioners diagnosing skin disorders based on ICD-10 coding. Stud Health Technol Inform 2019 Aug 21;264(834-8):834-838. [doi: 10.3233/SHTI190340] [Medline: 31438041]
9. Elsner P. Teledermatology in the time of COVID-19 – a systematic review. J Deutsche Derma Gesell 2020 Aug;18(8):841-845 [FREE Full text] [doi: 10.1111/ddg.14180]
10. Lee JJ, English JC. Teledermatology: a review and update. Am J Clin Dermatol 2018 Apr;19(2):253-260. [doi: 10.1007/s40257-017-0317-6] [Medline: 28871562]

11. Pearlman RL, Le PB, Brodell RT, Nahar VK. Evaluation of patient attitudes towards the technical experience of synchronous teledermatology in the era of COVID-19. Arch Dermatol Res 2021 Nov;313(9):769-772. [doi: 10.1007/s00403-020-02170-2] [Medline: 33403572]
12. Yeboah CB, Harvey N, Krishnan R, Lipoff JB. The impact of COVID-19 on teledermatology: a review. Dermatol Clin 2021 Oct;39(4):599-608. [doi: 10.1016/j.det.2021.05.007] [Medline: 34556249]
13. Trettel A, Eissing L, Augustin M. Telemedicine in dermatology: findings and experiences worldwide - a systematic literature review. J Eur Acad Dermatol Venereol 2018 Feb;32(2):215-224. [doi: 10.1111/jdv.14341] [Medline: 28516492]
14. Tensen E, van der Heijden JP, Jaspers MWM, Witkamp L. Two decades of teledermatology: current status and integration in national healthcare systems. Curr Dermatol Rep 2016;5(96-104):96-104. [doi: 10.1007/s13671-016-0136-7] [Medline: 27182461]
15. van der Heijden JP, de Keizer NF, Bos JD, Spuls PI, Witkamp L. Teledermatology applied following patient selection by general practitioners in daily practice improves efficiency and quality of care at lower cost. Br J Dermatol 2011 Nov;165(5):1058-1065. [doi: 10.1111/j.1365-2133.2011.10509.x] [Medline: 21729026]
16. Whited JD. Teledermatology. Med Clin North Am 2015 Nov;99(6):1365-1379. [doi: 10.1016/j.mcna.2015.07.005] [Medline: 26476258]
17. Ferrandiz L, Moreno-Ramirez D, Nieto-Garcia A, et al. Teledermatology-based presurgical management for nonmelanoma skin cancer: a pilot study. Dermatol Surg 2007 Sep;33(9):1092-1098. [doi: 10.1111/j.1524-4725.2007.33223.x] [Medline: 17760600]
18. van Sinderen F, Tensen E, Lansink RA, Jaspers MW, Peute LW. Eleven years of teledermoscopy in the Netherlands: a retrospective quality and performance analysis of 18,738 consultations. J Telemed Telecare 2024 Jul;30(6):1037-1046. [doi: 10.1177/1357633X221122113] [Medline: 36052405]
19. Wootton R, Bahaadinbeigy K, Hailey D. Estimating travel reduction associated with the use of telemedicine by patients and healthcare professionals: proposal for quantitative synthesis in a systematic review. BMC Health Serv Res 2011 Aug 8;11(185):21824388. [doi: 10.1186/1472-6963-11-185] [Medline: 21824388]
20. Whited JD, Warshaw EM, Kapur K, et al. Clinical course outcomes for store and forward teledermatology versus conventional consultation: a randomized trial. J Telemed Telecare 2013 Jun;19(4):197-204. [doi: 10.1177/1357633x13487116] [Medline: 23666440]
21. Romero G, de Argila D, Ferrandiz L, et al. Practice models in teledermatology in Spain: longitudinal study, 2009-2014. Actas Dermosifiliogr (Engl Ed) 2018 Sep;109(7):624-630. [doi: 10.1016/j.ad.2018.03.015] [Medline: 29807618]
22. Wootton R, Bloomer SE, Corbett R, et al. Multicentre randomised control trial comparing real time teledermatology with conventional outpatient dermatological care: societal cost-benefit analysis. BMJ 2000 May 6;320(7244):1252-1256. [doi: 10.1136/bmj.320.7244.1252] [Medline: 10797038]
23. van den Akker TW, Reker CH, Knol A, Post J, Wilbrink J, van der Veen JP. Teledermatology as a tool for communication between general practitioners and dermatologists. J Telemed Telecare 2001;7(4):193-198. [doi: 10.1258/1357633011936390] [Medline: 11506753]
24. Mohan GC, Molina GE, Stavert R. Store and forward teledermatology improves dermatology knowledge among referring primary care providers: a survey-based cohort study. J Am Acad Dermatol 2018 Nov;79(5):960-961. [doi: 10.1016/j.jaad.2018.05.006] [Medline: 29753059]
25. Cumsky HJL, Maly CJ, Costello CM, et al. Impact of standardized templates and skin cancer learning modules for teledermatology consultations. Int J Dermatol 2019 Dec;58(12):1423-1429. [doi: 10.1111/ijd.14437] [Medline: 30916785]
26. Thind CK, Brooker I, Ormerod AD. Teledermatology: a tool for remote supervision of a general practitioner with special interest in dermatology. Clin Exp Dermatol 2011 Jul;36(5):489-494. [doi: 10.1111/j.1365-2230.2011.04073.x] [Medline: 21507041]
27. van Sinderen F, Tensen E, van der Heijden JP, Witkamp L, Jaspers MWM, Peute LWP. Is teledermoscopy improving general practitioner skin cancer care? Stud Health Technol Inform 2019 Aug 21;264:1795-1796. [doi: 10.3233/SHTI190652] [Medline: 31438348]
28. Song E, Amerson E, Twigg AR. Teledermatology in medical and continuing education. Curr Derm Rep 2020 Jun;9(2):136-140. [doi: 10.1007/s13671-020-00304-3]
29. van der Heijden JP, Spuls PI, Voorbraak FP, de Keizer NF, Witkamp L, Bos JD. Tertiary teledermatology: a systematic review. Telemed J E Health 2010;16(1):56-62. [doi: 10.1089/tmj.2009.0020] [Medline: 20064068]
30. Muntz MD, Franco J, Ferguson CC, Ark TK, Kalet A. Telehealth and medical student education in the time of COVID-19-and beyond. Acad Med 2021 Dec 1;96(12):1655-1659. [doi: 10.1097/ACM.0000000000004014] [Medline: 35134026]
31. Loh CH, Ong FLL, Oh CC. Teledermatology for medical education in the COVID-19 pandemic context: a systematic review. JAAD Int 2022 Mar;6:114-118. [doi: 10.1016/j.jdin.2021.12.012] [Medline: 35036962]
32. Ladha MA, Lui H, Carroll J, et al. Medical student and resident dermatology education in Canada during the COVID-19 pandemic. J Cutan Med Surg 2021;25(4):437-442. [doi: 10.1177/1203475421993783] [Medline: 33593087]
33. Boyers LN, Schultz A, Baceviciene R, et al. Teledermatology as an educational tool for teaching dermatology to residents and medical students. Telemed J E Health 2015 Apr;21(4):312-314. [doi: 10.1089/tmj.2014.0101] [Medline: 25635528]

34. Mahmood F, Cyr J, Keely E, et al. Teledermatology utilization and integration in residency training over the COVID-19 pandemic. J Cutan Med Surg 2022;26(2):135-142. [doi: 10.1177/12034754211045393] [Medline: 34551623]
35. Harkemanne E, Baeck M, Tromme I. Training general practitioners in melanoma diagnosis: a scoping review of the literature. BMJ Open 2021 Mar 23;11(3):e043926. [doi: 10.1136/bmjopen-2020-043926] [Medline: 33757946]
36. Vermeir P, Vandijck D, Degroote S, et al. Communication in healthcare: a narrative review of the literature and practical recommendations. Int J Clin Pract 2015 Nov;69(11):1257-1267. [doi: 10.1111/ijcp.12686] [Medline: 26147310]
37. Scaioli G, Schäfer WLA, Boerma WGW, Spreeuwenberg PMM, Schellevis FG, Groenewegen PP. Communication between general practitioners and medical specialists in the referral process: a cross-sectional survey in 34 countries. BMC Fam Pract 2020 Mar 17;21(1):54. [doi: 10.1186/s12875-020-01124-x] [Medline: 32183771]
38. OECD. Health at a Glance 2021: OECD Indicators: OECD Publishing; 2021. [doi: 10.1787/ae3016b9-en]
39. Rübsam M, Esch M, Baum E, Bösner S. Dermatology in primary care: coordination, prevention, therapy, and educational needs. Z Allg Med 2015;91(5):227-232. [doi: 10.3238/zfa.2015.0227–0232]
40. Kringos DS, Boerma WGW. In: Hutchinson A, editor. Building Primary Care in a Changing Europe [Internet] Copenhagen (Denmark): European Observatory on Health Systems and Policies (Observatory Studies Series, No 38) 2015. URL: https://www.ncbi.nlm.nih.gov/books/NBK458728 [accessed 2025-04-30]
41. Brauns HJ, Loos W. Telemedicine in Germany. Status, barriers, perspectives. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2015 Oct;58(10):1068-1073. [doi: 10.1007/s00103-015-2223-5] [Medline: 26324096]
42. Koch R, Rösel I, Polanc A, et al. TELEDerm: implementing store-and-forward teledermatology consultations in general practice: results of a cluster randomized trial. J Telemed Telecare 2024 May;30(4):647-660. [doi: 10.1177/1357633X221089133] [Medline: 35578544]
43. Colombo MG, Joos S, Koch R. Implementing interprofessional video consultations with general practitioners and psychiatrists in correctional facilities in Germany: results from a mixed-methods study. BMC Health Serv Res 2023 Jun 5;23(1):578. [doi: 10.1186/s12913-023-09592-4] [Medline: 37277811]
44. Bundesärztekammer. Baden-Württemberg: Fernbehandlung bald für alle Patienten möglich. URL: https://www.bundesaerztekammer.de/presse/aktuelles/detail/baden-wuerttemberg-fernbehandlung-bald-fuer-alle-patienten-moeglich [accessed 2025-04-30]
45. Armstrong AW, Kwong MW, Ledo L, Nesbitt TS, Shewry SL. Practice models and challenges in teledermatology: a study of collective experiences from teledermatologists. PLoS One 2011;6(12):e28687. [doi: 10.1371/journal.pone.0028687] [Medline: 22194887]
46. Standardized Billing Scale - Status: Fourth quarter of 2020. National Association of Statutory Health Insurance Physicians. URL: https://www.kbv.de/documents/praxis/abrechnung/ebm/archiv/2020-4-ebm.pdf [accessed 2025-04-30]
47. Koch R, Polanc A, Haumann H, et al. Improving cooperation between general practitioners and dermatologists via telemedicine: study protocol of the cluster-randomized controlled TeleDerm study. Trials 2018 Oct 24;19(1):583. [doi: 10.1186/s13063-018-2955-2] [Medline: 30355358]
48. Ericsson KA, Simon HA. Protocol Analysis: Verbal Reports as Data (Revised Ed): Cambridge; 1993. [doi: 10.7551/mitpress/5657.001.0001]
49. Kolb DA. Experiential Learning: Experience as the Source of Learning and Development: Prentice-Hall; 1984.
50. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016 Nov;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]
51. Bruner J. Acts of Meaning: Harvard University Press; 1990:xvii-x179.
52. Wijnen-Meijer M, Brandhuber T, Schneider A, Berberat PO. Implementing Kolb´s experiential learning cycle by linking real experience, case-based discussion and simulation. J Med Educ Curric Dev 2022;9:23821205221091511. [doi: 10.1177/23821205221091511] [Medline: 35592131]
53. Gerstenberger JP, Hayes L, Chow CJ, Raaum S. Medical student experiential learning in telesimulation. J Med Educ Curric Dev 2023;10:23821205231216067. [doi: 10.1177/23821205231216067] [Medline: 38025030]
54. Sanseau E, Lavoie M, Tay KY, et al. TeleSimBox: a perceived effective alternative for experiential learning for medical student education with social distancing requirements. AEM Educ Train 2021 Apr;5(2):e10590. [doi: 10.1002/aet2.10590] [Medline: 33842815]
55. Keyserling K, Janetos E, Sprague C. Teaching telehealth during a pandemic and beyond: an intern's survival guide for virtual medicine. J Gen Intern Med 2021 Oct;36(10):3219-3223. [doi: 10.1007/s11606-021-07009-8] [Medline: 34287776]
56. Mayring P, Fenzl T. Qualitative Inhaltsanalyse. In: Baur N, Blasius J, editors. Handbuch Methoden der empirischen Sozialforschung [Book in German]: Springer Fachmedien Wiesbaden; 2019:633-648. [doi: 10.1007/978-3-658-21308-4_42]
57. Tuknayat A, Bhalla M, Dogar K, Thami GP, Sandhu JK. Clinical profile and diagnostic accuracy of patient-submitted photographs in teledermatology. J Clin Aesthet Dermatol 2023 Apr;16(4):21-25. [Medline: 37077931]
58. van der Heijden JP, Thijssing L, Witkamp L, Spuls PI, de Keizer NF. Accuracy and reliability of teledermatoscopy with images taken by general practitioners during everyday practice. J Telemed Telecare 2013 Sep;19(6):320-325. [doi: 10.1177/1357633X13503437] [Medline: 24163296]
59. Jalaboi R, Winther O, Galimzianova A. Explainable image quality assessments in teledermatological photography. Telemed J E Health 2023 Sep;29(9):1342-1348. [doi: 10.1089/tmj.2022.0405] [Medline: 36735575]

60. Augustin M, Wimmer J, Biedermann T. Practice of teledermatology. J Dtsch Dermatol Ges 2018 Jul;16 Suppl 5:6-57. [doi: 10.1111/ddg.13512] [Medline: 29998512]
61. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw 2015 Oct;67(1):48. [doi: 10.18637/jss.v067.i01]
62. Pinheiro J, Bates D, R Core Team. nlme: linear and nonlinear mixed effects models R package version 31-164. 2023. URL: https://www.rdocumentation.org/packages/nlme/versions/3.1-164 [accessed 2024-04-09]
63. Digitalisierung für Gesundheit - Ziele und Rahmenbedingungen eines dynamisch lernenden Gesundheitssystem [Report in German]. : Advisory Council on the Assessment of Developments in the Health Care System URL: https://www.svr-gesundheit.de/fileadmin/Gutachten/Gutachten_2021/SVR_Gutachten_2021.pdf [accessed 2024-02-01]
64. Verordnung zur Neuregelung der ärztlichen Ausbildung [Report in German]. : Federal Ministry of Health; 2023 URL: https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/3_Downloads/Gesetze_und_Verordnungen/GuV/A/AEApprO_RefE_ueberarbeitet.pdf [accessed 2024-02-01]

Abbreviations

AOK-BW: General Local Health Insurance Fund Baden-Württemberg (Allgemeine Ortskrankenkasse Baden-Württemberg)
BVDD: Federal Association of German Dermatologists
DDG: German Dermatological Society
GP: general practitioner
ICC: intraclass correlation coefficient
***ICD-10*:** *International Statistical Classification of Diseases and Related Health Problems, 10th Revision*
IP: intervention practice
OECD: Organisation for Economic Co-operation and Development
SaF: store-and-forward
TC: teledermatology consultation
TC-AT: teledermatology consultation assessment tool
TD: teledermatology

Edited by B Lesselroth; submitted 30.08.24; peer-reviewed by A Reedy-Cooper, M de Zwaan; revised version received 09.05.25; accepted 16.07.25; published 10.09.25.

Please cite as:
Polanc A, Roesel I, Feil E, Martus P, Joos S, Koch R
Investigating Learning Effects Through the Implementation of Teledermatology Consultations Among General Practitioners in Germany: Mixed Methods Process Evaluation
JMIR Med Educ 2025;11:e65915
URL: https://mededu.jmir.org/2025/1/e65915
doi:10.2196/65915

Impact of Motivational Interviewing Education on General Practitioners' and Trainees' Learning and Diabetes Outcomes in Primary Care: Mixed Methods Study

Isaraporn Thepwongsa[1], MD, MFM, PhD; Pat Nonjui[1], MD; Radhakrishnan Muthukumar[2], MBBS, MClinEmbryol, PhD; Poompong Sripa[3], MD

[1]Department of Community, Family and Occupational Medicine, Faculty of Medicine, Khon Kaen University, 123, Mittraparb Road, Khon Kaen, Thailand

[2]Academic Affairs, Faculty of Medicine, Khon Kaen University, Khon Kaen, Thailand

[3]Inverkeithing Medical Group, Inverkeithing, United Kingdom

Corresponding Author:
Isaraporn Thepwongsa, MD, MFM, PhD
Department of Community, Family and Occupational Medicine, Faculty of Medicine, Khon Kaen University, 123, Mittraparb Road, Khon Kaen, Thailand

Abstract

Background: Effective diabetes management requires behavioral change support from primary care providers. However, general practitioners (GPs) often lack training in patient-centered communication methods such as motivational interviewing (MI), especially in time-constrained settings. While brief MI offers a practical alternative, evidence on its impact among GPs and patient outcomes remains limited.

Objective: This study aimed to evaluate the effectiveness of a structured MI educational program for GPs and GP trainees on their MI knowledge and confidence, and its impact on clinical outcomes among patients with type 2 diabetes in primary care settings.

Methods: A mixed methods study was conducted using a before-and-after two-group design with quantitative assessments of GPs' knowledge and patients' biomarkers, supplemented by qualitative interviews. The intervention group (n=35) received a 4-hour interactive MI workshop, optional web-based modules, and brief MI guides. The control group received standard care. A total of 149 and 167 patients with diabetes were included in the study and control groups, respectively.

Results: A paired-sample *t* test was conducted to evaluate the impact of the MI course on the learners' knowledge. There was a statistically significant difference in the knowledge test scores from Time 1 (mean 11.46, SD 3.48) to Time 2 (mean 15.04, SD 2.35), t_{28}= –7.74; *P*<.001 (2-tailed). The mean increase in knowledge score was 3.57 (SD 2.44), with a 95% CI of 2.62 to 4.52, indicating a large and statistically significant effect. The eta-squared statistic indicated a large effect size (eta-squared=0.85). Patients in the intervention group had greater improvements in HbA_{1c} (mean difference= −0.50, 95% CI −0.91 to −0.09; *P*=.02) and diastolic blood pressure (mean difference= −5.96 mmHg, 95% CI −8.66 to −3.25; *P*<.001) compared to controls. Qualitative feedback highlighted the usefulness of brief MI, along with challenges in mastering advanced techniques and time constraints.

Conclusions: The MI educational program improved GP trainees' MI knowledge and patient outcomes. Brief MI appears feasible in primary care but requires ongoing support for skill development and implementation.

(JMIR Med Educ 2025;11:e75916) doi:10.2196/75916

KEYWORDS

motivational interviewing; general practitioners; general practitioner trainees; diabetes management; primary care; behavioral change; web-based learning; brief motivational interviewing

Introduction

Diabetes is a chronic illness with a rising prevalence worldwide [1]. Over the past two decades, the global prevalence of diabetes has significantly increased. In 2021, it was estimated that 536.6 million people were living with diabetes [1]. This number is projected to rise to 642.7 million (11.3%) by 2030 and 783.2 million (12.2%) by 2045 [1]. Furthermore, 3 out of 4 adults with diabetes reside in low- and middle-income countries [1]. Approximately 240 million people are living with undiagnosed diabetes globally, with 90% of these individuals residing in low- and middle-income countries [1]. Diabetes and its complications, such as heart disease, stroke, and chronic kidney

disease, are pressing public health challenges that are expected to worsen globally [2,3].

Patients with diabetes often face significant challenges in adopting and maintaining healthy lifestyle changes, which can hinder blood sugar control and increase the risk of complications. Research and systematic reviews have highlighted multiple barriers to sustained behavioral change, including low motivation, insufficient self-discipline, and limited adherence to recommended lifestyle practices [4-7]. These challenges are often compounded by a lack of personalized guidance and support for self-management from health care providers [8], as well as systemic constraints such as limited consultation time for effective behavioral counseling [9,10]. Many providers report difficulty engaging patients in meaningful discussions about lifestyle modification due to limited training in behavioral change techniques and patient-centered communication [9]. Addressing these issues requires strategies that not only convey information but also actively enhance patient motivation, strengthen self-efficacy, and support the development of sustainable health behaviors within routine care.

Motivational interviewing (MI), a counseling method widely recognized and developed by Miller and Rollnick [11], is a patient-centered approach [12]. MI focuses on the client's perspective, allowing them to express their feelings and motivations for change [12]. It empowers clients to take ownership of their change process, thereby reducing defensiveness and resistance [12]. MI consists of 4 key processes: engaging, focusing, evoking, and planning [11]. The foundational skills of MI—open-ended questions, affirmations, reflections, and summaries—are used to improve engagement and facilitate change talk [11]. Focusing is the process of finding or discussing a common interest topic for behavioral change between the client and counselor [11]. In the evoking process, the counselor aligns with the patient's talk (change vs sustain talk) to encourage change. The more the patient makes change talk, the closer she or he is to the change. Patients were guided, invited, and responded toward change talk, such as reflecting on past successes, envisioning future benefits, considering consequences, exploring extreme scenarios, identifying personal values, and addressing ambivalence [11]. During the planning process, the patient expresses more change talk and less sustain talk. In this process, potential obstacles to change and strategies to overcome them are explored [11]. This approach has been successfully applied to various lifestyle challenges and medical conditions [12,13]. Systematic reviews and meta-analyses have demonstrated efficacy of MI in addressing various health behaviors, including alcohol abuse, increasing physical activity, smoking cessation, pain management, and improving body weight and blood sugar control [14-17]. The application of MI varies across health care settings and among different clinicians. Reviews continue to support the positive impact of MI across diverse diseases, settings, and delivery modes, effectively delivered by various health care professionals [12,16,18].

General practitioners (GPs) or family physicians play a pivotal role in diabetes management [19]. Their involvement is essential for improving health outcomes, particularly by facilitating behavioral changes [20]. GPs are uniquely positioned to promote health and prevent disease through lifestyle counseling and guidance on managing risk factors [21,22]. However, previous studies have found that patient-centered, motivational techniques aimed at behavioral change are not routinely used during diabetes consultations [23,24].

Given their central role in supporting behavioral change, GPs are well-positioned to apply MI to promote lifestyle modification and improve outcomes in patients with type 2 diabetes. Previous studies have shown that MI delivered by health care professionals in primary care can improve diabetes management outcomes, including glycemic control, weight loss, and self-management [12,15,18,25,26]. When delivered specifically by GPs, a systematic review reported that MI may enhance patient outcomes; however, the effects are often modest and inconsistent, with variable results observed for HbA_{1c}, cholesterol levels, and physical activity [17,27]. In addition, a significant research gap remains regarding MI's impact on GPs themselves, particularly in terms of their knowledge, competencies, and barriers to implementation in routine diabetes care [17].

Despite its feasibility in daily general practice, the uptake of MI remains inconsistent. Common barriers include time constraints—especially in managing complex conditions like diabetes, which require detailed behavioral counseling—and the difficulty of balancing MI principles with competing clinical priorities during a single consultation [12,17]. Brief interventions have been suggested for use in routine diabetes care; however, these often lack the patient-centered techniques central to MI [20,28]. To address this, brief MI has been developed as an adapted form of MI for high-volume clinical settings [29]. It emphasizes core processes such as engaging and evoking within a condensed interaction, making it feasible within standard consultation times [29]. This adaptation enables GPs to facilitate patient-centered behavioral change without conducting a full-length MI session.

Brief MI may be a more practical approach for busy GPs in managing diabetes. Despite its potential, there is a notable lack of rigorous studies evaluating the effectiveness of brief MI in diabetes care. For instance, a study implemented brief MI but involved only 4 patients with diabetes and was delivered by providers who were not exclusively GPs [30]. Another study found that training brief MI for family medicine residents can enhance their counseling skills and improve patient self-management, as evidenced by the increased use of MI-adherent approaches and improved clinical outcomes [31]. Additional challenges to implementing brief MI include the complexity of the technique, which demands that GPs develop advanced clinical skills to apply MI effectively, and the lack of ongoing support systems to sustain and enhance these skills over time [17]. Addressing these barriers is critical for optimizing the feasibility and impact of brief MI in primary care settings.

Continuing medical education (CME) or the broader concept of continuing professional development (CPD) is widely used to maintain the quality of GPs' clinical practice [32] and for GP diabetes education [33]. Multifaceted educational approaches, such as interactive workshops, small-group discussions, outreach visits, audits and feedback, and reminders, have been shown to

effectively improve physician practices and clinical outcomes [34,35]. In contrast, traditional CME methods, such as lectures and educational material distribution, have a limited impact on behavioral change [34,35].

Web-based educational methods are increasingly popular due to their flexibility, convenience, and reduced travel costs, offering similar effectiveness to traditional methods [36,37]. While web-based CME methods may not replace traditional approaches, they can serve as complementary tools to enhance learning outcomes [36,38]. Previous studies have examined the effectiveness of web-based MI training across diverse diseases, settings, and health care professional groups [29,39-43]. These studies have shown the potential of web-based MI training to improve learners' knowledge, skills [39,40], and confidence [29,42]. However, evidence on its impact on clinical practice or patient outcomes is scarce [39]. In addition, the role of web-based educational methods in supporting MI skill development for GPs managing diabetes has not been rigorously tested.

This study addresses existing gaps by implementing a comprehensive educational program for GPs, combining an interactive workshop, structured web-based learning, and practical guides for brief MI. The primary research question was whether this multifaceted MI educational approach could improve GPs' learning and patient clinical outcomes. Therefore, this study aimed to evaluate the effectiveness of a structured educational program on MI for GPs and GP trainees in primary care settings. Specifically, it assessed whether a combination of interactive workshop training, web-based modules, and practical brief MI guides could improve participants' MI-related knowledge and confidence, and whether this translated into improved clinical outcomes among patients with type 2 diabetes in pilot communities, and exploring GPs' experiences, perceptions of the MI training, and their implementation of MI in routine consultations.

Methods

Study Components and Research Design

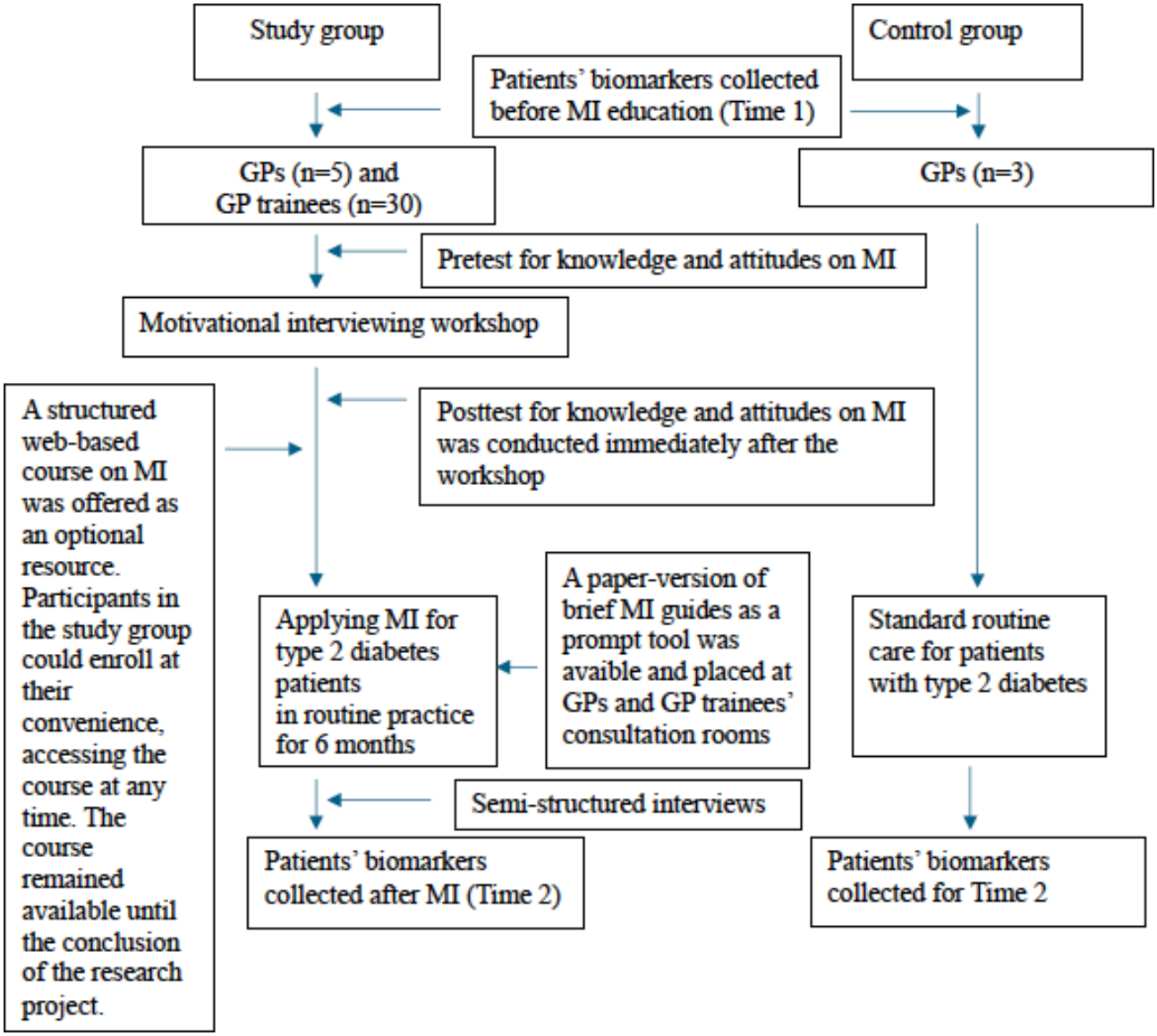
To address the overarching aim of evaluating the effectiveness and feasibility of MI training in primary care, this study used a mixed methods design composed of three components: (1) Evaluation of GP learning outcomes: A before-and-after study assessed changes in GPs' and GP trainees' knowledge and confidence related to MI through pre- and posttests; (2) Assessment of patient clinical outcomes: A quasi-experimental, nonrandomized 2-group design compared biomarker changes in patients with diabetes from intervention (MI-trained) and control (routine care) sites; and (3) Qualitative exploration of GP experiences: A descriptive qualitative study using semistructured interviews explored GPs' experiences in applying MI in routine consultations after training and their experiences in the MI education. GP interviewees were selected based on their confidence in applying MI in routine consultation after their completion of the MI courses. This integrated structure is justified by the implementation nature of the study, aiming to evaluate not only educational effects but also practical application and outcome relevance. The integration is intentional and appropriate for a pragmatic, real-world educational intervention. Improving GP knowledge without understanding its translation into patient care would offer limited insights. Conversely, changes in patient biomarkers without a provider perspective could misattribute causality. Including qualitative data allowed us to contextualize the quantitative results, explain variability in outcomes, and identify barriers and facilitators to implementation.

The sample size was calculated separately for the two primary components of the study: (1) the before-and-after evaluation of GPs' knowledge and (2) the comparison of patient biomarker outcomes between the intervention and control groups. For the knowledge assessment, the sample size was calculated using the G*Power (version 3.1.9.7; Heinrich-Heine-Universität Düsseldorf) based on the proportion of GP trainees enrolled in diabetes CME [44]. Assuming a power of 80%, a significance level of 0.05, and an effect size of 0.6, a total of 24 participants was sufficient to detect a significant difference in knowledge scores before and after completing the CME courses. For the patient outcome component, the sample size estimation was based on previously published research in which patients with diabetes received MI-guided behavioral change counseling [26]. To detect a 10% effect size with an alpha of .05 and a power of 80%, a total of 241 participants was required. This meant that 121 patients were needed for each of the study and control groups.

Participants and Recruitment

Figure 1 shows the flow of participant recruitment for the study. The study was conducted from January 2024 to January 2025.

Figure 1. Flow of recruitment. GP: general practitioner; MI: motivational interviewing.



GPs and GP trainees' recruitment

Study and control groups were recruited by convenience from primary care units, which had family doctors working full-time. Notably, in Thailand, the majority of family or primary care doctors work in the hospitals, some work part-time at primary care units, and the majority of primary care units do not have doctors but are operated by nurse practitioners.

For the study group, a primary care unit affiliated with a medical school was selected. Doctors working at this unit included 5 full-time GPs and 2 to 3 GP trainees rotating in each month (the total number of GPs was 5, and the number of GP trainees was 30). For GP trainees, each would rotate at this primary care unit for at least 3 months per year. This primary care unit covered the care of 10 communities with a total population of approximately 13,000. For the control group, 3 primary care units were selected. These units served 4 communities with a combined population of approximately 10,897 and were staffed by 3 family doctors. Notably, the study and control areas are located approximately 300 km apart.

Patients' Recruitment

The study used primary care catchment areas as the unit of allocation and analysis. This enabled the impact of the interventions to be examined at the population level and allowed the outcomes related to diabetes care for individuals to be examined, regardless of whether care was sought from a single or multiple GPs within their community. All patients with diabetes under care of the selected primary care units were included in the analysis as a unit of analysis. Therefore, 149 patients with diabetes in the study areas and 167 patients with diabetes in the control areas were included.

Educational Interventions

An interactive 4-hour workshop on MI was offered to GPs and GP trainees who worked at the primary care unit of the study group. The instructors (IT and PN) attended several MI web-based courses provided by William Miller, Stephen Rollnick, and Theresa Moyers. The GPs and GP trainees voluntarily attended the workshop. The content of the workshop included human behavior, introduction to MI, roles of GPs in MI technique, MI spirit, 4 processes of MI, brief MI for busy

GPs caring for people with diabetes, MI vs stage of change, and case studies using MI in each stage of change.

An additional educational course on MI was offered to GPs and GP trainees to self-study at their convenience. A structured web-based course on MI and stage of change was designed and developed to provide the foundation for GPs using MI in their routine consultations. The web-based course was developed and designed by incorporating several steps to create effective web-based learning activities, as recommended in a previous study [45]. The intervention, designed to provide GPs with knowledge of MI, was offered for a period of study. The content was designed as text-based plus interaction with text, that is, knowledge-based quizzes, self-practice, and self-reflection activities. The course comprised approximately 4 hours of learning activities for individuals to complete. The course content was developed in a manner similar to the workshop. Notably, participants could access other forms of MI education during the study period. The content of this course could be accessed through our medical school web-based platform free of charge.

A paper version of brief MI guides as a prompt tool for GPs to help people with diabetes change their behavior was available and placed in the GP consultation room at the primary care unit. These guides were developed based on "A Taste of MI" Exercise from the Motivational Interviewing–Foundational web-based course by William Miller, Stephen Rollnick, and Theresa Moyers (see the English translated version of brief MI guides in Multimedia Appendix 1).

Quantitative Component

Data Collection: GPs and GP Trainees

A before-and-after design was used to compare the knowledge and attitudes of the GPs and GP trainees who completed the MI workshop. The GPs and GP trainees self-selected by voluntarily choosing to enroll in the workshop and the web-based course and participating in the before-and-after study. The questionnaires were developed by the authors (IT and RM) based on the literature [17] and inputs from academic GPs and experts in the field. Each item was assessed thoroughly regarding the intention of the measurement, relevance, ambiguity, understandability, and necessity by 3 independent experts in medical education and academic GPs at our medical school to ensure the face and content validity of the questionnaires. The tests were conducted on the web and divided into 2 phases: pre- and postsurveys. These tests included questions in the form of multiple-choice, yes-no, open-ended, attitudinal, and knowledge questions. The questionnaires were piloted with 30 GPs and were revised accordingly. Cronbach α coefficient of the questionnaire was 0.95.

Data Collection: Patient Outcomes

Data on patients with diabetes were collected twice from the patient data records. First, before the GPs and GP trainees attended the MI workshop. Second, 6 months after the workshop completion. This means that GPs and GP trainees had time to apply MI for their patients for at least 6 months. Patients' biomarker data were obtained from the patients' medical records for the latest values collected for each patient, according to the routine collection of the primary care units. Patient data included blood pressure, body weight, and HbA_{1c} and lipid levels.

Qualitative Component—Data Collection: Interviews With GPs and GP Trainees

Semistructured interviews were also conducted. The GPs and GP trainees who completed the pretest were invited to participate in semistructured interviews. A small sample of 10 GPs and GP trainees was sought for convenience. A total of 5 GPs or GP trainees from each of the following groups were sought: GPs and GP trainees who reported having less confidence in applying MI in their routine consultation and those who reported having confidence in applying MI in their routine consultation. The interviews were conducted in September 2024 via Zoom (Zoom Communications) and were recorded for documentation. Each session lasted approximately 15 minutes. The audio recordings were transferred to a computer for repeated listening and detailed content review. The recordings were transcribed into Microsoft Word documents, and they were confirmed with participants to ensure accuracy and for further analysis.

Quantitative Analysis

IBM SPSS 19 for Windows (version 20.0) was used for statistical analysis, and a pairwise deletion strategy was applied to handle missing data. Paired-sample *t* tests were applied to assess within-group differences between pre- and postintervention measurements, and independent-sample *t* tests were used to compare outcomes between the study and control groups. Baseline characteristics were descriptively compared across study and control groups; however, formal statistical adjustments for potential confounders (eg, baseline differences in patient demographics or clinical characteristics) were not performed. This decision was based on the study's design focus on population-level comparisons and the available sample size, which limited the feasibility of multivariable analyses. Descriptive statistics were used to describe demographic data. The participants' responses on Likert scales to a 10-item questionnaire on their confidence in applying MI for diabetes management were tallied, creating a confidence score range of 10 to 50. A paired-samples *t* test was used to compare the mean differences in the knowledge scores between the two tests.

Qualitative Analysis

Interview transcripts were analyzed using conventional content analysis. An inductive coding approach was applied, whereby initial codes were generated from repeated reading of transcripts. These codes were then grouped into categories, and overarching themes were derived by identifying recurring patterns and relationships among the data. Coding and analysis were conducted manually by a single trained coder (IT). Although intercoder reliability was not assessed, the coding process was reviewed by members of the research team to enhance consistency and credibility. The codes were then organized into themes, identifying the relationships and connections between them.

Integration of Quantitative and Qualitative Data

The quantitative and qualitative findings were synthesized during the interpretation phase to provide a comprehensive

understanding of the intervention's impact. Quantitative data provided objective evidence of improvements in knowledge and patient outcomes, while qualitative insights offered contextual understanding of how GPs applied MI, the challenges encountered, and their suggestions for improvement. This integration supported triangulation of results and enhanced the depth and relevance of the study's conclusions.

Ethical Considerations

This study was approved by the Human Research Ethics Committee of Khon Kaen University (project HE631031). GPs and GP trainees were recruited through the researcher's assistant, who invited volunteers to participate. Before completing the questionnaires, participants were informed that participation was voluntary and that they could drop out of the study at any time. They were informed that their opinions were important for enhancing diabetes practice in primary care and were therefore encouraged to express them. All GPs and GP trainees voluntarily agreed to participate without receiving compensation. The participants' privacy and identity were protected, and confidentiality was assured. Informed consent was obtained from each participant who participated in the semi-structured interviews. The study objectives were explained to the participants and the study was conducted according to the academic ethical code.

For patients, individual consent was not required because all clinical outcome data were extracted from anonymized, routinely collected health records, in accordance with institutional ethical guidelines. No experimental procedures or additional data collection were imposed on patients. After the study concluded, the brief MI materials and web-based MI course were shared and offered to clinicians in the control group as part of standard CPD dissemination. This approach ensured that no participants were denied care and aligns with accepted ethical principles for real-world implementation research.

Results

Participants' Characteristics

The MI workshop was attended by 32 GPs and GP trainees out of a total 35 participants (91%). A total of 2 participants out of 38 had previous training in MI (n=2, 5%). While both GPs and GP trainees participated in the educational interventions (n=32), only 28 GP trainees (87%) completed the pre- and posttests to assess their knowledge, and later 10 GP trainees provided feedback through semistructured interviews, whereas no GPs participated in these assessments (see Table 1).

Table . Participating general practitioners (GPs) and GP trainees' demographic data.

Demographic data (N=38)[a]	Results
Sex, n (%)	
Males	18 (47)
Females	20 (53)
Age (years)	
Mean (SD)	28.5 (7.85)
Median (range)	28 (25-62)
Previous web-based learning, n (%)	
Yes	25 (66)
No	13 (34)
Previous web-based course completion	
Mean (SD)	0.86 (1.15)
Median (range)	1 (0-4)
Previous motivational interviewing course completion, n (%)	
Yes	2 (5)[b]
No	36 (95)
Diabetes patients in the catchment areas, n (%)	
Study areas (N=13,000)	149 (1.14)
Control areas (N=10,897)	167 (1.53)

[a]A total of 35 GPs and GP trainees participated in the study group, while 3 were from the control group.

[b]The general practitioners in the study group.

Data from the web-based learning management system (LMS; (updated on January 21, 2025) showed that there were 35 GPs and GP trainees (100%) enrolled in the structured web-based MI courses, 14/35 (40%) completed the courses, 5/35 (14%) had not yet started the lesson, and 16/35 (46%) had started but had not yet completed. The course had 21 lessons. Participants'

enrollment time in the course was a median of 3 minutes (range 0-416 min), with a mean of 68.54 (SD 108.82). The number of completed lessons by participants was a median of 2 lessons (range 0-21), with a mean of 6.46 (SD 8.79).

Changes in MI Knowledge and Confidence

A paired-sample *t* test was conducted to evaluate the impact of the MI course on learners' knowledge. There was a statistically significant improvement in knowledge test scores from Time 1 (mean 11.46, SD 3.48) to Time 2 (mean 15.04, SD 2.35), t_{28}=7.74; *P*<.001 (2-tailed). The mean increase in knowledge test score was 3.57 (SD 2.44), 95% CI 2.62-4.52. Notably, overlapping pre- and postintervention CIs do not invalidate the finding, as this was a within-subjects comparison. The eta-squared statistic indicated a large effect size (eta-squared=0.85). Table 2 shows participants' confidence levels in applying MI techniques after completing the course. The highest confidence was reported for the focusing process (mean 3.86, SD 0.65) and using affirmations (mean 3.79, SD 0.63). Other skills, such as using open-ended questions, planning, and inviting change talks, scored slightly lower but consistently reflected moderate to high confidence levels, indicating the course's effectiveness in enhancing practical MI skills.

Table . After the course completion, participants' level of confidence in applying motivational interviewing (MI) with patients (n=28).

Participants' level of confidence in applying MI with patients	Mean[a] (SD)
Focusing process	3.86 (0.65)
Using affirmation	3.79 (0.63)
Using open-ended questions	3.71 (0.76)
Planning process	3.71 (0.66)
Inviting change talks	3.68 (0.61)
Using summary	3.68 (0.67)
Using MI with stage of change	3.68 (0.61)
Using reflection	3.61 (0.63)
Responding to change talks	3.61 (0.74)
Recognizing change talks and sustain talks	3.57 (0.63)

[a]Mean was calculated using a 5-point Likert scale ranging from 1 (not at all confident), 2 (somewhat not confident), 3 (neutral), 4 (confident), and 5 (very confident).

Changes in Patient's Outcomes

Table 3 provides biomarker outcomes for patients with diabetes over 6 months after GP trainees completed the MI workshop. In the study group, there were significant improvements, including reductions in HbA_{1c} (mean difference 0.35, 95% CI 0.03-0.66; *P*=.03), body weight (mean difference 1.11, 95% CI 0.27-1.95; *P*=.01), triglyceride (mean difference 14.18, 95% CI 1.69-26.67; *P*=.02), systolic blood pressure (mean difference 5.14, 95% CI 1.06-9.03; *P*=.01) and diastolic blood pressure (mean difference 3.82, 95% CI 1.66-5.99; *P*=.001). Between-group analysis showed that the study group had significantly better outcomes for HbA_{1c} (mean difference −0.50, 95% CI −0.91 to −0.09; *P*=.02) and diastolic blood pressure levels (mean difference −5.96, 95% CI −8.66 to −3.25; *P*<.001) compared to the control group.

Table . The patients' biomarkers before and after 6 months of GPs and GP trainees' enrollment in the motivational interviewing (MI) workshop. The number of cases used in each calculation varies slightly across variables.

Patient outcomes	Study group (n=149)					Control group (n=167)					Between group			
	Pre-MI mean (SD)	Post-MI mean (SD)	Mean difference (SD)	95% CI	*P* value (within group)[a]	Time 1 mean (SD)	Time 2 mean (SD)	Mean difference (SD)	95% CI	*P* value (within group)[a]	Pre-MI		Post-MI	
											Mean difference (95% CI)	*P* value[b]	Mean difference (95% CI)	*P* value[b]
HbA_{1c}	8.07 (1.59)	7.71 (1.53)	0.35 (1.88)	0.03-0.66	.03	8.38 (1.97)	8.23 (2.10)	0.16 (1.74)	−.11 to .42	.25	−0.30 (−0.70 to 0.11)	.11	−0.50 (−0.91 to −0.09)	.02
Body weight	63.55 (12.33)	62.44 (12.89)	1.11 (4.70)	0.27-1.95	.01	59.83 (11.51)	58.41 (11.71)	1.42 (5.08)	0.64-2.20	<.001	3.08 (0.49-5.68)	.02	4.03 (1.16-6.89)	.006
Systolic blood pressure	138.94 (18.47)	133.79 (18.55)	5.14 (23.40)	1.06-9.03	.01	125.69 (12.59)	131.32 (14.33)	−5.62 (16.05)	−8.09 to −3.15	<.001	13.16 (9.60-16.73)	<.001	2.47 (−1.22 to 6.17)	.18
Diastolic blood pressure	74.32 (11.61)	70.50 (11.06)	3.82 (13.04)	1.66-5.99	.001	73.02 (8.88)	76.46 (12.74)	−3.43 (14.54)	−5.67 to −1.20	.003	1.13 (−1.18 to 3.45)	.33	−5.96 (−8.66 to−3.25)	<.001
Cholesterol	181.21 (37.66)	179.33 (43.69)	1.89 (40.66)	−5.11 to 8.88	.59	177.26 (41.60)	155.38 (37.49)	21.87 (35.19)	16.46-27.28	<.001	4.81 (−4.08 to 13.71)	.28	23.62 (14.47-32.77)	<.001
Triglyceride	154.98 (81.19)	140.79 (69.61)	14.18 (72.26)	1.69-26.67	.02	181.12 (124.59)	151.38 (72.65)	29.73 (119.01)	11.44-48.03	.002	−27.27 (−50.13 to −4.42)	.02	−10.35 (−26.62 to 5.90)	.21
LDL[c]	107.58 (33.43)	107.04 (37.37)	0.54 (38.61)	−6.03 to 7.11	.87	105.25 (31.97)	93.26 (31.26)	11.99 (30.27)	7.33-16.64	<.001	4.00 (−3.46 to 11.47)	.29	13.43 (5.49-21.36)	<.001
HDL[d]	51.40 (13.60)	51.92 (13.03)	−0.51 (9.02)	−2.07 to 1.04	.52	51.21 (13.22)	50.04 (13.12)	1.17 (11.56)	−0.60 to 2.94	.19	0.17 (−2.80 to 3.15)	.90	1.82 (−1.15 to 4.81)	.22

[a]Data were analyzed using paired-samples *t* test.

[b]Data were analyzed using independent-samples *t* test.

[c]LDL: low-density lipoprotein.

[d]HDL: high-density lipoprotein.

Learners' Experiences and Perceived Impact of the Course

Table 4 summarizes the participants' expectations before enrolling in the MI workshop and how these expectations were met after the workshop. Participants reported improvement in skills and knowledge about MI, with mean scores increasing from 3.96 to 4.07 and from 3.93 to 4.14, respectively. In addition, the learning methods, venue, and learning period received positive ratings postcourse, highlighting the course's effectiveness in meeting expectations.

Table . Expectations for learning in this motivational interviewing (MI) course before and how they were met after the participants' enrollment in the MI workshop (n=28).

Participants' expectations before enrolling in MI courses and reported their expectations were met after the course completion	Mean[a] (SD)	
	Before enrollment in the course	After the course completion
Having skills in MI	3.96 (0.69)	4.07 (0.86)
Having knowledge about MI	3.93 (0.60)	4.14 (0.80)
Promptly applying MI techniques with patients	3.64 (0.83)	3.79 (0.79)
After applying MI with patients, the patients can control their diseases	3.93 (0.77)	—[b]
After applying MI with patients, the patients can change their behaviors	3.96 (0.79)	—
Teaching and learning methods for MI	—	3.93 (0.77)
Place or venue for learning	—	3.96 (0.88)
Period of learning	—	3.86 (0.89)

[a]Mean was calculated using a 5-point Likert scale ranging from 1 (strongly disagree), 2 (disagree), 3 (neutral), 4 (agree), and 5 (strongly agree).

[b]Not available.

Insights From Participant Interviews

Table 5 summarizes the interviews with participating GP trainees regarding their experience in implementing MI during their routine consultation. The GP trainees found MI useful for engaging patients with diabetes, especially in brief MI formats. All participants reported that MI techniques were used in patient interactions, particularly in cases of diabetes, hypertension, and dyslipidemia. The engaging and evoking processes were most frequently used, while planning and responding to sustain talk were less commonly applied.

Table . General practitioner (GP) trainees' experiences and challenges in implementing motivational interviewing (MI) in their routine consultation.

General practice trainees' experiences in and perceptions toward the use of MI with patients	Comments
Experience with applying MI in practice. Contexts and types of cases where MI is used.	
• All participants reported using MI techniques in their patient interactions. They primarily applied MI in cases involving noncommunicable diseases, such as diabetes, hypertension, and dyslipidemia. In addition, all participants attempted to use brief MI, particularly in diabetes cases (T1-T5[a] and P1-P5[b]).	• "Most of the cases I applied MI are NCDs" (T3). • "I tried to apply MI with my patients to help them with lifestyle modification, including for NCDs, diabetes, hypertension, dyslipidemia, but just some processes of MI such as empathic listening" (P5). • "I have only experienced using brief MI as a guide for approaching patients, and all the cases were diabetes, so I have not applied the full MI process with my patients" (T5). • "While I was rotating in Internal Medicine, I applied MI with patients to help them lose weight and with the issues of nutrition management. During my rotation at primary care practice, I used brief MI for only diabetes patients" (T1).
Most frequently used MI processes (eg, engaging, focusing, evoking, and planning)	
• For full MI, the majority of participants most frequently used the engaging process, as they were familiar with its fundamental skills, such as open-ended questions, affirmations, reflections, and summaries (T2-T3 and P2-P5). A participant reported primarily using the focusing process (T5), while others frequently applied the evoking process (T1, T4, and P1). • For those using brief MI, all MI processes were experienced; however, they might not have explicitly recognized the distinct processes involved (T1-T5 and P1-P5).	• "Engaging is easy to use. It is like a small talk before starting the consultation. I used open-ended questions quite often as well as reflection" (P3). • "I used only brief MI with all diabetes cases,… engaging with fundamental skills and evoking with important ruler" (P2). • "Evoking with important ruler is easy to use" (T1). • "Engaging since it is like we are getting to know the patient like a small talk before starting consultation, so I'm familiar with OARS and open-ended questions were used the most" (P4).
Least used MI processes and reasons behind it.	
• The evoking process (T3, T5, and P2-P5) and the planning process (T1, T2, and T4) were reported as the least frequently used by participants. The primary reasons included not yet reaching these stages in their patient interactions, uncertainty about how to transition to these processes, or a lack of opportunity to follow-up with patients until the final stage of MI due to rotations in their training programs.	• "I am not yet good at using evoking and also planning since I have not yet took the patient to this process yet and I did not have a chance to follow-up the same patient because of changing rotation" (T3). • "Planning process was used the least since I have not reached this process yet and sometimes was not sure how to go next" (T1). • "I have not reached this process yet and I am not familiar of how to use it" (T2). • "Evoking is the least. I tried to use important ruler to invite change talks however, it seemed like patient did not understand particularly those aged above 60." (T5). • "I tried to all the processes, but each was like little here and there" (P1) • "I'm not yet familiar with evoking. To start evoking, you may need to have good relationship with patients and important ruler is hard to understand" (P4).
Challenges faced when applying specific MI processes	
• All the participants reported that evoking process is the most difficult to use in all aspects of evoking including recognizing, inviting, and responding to change talks (T1-T5 and P1-P5).	• "To deal with sustain talk, how to respond to sustain talk is the most difficult for me" (T1). • "Evoking, particularly inviting patient for change talk, it is difficult to do" (P1). • "In inviting change talks, patient did not understand the question, particularly using important rulers and it felt like it was not flow naturally" (T3). • "Evoking is difficult particularly, recognizing and inviting change talks since I do not have much experience in applying full MI sessions but for brief MI, there is check lists to help how to proceed" (P2). • "Recognizing change talk is the hardest part because it's difficult to identify which statements are actually change talk. In some cases, patients didn't say much, and when they did, their speech was not sequential, making it hard to distinguish between change talk and sustain talk" (P4).
Obstacles encountered during the use of MI in routine consultations	

General practice trainees' experiences in and perceptions toward the use of MI with patients	Comments
• All participants (T1-T5 and P1-P5) agreed that incorporating MI into routine consultations adds extra time to each patient encounter. They noted that if the daily patient load were lower, MI could be applied more comfortably. While brief MI may offer a practical solution, participants acknowledged the need to first master the technique to use it effectively without extending consultation time. • Another challenge is limited experience, which hinders skill development and prevents MI from becoming a natural part of routine practice. In addition, for GP trainees, structured training rotations often limit their ability to follow-up with patients through the full MI process.	• "Time and number of patients seeing in routing consultation per day. If small number of patients like if less than 60 a day, MI can be used but if not, it would increase time for each case and ended up with delay waiting time for the next patient" (P2). • "Elderly patients do not understand the questions may be because doctor could not make the question flow naturally or even do not know how to make it easy to answer. I, myself, not yet good at execute MI" (P2). • "MI takes longer consultation time. It matches with the setting that have a smaller number of patients or have many doctors in the same practice or it needs to separate MI session out of routine consultation, but this would be solved with brief MI" (T4). • "I used brief MI, it helps a lot but since I' m not yet mastered the whole processes of MI, so it is not yet felt naturally in applying it routinely" (P5). • "Because of trainee rotation, I cannot follow the same patient till the final process of MI" (T1). • "I have less experience in using MI but when start using it I feel good and better when dealing with patient's lifestyle modification, so I need to practice more" (T1). • "In the form of brief MI, it helps shorten the time but overall, still adds time to the consultation and my MI skills are not yet ready" (T3). • "MI needs time to explore patients to get to know them and sometimes they did not know how to respond or say, so it takes time" (P3). • "Spending longer time and together with "I'm not yet good at using MI," sometimes I need to stop thinking what is next because this also adds up time" (P4).
Perceived benefits of using MI in routine consultations	
• All participants (T1-T5 and P1-P5) highlighted the advantages of MI, emphasizing its foundation in the patient-centered approach. They noted that MI empowers patients to decide what and how to change, while the doctor provides guidance and support to help them achieve their goals.	• "Patient selects by themselves what they want to change so this will eventually lead to actual change" (T2). • "Concepts and principles of MI are great, unlike the traditional approach of ordering or telling patients to change but MI is to guide and help them change with the direction" (T3). • "We as a doctor understand patient more and respect them for whom they are" (T4). • "There is a direction in MI, making it like a framework that guides us toward the goal. The process consists of steps or pathways that help us navigate patient interactions and continue follow-ups from where we left off in the last visit. Otherwise, we would end up repeating the same advice at each visit, providing a generic package of behavior change recommendations, which would ultimately lead to unsuccessful outcomes." (P2).
Potential drawbacks or limitations of using MI in consultations	
• Participants identified several limitations in implementing MI in routine consultations. The primary challenge was a lack of skill, largely due to limited experience in applying MI in daily practice (P1-P3). While brief MI offers a promising approach, it still requires practice to master and integrate naturally into consultations without extending their duration (T1, T4, and T5). Since participants had not yet fully mastered MI, their consultations often took longer than usual (P1 and P2).	• "Full MI taking long time and may not suit particularly in the setting of outpatients where each doctor needs to see 100 cases a day" (T5) • "If you are not familiar or not good at it, it will take more time in consultation. In some cases, we cannot use this method such as dementia" (P2). • "I'm not yet confident in using the MI process. Those who want to apply MI should have adequate knowledge and skills, become familiar with the process, and integrate it into their routine practice" (P3). • "I could not yet find any limitation or disadvantage" (T3) • "Take longer time in consultation and need to be cautious using some questions which may affect the doctor-patient relationship" (P4).
Areas for further development in MI	

General practice trainees' experiences in and perceptions toward the use of MI with patients	Comments
• Some participants need to improve all knowledge and skills to gain their confidence in using MI (T2-T5, P4, and P5). Some mentioned have sufficient knowledge but need more practice to improve skills and particularly for evoking process (P1-P3).	• "I feel a lack of knowledge and skills in MI. I think I know and can do only 50%. Brief MI works for me — no need for more time in consultation. It is practical, easy to use, and can be applied in real practice" (Participant 5). • "Knowledge and skills in evoking are still limited, and I also have little experience using MI. Sometimes I have to focus too much when using MI, as I need to practice empathic listening, be able to reflect back what the patient is saying, and also summarize parts of the conversation" (P4). • "Since I have not yet become good at using MI and do not use it regularly, only with some diabetes patients using brief MI, I need more practice to build my confidence" (P3). • "My knowledge is ok but I need to improve my skills, particularly for evoking process" (P2). • "I am not yet confident in my overall MI knowledge and skills, so I can not apply them naturally in practice. In the evoking process, for inviting change talk, there are plenty of ways to do it, so I need to go back and practice more of those options" (T5). • "I need to update all MI knowledge and skills" (T4).
Need for additional training or support in MI techniques	
• Majority of the participants need more observations to become experts in using MI in routine practice and, if possible, need coaching (T1, T3-T5, and P2-P5). One mentioned role-play sessions with colleagues will also help to gain more confidence in applying MI in their routine consultation (P1). Additional video clips would also help in self-learning (P1). Others needed to practice brief MI more (T2 and P2).	• "I need to keep practicing using MI in my routine consultations, and share cases that were successful after using MI. I also need observation sessions with experts or supervisors, as well as coaching" (P4). • "I need more observation sessions to see experts using this technique in real practice, followed by coaching to get feedback on whether I'm doing it right" (P3). • "I need more role plays in class so all my colleagues can practice together and learn which parts we do right or wrong, and practice the whole process step by step. Other helpful methods would be coaching or observation in real practice settings. Video clips of different situations applying MI in routine consultations would also be great" (P1). • "I want more observation from expert or coaching" (T3) • "I need to practice brief MI more and use more" (T2) • "I need more examples of how to use MI in a variety of situations—what to say and how to say it, and how to apply MI in each situation. These would help me get a clearer picture of applying MI in real consultations. Brief MI is great and I can use it in my routine, but I need more observation classes to use it naturally rather than relying on checklists or guidelines all the time" (P2).
Impact of MI on practice and likelihood of continued use	

General practice trainees' experiences in and perceptions toward the use of MI with patients	Comments
• Initially, MI has changed their practice, getting to know more for their own patients. All participants mentioned further practice in MI and will continue using it in their routine consultation (T1-T5, P1-P5).	• "I will definitely continue to use MI in my routine practice, particularly brief MI, as it does not take more time than a regular consultation. I will use it as part of my routine work because even when there are many patients, brief MI can still be applied within 5 - 6 minutes as part of usual care" (P2). • "There are many advantages to applying MI in routine consultations, particularly for DM and CKD patients, since these techniques can be used before they become sick, while they are sick, and also in the process of tertiary prevention for disease complications" (T5). • "I will definitely continue practicing until I master MI and will integrate it into my routine practice" (P5). • "The process of MI is very interesting; it helps you have something to talk about with patients and understand them more. I need more practice to become good at it and use it in my routine practice" (T4). • "Empathic listening and important rulers help you get clear answers and lead to change" (T2). • "I will use it more, since if the patient selects the option themselves—what and how to change—it is much better than having someone else tell them what to change" (T1). • "I will definitely continue using MI so the way I help my patient with their lifestyle change is now not the same" (T3).

[a]T: Participants who reported having confidence in applying MI in routine consultations, based on pretest survey data.

[b]P: Participants who reported having less confidence in applying MI in routine consultations.

> *I feel confident using open-ended questions and affirmations because they make patients open up more, and I can understand their situation better.* [P3]
>
> *Even with brief MI, I tried to engage and ask about their goals… patients seem more open, and it helps me decide what to suggest next.* [T2]

However, many participants noted difficulty in applying the evoking and planning processes.

> *I know the steps of MI, but the evoking part—like using importance rulers or asking about change—is really hard to apply when the patient doesn't understand the question.* [T4]
>
> *Planning is hard when we don't see the patient again. During rotations, we only see them once, so it's difficult to finish the full MI cycle.* [P5]

Challenges also included time constraints and unfamiliarity with the MI approach.

> *MI takes longer time in consultation and my skills are not yet good enough, so I need to stop and think before responding.* [P1]

Despite these obstacles, participants valued the patient-centered nature of MI.

> *Using MI makes me feel like I really listen to my patients. They appreciate it more, and some of them even said they feel more motivated to change.* [T1.]
>
> *Brief MI is a practical tool I can use in real life—it gives structure to the conversation without taking too much time.* [P2.]

Discussion

Principal Findings

This study evaluated the implementation of a combined educational intervention on MI for GPs and GP trainees involved in type 2 diabetes care. The findings revealed significant improvements in participants' MI knowledge, confidence, and related patient outcomes. Furthermore, participants reported positive experiences with the training and perceived enhancements in their communication skills and patient care practices.

All GPs and GP trainees in the study group completed the interactive MI workshop, but only GP trainees participated in knowledge pre- and posttests, and the latter half of GP trainees completed the structured web-based courses. The study demonstrated that MI education provided for GPs and GP trainees in interactive workshop formats significantly improved their knowledge, and together with additional self-learning through structured web-based courses and brief MI guides, had a measurable impact on the patient outcomes. In terms of patient outcomes, the MI-trained group showed improvements in key health metrics, such as HbA_{1c} and diastolic blood pressure, compared to the control group, emphasizing the practical benefits of MI training in clinical settings. Semistructured interviews revealed that GP trainees valued MI for its patient-centered approach but faced challenges related to its time demands and their own proficiency in its processes. However, this study was conducted in a single study area; therefore, the results should be interpreted with caution.

This study contributes to the literature on CME and the application of behavioral science in clinical practice. It addresses the gap between theoretical knowledge of MI and its practical application by GPs and trainees in routine care. Through the

design of a structured, multifaceted CME intervention—comprising an interactive workshop, web-based learning, and practical tools—this study illustrates how behavioral science principles can be effectively translated into real-world practice. The findings may also guide the development of future training programs to enhance patient-centered communication among healthcare providers.

Effects of MI Training on GP Trainees' Knowledge and Confidence

GP trainees who underwent MI training demonstrated a significant improvement in their knowledge scores. The educational program successfully met its objective of enhancing knowledge among GP participants, aligning with previous studies that highlight the potential of MI education to improve both knowledge and confidence in delivering behavior-change interventions [17,31].

Knowledge assessments were conducted before and immediately after the workshop, indicating an immediate knowledge gain. This finding is consistent with previous studies on the effectiveness of MI training in primary care, which have demonstrated improvements in both knowledge acquisition and skill application [17,46,47]. Given that knowledge was assessed immediately after the workshop, the observed improvement could be attributed primarily to the interactive workshop component. However, following the workshop, participants also had access to web-based resources and MI guides. The combination of these educational methods aligns with existing evidence demonstrating that multifaceted educational interventions effectively enhance not only knowledge but also clinical practices and patient outcomes [34,35].

The educational content covered MI processes and their practical application through case studies in accordance with the frameworks established by Miller and Rollnick [11]. Notably, all participants who took the knowledge pre- and posttests had no previous MI training, which aligned with their self-reported learning expectations from the pretraining questionnaire. This targeted approach contributed to the successful development and implementation of an effective educational program as it addressed learners' specific needs [44].

Participants' feedback indicated that fundamental MI processes such as engaging, evoking, and planning were effectively taught. However, more advanced skills, particularly responding to change talk, remain challenging, an issue frequently noted in MI training [30,48]. Participants also reported that this was their first exposure to MI, and that the combination of workshop-based learning, structured web-based modules, and brief MI guides was sufficient for acquiring foundational MI knowledge.

In addition to knowledge acquisition, the training significantly enhanced participants' confidence in applying MI techniques, particularly in the practical aspects of brief MI. As shown in Table 3, participants reported the highest confidence in performing the focusing process and using affirmations, both essential elements of patient-centered communication. Other MI skills—including open-ended questioning, planning, and inviting change talks—also received moderate to high confidence ratings. These findings support the effectiveness of the combined educational interventions in building participants' self-efficacy for using MI in clinical practice. The pattern of reported confidence aligns with the stages of skill development, where foundational processes such as engaging and focusing are typically mastered earlier, while more advanced tasks like evoking and planning require further practice and reinforcement. This confidence boost is crucial, as self-efficacy is strongly linked to successful behavioral change counseling and continued use of MI techniques in routine care.

Impact of MI Training on the Patient Outcomes

Patient outcomes in this study were analyzed at the unit level. Patients with diabetes treated by MI-trained GPs and GP trainees demonstrated significant improvements in HbA_{1c} levels, body weight, triglyceride levels, systolic blood pressure, and diastolic blood pressure when comparing pre- and posteducational interventions and MI implementation. However, only HbA_{1c} levels and diastolic blood pressure showed statistically significant differences between the study and control groups after the educational interventions and MI implementation. Notably, although the study was powered to detect a 10% change in key clinical outcomes, the observed changes in most patient biomarkers—including HbA_{1c}, blood pressure, and lipid levels—were smaller than this threshold. While some changes reached statistical significance, the study may have been underpowered to detect modest effects, and tests of significance should be interpreted with caution. These findings suggest that knowledge gained from a single training session, supplemented by self-learning through the provided resources, can be effectively applied in clinical practice, leading to measurable improvements in diabetes patient outcomes.

Although few studies have examined MI delivered by GPs for diabetes patient outcomes, this study aligns with previous research demonstrating MI's mixed but promising efficacy in GP-led diabetes management [17]. Similar to the current findings, previous studies have reported improvements in HbA_{1c} levels between the study and control groups [49,50], and reductions in diastolic blood pressure [49].

This study found improvements in systolic blood pressure within both groups, but no statistically significant difference between them, which is consistent with previous findings [51]. A similar pattern was observed for triglyceride levels [51]. Notably, the study also found a positive effect on body weight in both groups; however, the control group showed a statistically greater reduction than the study group. In contrast, a previous study reported no significant impact on body weight [52]. Despite the variability in effects of MI on patient outcomes, these findings contribute to a growing body of evidence supporting the potential of MI when implemented by GPs in diabetes care.

Semistructured interviews with GP trainees in this study reinforced the quantitative findings, as participants reported that MI facilitated better patient engagement and empowered individuals to make sustainable behavioral changes. Although many GP trainees did not complete all 4 MI processes with patients, the majority implemented brief MI guides during diabetes consultations. Participants most frequently applied the

engaging and evoking processes, which may help explain the observed improvements in HbA_{1c} and diastolic blood pressure—both outcomes known to be influenced by enhanced patient motivation and communication. Even limited MI skills, such as empathic listening and goal-focused dialog, have been perceived as effective in fostering patient engagement, supporting the feasibility of brief MI formats in high-volume primary care settings. However, some clinical metrics, such as triglyceride levels and systolic blood pressure, have shown mixed results, highlighting the need for further research on MI uptake and its long-term impact, particularly regarding the sustained effects of brief interventions. This variability aligns with a systematic review that identified mixed yet promising results in MI interventions, underscoring the importance of more consistent training and implementation to improve outcomes [17].

While the observed improvements in HbA_{1c} and diastolic blood pressure among patients in the MI-trained group are promising, it is important to interpret these findings with caution. HbA_{1c} is a valid and widely used clinical outcome; however, it can be influenced by a variety of factors beyond the educational intervention, such as medication adherence, lifestyle changes, appointment attendance, or changes in health service delivery. Since we did not measure intermediate outcomes such as patient adherence, satisfaction, or visit frequency, the attribution of clinical improvements solely to MI training may be limited. Future studies should incorporate additional process measures to more fully explore the mechanisms by which MI training may impact patient outcomes. We also acknowledge limitations related to the interpretation of lipid profile outcomes. Variability in LDL and total cholesterol levels across groups may reflect unmeasured factors such as baseline statin use, medication adherence, and dietary or genetic differences. These variables were not adjusted for in this analysis. We recommend that future research include medication adherence monitoring and baseline lipid-lowering therapy documentation to better account for such confounders in lipid-related outcomes.

In addition to statistical significance, the improvements observed in certain outcomes appear to be clinically meaningful. Patients managed by MI-trained GPs and GP trainees demonstrated significant reductions in HbA_{1c} and diastolic blood pressure compared to the control group. These changes are important in diabetes management, as even modest reductions in HbA_{1c} and blood pressure have been linked to decreased risk of microvascular and cardiovascular complications. While improvements in other parameters, such as systolic blood pressure and triglycerides, were mixed or nonsignificant between groups, the observed HbA_{1c} and diastolic blood pressure reductions suggest that gains in MI-related knowledge and confidence may have translated into tangible patient benefits. This underscores the potential for brief MI training, when implemented in routine primary care, to yield clinically relevant outcomes.

GP Trainees' Perspectives and Experiencers Implementing MI

MI requires complex skills [53]. Effective MI implementation necessitates comprehensive training programs that integrate didactic presentations, experiential exercises, and role-playing to develop both basic and advanced MI skills. This study used these combined educational methods. In addition, these educational methods encourage practitioners to progress toward becoming trainers themselves [54]. Evidence suggests that practical applications and expert feedback are crucial for developing MI proficiency [46]. Participants in this study expressed a strong need for coaching or observational support from experts who could demonstrate MI in real consultations, aligning with research indicating that mentorship significantly enhances MI skills [55].

The GPs and GP trainees in this study were trained in the full MI process for comprehensive consultations. However, time constraints in Thai primary care settings make full-session MI difficult to implement. This issue has been noted in other health care settings, where time limitations hinder MI adoption [17]. To address this, brief MI guides were introduced, enabling GPs and GP trainees to apply MI techniques to patients with diabetes within typical consultation durations. Previous studies have suggested that brief MI can be an effective alternative in high-volume practices while maintaining its patient-centered benefits [30]. Participants from this study who had foundational MI knowledge from the workshop found brief MI guides particularly useful for integrating MI into clinical practice, aligning with research that highlights the importance of initial training before adopting streamlined approaches [26].

Qualitative data revealed mixed experiences with the implementation of MI. While GP trainees valued MI's structured, patient-centered approach, they faced barriers such as time constraints, difficulty in mastering advanced techniques, and limited follow-up opportunities due to rotational training schedules. The engaging and evoking phases were widely used, but many GP trainees struggled with the planning phase, a challenge documented in previous MI research [17]. Participants also noted that MI's emphasis on empathic listening and reflective questioning extended consultation times, making it difficult to sustain in high-volume clinical settings [30]. Although brief MI guides were designed to fit within standard consultation durations, some GP trainees reported that a lack of confidence in advanced MI techniques and incomplete mastery of MI skills could inadvertently prolong consultations. This aligns with findings from previous studies, which highlight that practitioners in the early MI training phases may take longer to implement MI effectively [47,56].

Despite these challenges, all participants acknowledged the benefits of MI, particularly brief MI, which they found more manageable within their workloads. Most reported using brief MI in nearly all encounters with patients with diabetes, while recognizing the need for continued practice to enhance their confidence and further refine their skills.

Our findings align with previous literature that supports the potential of MI to improve clinical outcomes in patients with type 2 diabetes when delivered by primary care providers

[17,49,50]. Similar to previous studies, we observed a significant reduction in HbA_{1c} and diastolic blood pressure following MI training [17,49], reinforcing the value of even brief MI interventions when embedded in routine consultations. However, as reported in earlier work [17,30,48], our participants also faced challenges in mastering advanced MI processes such as evoking and planning. This underscores the need for structured, staged training that starts with foundational MI concepts and provides opportunities for progressive skill development.

In line with recommendations from systematic reviews on effective MI training [13,53,55], our study supports the implementation of multifaceted educational strategies. These include combining in-person workshops with asynchronous web-based learning, distributing practical tools (eg, brief MI guides), and offering ongoing peer coaching or expert feedback. Such approaches help accommodate varying learning styles and time constraints, particularly in high-volume clinical settings. Based on our findings, we recommend that MI training programs for GPs incorporate brief MI frameworks early in the training, followed by reinforcement through mentorship, peer observation, or reflective practice sessions to sustain skill application over time. Tailoring training to local practice constraints—such as patient volume, consultation duration, and provider rotation—can enhance feasibility and long-term adoption.

Strengths and Weaknesses

This study has several strengths and limitations. The study design combining quantitative and qualitative approaches enriched the understanding of MI's implementation in real-world practice and its impact on GPs and their patients, offering both measurable outcomes and contextual insights. The study's focus on routine practice settings enhances the relevance and applicability of the findings to primary care environments. In addition, the pre- and posttraining assessments provided robust evidence of the effectiveness of training in improving knowledge of GPs.

While this study incorporated a control group to compare patient outcomes, we acknowledge that the nonrandomized design limits the ability to draw strong causal inferences regarding the effectiveness of the MI intervention. The control group consisted of primary care units in another province where GPs did not receive MI training and thus served as a comparator for usual care. Although this allowed for between-group comparisons, differences in context, provider characteristics, or patient demographics may have introduced selection bias or residual confounding. To minimize these biases, control sites were selected based on similarities in service structure, population size, and staffing (ie, family doctors in community-based settings). In addition, the use of population-level data from primary care catchment areas helped reduce provider-specific effects and reflect routine care delivery more accurately. However, unmeasured differences—such as provider motivation, local health system factors, or patient case mix—may have influenced the results. Therefore, findings should be interpreted with caution and viewed as indicative of association rather than definitive evidence of causality. Future studies employing randomized controlled designs or matched comparison groups with standardized training exposures and comparing the observed changes with broader regional trends in diabetes outcomes are recommended to strengthen causal conclusions and improve generalizability.

This study has many limitations. First, the study did not include a control group among GPs or GP trainees for the educational intervention, limiting our ability to attribute changes in learning outcomes solely to the MI training. Comparisons were made only within the intervention group before and after the training. Second, the sample size of GP trainees who completed both pre- and postassessments was modest, which may affect statistical power and limit the robustness of subgroup analyses. Third, the use of self-reported measures for MI confidence and qualitative interviews introduces potential bias, as participants may have overestimated their skills or provided socially desirable responses. Fourth, the relatively short follow-up period of 6 months for assessing clinical outcomes such as HbA_{1c} and blood pressure. While this timeframe allowed us to evaluate initial effects of MI training and implementation, it may be insufficient to capture longer-term behavioral change and sustained clinical benefits. In addition, while the study demonstrated statistically significant improvements in certain diabetes outcomes (eg, HbA_{1c} and diastolic blood pressure), the magnitude of these changes was modest. This may be due in part to the relatively short duration and intensity of the MI training provided. A single 4-hour workshop, even when supplemented by optional web-based modules, may not be sufficient to fully develop and sustain advanced MI competencies. Furthermore, qualitative findings indicated that GPs and trainees faced several implementation barriers, including time constraints, limited patient continuity, and lack of confidence in using more advanced MI processes such as evoking and planning. These challenges likely impacted the consistency and depth of MI delivery. Fifth, despite having a control group of patients with diabetes, the study sample consisted of a homogeneous group of GPs, GP trainees, and patients from a single area, potentially limiting the generalizability of the findings to broader populations. Another limitation is that no formal statistical adjustments were made for potential confounders, such as baseline differences in patient characteristics between the study and control groups. Although descriptive comparisons were conducted, the absence of adjusted analyses means residual confounding cannot be ruled out. Future studies using larger sample sizes and multivariable analytical approaches are warranted to strengthen the robustness of causal inferences. Another limitation is the absence of data on intermediate behavioral indicators—such as patient adherence to lifestyle recommendations, medication compliance, frequency of follow-up visits, or satisfaction with care—which may mediate or moderate the observed effects on HbA_{1c} and blood pressure. Including such measures in future studies would provide a more comprehensive understanding of the impact and mechanisms of MI in real-world primary care settings. Sixth, a further limitation relates to the qualitative analysis process. Thematic derivation was based on conventional content analysis using an inductive coding approach, whereby initial codes were generated from repeated readings of the transcripts and subsequently organized into categories and overarching themes.

However, coding was undertaken by a single trained coder, and intercoder reliability was not formally assessed. Although the coding process followed established qualitative research methods and was reviewed for internal consistency by members of the research team, the absence of formal intercoder reliability testing may have introduced subjectivity in theme interpretation. Finally, GP trainees' inability to follow-up with the same patients owing to rotational schedules reduced the opportunity to apply MI processes comprehensively.

Implications of Findings for Practice

MI may assist in addressing behavioral and relational barriers to diabetes care—such as patient ambivalence, low motivation, or resistance to change—we acknowledge that MI alone cannot resolve the full range of complex challenges associated with chronic disease management. Structural issues, resource constraints, and system-level determinants also contribute significantly to outcomes. Thus, MI should be viewed as a single component of a broader multifaceted strategy aimed at enhancing communication, empowering patients, and supporting behavioral change.

While MI is often perceived as time-consuming, the use of brief MI, as emphasized in our training, can actually help structure consultations more efficiently. Several participants reported that once familiar with brief MI, they were able to guide patient discussions more purposefully, avoid redundant advice-giving, and promote greater patient involvement in setting goals. Over time, this may reduce the burden on clinicians by improving communication flow and minimizing repeated discussions of unresolved issues in future visits. Thus, brief MI may mitigate, rather than exacerbate, the challenges posed by time-limited practice environments. The development of brief MI protocols tailored to high-volume clinical settings could enhance feasibility without compromising effectiveness. MI's emphasis on patient autonomy and behavioral change aligns with modern health care priorities, making it a valuable tool in chronic disease management. To support GPs in integrating MI into their daily practices, it is essential to provide continued education, peer coaching, and expert supervision. Policymakers should consider supporting MI training as part of CPD programs, given its demonstrated benefits in improving patient outcomes.

Further studies should explore the key components of brief MI that can retain the effectiveness of full-process MI while remaining practical in busy clinical environments. In addition, future studies should include longer follow-up periods to assess the sustained impact of MI on both GPs' skills retention and patient outcomes over time. Research should also focus on optimizing MI training approaches by identifying effective educational methods suited to GPs, as well as strategies to increase uptake and sustained implementation of MI in routine practice. The potential of digital tools, such as mobile apps or web-based coaching platforms, to support MI training and delivery warrants further exploration to enhance accessibility and long-term integration into primary care.

Conclusions

This study underscores the transformative potential of MI in primary care, particularly in diabetes management. MI training not only enhanced GP trainees' knowledge, but also translated into tangible improvements in patient health metrics. Brief MI is an alternative if one cannot perform a full session of MI and better suits the busy schedule of GPs. However, challenges such as time demands, skill mastery, and patient continuity must be addressed to maximize its usability. By refining training methods, supporting GPs in their implementation efforts, and exploring innovative applications of MI, health care systems can better harness its benefits to improve patient outcomes and advance the quality of care.

Acknowledgments

We sincerely thank the Production House teams at our school for their invaluable support in developing and producing a structured web-based course on MI and stage of change. This study was supported by the Faculty of Medicine, Khon Kaen University, Thailand, under grant IN63237.

Data Availability

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

Conflicts of Interest

None declared.

Multimedia Appendix 1
Brief motivational interviewing guides.
[DOCX File, 21 KB - mededu_v11i1e75916_app1.docx]

References

1. IDF Diabetes Atlas, 6th edition: International Diabetes Federation; 2021. URL: https://www.ncbi.nlm.nih.gov/books/NBK581934/ [accessed 2025-09-12]
2. Yameny AA. Diabetes mellitus overview 2024. J Biosci Appl Res 2024 Sep 28;10(3):641-645. [doi: 10.21608/jbaar.2024.382794]

3. Rawal LB, Tapp RJ, Williams ED, Chan C, Yasin S, Oldenburg B. Prevention of type 2 diabetes and its complications in developing countries: a review. Int J Behav Med 2012 Jun;19(2):121-133. [doi: 10.1007/s12529-011-9162-9] [Medline: 21590464]
4. Mahmoodi MR, Khanjani N. Barriers and limitations to obstacle diabetes self-management with a focus on nutritional literacy: solutions and opportunities. Critical review and research synthesis. CCB 2020. [doi: 10.18502/ccb.v1i1.2870]
5. Booth AO, Lowis C, Dean M, Hunter SJ, McKinley MC. Diet and physical activity in the self-management of type 2 diabetes: barriers and facilitators identified by patients and health professionals. Prim Health Care Res Dev 2013 Jul;14(3):293-306. [doi: 10.1017/S1463423612000412] [Medline: 23739524]
6. Chen CC, Chen CL, Ko Y. The misconceptions and determinants of diabetes knowledge in patients with diabetes in Taiwan. J Diabetes Res 2020;2020:2953521. [doi: 10.1155/2020/2953521] [Medline: 32656263]
7. Lakerveld J, Palmeira AL, van Duinkerken E, Whitelock V, Peyrot M, Nouwen A. Motivation: key to a healthy lifestyle in people with diabetes? Current and emerging knowledge and applications. Diabet Med 2020 Mar;37(3):464-472. [doi: 10.1111/dme.14228] [Medline: 31916283]
8. Kapur K, Kapur A, Ramachandran S, et al. Barriers to changing dietary behavior. J Assoc Physicians India 2008 Jan;56(27):27-32. [Medline: 18472496]
9. Kahan S, Manson JAE. Nutrition counseling in clinical practice: how clinicians can do better. JAMA 2017 Sep 26;318(12):1101-1102. [doi: 10.1001/jama.2017.10434] [Medline: 28880975]
10. Micha R, Peñalvo JL, Cudhea F, Imamura F, Rehm CD, Mozaffarian D. Association between dietary factors and mortality from heart disease, stroke, and type 2 diabetes in the United States. JAMA 2017 Mar 7;317(9):912-924. [doi: 10.1001/jama.2017.0947] [Medline: 28267855]
11. Miller W, Rollnick S. The method of motivational interviewing. In: Motivational Interviewing: Preparing People for Change: Guilford Press; 2013:25-36 URL: https://bluepeteraustralia.wordpress.com/wp-content/uploads/2012/12/motivational-interviewing.pdf [accessed 2025-09-12]
12. Rubak S, Sandbaek A, Lauritzen T, Christensen B. Motivational interviewing: a systematic review and meta-analysis. Br J Gen Pract 2005 Apr;55(513):305-312. [Medline: 15826439]
13. Söderlund LL, Madson MB, Rubak S, Nilsen P. A systematic review of motivational interviewing training for general health care practitioners. Patient Educ Couns 2011 Jul;84(1):16-26. [doi: 10.1016/j.pec.2010.06.025] [Medline: 20667432]
14. Zhu S, Sinha D, Kirk M, et al. Effectiveness of behavioural interventions with motivational interviewing on physical activity outcomes in adults: systematic review and meta-analysis. BMJ 2024 Jul 10;386:e078713. [doi: 10.1136/bmj-2023-078713] [Medline: 38986547]
15. West DS, DiLillo V, Bursac Z, Gore SA, Greene PG. Motivational interviewing improves weight loss in women with type 2 diabetes. Diabetes Care 2007 May;30(5):1081-1087. [doi: 10.2337/dc06-1966] [Medline: 17337504]
16. Frost H, Campbell P, Maxwell M, et al. Effectiveness of motivational interviewing on adult behaviour change in health and social care settings: a systematic review of reviews. PLoS ONE 2018;13(10):e0204890. [doi: 10.1371/journal.pone.0204890] [Medline: 30335780]
17. Thepwongsa I, Muthukumar R, Kessomboon P. Motivational interviewing by general practitioners for type 2 diabetes patients: a systematic review. Fam Pract 2017 Aug 1;34(4):376-383. [doi: 10.1093/fampra/cmx045] [Medline: 28486622]
18. Christie D, Channon S. The potential for motivational interviewing to improve outcomes in the management of diabetes and obesity in paediatric and adult populations: a clinical review. Diabetes Obes Metab 2014 May;16(5):381-387. [doi: 10.1111/dom.12195] [Medline: 23927612]
19. Kushner PR, Cavender MA, Mende CW. Role of primary care clinicians in the management of patients with type 2 diabetes and cardiorenal diseases. Clin Diabetes 2022;40(4):401-412. [doi: 10.2337/cd21-0119] [Medline: 36381309]
20. Sturgiss E, Advocat J, Ball L, Williams LT, Prathivadi P, Clark AM. Behaviour change for type 2 diabetes: perspectives of general practitioners, primary care academics, and behaviour change experts on the use of the 5As framework. Fam Pract 2022 Sep 24;39(5):891-896. [doi: 10.1093/fampra/cmab182] [Medline: 35079780]
21. John NA, John J, Tarnikanti M, et al. Implications of lifestyle medicine in medical practice. J Family Med Prim Care 2023;12(2):208-212. [doi: 10.4103/jfmpc.jfmpc_1587_22]
22. Abdullah MY, Alshehri SA, Mahnashi HA, et al. Role of primary care physician in health promotion and education. Int J Community Med Public Health 2022;9(12):4705-4709. [doi: 10.18203/2394-6040.ijcmph20223234]
23. Kaczmarek T, Kavanagh DJ, Lazzarini PA, Warnock J, Van Netten JJ. Training diabetes healthcare practitioners in motivational interviewing: a systematic review. Health Psychol Rev 2022 Sep;16(3):430-449. [doi: 10.1080/17437199.2021.1926308] [Medline: 33970799]
24. Moran J, Bekker H, Latchford G. Everyday use of patient-centred, motivational techniques in routine consultations between doctors and patients with diabetes. Patient Educ Couns 2008 Nov;73(2):224-231. [doi: 10.1016/j.pec.2008.07.006] [Medline: 18701234]
25. Kıral MA, Cansu GB. Glycemic regulation in patients with type 2 diabetes mellitus: effects of motivational interviewing. Ankara Med J 2022;22(3):336-346. [doi: 10.5505/amj.2022.87854]

26. Li Z, Chen Q, Yan J, Liang W, Wong WCW. Effectiveness of motivational interviewing on improving care for patients with type 2 diabetes in China: a randomized controlled trial. BMC Health Serv Res 2020 Jan 23;20(1):57. [doi: 10.1186/s12913-019-4776-8] [Medline: 31973759]
27. Charles M, Bruun NH, Simmons R, et al. The effect of training GPs in motivational interviewing on incident cardiovascular disease and mortality in people with screen-detected diabetes. Results from the ADDITION-Denmark randomised trial. BJGP Open 2020;4(1):bjgpopen20X101012. [doi: 10.3399/bjgpopen20X101012] [Medline: 32071038]
28. Thatcher R, Gregory N, Cheung WY, et al. Brief lifestyle interventions for prediabetes in primary care: a service evaluation. BMC Prim Care 2022 Mar 14;23(1):45. [doi: 10.1186/s12875-022-01658-2] [Medline: 35282823]
29. Fontaine G, Cossette S, Heppell S, et al. Evaluation of a web-based e-learning platform for brief motivational interviewing by nurses in cardiovascular care: a pilot study. J Med Internet Res 2016 Aug 18;18(8):e224. [doi: 10.2196/jmir.6298] [Medline: 27539960]
30. Adams CM. Evaluating the Feasibility and Efficacy of a Brief Motivational Interviewing Nutrition Intervention for Women with Type 2 Diabetes in Primary Care: University of Kentucky; 2023. URL: https://uknowledge.uky.edu/cgi/viewcontent.cgi?article=1465&context=dnp_etds [accessed 2025-09-12]
31. Nightingale B, Gopalan P, Azzam P, Douaihy A, Conti T. Teaching brief motivational interventions for diabetes to family medicine residents. Fam Med 2016 Mar;48(3):187-193. [Medline: 26950907]
32. Davis DA, Barnes BE, Fox RD. The Continuing Professional Development of Physicians: From Research to Practice: AMA Press; 2003.
33. Thepwongsa I, Kirby C, Schattner P, Shaw J, Piterman L. Type 2 diabetes continuing medical education for general practitioners: what works? A systematic review. Diabet Med 2014 Dec;31(12):1488-1497. [doi: 10.1111/dme.12552] [Medline: 25047877]
34. Marinopoulos SS, Dorman T, Ratanawongsa N, et al. Effectiveness of continuing medical education. Evid Rep Technol Assess (Full Rep) 2007 Jan(149):1-69. [Medline: 17764217]
35. Thepwongsa I. Education of rural and remote general practitioners (gps) in australian on type 2 diabetes: impact of online continuing medical education on gps' knowledge, attitudes and practices and barriers to online learning. : Monash University; 2017 URL: https://bridges.monash.edu/articles/thesis/Education_of_rural_and_remote_general_practitioners_GPs_in_Australia_on_type_2_diabetes_impact_of_online_continuing_medical_education_on_GPs_knowledge_attitudes_and_practices_and_barriers_to_online_learning/4684069?file=16443299 [accessed 2025-09-04]
36. Thepwongsa I, Kirby CN, Schattner P, Piterman L. Online continuing medical education (CME) for GPs: does it work? A systematic review. Aust Fam Physician 2014 Oct;43(10):717-721. [Medline: 25286431]
37. Raumer-Monteith L, Kennedy M, Ball L. Web-based learning for general practitioners and practice nurses regarding behavior change: qualitative descriptive study. JMIR Med Educ 2023 Jul 27;9(1):e45587. [doi: 10.2196/45587] [Medline: 37498657]
38. Muthukumar R, Thepwongsa I, Sripa P, Jindawong B, Jenwitheesuk K, Virasiri S. Preclinical medical students' perspectives and experiences with structured web-based english for medical purposes courses: cross-sectional study. JMIR Med Educ 2025 Mar 27;11(1):e65779. [doi: 10.2196/65779] [Medline: 40177857]
39. Oster C, Leibbrandt R, Schoo A, et al. A feasibility study of teaching motivational interviewing in a fully online environment using a virtual client. Int J Health Promot Educ 2022. [doi: 10.1080/14635240.2022.2047095]
40. Lukaschek K, Schneider N, Schelle M, et al. Applicability of motivational interviewing for chronic disease management in primary care following a web-based e-learning course: cross-sectional study. JMIR Ment Health 2019 Apr 29;6(4):e12540. [doi: 10.2196/12540] [Medline: 31033446]
41. Chawla N, Gyawali S, Sharma P, Balhara YPS. Internet-based learning for professionals in addiction psychiatry: a scoping review. Indian J Psychol Med 2022 Jul;44(4):325-331. [doi: 10.1177/02537176221082897] [Medline: 35949641]
42. Ng WCE, Koura R, Khaider KB, Chew CSE, Davis C. Effectiveness of a hybrid, obesity-specific counselling programme in improving medical students' self-efficacy and motivational interviewing skills for paediatric obesity counselling. BMC Med Educ 2025 Feb 11;25(1):224. [doi: 10.1186/s12909-024-06589-3] [Medline: 39934793]
43. Schaper K, Woelber JP, Jaehne A. Can the spirit of motivational interviewing be taught online? A comparative study in general practitioners. Patient Educ Couns 2024 Aug;125:108297. [doi: 10.1016/j.pec.2024.108297] [Medline: 38728998]
44. Thepwongsa I, Muthukumar R, Sripa P, Piterman L. Uptake and effectiveness of online diabetes continuing education: The perspectives of Thai general practitioner trainees. Heliyon 2023 Feb;9(2):e13355. [doi: 10.1016/j.heliyon.2023.e13355] [Medline: 36755621]
45. Cook DA, Dupras DM. A practical guide to developing effective web-based learning. J Gen Intern Med 2004 Jun;19(6):698-707. [doi: 10.1111/j.1525-1497.2004.30029.x] [Medline: 15209610]
46. Rubak S, Sandbaek A, Lauritzen T, Borch-Johnsen K, Christensen B. General practitioners trained in motivational interviewing can positively affect the attitude to behaviour change in people with type 2 diabetes. One year follow-up of an RCT, ADDITION Denmark. Scand J Prim Health Care 2009;27(3):172-179. [doi: 10.1080/02813430903072876] [Medline: 19565411]

47. Rubak S, Sandbaek A, Lauritzen T, Borch-Johnsen K, Christensen B. An education and training course in motivational interviewing influence: GPs' professional behaviour--ADDITION Denmark. Br J Gen Pract 2006 Jun;56(527):429-436. [Medline: 16762124]
48. Swanson V, Maltinsky W. Motivational and behaviour change approaches for improving diabetes management. Practical Diabetes 2019 Jul;36(4):121-125. [doi: 10.1002/pdi.2229]
49. Racic M, Katic B, Joksimovic BN. Impact of motivational interviewing on treatment outcomes in patients with diabetes type 2: a randomized controlled trial. J Fam Med 2015;2(2):1020-1021 [FREE Full text]
50. De Greef K, Deforche B, Tudor-Locke C, De Bourdeaudhuij I. Increasing physical activity in Belgian type 2 diabetes patients: a three-arm randomized controlled trial. Int J Behav Med 2011 Sep;18(3):188-198. [doi: 10.1007/s12529-010-9124-7] [Medline: 21052886]
51. Rubak S, Sandbæk A, Lauritzen T, Borch-Johnsen K, Christensen B. Effect of "motivational interviewing" on quality of care measures in screen detected type 2 diabetes patients: a one-year follow-up of an RCT, ADDITION Denmark. Scand J Prim Health Care 2011 Jun;29(2):92-98. [doi: 10.3109/02813432.2011.554271] [Medline: 21306296]
52. Christian JG, Bessesen DH, Byers TE, Christian KK, Goldstein MG, Bock BC. Clinic-based support to help overweight patients with type 2 diabetes increase physical activity and lose weight. Arch Intern Med 2008 Jan 28;168(2):141-146. [doi: 10.1001/archinternmed.2007.13] [Medline: 18227359]
53. Madson MB, Loignon AC, Lane C. Training in motivational interviewing: a systematic review. J Subst Abuse Treat 2009 Jan;36(1):101-109. [doi: 10.1016/j.jsat.2008.05.005] [Medline: 18657936]
54. Polley SJ, Hunter SR, McBride A, Deister D, Heward BJ. Motivational interviewing: an introduction to spirit and skills. J Am Acad Child Adolesc Psychiatry 2023 Oct;62(10):S128. [doi: 10.1016/j.jaac.2023.07.554]
55. Gabarda A, Butterworth S, Liang Q, Beckjord E. Pilot study of a motivational interviewing training on practitioners' skill set for patient centered communication. Am J Health Promot 2023 Nov;37(8):1070-1077. [doi: 10.1177/08901171231191130] [Medline: 37494296]
56. Boom SM, Oberink R, Zonneveld AJE, van Dijk N, Visser MRM. Implementation of motivational interviewing in the general practice setting: a qualitative study. BMC Prim Care 2022 Jan 28;23(1):21. [doi: 10.1186/s12875-022-01623-z] [Medline: 35172737]

Abbreviations

CME: continuing medical education
CPD: continuing professional development
GP: general practitioner
MI: motivational interviewing

Edited by D Chartash; submitted 13.04.25; peer-reviewed by AN Ali, N Mor; revised version received 17.08.25; accepted 20.08.25; published 16.09.25.

Please cite as:
Thepwongsa I, Nonjui P, Muthukumar R, Sripa P
Impact of Motivational Interviewing Education on General Practitioners' and Trainees' Learning and Diabetes Outcomes in Primary Care: Mixed Methods Study
JMIR Med Educ 2025;11:e75916
URL: https://mededu.jmir.org/2025/1/e75916
doi:10.2196/75916

Using Web-Based Continuing Education to Improve New Diagnoses of Alzheimer Disease in Claims Data: Retrospective Case-Control Study

Katie Lucero[1], PhD; Thomas Finnegan[1], PhD; Soo Borson[2], MD

[1]Medscape, 283-299 Market Street, 2 Gateway Center - 4th Floor, Newark, NJ, United States

[2]Keck School of Medicine, University of Southern California, Los Angeles, CA, United States

Corresponding Author:
Katie Lucero, PhD
Medscape, 283-299 Market Street, 2 Gateway Center - 4th Floor, Newark, NJ, United States

Abstract

Background: Alzheimer disease (AD) presents significant challenges to health care systems worldwide. Early and accurate diagnosis of AD is crucial for effective management and care to enable timely treatment interventions that can preserve cognitive function and improve patient quality of life. However, there are often significant delays in diagnosis. Continuing medical education (CME) has enhanced physician knowledge and confidence in various medical fields, including AD. Notably, web-based CME has been shown to positively influence physician confidence, which can lead to changes in practice and increased adoption of evidence-based treatment selection.

Objective: This study investigated the impact of a targeted, web-based CME intervention on health care providers' confidence, competence, and real-world outcomes in diagnosing early AD.

Methods: The study employed a 2-phase design. Phase I used a pre-post assessment to evaluate immediate changes in knowledge and confidence before and after CME participation. Phase II involved a retrospective, matched case-control study to examine the impact of CME on AD diagnoses in claims data.

Results: A 1-way ANOVA showed a significant effect of CME regarding change in the volume of AD diagnoses (F_{1900}=5.50; P=.02). Compared to controls, CME learners were 1.6 times more likely to diagnose AD, resulting in an estimated net increase of 7939 new diagnoses annually. Post-CME confidence was associated with a greater likelihood of diagnosing AD (odds ratio 1.64; 95% CI 0.92-2.92; P=.09; n=219).

Conclusions: Web-based CME participation is associated with increased real-world AD diagnoses. Findings offer a mechanism to explain the changes in clinical practice seen as a result of the CME intervention, which improves skills and confidence.

(JMIR Med Educ 2025;11:e72000) doi:10.2196/72000

KEYWORDS

Alzheimer disease diagnosis; continuing medical education; real-world outcomes; physician confidence; web-based CME; CME; self-efficacy

Introduction

Alzheimer disease (AD) is a progressive neurodegenerative disorder that poses significant challenges to health care systems worldwide. AD affects more than 6.0 million persons in the United States, 7.9 million in Europe, and at least 50 million people worldwide [1,2]. The risk of AD increases with age. By 2050, the number of affected persons 65 years and older is expected to reach 12.7 million in the United States and over 152 million worldwide [1]. As the global population ages, the prevalence of AD has risen dramatically and will continue to rise, bringing new urgency to addressing the widespread lags in diagnosis that impede effective patient management and care.

Early diagnosis is vital for maintaining quality of life, delaying institutionalization, and improving treatment outcomes. Monoclonal antibodies that target plaque work best in the early stages of AD when pathologic changes are still relatively mild [3,4]. Early diagnosis allows for early initiation of treatment, which can help preserve patients' functional abilities and cognitive function, thereby improving quality of life [5]. Early diagnosis can also reduce caregiver burden by helping patients and caregivers access culturally competent care and support services to improve quality of life [6].

Significant delays are common in the diagnosis and management of patients with AD. Physician practice patterns across several countries, including the United States, reveal that while approximately half of patients globally receive an AD diagnosis

within 6 months of initial presentation, a significant number of patients remain undiagnosed for several months after initially presenting to a physician [7-9]. Misdiagnosis of AD in primary care settings is exceptionally high, with as many as two-thirds of patients being misdiagnosed [6]. While primary care physicians (PCPs) are typically first to see patients with mild cognitive impairment (MCI) and early AD [10], physician suspicion accounts for only 20% of AD diagnoses globally, with caregivers often serving as the primary impetus for seeking medical attention [8]. Overall, referral rates for specialist care are also low (14% - 23%) [8].

Recent updates to diagnostic and staging criteria for AD are based on biological indicators versus clinical syndrome [11]. In practice, AD is often diagnosed through the evaluation of cognitive symptoms, which is highly dependent on a clinician's experience and skill [6]. In the absence of a single diagnostic test for AD, physicians rely on physical and neurological examination, mental status tests, imaging, and biomarkers for diagnostic purposes [12]. However, by the time a patient starts showing signs of cognitive impairment, underlying pathologic changes have likely been happening for a decade or longer [13].

Gaps in physician knowledge and insufficient specialized training partly drive challenges in the diagnostic process [12]. Physicians often struggle to distinguish normal aging from dementia, and between various types of dementia [14], and demonstrate limited awareness of early cognitive impairment indicators [8]. Notably, many physicians lack self-efficacy in diagnostic abilities, including their skills to detect signs of MCI and differentiate MCI from AD [12]. Physicians also lack self-efficacy to use and interpret cognitive testing and neuroimaging [5]. While specialists are more likely to use magnetic resonance imaging, PCPs often rely on computed tomography scans, which are less informative [12]. This lack of self-efficacy can lead to delayed diagnoses, hindering timely interventions and optimal patient outcomes.

Continuing medical education (CME), including web-based CME, has shown promise in improving physician knowledge and self-efficacy across various medical domains, including in AD diagnosis [15-17]. However, the relationship between improving knowledge, competence, self-efficacy, and real-world outcomes (RWOs) in AD diagnosis remains understudied [10,16-18]. Improving knowledge does not guarantee its application in practice. Rather, improving self-efficacy is an essential intermediary between knowledge and practice change. Self-efficacy, a motivational construct also known as confidence, empowers physicians to act upon their knowledge and implement learned skills (also known as competence) [19]. However, the relationship is not strictly linear. Improvements in and reinforcement of knowledge and competence can also increase self-efficacy, which in turn influences practice change [20,21]. These relationships suggest that clinicians with a greater sense of self-efficacy following CME activities demonstrate a stronger intention to change their practice, regardless of whether they improved their knowledge [22,23].

This study investigates whether a targeted, web-based CME intervention can improve physicians' self-efficacy, competence, and RWOs in diagnosing early AD. The study addressed the following hypotheses: (1) competency scores for HCPs will increase; (2) the proportion of HCPs who are confident will increase; and (3) the volume of new AD diagnoses will increase for the CME group compared with the matched control group.

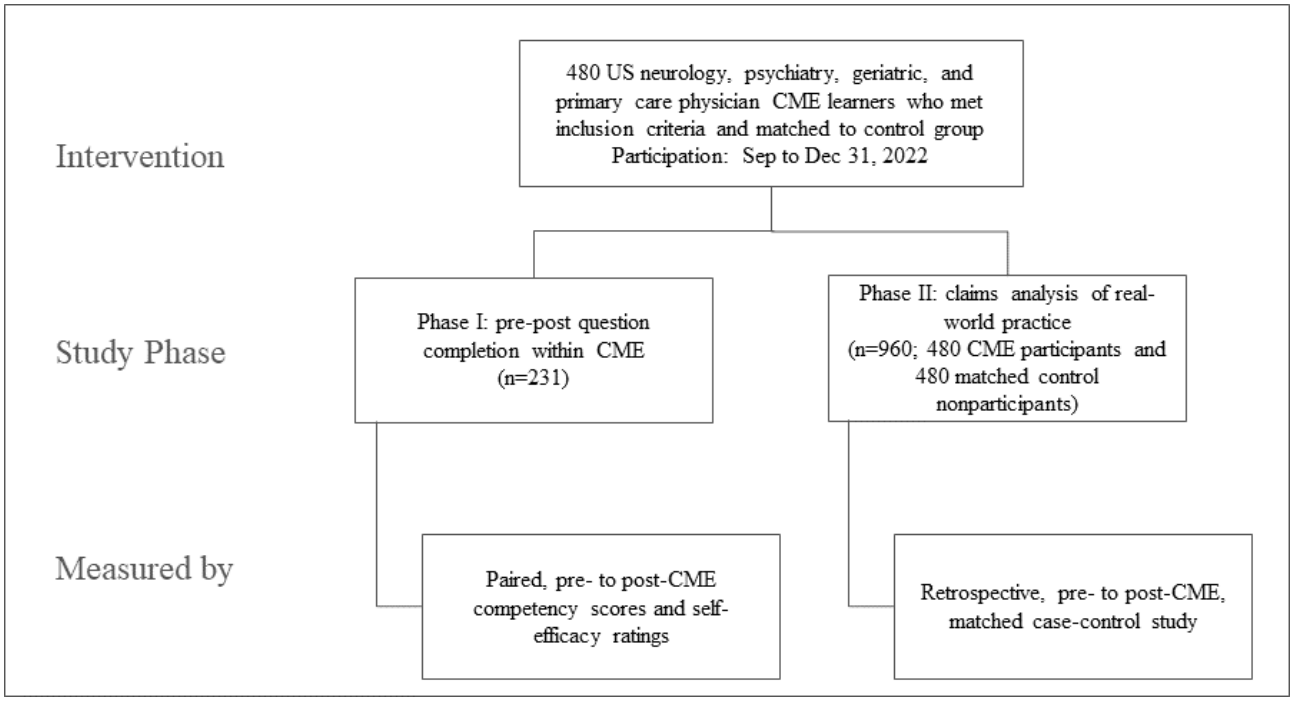
Methods

Study Design

Overview

We conducted 2 study phases using the outcomes assessment framework by Moore et al [24] to assess leading indicators (changes in competency scores and confidence ratings) and lagging indicators of success (changes in real-world performance, specifically the volume of new AD diagnoses). Phase I focused on AD diagnosis within CME activities. Phase II focused on AD diagnosis in the real world (Figure 1).

Figure 1. Study design. CME: continuing medical education.



Phase I: Educational Assessment

We employed a paired, pre-post design to assess the impact of CME activities on knowledge and competency scores and confidence ratings immediately before and after the point of learning in the activity for learners from September 13 to December 31, 2022.

Phase II: Real-World Outcomes

We conducted a retrospective, matched case-control study from March 2022 to June 2023 to evaluate the impact of CME activities on diagnosing patients with AD. The intervention period spanned from September 13 to December 31, 2022, with data collection extending from March 2022 to June 2023, assessing practice 6 months before and 6 months after the CME participation date (“index date”).

CME Intervention

The intervention consisted of a web-based CME initiative for PCPs and neurologists designed to improve competence and confidence in early recognition and diagnosis of AD. The first activity focused on best practices in delivering care for patients with AD in primary care and neurology (released September 13, 2022, through September 13, 2023) [25]. The activity was valid for a maximum of 0.50 American Medical Association Physician Recognition Award (AMA PRA) Category 1 Credit. Topics included triaging and assessing patients with cognitive impairment in primary care, coordinating with neurologists, and treatment goals. A subsequent series of 3 simulated online office visits focused on increasing clinicians’ ability to identify and communicate with patients experiencing early cognitive impairment (released October 14, 2022, through October 14, 2023, valid for a maximum of 0.25 AMA PRA Category 1 Credit) [26]. Each office visit centered around an interactive patient-physician vignette for which a decision on screening and evaluation was required. Vignettes included White and Black patients with cognitive impairment due to MCI, cognitive impairment due to early AD, and cognitive impairment not due to dementia. Participants previewed the chief concern for each patient vignette via a landing page with an interactive, graphic table of contents. The activity required participants to investigate cognitive complaints and select diagnostic tests. Faculty feedback was included in each activity.

Inclusion Criteria

Physicians were included if they participated (that is, viewed the content, after the front matter and disclosures) in at least 1 activity in the study period, practiced in the United States, had at least 1 patient who met the inclusion criteria, and had complete claims data available for the study period. Patients were included if they were at least 60 years of age and saw the learner during the study period as evidenced by at least 1 *International Statistical Classification of Diseases and Related Health Problems, Tenth Revision* (*ICD-10*) code or prescription by the learner.

Sample

Medscape member registration provides the country of residence, profession, specialty, and National Provider Identifier (NPI; if applicable) number for health care providers (HCPs). A total of 1725 US physicians in the target specialties with a valid NPI were learners between September 13 and December 31, 2022. Of these, 1310 had matches with claims data with at least 1 patient who met the inclusion criteria. The RWO intervention group (RWO learners) comprised 480 physicians who met the full inclusion criteria, participated in the CME activities during the intervention period, and were matched to a nonparticipant control (287/480, 60% PCPs; 100/480, 21% neurologists; 71/480, 15% psychiatrists; and 22/480, 5% geriatric specialists). Of the 480, 231 fully completed an activity.

Matching Process

An equal number of 480 HCPs who did not participate in CME served as the control group. We used a 1:1 matching ratio to pair cases with controls. The matching criteria included: (1) number of patients with AD, (2) number of patients diagnosed with AD by the HCP, (3) profession, (4) specialty, and (5) the first 2 digits of the HCP's ZIP code. Match-It in R was used to match the propensity score by the number of patients with AD and the number of patients diagnosed with AD by the HCP. Exact matching was used for profession, specialty, and the first 2 digits of the ZIP code. An independent samples *t* test showed no statistically significant difference in the number of patients with AD seen in the preperiod by the CME and matched control groups (t=−0.24; P=.81).

Real-World Data Collection

We sourced real-world data from Medscape licensed claims data, accessed in November 2023. This dataset provided comprehensive information on patient visits, diagnoses, and procedures performed by the participating HCPs. Data were aggregated at the patient level, and patient counts were aggregated at the HCP level. See Multimedia Appendix 1 for the codes accepted as indicators of diagnosis.

Measures

Three primary outcome measures assessed the effectiveness of the CME intervention: (1) competency score, (2) confidence rating, and (3) number of patients newly diagnosed with AD by the HCP. Additionally, we examined whether there was an association between being confident following CME and real-world diagnoses.

Competency Score

Competency scores were assessed by asking 3 case-based vignette questions pre- and postpoint of learning in the CME activity. Scores were aggregated at either the learning topic or activity level and then to the HCP. Scores ranged from 0% to 100%. Questions were developed to assess learning against learning objectives related to cognitive assessment or differentiation between MCI and AD. Questions were developed by content experts and reviewed by an expert AD HCP, an outcomes assessment specialist, and a copyeditor.

Self-Efficacy Rating

A question assessed self-efficacy on a 5-point Likert-type scale, with higher scores indicating greater self-efficacy (eg, "How confident are you right now in your ability to assess patients for cognitive impairment?"). Learners were deemed "confident" if they rated themselves as a 4 or 5. We used the term "confident" because it is more easily understood by the respondent than self-efficacy and use "confident" in the methodology and results for ease of interpretation when referring to those who select a 4 or 5.

New AD Diagnoses

New AD diagnoses were assessed by examining patient-level *ICD-10* data. Patients with an AD *ICD-10* code by the HCP of interest in the timeframe of interest were identified. Then, the patient's history 2.5 years prior was examined to assess whether they had previously received any AD *ICD-10* codes by any HCP. If there was no history of receiving an AD *ICD-10* code previously, the patient's *ICD-10* code was considered a new AD diagnosis by the HCP of interest. The count of patients who met these criteria was aggregated at the HCP level.

Statistical Analysis

Phase I: Educational Assessment

We assessed immediate changes in learner competency scores and confidence via 4 matched pair questions before and after CME participation. The McNemar test evaluated change in the proportion of learners who rate themselves as confident. The McNemar test was chosen because it measures differences in paired proportions against the null hypothesis. A paired samples *t* test was conducted to measure mean differences in paired samples. Overall competency changes were assessed using paired samples *t* tests. Statistical significance was set at P<.05 for all tests.

Phase II: Real-World Outcomes

The relationship between CME participation and the postintervention volume of new AD diagnoses was assessed with a 1-way ANOVA. One-way ANOVA was chosen because we wanted to examine the change in volume of AD diagnosis from pre– to post–index date (dependent variable) and whether being a learner was associated with this change (independent variable with 2 independent groups). The dependent variable was the change in the volume of AD diagnoses from pre– to post–index date, and the independent variable was CME participation (versus control). In a secondary analysis, we explored the association between postintervention confidence (confident=1, not confident=0) and AD diagnosis (diagnoser=1, nondiagnoser=0) via logistic regression.

The association between being confident and diagnosing AD was explored because the more we know about mechanisms for change that we can immediately measure at scale within a web-based CME activity, the more effective our education can be. A dichotomous independent variable and dichotomous dependent variable were selected because the research question focused on whether confidence predicts being a diagnoser. More confidence should not equate to diagnosing more because how many patients get diagnosed depends on the types of patients an HCP sees. Previous research used this same dichotomy and found similar results [25].

Statistical significance was set at P<.05 for all tests, and analyses were performed using SAS version 9.4 (SAS Institute).

Ethical Considerations

The Sterling Institutional Review Board deemed this study exempt under the terms of the US Department of Health and Human Service's Policy for Protection of Human Research Subjects at 45 CFR §46.104(d) [27]. The ethical standards of the Declaration of Helsinki were applied to all research procedures. As the study was exempt, there was no requirement for informed consent. The institutional review board approval covered secondary analysis without additional consent. The data were deidentified prior to analysis to safeguard participant information. No compensation was provided to participants.

Results

Phase I: Competency and Confidence

"Completers" answered all linked questions within at least 1 of the CME activities, representing 48% (231/480) of the larger learner population. After participation, RWO learners demonstrated a 34 percentage point increase in correct answers for competency in the diagnosis of AD (33% prescore to 67% postscore; *P*=.008) and a 16 percentage point pre- to postactivity increase in the proportion of those who were confident in assessing cognitive function and diagnosing AD (75/231 preactivity to 99/231 postactivity; *P*<.001).

Phase II: New AD Diagnoses

ANOVA showed a significant effect of CME regarding change in the volume of AD diagnoses (F_{1900}=5.50; *P*=.02). The 6-month postactivity increase in new AD diagnoses was 160% greater for the CME group than the control group, as verified by claims data. RWO learners diagnosed 239 more patients after education (487 diagnoses pre-education vs 726 diagnoses posteducation). Control-group learners diagnosed 91 more patients after education (517 diagnoses pre-education vs 608 diagnoses posteducation). Neurologists had the highest increase in new AD diagnoses (1.58 per neurologist), while psychiatrists had the lowest (0.10 per psychiatrist). The logistic regression model showed a trend within the CME group toward a significant positive relationship between being confident in AD assessment post-CME and diagnosing AD in the real world in the 6 months following CME (odds ratio [OR] 1.64, 95% CI 0.92-2.92; *P*=.09; n=219). Table 1 summarizes the RWO learners and the matched control group on key outcomes from claims data.

Table . Number of patients with Alzheimer disease (AD) and number of patients newly diagnosed with AD before and 6 months after the activity.

	Patients with AD, mean (SD)		Patients newly diagnosed with AD, mean (SD)	
	Pre	Post	Pre	Post
CME[a] (n=480)	2.16 (7.51)	2.33 (8.35)	1.01 (3.60)	1.51 (5.59)
Geriatric specialists (n=22)	2.27 (5.23)	2.64 (5.95)	1.41 (2.92)	2.05 (4.36)
Neurologists (n=100)	6.50 (15.03)	7.14 (16.79)	3.04 (7.17)	4.62 (11.29)
PCPs[b] (n=287)	1.12 (2.38)	1.14 (2.43)	0.51 (1.14)	0.71 (1.52)
Psychiatrists (n=71)	0.20 (0.60)	0.27 (0.81)	0.10 (0.38)	0.20 (0.62)
Control (n=480)	2.05 (6.27)	1.92 (6.03)	1.08 (3.33)	1.27 (4.07)
Geriatric specialists (n=22)	1.95 (4.36)	1.86 (2.92)	0.82 (1.62)	1.00 (1.69)
Neurologists (n=100)	5.94 (12.11)	5.50 (11.40)	3.24 (6.40)	3.71 (7.46)
PCPs (n=287)	1.15 (2.56)	1.08 (3.03)	0.57 (1.42)	0.68 (2.31)
Psychiatrists (n=71)	0.24 (0.60)	0.31 (1.04)	0.17 (0.48)	0.27 (1.03)

[a]CME: continuing medical education.

[b]PCP: primary care physician.

Discussion

Principal Findings

This matched case-control study examined the impact of a web-based, vignette-based CME on participants' knowledge, competence, self-efficacy, and RWOs in diagnosing early AD. Participation in CME was associated with a significant (*P*=.02) increase in the diagnosis of early AD. RWO learners were more likely to be diagnosers than control-group physicians, with a magnitude of increase in AD diagnoses that was 1.6 times higher for RWO learners than control-group physicians. The estimated net increase of 7939 in new AD diagnoses in the year following participation for CME learners through the expiration of the activities for credit indicates a substantial positive impact of education on AD diagnosis rates. RWO learners also improved their confidence in identifying early forms of AD (*P*<.001). When HCPs were confident after CME, they had a 1.64 greater odds of diagnosing AD.

Comparison with Prior Work

Research suggests that CME can effectively improve physician knowledge, self-efficacy, and competence regarding dementia care in general. A large study in Australia evaluated an accredited CME program on the diagnosis and management of dementia in primary care. Participants who completed the program reported feeling significantly more confident in their knowledge, skills, and ability to provide care for people with dementia [15,16]. Our study not only affirms the impact of CME on real-world AD diagnoses but also offers a mechanism to explain the changes in real-world practice seen as a result of the CME intervention. Previous research shows that improvements in knowledge and competence following CME participation are associated with increased self-efficacy, and posteducation self-efficacy mediates the relationship between knowledge and competence and intention to change [20,21]. A recent secondary analysis of knowledge, competency, self-efficacy, and clinical practice using pre- and postparticipation data from web-based CME interventions in 3

different therapeutic areas combined with medical claims data examined the relationship between knowledge, competency, self-efficacy, and real-world clinical practice [23]. Knowledge and competency (P=.08; OR 1.515, 95% CI 0.953-2.410) and self-efficacy (P<.001; OR 2.768, 95% CI 1.705-4.492) were significant predictors of clinical practice. However, the effect size for self-efficacy was larger, suggesting that clinicians confident in their abilities were more likely to utilize evidence-based treatments. These results suggest that self-efficacy plays a significant mediating role in influencing clinical practice.

Reinforcement of existing knowledge also appears to influence clinical practice. A study that examined the relationships among knowledge, competence, self-efficacy, and intention to change across 57 online oncology-certified education programs published from 2018 to 2020 found that both improvements in and reinforcement of knowledge and competence are significant predictors of changes in self-efficacy [20]. Lucero et al [28] supported this finding. They found that participants who reinforced their knowledge had higher posteducation confidence ratings than participants who improved their knowledge after controlling for posteducation scores. Reinforcement of knowledge also likely explains why neurologists demonstrated the most significant increase in the number of new AD diagnoses.

Limitations

Potential confounding factors could affect the relationship between CME participation and increased AD diagnoses. Physicians who participated in CME may have been more motivated to improve their practice, potentially leading to increased diagnoses regardless of the CME content. Three activities were case-based simulated patient visits and 1 was a video-based discussion on cases. We did not tease out which activities might have been more or less impactful and whether participation in multiple activities was associated with practice change. Concurrent initiatives, such as other AD awareness campaigns or participation in non–study-related CME activities focused on AD or cognitive disorders during the study period, could also have confounded results. The control group also saw an increase of 91 more diagnoses postintervention, suggesting some external factors may have influenced diagnosis rates. Professional learning occurs in many places, given the demands of clinical practice and the requirement to maintain licensure. While changes for the control group were anticipated, matching based on demographic and practice factors helps reduce biases associated with those factors such as opportunity to diagnose, training, types of patients seen, and environment in which one practices.

Despite these limitations, the comprehensive matching strategy minimizes potential confounding factors and ensures group comparability. By matching profession, specialty, and ZIP code, the study controls for some differences in baseline knowledge, experience, and practice patterns associated with different medical specialties, as well as variations in patient demographics and health care access. Matching based on the number of patients with AD controlled for differences in patient population and exposure to AD cases. Using a time-aligned control group helped to control temporal factors, such as concurrent initiatives, that would affect both groups equally. Results were assessed by counting claims nested in patients to better tease out patients with their first AD diagnosis from a learner versus a nonlearner physician. Using paired pre- and postintervention data for individual learners enhances the statistical precision of the analysis, reducing sampling error and providing a robust assessment of the education's impact. Including the matched control group at follow-up increases confidence that changes are associated with education. Future research should explicitly examine how CME interventions affect AD diagnosis rates across different racial and ethnic groups and identify with more detail the mechanisms for change. We identified self-efficacy as a mechanism for practice change, but we should further understand which components in the CME influenced self-efficacy.

Conclusion and Significance

Diagnostic delays contribute to suboptimal patient outcomes in AD. By using a matched case-control design and assessing both immediate educational outcomes and subsequent changes in diagnostic behavior, this study provides evidence for the potential of CME as a tool to increase AD diagnosis. This web-based CME intervention increased participant likelihood of diagnosing AD, led to a greater number of new AD diagnoses than the control group, and fostered a positive relationship between postintervention confidence and diagnosis rates. Building self-efficacy should be a key objective in education interventions with practice-changing potential, alongside improving, reinforcing, and validating existing knowledge. Overall, this study shows the power of real-world data in demonstrating the impact of CME on clinical behavior and offers a first step in identifying CME's impact on dementia care. We are currently conducting a second phase of this initiative. Future directions could include a breakdown of CME engagement levels and learning outcomes by specialty to clarify which provider groups benefit most from this intervention.

Acknowledgments

We thank Garimesh Kumar for his work on pulling and analyzing data; to continuing medical education content faculty Charles Vega, MD; SB; Frances McFarland, MA, PhD; and Sharon Cohen, MD; and Alexandra Howson, PhD, for written contributions. The study was funded by an Independent Medical Education Grant from Lilly.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to Medscape member privacy. Access to the data is restricted to Medscape employees who have permission. Statistical code and study protocol are available from the corresponding author on reasonable request.

Conflicts of Interest

TF and KL are employees of Medscape, LLC. SB receives funding from the Centers for Disease Control, National Institute on Aging, and National Institute of Neurological Disease and Stroke; honoraria as deputy editor of the *Journal of the American Geriatrics Society*; consulting fees for service on clinical and scientific advisory boards for Biogen, Eisai, Novo Nordisk, Abbvie, Lilly, and Linus Health, and as a speaker and content consultant from Medscape.

Multimedia Appendix 1
Codes used for the study.
[DOCX File, 13 KB - mededu_v11i1e72000_app1.docx]

References

1. Alzheimer's Association. 2022 Alzheimer's disease facts and figures. Alzheimer's & Dementia 2022 Apr;18(4):700-789. [doi: 10.1002/alz.12638]
2. Dementia. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/dementia [accessed 2024-10-28]
3. Mintun MA, Lo AC, Duggan Evans C, et al. Donanemab in early Alzheimer's disease. N Engl J Med 2021 May 6;384(18):1691-1704. [doi: 10.1056/NEJMoa2100708] [Medline: 33720637]
4. van Dyck CH, Swanson CJ, Aisen P, et al. Lecanemab in early Alzheimer's disease. N Engl J Med 2023 Jan 5;388(1):9-21. [doi: 10.1056/NEJMoa2212948] [Medline: 36449413]
5. Arvanitakis Z, Shah RC, Bennett DA. Diagnosis and management of dementia: review. JAMA 2019 Oct 22;322(16):1589-1599. [doi: 10.1001/jama.2019.4782] [Medline: 31638686]
6. Sluder KM. Acknowledging disparities in dementia care for increasingly diverse ethnoracial patient populations. Fed Pract 2020 Feb;37(2):69-71. [Medline: 32269468]
7. Judge D, Roberts J, Khandker RK, Ambegaonkar B, Black CM. Physician practice patterns associated with diagnostic evaluation of patients with suspected mild cognitive impairment and Alzheimer's disease. Int J Alzheimers Dis 2019;2019(4942562):4942562. [doi: 10.1155/2019/4942562] [Medline: 30937189]
8. Podhorna J, Winter N, Zoebelein H, Perkins T, Walda S. Alzheimer's diagnosis: real-world physician behavior across countries. Adv Ther 2020 Feb;37(2):883-893. [doi: 10.1007/s12325-019-01212-0] [Medline: 31933051]
9. Bouldin E. BRFSS data on cognitive decline and caregiving: national and state implications. Innov Aging 2024 Dec 31;8(Supplement_1):477-477. [doi: 10.1093/geroni/igae098.1556]
10. Berenbaum R, Dresner J, Maaravi Y, Erlich B, Pivko N, Tziraki C. Translating knowledge into practice at the local level: evaluation of a pilot CME for primary care physicians on dementia early diagnosis and management. Int Psychogeriatr 2020 Dec;32(12):1469-1470. [doi: 10.1017/S1041610219000097] [Medline: 30782229]
11. Jack CR Jr, Andrews JS, Beach TG, et al. Revised criteria for diagnosis and staging of Alzheimer's disease: Alzheimer's Association Workgroup. Alzheimers Dement 2024 Aug;20(8):5143-5169. [doi: 10.1002/alz.13859] [Medline: 38934362]
12. Bernstein A, Rogers KM, Possin KL, et al. Dementia assessment and management in primary care settings: a survey of current provider practices in the United States. BMC Health Serv Res 2019 Nov 29;19(1):919. [doi: 10.1186/s12913-019-4603-2] [Medline: 31783848]
13. Pais M, Martinez L, Ribeiro O, et al. Early diagnosis and treatment of Alzheimer's disease: new definitions and challenges. Braz J Psychiatry 2020 Aug;42(4):431-441. [doi: 10.1590/1516-4446-2019-0735] [Medline: 31994640]
14. Prins A, Hemke F, Pols J, Moll van Charante EP. Diagnosing dementia in Dutch general practice: a qualitative study of GPs' practices and views. Br J Gen Pract 2016 Jun;66(647):e416-e422. [doi: 10.3399/bjgp16X685237] [Medline: 27114209]
15. Lathren CR, Sloane PD, Hoyle JD, Zimmerman S, Kaufer DI. Improving dementia diagnosis and management in primary care: a cohort study of the impact of a training and support program on physician competency, practice patterns, and community linkages. BMC Geriatr 2013 Dec 10;13(134):24325194. [doi: 10.1186/1471-2318-13-134] [Medline: 24325194]
16. Schütze H, Shell A, Brodaty H. Development, implementation and evaluation of Australia's first national continuing medical education program for the timely diagnosis and management of dementia in general practice. BMC Med Educ 2018 Aug 10;18(1):194. [doi: 10.1186/s12909-018-1295-y] [Medline: 30097036]
17. Casey AN, Islam MM, Schütze H, et al. GP awareness, practice, knowledge and confidence: evaluation of the first nation-wide dementia-focused continuing medical education program in Australia. BMC Fam Pract 2020 Jun 10;21(1):104. [doi: 10.1186/s12875-020-01178-x] [Medline: 32522153]

18. Moehead A, DeSouza K, Walsh K, Pit SW. A web-based dementia education program and its application to an Australian web-based dementia care competency and training network: integrative systematic review. J Med Internet Res 2020 Jan 22;22(1):e16808. [doi: 10.2196/16808] [Medline: 32012077]
19. Bandura A. Human agency in social cognitive theory. Am Psychol 1989 Sep;44(9):1175-1184. [doi: 10.1037/0003-066x.44.9.1175] [Medline: 2782727]
20. Lucero KS, Chen P. What do reinforcement and confidence have to do with it? A systematic pathway analysis of knowledge, competence, confidence, and intention to change. J Eur CME 2020 Oct 12;9(1):1834759. [doi: 10.1080/21614083.2020.1834759] [Medline: 33133769]
21. Spyropoulos J, Boutsalis G, Lucero K, Waskelo J, Wilson K, Anderson DR. Improving appropriate use of omega-3 fatty acids for patients with dyslipidemia: effect of online CME. Crit Pathw Cardiol 2021 Dec 1;20(4):208-212. [doi: 10.1097/HPC.0000000000000265] [Medline: 34431820]
22. Lucero KS, Larkin A, Zakharkin S, Wysham C, Anderson J. The impact of web-based continuing medical education using patient simulation on real-world treatment selection in type 2 diabetes: retrospective case-control analysis. JMIR Med Educ 2023 Aug 29;9:e48586. [doi: 10.2196/48586] [Medline: 37642994]
23. Lucero KS, Moore DE Jr. A systematic investigation of assessment scores, self-efficacy, and clinical practice: are they related? J CME 2024;13(1):2420373. [doi: 10.1080/28338073.2024.2420373] [Medline: 39498264]
24. Moore DE Jr, Green JS, Gallis HA. Achieving desired results and improved outcomes: integrating planning and assessment throughout learning activities. J Contin Educ Health Prof 2009;29(1):1-15. [doi: 10.1002/chp.20001] [Medline: 19288562]
25. Best practices in the delivery of care for Alzheimer's disease: from primary care to neurology. Medscape. URL: https://www.medscape.org/viewarticle/980622 [accessed 2025-05-09]
26. Waiting room for patients with cognitive impairment. Medscape. URL: https://www.medscape.org/sites/pwr/cognitive-impairment [accessed 2025-05-09]
27. 45 CFR 46.104 -- exempt research. Code of Federal Regulations. URL: https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46/subpart-A/section-46.104 [accessed 2025-05-13]
28. Lucero KS, Williams B, Moore DE Jr. The emerging role of reinforcement in the clinician's path from continuing education to practice. J Contin Educ Health Prof 2023 Nov 14;44(2):143-146. [doi: 10.1097/CEH.0000000000000541] [Medline: 37962911]

Abbreviations

AD: Alzheimer disease
AMA PRA: American Medical Association Physician Recognition Award
CME: continuing medical education
HCP: health care provider
***ICD-10*:** *International Statistical Classification of Diseases and Related Health Problems, Tenth Revision*
MCI: mild cognitive impairment
NPI: National Provider Identification
OR: odds ratio
PCP: primary care physician
RWO: real-world outcome

Edited by T Gladman; submitted 31.01.25; peer-reviewed by C Br, MJ Ebadi, RS Gomaa Mahmoud, SM Shaffi; revised version received 03.04.25; accepted 28.04.25; published 22.05.25.

Please cite as:
Lucero K, Finnegan T, Borson S
Using Web-Based Continuing Education to Improve New Diagnoses of Alzheimer Disease in Claims Data: Retrospective Case-Control Study
JMIR Med Educ 2025;11:e72000
URL: https://mededu.jmir.org/2025/1/e72000
doi:10.2196/72000

Original Paper

Exploring Health Care Professionals' Perspectives on Education, Awareness, and Preferences for Digital Educational Resources to Support Transgender, Nonbinary, and Intersex Care: Interview Study

Sravya Katta[1], MSc; Nadia Davoody[1], MSc, PhD

Health Informatics Centre, Department of Learning, Informatics, Management, and Ethics, Karolinska Institutet, Stockholm, Sweden

Corresponding Author:
Nadia Davoody, MSc, PhD
Health Informatics Centre
Department of Learning, Informatics, Management, and Ethics
Karolinska Institutet
Tomtebodavägen 18 A
S-17177
Stockholm
Sweden
Phone: 46 (0)8 524 864
Email: nadia.davoody@ki.se

Abstract

Background: Health care professionals often face challenges in providing affirming and culturally competent care to transgender, nonbinary, and intersex (TNBI) patients due to a lack of understanding and training in TNBI health care. This gap highlights the opportunity for tailored educational resources to enhance health care professionals' interactions with TNBI individuals.

Objective: This study aimed to explore health care professionals' perspectives on education and awareness of health issues related to TNBI individuals. Specifically, it aimed to identify their needs, challenges, and preferences in accessing and using digital educational resources to enhance their knowledge and competence in providing inclusive and effective care for this population.

Methods: A qualitative research approach was used in this study. In total, 15 health care professionals were recruited via convenience sampling to participate in semistructured interviews. Thematic analysis was applied to identify recurring codes and themes.

Results: The study identified several themes and subthemes related to gender diversity awareness, inclusive communication and understanding the needs of TNBI individuals, societal and structural challenges, regulatory gaps in training and support infrastructure, education and training needs for health care professionals on TNBI care, educational resources and training tools for TNBI care, challenges and design considerations for eHealth tools integrations, and evaluating eHealth impact. Participants identified communication barriers, the need for health care providers to use inclusive language, and gaps in both health care system infrastructure and specialized training for gender-affirming care. In addition, participants expressed a need for comprehensive education on transgender and nonbinary health issues, resources for mental health professionals, user-friendly design, and accessibility features in eHealth tools.

Conclusions: The study revealed substantial deficiencies in health care professionals' knowledge of gender diversity, cultural competency, and the importance of inclusive communication. Addressing the identified barriers and challenges through targeted interventions, such as providing training and support for health care professionals, investing in user-friendly design and data security, and promoting cultural competence in TNBI health care, is essential. Despite integration challenges, eHealth tools have the potential to improve patient–health care professional relationships and access to care.

(JMIR Med Educ 2025;11:e67993) doi:10.2196/67993

KEYWORDS

health care professionals; transgender, nonbinary, and intersex; communication challenges; systematic barriers; information and communication technology

Introduction

Background

The term transgender refers to individuals whose gender identity or expression differs from the sex they were assigned at birth [1]. The concept has expanded over time to include a wide range of gender identities, such as transmen, transwomen, nonbinary individuals, and those who are gender nonconforming [2,3]. Nonbinary individuals may not exclusively identify as male or female. Their gender identity can be fluid, agender, or fall outside the binary spectrum [1,4]. The transgender community is highly diverse, and the understanding of transgender identity varies across different cultures [2]. Intersex individuals are those whose physical sex characteristics do not conform to the traditional binary classification of bodies as strictly male or female [4]. Previous epidemiological and clinic-based investigations have suggested that approximately 0.1% to 2% of the population identifies as transgender or with other noncisgender identities [5-7].

The literature reveals that transgender, nonbinary, and intersex (TNBI) individuals experience disproportionate levels of human rights violations and adverse health outcomes, largely attributed to intersecting forms of social marginalization and legal exclusion. Particularly transgender individuals, especially those from minority ethnic groups, are disproportionately impacted by gender-based hate crimes [8]. In addition, TNBI individuals face multiple challenges in accessing adequate health care services [9-12]. The challenges faced by TNBI individuals in accessing health care services are often rooted in systemic barriers that perpetuate stigma, discrimination, and lack of understanding. These systemic barriers contribute to disparities in health care access, quality, and outcomes for these individuals [9-12]. These barriers may include policies that fail to recognize their gender identity and social factors such as discrimination and stigma from health care providers [9-13]. TNBI individuals often face a range of health disparities, largely driven by a lack of awareness regarding their unique health needs. These disparities manifest in higher rates of mental health issues, such as depression and anxiety, as well as increased risks of sexually transmitted infections, including HIV, and other diseases such as cancer, smoking-related conditions, and cardiovascular disease [14,15]. The elevated risk of acquiring sexually transmitted infections among TNBI individuals is influenced by factors such as limited access to comprehensive sexual health education, prevention measures, and adequate preventive care [16]. Furthermore, TNBI individuals often face significant barriers to accessing gender-affirming health care, which can lead to poorer overall health outcomes [13].

In addition to these health care challenges, TNBI individuals frequently experience widespread stigma, discrimination, and prejudice in various facets of life, including employment, education, housing, and health care. This systemic stigma surrounding gender diversity and nonconformity creates a hostile environment for TNBI individuals even within health care systems [6,17-20]. In a sexual health seminar held in Minnesota, a sample of 181 transgender participants revealed that 66% reported experiencing discrimination based on their gender identity or presentation [21]. Consequently, the denial of essential health care services can intensify feelings of dysphoria and distress among transgender individuals. While TNBI individuals face significant challenges in accessing care, health care professionals also encounter numerous obstacles that hinder their ability to provide effective support and services to these individuals.

Challenges Faced by Health Care Professionals

Health care professionals play a crucial role in caring for patients, acting as key facilitators of essential health care services and resources. This highlights the responsibility of health care providers to promote the health and overall welfare of the populations including TNBI [22]. However, health care professionals face challenges in treating and communicating with TNBI individuals, as many of them receive minimal or no training during their medical education and professional development regarding hormone therapy, gender-affirming surgeries, and mental health support—key aspects that are aimed at aligning an individual's physical appearance and gender identity with their affirmed gender during medical education and professional development [10,23].

In addition, health care professionals often encounter challenges when interacting with TNBI individuals, including issues related to cultural competency, communication barriers, and a lack of knowledge about appropriate care practices [24]. These challenges can result in disparities in health care access and quality for TNBI individuals, leading to negative health outcomes and experiences of discrimination [25]. Without adequate training, health care professionals may lack the necessary knowledge and skills to provide competent and affirming care to TNBI patients. This can result in misdiagnosis and inappropriate treatment for patients [26]. Moreover, health care professionals who are unfamiliar with the specific health needs of TNBI individuals may inadvertently overlook important aspects of care and may fail to provide appropriate interventions. Increasing awareness and understanding of TNBI health issues through education, professional development, and exposure to diverse patient populations can help health care professionals better meet the needs of their TNBI patients [27].

In this study, we focused on 4 countries—Sweden, India, the United Kingdom, and the United States—due to their diverse sociocultural context, legal frameworks, and health care systems. This selection provides valuable insights into the varying approaches to TNBI health care and medical education within each country, which allows for a broad examination of how medical education on TNBI individuals can be improved globally. Sweden's progressive health care policies and a strong emphasis on human rights, including TNBI rights, make it an ideal setting to explore advanced practices and identify areas

for improvement. India, with a complex sociocultural landscape and a vast and diverse population, presents unique challenges and opportunities in providing health care to TNBI individuals. Understanding health care professionals' perspectives in India can reveal the specific needs and barriers faced by TNBI individuals in a resource-limited setting. The United Kingdom has recently experienced significant sociopolitical changes affecting health care policies for TNBI individuals. The United States' diverse health care environment regarding TNBI care, with some states enacting progressive policies and others imposing restrictions on gender-affirming treatments, offers insights into the challenges and successes in providing care to TNBI individuals in such varied regulatory settings. By understanding these varied contexts, we aimed to identify gaps in education and awareness as well as potential best practices to address health care professionals' needs, challenges, and preferences in accessing and using educational resources. This insight will help enhance their knowledge and competence in providing inclusive and effective care for TNBI individuals. The medical education systems in each of these countries present unique structures but share a common challenge: a lack of formal education on TNBI health care needs. In Sweden, despite its progressive stance on gender equality, medical curricula still rarely address the specific health care concerns of gender-diverse individuals. India's medical education, influenced by traditional values and diverse cultural perspectives, similarly lacks comprehensive training on gender diversity, despite growing awareness of TNBI rights [28]. The United Kingdom and the United States have made strides in addressing health care inequalities, yet many health care professionals report limited exposure to TNBI health topics during their training [29,30].

The Importance of Educational Tools and the Role of Information and Communication Technology

Lack of necessary training can result in misunderstandings, misgendering, and insensitive or inappropriate interactions that undermine trust and rapport between patients and health care professionals [31]. To address these challenges, it is essential to explore and develop tailored educational resources and interventions that provide health care professionals with the knowledge and skills needed to interact effectively with TNBI individuals [32]. Promoting equitable health care access for all individuals requires innovative solutions that empower health care professionals to provide supportive and inclusive care to TNBI patients [32-36]. Incorporating information and communication technology (ICT) into training and education has proven highly beneficial, as it enhances learning opportunities, improves communication, and increases accessibility to educational resources [37]. While there are other strategies for training and education, ICT offers unique advantages such as personalized learning, interactive content, and the ability to reach a wider audience [38]. These benefits make ICT a more effective approach for improving various aspects of education and training. ICT plays a transforming role in health education by providing innovative and accessible solutions to train health care professionals effectively. It also enhances learning flexibility, promotes collaborative opportunities, and ensures scalability to meet diverse educational needs in health care settings [39].

In health care, ICT can be leveraged through digital tools to support training by providing diverse and interactive educational resources, facilitating remote learning, and enabling real-time access to up-to-date medical information and best practices. The World Health Organization defines eHealth as the "cost-effective and secure use of information and communications technologies in support of health and health-related fields, including health-care services, health surveillance, health literature, and health education, knowledge and research" [40]. In this study, eHealth tools refer specifically to the digital health tools (eg, mobile apps and web-based platforms) used by health care professionals for health education purposes. By fostering inclusive practices, these tools can enhance patient trust, reduce discrimination, and ultimately lead to better health outcomes for these populations. Given the limited research on this topic for health care professionals [41], understanding health care professionals' needs, challenges, and preferences is vital for developing effective, targeted educational resources that promote more inclusive and effective care for TNBI individuals.

Aim of the Study

This study aimed to explore health care professionals' perspectives on education and awareness of health issues related to TNBI individuals. Specifically, it aimed to identify their needs, challenges, and preferences in accessing and using digital educational resources to enhance their knowledge and competence in providing inclusive and effective care for this population.

Methods

Study Design

A qualitative research approach was chosen to investigate health care professionals' perspectives on education, awareness, and preferences for digital educational resources to support TNBI care. This method aligns with the study's aim by comprehensively understanding their experiences, needs, and challenges in accessing and using educational resources. To further enrich this exploration, a sociotechnical framework [42] was applied, as it provides a structured perspective to examine how social, cultural, and technological factors intersect and influence the delivery of inclusive and effective care for TNBI populations. This framework has been applied to examine how health care professionals engage with TNBI individuals and the potential of eHealth educational tools to enhance these interactions. The strengths of qualitative research lie in the ability to gain profound insights into a problem or necessity by directly engaging with individuals and their contexts where the issue arises [43].

Study Setting and Participants

In total, 15 health care professionals with various health care backgrounds were recruited to participate in this study. The data were collected in Stockholm County, Sweden. While most participant interviews were held via Zoom (Zoom Video Communications), 2 were conducted in person at the participants' workplaces. The participants were selected through convenience sampling, primarily via social media platforms.

The inclusion criteria required participants to be health care professionals with experience in using digital tools in their practice, aged ≥18 years, and proficient in written and spoken English. To capture diverse perspectives, we included professionals across 9 disciplines (physiotherapy, dentistry, pediatrics, general practice, general surgery, gynecology, oncology, psychology, and cosmetic surgery), representing a broad spectrum of patient care. Recruiting participants from different disciplines and countries with varying acceptance, health care system diversity, legal recognition, and social and cultural attitudes toward TNBI individuals enabled us to gather varied insights on the challenges that health care professionals face in interacting with TNBI individuals. Each group of health care professionals in this study plays a crucial role in different aspects of health care, from initial assessments and referrals to specialized care, ongoing support, and mental health services. The inclusion of these varied perspectives was essential for capturing the complexity of care required for TNBI individuals. However, the diversity in the respondent pool also presents challenges, as it can make it more difficult to maintain focused discussions and reach consensus. The exclusion criteria were non–English-speaking health care professionals and individuals without any health care background. These criteria ensured that participants had relevant and thorough experiences, insights, and recommendations related to the research topic, thereby preserving the quality and validity of the study's findings.

In qualitative research, the number of participants is typically determined by reaching data saturation, meaning that further data collection does not yield new insights [44]. In this study, participants were interviewed until no new insights were generated from the interviews. The participants' characteristics are presented in Table 1. The participants had a mean age of 40 (SD 7.3) years and worked in various fields within the health care sector. The study included 4 participants from Sweden, 5 participants from India, and 3 participants each from the United States and the United Kingdom. The variation in participant numbers across countries was due to the use of convenience sampling and the differing availability of health care professionals in each discipline and location. Among the 15 participants, 8 (53%) were female, 6 (40%) were male, and 1 (6%) participant did not disclose their sex. In total, 6 (40%) out of 15 participants had no prior experience working with TNBI individuals.

Table 1. Participants' characteristics.

Participant ID	Age range (y)	Geographic location	Occupation	Experience in health care (y)	Experience working with TNBI[a] individuals
Participant 1	35-45	Sweden	Physiotherapist	5-10	No
Participant 2	30-40	Sweden	Dentist	1-5	No
Participant 3	30-40	Sweden	Pediatrician	5-10	No
Participant 4	30-40	Sweden	Dentist	1-5	Yes
Participant 5	30-40	India	General physician	5-10	Yes
Participant 6	40-50	India	General surgeon	10-15	Yes
Participant 7	50-60	India	Gynecologist	15-20	Yes
Participant 8	50-60	India	Oncologist	1-5	Yes
Participant 9	30-40	India	General practitioner	5-10	No
Participant 10	40-50	United Kingdom	General practitioner	5-10	Yes
Participant 11	30-40	United Kingdom	General surgeon	1-5	No
Participant 12	25-35	United Kingdom	Surgical intern	1-5	No
Participant 13	30-40	United States	Psychologist	5-10	Yes
Participant 14	35-45	United States	General practitioner	5-10	Yes
Participant 15	40-50	United States	Cosmetic surgeon	10-15	Yes

[a]TNBI: transgender, nonbinary, and intersex.

The selected professions were included because they are likely to engage TNBI individuals in their practice. This diversity allows us to capture a wide range of insights and experiences, which are crucial for understanding the multifaceted needs of these patients. Although 6 (40%) out of 15 participants had no prior experience working with TNBI individuals, their contributions were insightful and added significant value to the findings. By including a variety of roles, we aimed to identify common themes and differences across different medical specialties, which can inform more inclusive health care practices.

Data Collection

Overview

Data were collected through semistructured interviews. To formulate our interview script, we followed a systematic approach grounded in the sociotechnical framework by Sittig and Singh [42]. This framework ensures that we comprehensively address the intersection of social, cultural, and technical factors and their impact on providing inclusive and effective care for the TNBI population.

Development of the Interview Schedule

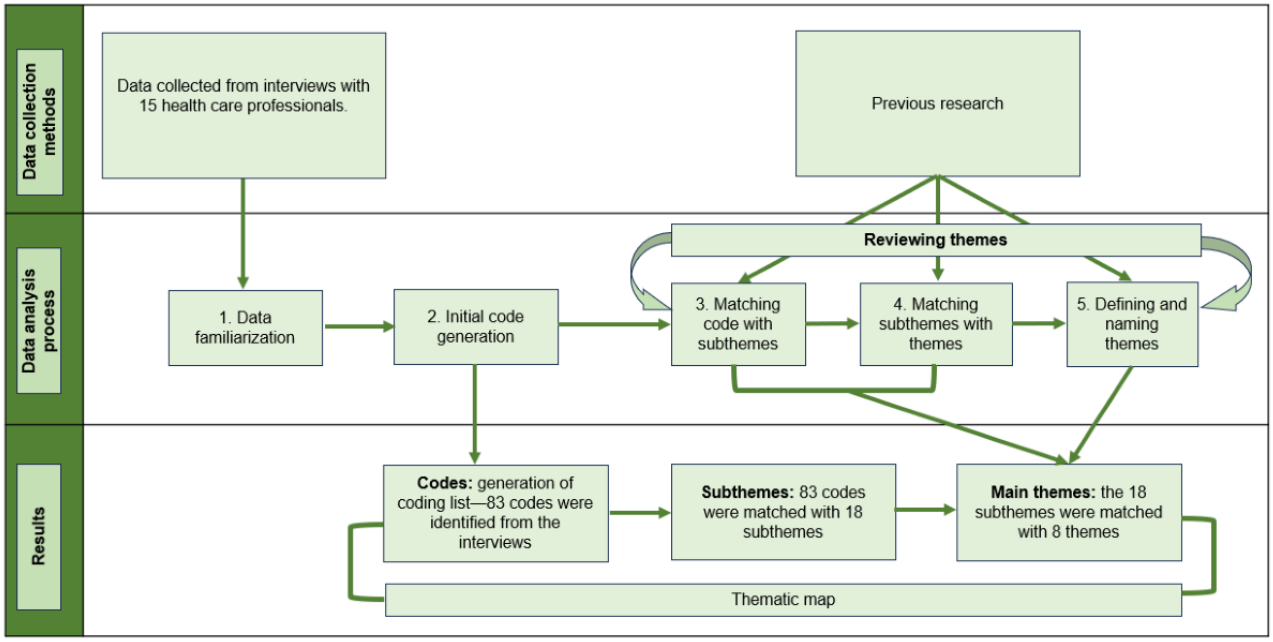
The interview schedule (Multimedia Appendix 1) was developed based on the 8 key dimensions of the sociotechnical framework: hardware and software; clinical content; human-computer interface; people; workflow and communication; internal policies, procedures, and culture; external rules and regulations; and measurement and monitoring [42]. We mapped the objectives of our study to the relevant dimensions of the framework to ensure comprehensive coverage: understanding current interactions and challenges (people and workflow and communication), identifying gaps in resources and training (internal policies, procedures, and culture and clinical content), exploring the potential of eHealth tools (hardware and software and human-computer interface), ensuring usability and integration (human-computer interface and communication), considering regulatory and organizational factors (external rules and regulations and internal policies, procedures, and culture), and evaluating effectiveness and feedback mechanisms (measurement and monitoring). We then continued with developing specific questions. For each dimension, we developed specific questions that align with our research objectives. The final interview script was designed to comprehensively address all aspects of the sociotechnical framework, ensuring the collection of detailed data on health care professionals' perspectives regarding education and awareness of health issues related to TNBI individuals and the potential of digital educational resources to enhance their care delivery. The interview guide was iteratively refined during early interviews to ensure its validity and relevance to the study's aim.

By using semistructured interviews, we could get a deeper understanding of what the participants had experienced during their clinical practice and their thoughts and feelings related to the need for digital educational resources. Semistructured interviews also allowed for follow-up questions to further deepen the understanding of their thoughts. Each interview lasted between 40 and 60 minutes. All the interviews were conducted in English.

Data Analysis

The interviews were audio recorded, transcribed, and analyzed using thematic analysis according to the guidelines by Braun and Clarke [45]. The audio recordings were transcribed verbatim using Microsoft Word and then reviewed for accuracy. The transcribed data were first thoroughly reviewed to gain a comprehensive understanding of the data. The next phase involved coding the text where codes are segments of the text that are linked by content and context, allowing a deeper exploration of the underlying themes and concepts. Descriptive coding was used to capture and summarize the main topics of the text. To build on this, pattern coding was applied to condense meaning units into overarching patterns, grouping the initial codes into broader themes. This approach facilitated the identification and understanding of larger patterns and relationships within the data [46]. SK conducted the initial coding by identifying meaning units, condensing them, and assigning relevant codes. Both authors reviewed the initial codes and grouped them into clusters reflecting emerging subthemes and then examined the relationships between the subthemes. They identified patterns and combined related subthemes into broader themes. The data analysis process was an iterative process in which discrepancies were resolved through several meetings and discussions. The authors analyzed the collected data and continuously revisited and refined the codes, subthemes, and themes as needed to ensure a comprehensive understanding of the data. The data analysis process is presented in Figure 1.

Figure 1. Data analysis process.



Ethical Considerations

This research was conducted in Sweden. According to the Swedish Ethical Review Act, this study does not require ethics approval as it does not handle sensitive personal information (as defined by the European General Data Protection Regulation). However, ethical requirements still apply. Participants were recruited from Sweden, India, the United States, and the United Kingdom. No sensitive personal information (eg, health status, political opinion, or racial or ethnic background) was collected. Prospective study participants were provided with comprehensive information regarding the study's objectives, methodologies, potential risks and benefits, and their right to withdraw at any time. This information was conveyed through both a written consent form, which participants were required to sign, as well as verbal explanations provided by the first author. To maintain confidentiality and privacy, the collected data were anonymized. In addition, any personal or sensitive information shared by participants was excluded from the study. The participants were guaranteed confidentiality and informed about how their data would be handled. Although the study was conducted in Sweden, ethical principles such as confidentiality and respect for participants' autonomy were upheld in line with international research standards, including those applicable in the United States (eg, Common Rule) [47], the United Kingdom (eg, Health Research Authority guidelines) [48], and India (eg, Indian Council of Medical Research guidelines) [49]. These measures ensured the ethical integrity of the research across all participant demographics.

Results

Overview

The analysis of the interviews resulted in 8 themes: gender diversity awareness, inclusive communication and understanding of the needs of TNBI individuals, societal structural challenges, regulatory gaps in training and support infrastructure, education and training needs for health care professionals on TNBI care, educational resources and training tools for TNBI care, challenges and design considerations for eHealth tools integration, and evaluating eHealth impact. In addition to the 8 themes, 18 subthemes and 83 codes were formulated. An overview of the sociotechnical aspects, themes, and subthemes is presented in Table 2.

In this study, the participants predominantly used the term *transgender* as an umbrella term to refer broadly to TNBI individuals.

Table 2. An overview of the subthemes and themes.

Sociotechnical aspects	Subthemes	Themes
People	• Limited understanding of gender diversity	Gender diversity awareness
Workflow and communication	• Acknowledgment of communication barriers and the need for inclusive language • Lack of understanding of TNBI[a] individuals' needs	Inclusive communication and understanding of the needs of TNBI individuals
Internal organizational policies, procedures, and culture	• Suppression of identity due to societal stigma, cultural norms, and societal pressures • Vulnerability arising from societal and political oppression • Limited research on TNBI health issues	Societal and structural challenges
External rules and regulations and pressures	• Lack of awareness among health care professionals and inadequate mental health support • Gaps in specialized training and guidelines for gender-affirming care • Weakness in health care infrastructure for TNBI individuals	Regulatory gaps in training and support infrastructure
Clinical content	• Inadequate training in cultural competency regarding gender diversity • Importance of education on TNBI health issues for health care professionals • Need for tailored resources and training modules designed for mental health professionals	Education and training needs for health care professionals on TNBI care
Hardware and software	• Interactive case studies and peer support forums • Comprehensive training modules workshops and web-based courses • Resources about gender-affirming therapy, trauma, and intersectionality	Educational resources and training tools for TNBI care
Human-computer interface	• Challenges in integrating eHealth tools into regular health care practice, including time constraints and cultural change • Emphasis on user-friendly design, accessibility features, and data security in health care tools	Challenges and design considerations for eHealth tools integration
System measurement and monitoring	• Expectations of improved patient–health care professional relationships using eHealth tools	Evaluating eHealth impact

[a]TNBI: transgender, nonbinary, and intersex.

Sociotechnical Aspect: People (Theme 1: Gender Diversity Awareness, Subtheme: Limited Understanding of Gender Diversity)

A hierarchical structure with the themes at the top, subthemes in the middle, and the corresponding codes at the base is illustrated in Figure 2.

A limited understanding of gender diversity may result in inadequate screening and assessment practices for TNBI patients' health care needs. Health care professionals discussed misunderstanding the unique health risks and concerns facing TNBI individuals, leading to delays in diagnosis, inappropriate treatment recommendations, or suboptimal care outcomes:

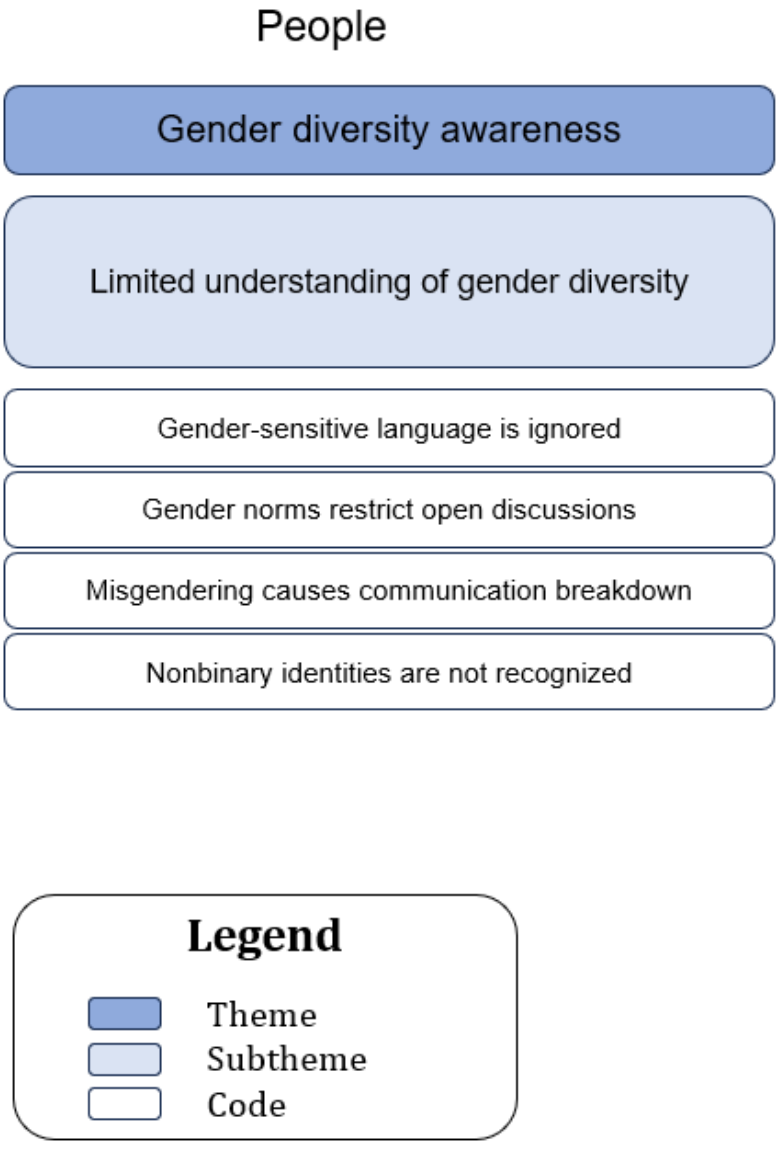
> *When I worked as an intern, we usually strengthened the biological sex or gender, ignoring the psychological stuff, because most of us in our education, are taught to learn the difference between male and female.* [Participant 10]

Some participants expressed that a limited understanding of gender diversity could create institutional barriers to accessing gender-affirming care, further exacerbating disparities in health care access and outcomes for TNBI individuals:

> *From my clinical work, outdated regulations often obstruct the care of transgender patients, a sign of the healthcare system's failure to fully grasp the nuances of gender diversity.* [Participant 11]

> *In my clinical experience, I've observed bureaucratic hurdles, such as outdated policies and procedures that fail to accommodate the unique needs of transgender individuals. These stemmed from a limited understanding of gender diversity.* [Participant 6]

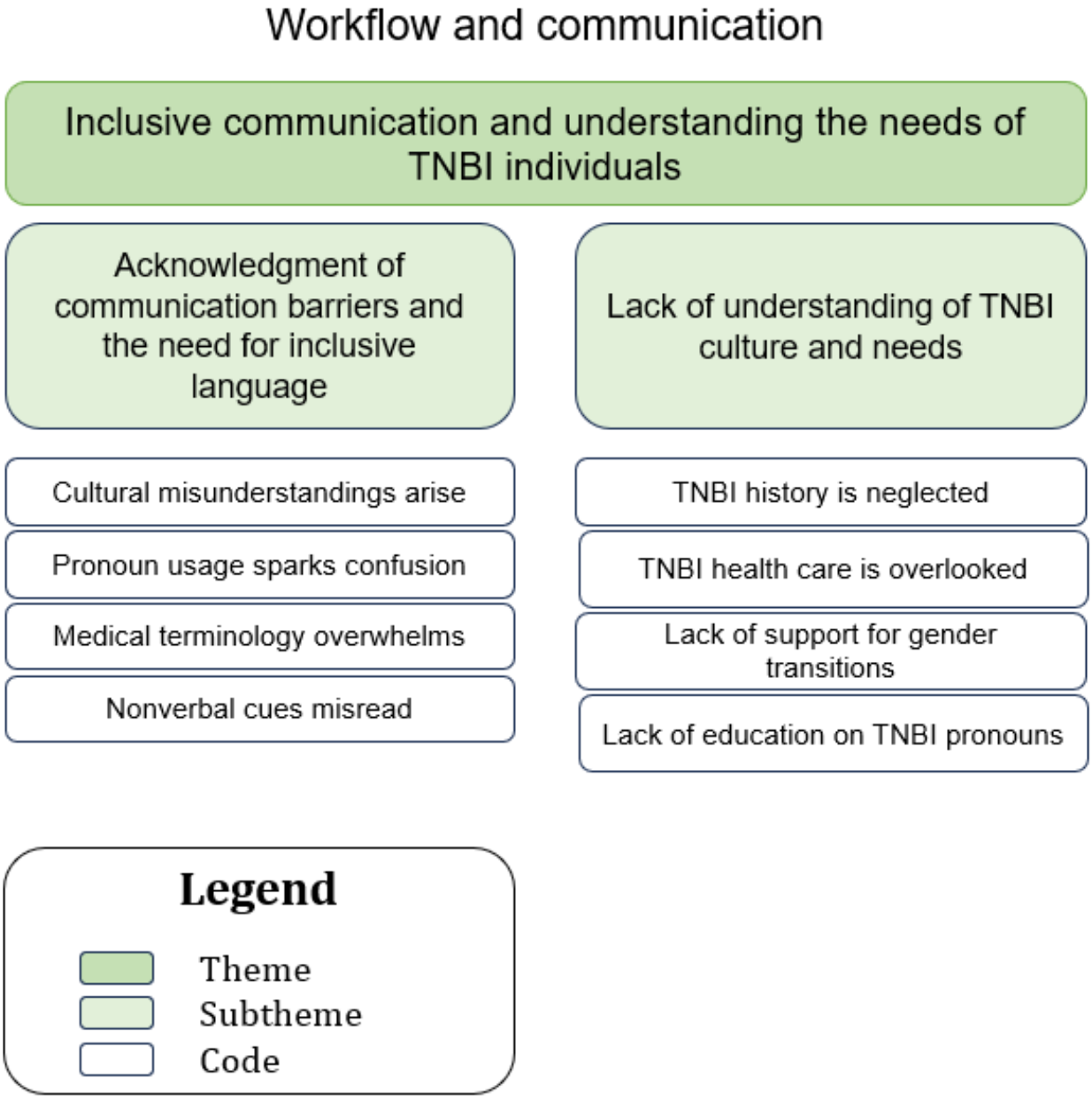
Figure 2. The relationship between the codes and subthemes for theme 1—gender diversity awareness.



Sociotechnical Aspect: Workflow and Communication (Theme 2: Inclusive Communication and Understanding the Needs of TNBI Individuals)

A hierarchical structure with the themes at the top, subthemes in the middle, and the corresponding codes at the base is illustrated in Figure 3.

Figure 3. The relationship between the codes and subthemes for theme 2—inclusive communication and understanding the needs of transgender, nonbinary, and intersex (TNBI) individuals.



Acknowledgment of Communication Barriers and the Need for Inclusive Language

Participants felt an existing lack of awareness about TNBI health issues, including appropriate language and communication strategies. They also stated that without adequate education on TNBI terminology and cultural competency, they may unintentionally use insensitive or outdated language, leading to misunderstandings and discomfort for TNBI patients:

> *I used inappropriate language for communication with transgender patients and they felt bad for that and some of them were even frustrated for not having adequate knowledge of terms that should be used.* [Participant 7]

> *I've unintentionally used terms that were not affirming, which caused discomfort for my transgender patients, and this has shown me the urgent need for proper education on inclusive language.* [Participant 5]

Most of the participants presented fear of inadvertently misgendering them. This fear of causing harm or disrespect can lead to hesitation or avoidance of discussions related to gender identity, which can hinder effective communication and rapport building with TNBI patients:

Sometimes I did avoid discussions regarding gender identity, as I was not sure about it, due to which rapport with the patients was not constructive. [Participant 4]

Some of the participants mentioned that there are limited resources and guidelines available to health care professionals on best practices for communication with TNBI patients. In the absence of clear guidance, they may struggle to navigate conversations about gender identity and may rely on personal assumptions or biases, which can contribute to communication barriers and misunderstandings:

There were moments when my judgment was clouded by my assumptions leading to uncomfortable situations. It is a mistake I have learned from. [Participant 15]

Struggled a lot and also felt embarrassed due to my personal assumptions that led to misunderstandings in a peculiar situation and never did it again. [Participant 6]

Lack of Understanding of TNBI Culture and Needs

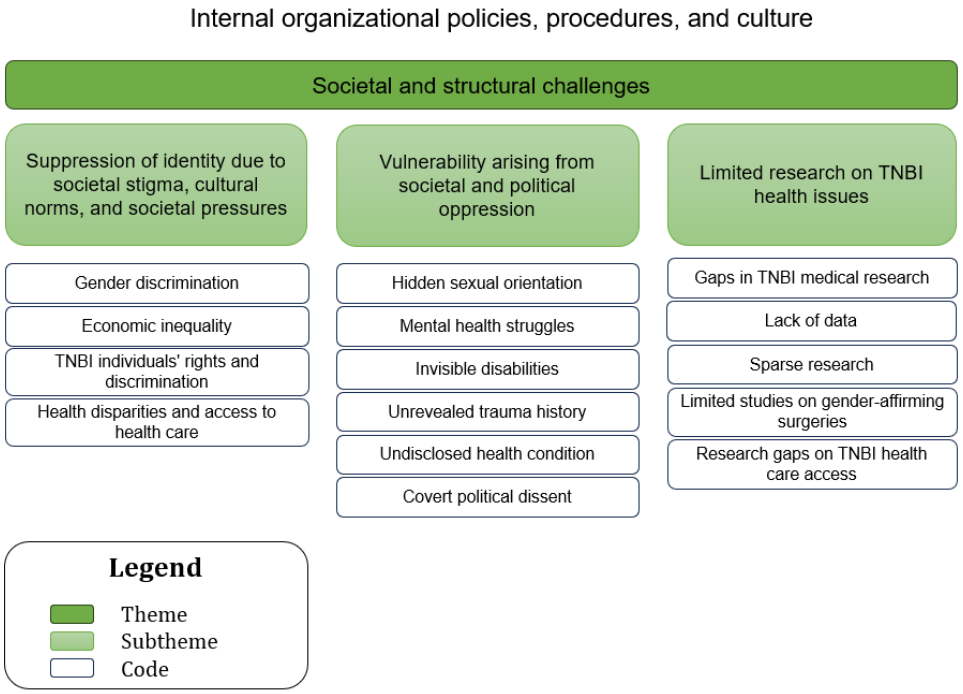
Participants also expressed that they struggled to establish trusting relationships due to a lack of understanding regarding their cultural identities and needs. Therefore, they observed a hindrance in disclosing patient's gender identity, expressing their health care concerns, and seeking support for their health and well-being:

In fact, I have no experience of working with transgender, hence building rapport with these patients is tough for me as I don't have a proper idea regarding their needs. [Participant 2]

Sociotechnical Aspect: Internal Organizational Policies, Procedures, and Culture (Theme 3: Societal and Structural Challenges)

A hierarchical structure with the themes at the top, subthemes in the middle, and the corresponding codes at the base is illustrated in Figure 4.

Figure 4. The relationship between the codes and subthemes for theme 3—societal and structural challenges. TNBI: transgender, nonbinary, and intersex.



Suppression of Identity Due to Societal Stigma, Cultural Norms, and Societal Pressures

Participants highlighted that experiences of discrimination, prejudice, and social rejection increased the vulnerability of TNBI individuals to mental health conditions, such as depression, anxiety, and posttraumatic stress disorder. They also believe that internalizing negative stereotypes and beliefs about their identity can further exacerbate mental health issues, causing feelings of shame, self-doubt, and identity concealment:

During my experiences in clinical practice, I have encountered transgender patients who have expressed fear, shame, and hesitation in disclosing their gender identity to healthcare providers. [Participant 10]

One of the participants said that traumatic experiences such as hate crimes, physical violence, or verbal abuse based on gender identity can lead to symptoms of posttraumatic stress disorder, including intrusive thoughts, hypervigilance, and avoidance behaviors:

They cannot say how they recognize themselves as transgender or non-binary in the workplace or even to their families. It's a kind of a secret for them. So, they are emotionally vulnerable and sensitive at the same time. [Participant 13]

Many transgender individuals feel they must conceal their identities both professionally and personally,

> *which places them under immense emotional strain and leaves them feeling isolated.* [Participant 2]

In addition to that, they also focused on the barriers to accessing health care faced by TNBI individuals because of societal stigma. Participants also revealed that past negative experiences or stories of discrimination within health care settings may lead to mistrust and reluctance to seek medical care.

Another point they mentioned is that the fear of being misgendered, invalidated, or subjected to invasive questioning can create significant barriers to accessing necessary health care services. In addition to that, internalized stigma and shame may also contribute to reluctance to seek health care services, as TNBI individuals may feel unworthy of receiving care or fear being perceived as *different*:

> *The first I think is the culture because trans people and also non-binary people, have their own culture, different than the majority culture and we, health care professionals didn't know that.* [Participant 5]

Some of the participants mentioned that using correct pronouns and respecting chosen names can help mitigate the effects of societal stigma and foster a sense of validation and belonging for TNBI patients:

> *It reminds me of a transgender patient, who did not behave like the gender that person looked like and that causes some confusion. So, it's necessary to ask them first how you recognize yourself and respect their social gender identification.* [Participant 5]

> *Gender identity is personal, and as clinicians, we should always lead with questions, not assumptions to ensure we are providing care that respects who they are.* [Participant 8]

Vulnerability Arising From Societal and Political Oppression

For TNBI individuals, intersectionality exacerbates vulnerability, particularly for those who belong to marginalized racial, ethnic, or socioeconomic groups. This means that TNBI individuals may face compounded discrimination and barriers to health care access due to overlapping forms of oppression:

> *I have seen many cases, who faced added discrimination to healthcare access due to political oppression...I believe that it's crucial to understand how multiple forms of discrimination intersect for transgender people, making access to healthcare even more challenging.* [Participant 6]

> *Political and social discrimination compounds the barriers transgender individuals face in accessing care, highlighting the need to address overlapping layers of bias in healthcare systems.* [Participant 11]

Limited Research on TNBI Health Issues

Participants said that they rely on evidence-based practices to guide their clinical decision-making and provide quality care to patients. The limited research on TNBI health issues means that there may be a lack of robust evidence to inform best practices in the diagnosis, treatment, and management of health conditions:

> *Without sufficient research, we are not only constrained in delivering optimal care, but it is also difficult to fully trust the treatments we prescribe to our patients.* [Participant 9]

> *Lack of research not only hampers our ability to deliver effective care but also undermines our confidence in the treatments and interventions we provide [...] As a result, we often find ourselves relying on anecdotal evidence, expert opinions, and extrapolations from related fields to inform our clinical decisions.* [Participant 4]

They also reported the lack of data on prevalence, risk factors, and outcomes of health conditions within these populations as a hindrance to their ability to assess and address their health care needs accurately. Without this information, it is challenging to determine the scope and magnitude of health disparities and to allocate resources effectively to address them:

> *Without accurate data, it's challenging to develop informed strategies for promoting the health and well-being of transgender and non-binary communities.* [Participant 3]

> *The absence of prevalence data makes it difficult to gauge the extent of health disparities within transgender and non-binary populations.* [Participant 13]

In addition, the lack of data on risk factors means that health care professionals may struggle to identify and mitigate factors that contribute to adverse health outcomes among TNBI individuals. Furthermore, the absence of data on outcomes of health conditions within TNBI populations hampers efforts to evaluate the effectiveness of interventions and treatments. Without data on treatment outcomes, health care professionals may be limited in their ability to tailor interventions to the unique needs of TNBI individuals and to optimize their health outcomes:

> *We struggle to allocate resources effectively and prioritize interventions without a clear understanding of the prevalence and severity of health conditions among transgender and non-binary individuals.* [Participant 5]

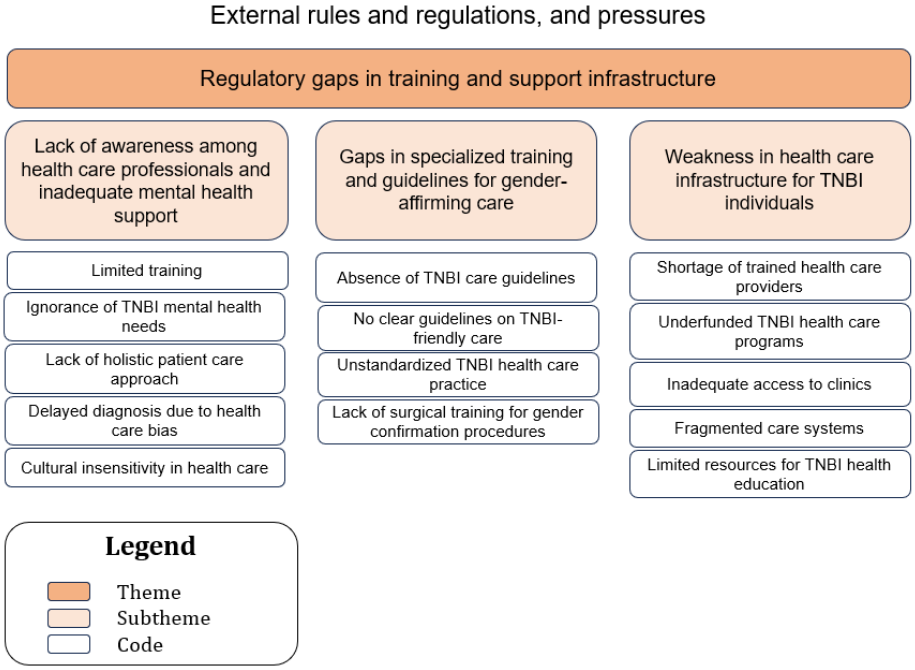
Participants also recognized the importance of investing in research initiatives focused on TNBI health to fill existing knowledge gaps and improve health care delivery:

> *It's necessary that we prioritize funding and resources for research aimed at filling existing knowledge gaps, addressing disparities, and promoting health equity among transgender and non-binary populations in transgender and non-binary healthcare.* [Participant 4]

Sociotechnical Aspect: External Rules and Regulations and Pressures (Theme 4: Regulatory Gaps in Training and Support Infrastructure)

A hierarchical structure with the themes at the top, subthemes in the middle, and the corresponding codes at the base is illustrated in Figure 5.

Figure 5. The relationship between the codes and subthemes for theme 4—regulatory gaps in training and support infrastructure. TNBI: transgender, nonbinary, and intersex.



Lack of Awareness Among Health Care Professionals and Inadequate Mental Health Support

Significant gaps discussed by the participants include a lack of awareness and understanding of TNBI health issues. They also expressed feelings of being frustrated by the systemic gaps in education and training on TNBI health issues within their profession. They were also concerned about the limited availability of culturally competent and affirming mental health services for TNBI patients, recognizing the detrimental impact of untreated mental health conditions on their overall well-being:

> *I feel ill-equipped to address their unique healthcare needs, leading to challenges in providing culturally competent and affirming care.* [Participant 3]

Gaps in Specialized Training and Guidelines for Gender-Affirming Care

The participants' identification of gaps in specialized training and guidelines for gender-affirming care highlights a critical challenge within health care systems. These gaps stem from limited education and training on transgender health issues and gender-affirming care during both formal education and professional training programs. They also expressed concerns about the lack of comprehensive education and training on transgender health topics throughout their academic and professional journeys:

> *I feel the lack of preparation and education on transgender health issues is a systemic issue that requires immediate attention and action from healthcare institutions and educational programs.* [Participant 3]

Many reported minimal exposure to transgender health issues and gender-affirming care protocols during their formal education, which left them feeling ill-prepared to provide culturally competent and affirming care to transgender patients.

In contrast, they expressed that they are encountering inconsistencies in clinical protocols, treatment approaches, and referral criteria for TNBI patients, resulting in variations in care quality and patient experiences due to the lack of standardized guidelines within health care institutions and disciplines.

Even though some of the participants were interested in specializing in transgender health, they said that they were facing challenges in accessing advanced training opportunities to develop expertise in this area, leading to a scarcity of trained specialists within the health care workforce:

> *Even with a strong interest in transgender healthcare, I'm struggling to access the specialized training necessary to develop my skills in this field.* [Participant 12]

> *Despite my interest in specializing in transgender health, I am encountering challenges in accessing advanced training opportunities.* [Participant 13]

Weaknesses in Health Care Infrastructure for TNBI Individuals

The participants also stated that they observed a lack of specialized health care facilities equipped to provide gender-affirming care for TNBI individuals. In addition, they also reported the absence of dedicated clinics or centers specializing in transgender health, which limits access to competent care and contributes to disparities in health outcomes:

> *There's a noticeable absence of dedicated clinics specializing in transgender health within our healthcare system, due to which patients are facing difficulties in receiving the affirming care they need.* [Participant 6]

> *The scarcity of transgender-focused healthcare facilities means many patients struggle to find affirming care, which compromises their overall well-being.* [Participant 4]

They have also mentioned the challenges they were facing in accessing gender-affirming treatments, such as hormone therapy and gender-affirming surgeries, for TNBI individuals, including a shortage of trained health care providers and long waiting times for appointments. Apart from the aforementioned challenges, they also recognized the need for expanded access to mental health support services for TNBI individuals:

> *The limited availability of trained providers is impeding my ability to provide timely and comprehensive gender-affirming care.* [Participant 4]

> *The shortage of providers skilled in transgender care significantly delays access to the affirming treatments many patients require.* [Participant 8]

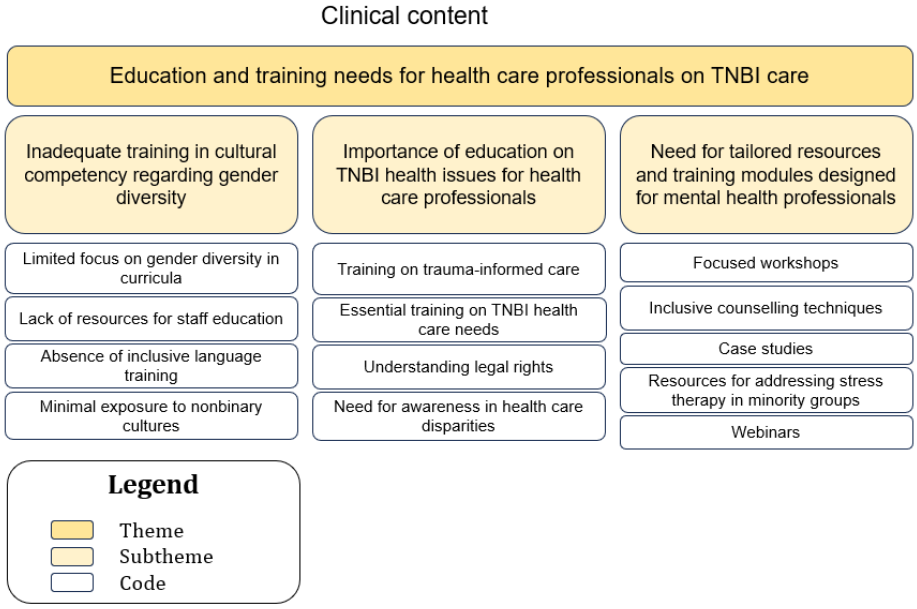
> *For transgender individuals, mental health support is essential for comprehensive care, yet it remains frustratingly inadequate in many healthcare settings.* [Participant 9]

> *Access to mental health support services is crucial for the holistic well-being of transgender individuals, yet it remains limited in many healthcare settings.* [Participant 6]

Sociotechnical Aspect: Clinical Content (Theme 5: Education and Training Needs for Health Care Professionals on TNBI Care)

A hierarchical structure with the themes at the top, subthemes in the middle, and the corresponding codes at the base is illustrated in Figure 6.

Figure 6. The relationship between the codes and subthemes for theme 5—education and training needs for health care professionals on transgender, nonbinary, and intersex (TNBI) care.



Inadequate Training in Cultural Competency Regarding Gender Diversity

Participants reported a lack of comprehensive instructions on gender diversity and cultural competency during their formal education and professional training. Many indicated that TNBI health topics were either inadequately covered or entirely omitted from their curriculum. Without proper education and training in this area, health care professionals find themselves inadequately prepared to navigate the complex landscape of gender identity and expression:

> *Without adequate training in transgender health, many of us enter practice unsure of how to approach the nuanced aspects of gender identity and provide affirming care to all patients.* [Participant 9]

> *The limited coverage of transgender health topics in our curriculum left us feeling ill-equipped to navigate the nuances of gender identity and expression in clinical practice.* [Participant 4]

> *Our training offered minimal focus on transgender health, leaving us underprepared to address the complexities of gender identity in practice.* [Participant 5]

Participants also expressed their struggles to provide patient-centered care that respects and affirms the diverse identities and needs of transgender patients, resulting in disparities in health care access, quality, and satisfaction:

> *I can say that I have experienced difficulties in tailoring care approaches to align with the diverse identities and needs of transgender patients. My*

> *attempts to provide affirming care to transgender patients have revealed gaps in my understanding and implementation of patient-centered principles.* [Participant 6]

Importance of Education on TNBI Health Issues for Health Care Professionals

Participants felt that education on TNBI health fosters the creation of inclusive health care environments where all patients feel respected, affirmed, and understood. Health care professionals who received training on transgender health issues are better equipped to provide culturally competent care, use affirming language, and create safe spaces for transgender patients to access health care without fear of discrimination or mistreatment:

> *I strongly believe education on transgender health is a crucial step towards building a healthcare system that is truly inclusive and affirming of all gender identities.* [Participant 5]

> *I assume by investing in education on transgender health, healthcare institutions can promote inclusivity and reduce disparities in healthcare access and outcomes for transgender individuals.* [Participant 7]

> *Investing in education about transgender health is a critical step toward fostering inclusivity and addressing disparities in care outcomes.* [Participant 9]

They also felt that education regarding TNBI health issues empowers them to recognize and address barriers, including delayed diagnosis and inappropriate treatment, ensuring that TNBI individuals receive timely, appropriate, and affirming health care services that meet their unique needs:

> *By educating ourselves on transgender health issues, we can break down barriers and create more inclusive healthcare environments and we can also work towards reducing disparities in healthcare access and outcomes for transgender individuals.* [Participant 10]

> *Through education on transgender health issues, we can eliminate obstacles to care and strive to reduce healthcare disparities for transgender individuals, ensuring better outcomes for all.* [Participant 12]

Participants were also interested in learning about the specific health care needs of TNBI patients, including hormone therapy, gender-affirming surgeries, preventive care, and mental health support, enabling them to deliver comprehensive and evidence-based care:

> *At least, I am eager to gain insights into the unique healthcare needs of transgender patients, including understanding nuances of hormone therapy and this interest in learning about the healthcare needs of transgender patients reflects our commitment to providing inclusive and affirming care within our practice.* [Participant 4]

Need for Tailored Resources and Training Modules Specifically Designed for Mental Health Professionals

The need for specialized training for mental health professionals to understand the unique mental health challenges faced by TNBI individuals has been observed by some participants, as these individuals are at increased risk of mental health disorders such as depression, anxiety, suicidality, and gender dysphoria due to societal stigma, discrimination, and identity-related stressors.

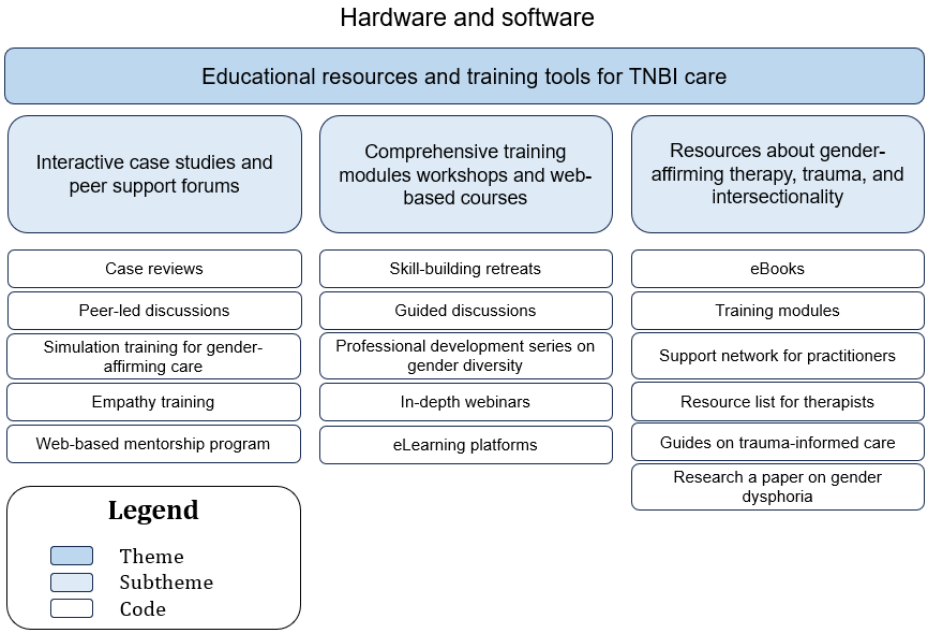
One of the participants, who was a mental health professional, was also expecting training in gender-affirming care principles to provide affirming and culturally competent mental health services to transgender and gender-diverse clients. The participant feels that tailored resources and training modules equip mental health professionals with strategies for creating affirming therapeutic environments and implementing gender-affirming interventions:

> *As per my observation, through education and training, mental health professionals can gain the knowledge and skills needed to provide competent and compassionate care to transgender and gender-diverse clients.* [Participant 13]

Sociotechnical Aspect: Hardware and Software (Theme 6: Educational Resources and Training Tools for TNBI Care)

A hierarchical structure with the themes at the top, subthemes in the middle, and the corresponding codes at the base is illustrated in Figure 7.

Figure 7. The relationship between the codes and subthemes for theme 6—educational resources and training tools for transgender, nonbinary, and intersex (TNBI) care.



Interactive Case Studies and Peer Support Forums

The participants generally perceived interactive case studies and peer support forums as valuable educational resources that promote active learning, cultural competency, professional networking, and skill development in transgender health.

Some of them felt that interactive case studies allow them to engage actively in the learning process by analyzing real-life scenarios, making decisions, and receiving immediate feedback:

> *My personal preference is to have interactive case studies, as they enable active participation and engagement in the learning process.* [Participant 7]

Others were interested in peer support forums that provide opportunities to learn from each other's experiences, perspectives, and insights, fostering a collaborative learning environment. They strongly believed that peer support forums facilitate knowledge sharing, discussion of clinical challenges, and exchange of best practices among peers, leading to enhanced learning outcomes and professional development.

They also felt that they could access these resources at their convenience to refresh their knowledge, stay updated on emerging practices, and enhance their clinical competencies in transgender health:

> *I agree that these resources encourage reflection, critical thinking, and dialogue among health care professionals, promoting continuous improvement in transgender healthcare delivery.* [Participant 3]

Comprehensive Training Modules, Workshops, and Web-Based Courses Focusing on Gender Diversity

To have an in-depth understanding of gender diversity, including the spectrum of gender identities and expressions, health care professionals were in favor of comprehensive training modules, workshops, and web-based courses.

They expect the training modules, workshops, and web-based courses to facilitate interdisciplinary collaboration among themselves from different specialties and disciplines, leading them to work collaboratively as part of multidisciplinary care teams to address the complex health care needs of transgender and gender-diverse patients, promoting coordinated and holistic care approaches:

> *The opportunity to exchange knowledge and best practices among health care professionals is vital for advancing our shared understanding of transgender health issues.* [Participant 11]

Some participants felt that training modules, workshops, and web-based courses focusing on gender diversity contribute to advancing health equity and social justice for transgender and nonbinary gender-diverse individuals. On the contrary, one of the participants expressed concern regarding time management:

> *In my opinion, it's essential to tackle issues regarding time management and resource allocation to ensure health care professionals can actively participate in gender diversity training and play their part in promoting health equity and social justice.* [Participant 2]

Resources About Gender-Affirming Therapy, Trauma, and Intersectionality

Health care professionals are expected to have resources that cover principles and practices of gender-affirming care, including approaches to hormone therapy, gender-affirming surgeries, and psychotherapy, enabling them to support transgender patients in aligning their bodies with their gender identities:

> *I expect to have access to resources on gender-affirming care will equip us with the knowledge and skills needed to navigate complex healthcare decisions and provide holistic support to*

> *transgender patients throughout their transition journey.* [Participant 3]

In addition, they were willing to have and learn about trauma-sensitive approaches to care delivery, recognizing the impact of past traumas on patients' mental health and well-being and creating safe and supportive environments for survivors of trauma to access health care services:

> *I am open to learning and implementing trauma-sensitive approaches into our practice, as I believe they are essential for providing patient-centered and holistic care to individuals affected by trauma.* [Participant 13]

> *As health care professionals, we recognize the importance of ongoing education and training in trauma-sensitive care to ensure that we can effectively meet the needs of patients who have experienced trauma.* [Participant 10]

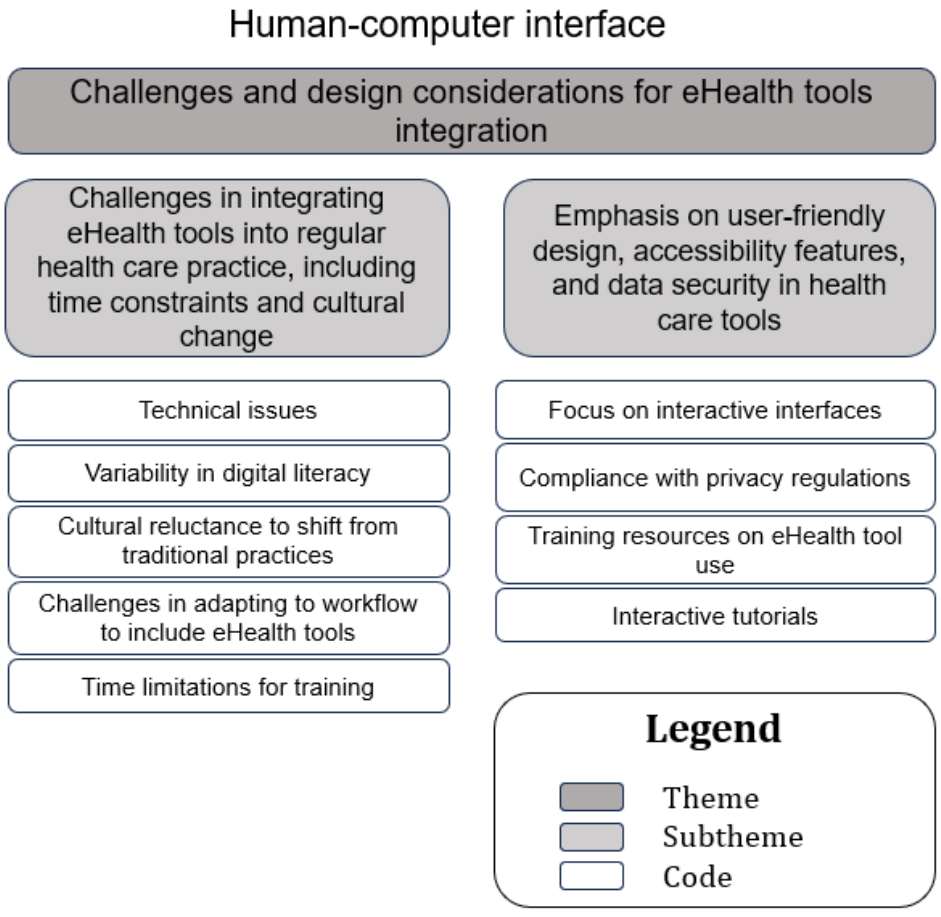
Health care professionals were interested in gaining an understanding of how multiple intersecting identities influence individuals' experiences of discrimination, access to health care, and health outcomes, enabling them to provide more nuanced and inclusive care to diverse transgender communities:

> *My priority is having education about intersectionality to work towards creating more equitable and inclusive healthcare systems that address the complex needs of all patients.* [Participant 13]

Sociotechnical Aspect: Human-Computer Interface (Theme 7: Challenges and Design Considerations for eHealth Tools Integration)

A hierarchical structure with the themes at the top, subthemes in the middle, and the corresponding codes at the base is illustrated in Figure 8.

Figure 8. The relationship between the codes and subthemes for theme 7—challenges and design considerations for eHealth tools integration.



Challenges in Integrating eHealth Tools Into Regular Health Care Practice, Including Time Constraints and Cultural Change

Several challenges were recognized in integrating eHealth tools into regular health care practice, including time constraints and cultural change. Participants said that they often face time constraints due to busy schedules, heavy workloads, and competing priorities in clinical practice.

They recognized that integrating eHealth tools could require additional time for training, learning new technologies, documentation, and troubleshooting technical issues. This demand may strain already limited time resources and disrupt workflow efficiency. Participants acknowledged that integrating eHealth tools requires a cultural shift in attitudes, behaviors, and practices within health care organizations and among health care professionals:

> *With our heavy workloads, it can be challenging to dedicate time to training and troubleshooting the technical issues that arise with eHealth tools.* [Participant 14]

> *Require additional time for documentation and adapting to new workflows, which can strain our already limited time resources.* [Participant 7]

They recognized that resistance to change, skepticism about technology, and concerns about the impact on traditional care delivery models may impede the adoption and acceptance of eHealth tools, necessitating cultural change initiatives, leadership support, and stakeholder engagement to foster a culture of innovation and digital transformation:

> *In the place I work, to overcome resistance to change, proactive efforts are required to educate, train, and support healthcare staff in adopting new technologies and workflows.* [Participant 7]

One participant anticipated that integrating eHealth tools into regular health care practice may disrupt existing workflows and processes, leading to initial challenges in adaptation and implementation. They expressed concerns about potential workflow inefficiencies, disruptions in patient flow, and coordination issues among health care team members, particularly during the transition phase, when adapting to new technologies and integrating them into clinical routines:

> *It would be difficult for most of us during the transition phase, as there may be a need for additional training, support, and resources to help us navigate the changes and overcome implementation barriers.* [Participant 13]

Participants expressed that they even encounter technical and logistical barriers, such as inadequate infrastructure, limited access to technology, interoperability challenges, and data security concerns, which hinder the seamless integration of eHealth tools into regular health care practice. They recognized the need for investment in IT infrastructure, resources for training and support, and adherence to regulatory requirements and standards to address these barriers effectively and ensure the successful implementation and use of eHealth tools:

> *Inadequate infrastructure and limited access to technology might hinder the seamless integration of eHealth tools into our practice.* [Participant 3]

They recognized that patients may have varying levels of comfort, literacy, and access to technology, which can influence their willingness and ability to engage with eHealth tools. Health care professionals emphasized the need for patient education, support, and empowerment to promote successful integration and use of eHealth tools in regular health care practice:

> *As health care professionals, we play a critical role in facilitating patient engagement and empowerment in the use of eHealth tools by providing guidance, encouragement, and ongoing support.* [Participant 15]

Emphasis on User-Friendly Design, Accessibility Features, and Data Security in eHealth Tools

Health care professionals prioritized user-friendly design in eHealth tools to ensure ease of use, intuitive navigation, and efficient workflow integration. They recognize that user-friendly interfaces enhance usability, minimize user errors, and promote acceptance and adoption among health care professionals, ultimately improving efficiency and productivity in clinical practice.

Health care professionals prioritized data security and privacy in eHealth tools to protect sensitive patient information, maintain confidentiality, and comply with regulatory requirements:

> *I prefer the user-friendly design of eHealth tools to ensure that we can easily navigate and utilize these technologies in our daily practice.* [Participant 4]

Health care professionals recognized that user-friendly design and accessibility features in eHealth tools contribute to patient engagement and empowerment. They believed that tools that are easy to use, accessible across devices, and customizable to individual preferences encourage active participation, self-management, and shared decision-making among patients, leading to improved health outcomes and patient satisfaction. Health care professionals' confidence and satisfaction with eHealth tools are influenced by the design, usability, and security features of these tools.

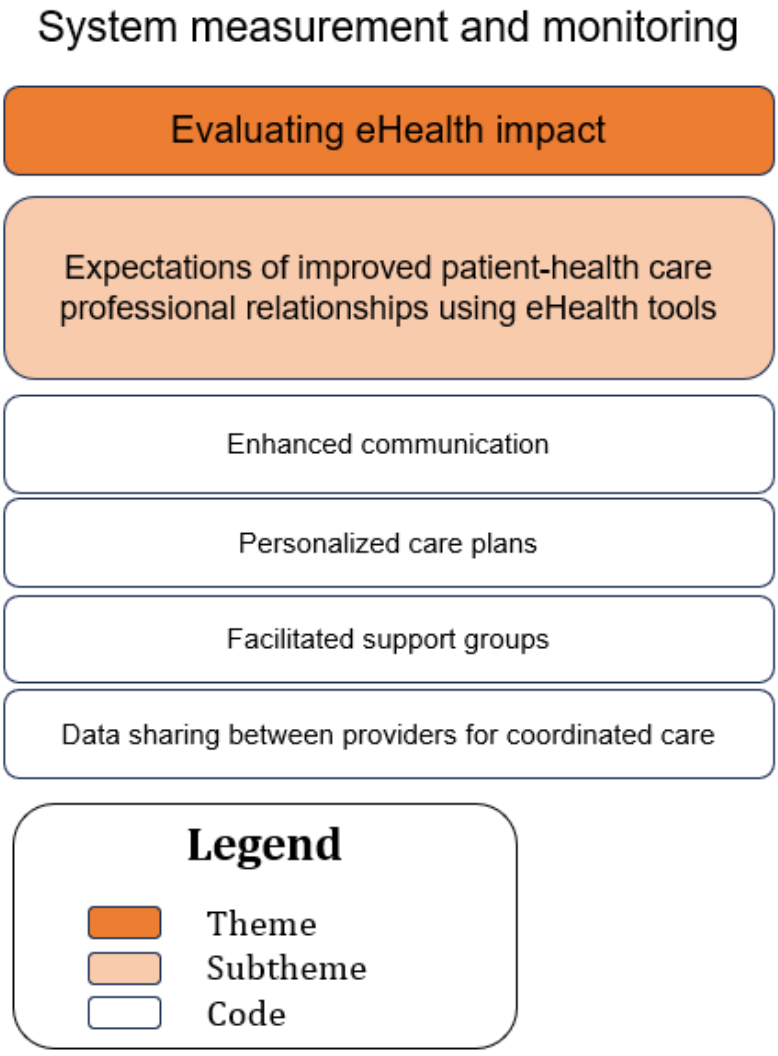
They expressed a preference for tools that are intuitive, reliable, and secure, enabling them to focus on patient care rather than technical frustrations or concerns about data integrity, thereby enhancing their overall professional experience and job satisfaction:

> *I believe that patient-centric design principles, such as intuitive interfaces and clear navigation, are essential for promoting patient autonomy and engagement in their healthcare.* [Participant 2]

Sociotechnical Aspect: System Measurement and Monitoring (Theme 8: Evaluating eHealth Impact, Subtheme: Expectations of Improved Patient–Health Care Professional Relationships Using eHealth Tools)

A hierarchical structure with the themes at the top, subthemes in the middle, and the corresponding codes at the base is illustrated in Figure 9.

Figure 9. The relationship between the codes and subthemes for theme 8—evaluating eHealth impact.



Health care professionals anticipated that eHealth tools would enhance communication, enabling more frequent and accessible contact with patients. This, in turn, could strengthen the patient–health care professional relationship by fostering trust, engagement, and continuity of care:

> *I still work in a healthcare setting, where there is no eHealth tools included in the regular practice. Through the integration of eHealth tools into our practice, we aim to create a healthcare environment that values transparency, accessibility, and patient engagement.* [Participant 7]

Discussion

Principal Findings and Comparison With Prior Work

Overview

The overall theme of this study is enhancing health care professionals' education and awareness of inclusive TNBI care. The analysis of health care professionals' interviews revealed significant insights into the challenges they face in providing care to TNBI individuals and the gaps in health care systems regarding these individuals' health. The analysis also highlighted the pivotal role of eHealth educational tools in addressing various challenges faced by health care professionals in providing inclusive and affirming care to transgender and gender-diverse individuals. Through its features and functionalities, an educational tool aims to bridge gaps in knowledge, communication, and access to resources, ultimately enhancing health care delivery and patient outcomes. In total, 8 themes were identified, highlighting the multifaceted nature of TNBI health care and the complex interplay among societal, cultural, regulations, institutional, technical, and individual factors.

Societal stigma and structural challenges emerged as a pervasive issue affecting the health and well-being of TNBI individuals. Participants highlighted the detrimental effects of discrimination, prejudice, and social rejection on mental health, emphasizing the need for culturally competent and affirming care to mitigate these challenges. In countries such as the United Kingdom and the United States, social acceptance and legal protections for TNBI individuals have progressed significantly, with antidiscrimination laws and health care rights actively supporting gender diversity. However, unlike in India, where acceptance remains complex due to strong cultural and religious norms, participants from these Western countries still reported that many patients experience fear, shame, and hesitation when disclosing their gender identity—whether to health care

professionals, in the workplace, or even within their families. The fear of being misgendered or invalidated, coupled with past traumas and societal pressures, creates significant barriers to accessing health care services, underscoring the importance of creating safe and supportive environments within health care settings.

Regarding external rules and regulatory gaps, although the United Kingdom and the United States are generally more open in terms of social and cultural norms, the recent sociopolitical shift in both countries has negatively affected TNBI care. For example, the United Kingdom has recently removed gender-affirming care with puberty blockers for those aged <18 years [50]. In the United States, >20 states have aimed to limit or ban access to gender-affirming care, especially for minors. These measures also included restrictions on access to mental health support and educational resources for TNBI individuals. Such policies reflect a growing sociopolitical climate that not only affects the patients directly but also impacts the engagement of health care professionals in providing gender-affirming care. These changes can lead to decreased accessibility to necessary treatments and negatively affect health outcomes for TNBI individuals [51].

Gender diversity awareness, inclusive communication, and understanding the needs of TNBI individuals have revealed a lack of awareness and understanding of TNBI health issues among health care professionals. Communication barriers, limited education on transgender and nonbinary terminology, and inadequate training in cultural competency contribute to misunderstandings and discomfort for these individuals. In addition, the limited understanding of gender diversity and institutional barriers further hinder effective care delivery, highlighting the need for comprehensive education and training on transgender and nonbinary health topics. Participants in this study noted that despite social acceptance, a lack of clear guidance and limited resources leaves them struggling with issues around gender identity. They expressed a fear of misgendering individuals and often avoided discussions on gender identity. Participants mentioned that their training has emphasized focusing on biological sex in treatment, often overlooking the psychological aspects of gender identity.

Various systemic issues, including the lack of awareness among health care professionals, limited mental health support, and gaps in research and specialized training for gender-affirming care, were identified as substantial barriers to providing comprehensive care for TNBI individuals. All participants emphasized the need for increased awareness, research, and resources to address these gaps and promote health equity and social justice for TNBI individuals.

Education and training needs for health care professionals on TNBI care highlighted the importance of comprehensive education on gender diversity and transgender and nonbinary health issues for health care professionals. Participants from all 4 countries expressed a desire for tailored resources and training modules specifically designed to enhance cultural competency and provide gender-affirming care. Moreover, the lack of specialized training and guidelines for gender-affirming care underscores the need for systemic changes within health care institutions to support the professional development of health care professionals in this area.

Educational resources and training tools for TNBI care identified interactive case studies, peer support forums, and comprehensive training modules as valuable tools for promoting active learning and skill development in transgender and nonbinary health. These resources facilitate knowledge sharing, collaboration, and reflection among health care professionals, ultimately enhancing the quality of care for these individuals.

Challenges and design considerations for eHealth tools integration and evaluating eHealth impact presented both opportunities and challenges in health care delivery. While eHealth tools have the potential to streamline communication and improve patient–health care professional relationships, participants identified various barriers, including time constraints, cultural change, and technical issues. Overcoming these barriers requires proactive efforts to address resistance to change, invest in IT infrastructure, and prioritize user-friendly design and data security.

Subanalysis of Participants With Experience in Providing Health Care to TNBI Individuals and Those Without Such Experience

Most of the participants without experience in working with TNBI individuals felt ill-equipped to address their unique health care needs. They struggled to build trusting relationships due to a limited understanding of cultural identities and health care requirements, citing that their education focused solely on binary male and female individuals' perspectives. Participants from Sweden noted that, despite living in an open society, gaps in transgender and nonbinary-specific training persist.

Both groups emphasized that delivering trustworthy care requires sufficient research and evidence-based practice to guide clinical decisions. There was no perceived difference in mental health support for TNBI individuals across Sweden, India, the United States, and the United Kingdom. All participants—regardless of prior experience—reported a lack of comprehensive training and instructions on TNBI health, with a shared belief that education is crucial for fostering an inclusive and affirming health care system. Participants across countries agreed that exchanging knowledge, ongoing training, and access to gender-affirming resources are essential for equipping health care professionals. While many were optimistic about eHealth tools improving communication with patients and continued education, they noted challenges such as additional training requirements, adapting to new workflows, and the importance of user-friendly designs.

The findings of this study align with and extend upon existing research in the field of health care education, particularly regarding the communication needs of health care professionals when interacting with TNBI individuals.

Hughto et al [33] highlighted the importance of addressing societal stigma and vulnerability in health care settings, particularly for TNBI individuals. Similarly, our research underscores the detrimental effects of societal stigma and cultural norms on TNBI mental health and well-being, as well

as the barriers to accessing health care services due to fear of discrimination and misgendering.

Furthermore, our study supports previous research conducted by Grant et al [22] on the challenges faced by health care professionals in delivering inclusive and affirming care to transgender patients. Consistent with these findings, our participants expressed concerns about communication barriers, limited understanding of gender diversity, and gaps in specialized training and guidelines for gender-affirming care. These challenges highlight the need for comprehensive education and training initiatives to enhance health care professionals' cultural competency and sensitivity to transgender and nonbinary health issues.

A scoping review of transgender health training in internal medicine and subspecialty residency programs identified significant gaps in medical education, emphasizing the need for clearly defined objectives to prepare health care professionals for competent and affirming transgender care [52]. Similarly, a systematic review of educational interventions for medical students and residents working with sexual and gender-minority patients demonstrated the effectiveness of structured programs in improving knowledge, attitudes, confidence, and skills, highlighting the importance of implementing comprehensive training to bridge these gaps [53]. Davidge-Pitts et al [54] studied the importance of comprehensive training and educational resources to address the gaps in health care professional's knowledge and skills related to TNBI health issues. Similarly, our study underscores the significance of tailored educational tools in enhancing health care professional's understanding and competency in providing gender-affirming care. In addition, the Transgender Education for Affirmative and Competent HIV and Healthcare Program further highlights the impact of structured educational initiatives in fostering gender-affirming knowledge, perceived competency, and inclusive practice behaviors among health care providers [55]. These findings align with our study's recommendation for interactive, solution-oriented tools to promote ongoing skill development and professional collaboration. In addition, research on barriers to transgender and gender-diverse care highlights several challenges, including the absence of clear guidelines; extended waiting times; a shortage of specialist centers; insufficient training in transgender and gender-diverse health; and technical, cultural, and social obstacles [26]. These findings align with and reinforce the results of our study.

Our study supports the notion that interactive case studies and peer support forums, integrated within an educational tool, promote active learning, cultural competency, and professional networking among health care professionals. These features facilitate ongoing learning and skill development, fostering a collaborative environment conducive to improving TNBI health care delivery. Moreover, the challenges identified in the integration and use of eHealth tools, as discussed in our study, echo the findings of previous research by Light et al [56]. Time constraints, cultural barriers, and technical issues have consistently been recognized as barriers to the adoption of digital health technologies in clinical practice by the participants.

In contrast to earlier studies conducted by Mansh et al [57], which primarily focused on identifying gaps and challenges in TNBI health care education, our study advances the field by proposing a solution-oriented approach. This study provides actionable recommendations for improving TNBI health care education and training initiatives. This shift toward solution-focused research aligns with the broader goal of enhancing health care delivery and patient outcomes through innovative educational interventions. The findings of this study contribute to the growing body of literature on educational interventions for health care professionals and their implications for TNBI health.

Strengths and Limitations of the Study

The main strength of this study is the use of a qualitative research approach, which allowed for an in-depth exploration of the challenges faced by health care professionals when communicating with TNBI individuals, facilitating a comprehensive understanding of the nuances and complexities surrounding this topic.

In addition, the study included participants from various health care backgrounds and geographic locations, enhancing the richness and diversity of perspectives gathered. This diversity contributed to a more comprehensive analysis of the educational eHealth tools' requirements and potential impact across different health care contexts. The diverse backgrounds of the participants offered a wide range of experiences and provided a more comprehensive understanding of the issues. While our study provides a broad overview of the needs across different medical professions, we recognize the importance of conducting more in-depth studies for each profession. Future research should focus on the specific needs and challenges faced by each medical specialty when working with TNBI individuals. As qualitative studies do not guarantee the generalizability of the results, we argue that the results of our study are transferable to other contexts.

We believe that the number of interviews conducted with health care professionals was sufficient to achieve data saturation across the overall sample. This indicates that we obtained a sufficient breadth and depth of information to address the research objectives. This ensured that a thorough exploration of the topic was achieved without the need for additional participants [37]. It is also important to note that due to the diversity of countries and health care professions in the sample, combined with the relatively low number of participants from each profession and country, data saturation specific to each subgroup may not have been fully achieved. This limitation could affect the transferability of our findings, particularly for specific roles or cultural contexts. Future research could benefit from a more focused examination of these subgroups to deepen the understanding of context-specific nuances. In addition, conducting additional interviews with health care professionals from settings not represented in this study could enrich findings within the subject.

Another limitation of the study is that the interview guide was not pilot-tested. However, the authors carefully designed the guide and conducted multiple meetings to review and refine it. In addition, during the interviews, we maintained a flexible and

adaptive approach, making minor adjustments needed to better capture health care professionals' experiences.

Conclusions

The study aimed to explore health care professionals' perspectives on education, awareness, and preferences for digital educational resources to support TNBI care. The results provided valuable insights into the barriers health care professionals encounter when providing care to TNBI. The study identified key gaps in health care professionals' understanding of gender diversity, cultural competency, and the need for inclusive communication. In addition, the study emphasized the importance of specialized training and the integration of user-friendly eHealth tools to improve the relationships between health care professionals with TNBI individuals.

eHealth tools play a significant role in enhancing patient–health care professional relationships, improving access to care, and promoting patient engagement in health care. Despite the challenges associated with their integration, health care professionals acknowledged their potential to facilitate more efficient, patient-centered care delivery.

Addressing the identified barriers and challenges through targeted interventions, such as providing training and support for health care professionals, investing in user-friendly design and data security, and promoting cultural competence in providing health care for TNBI individuals, is essential.

In conclusion, this study contributes to the growing literature on eHealth interventions in TNBI health care and sets the stage for future research and practice initiatives aimed at leveraging technology to improve health outcomes and reduce health disparities for these individuals.

Acknowledgments

The authors would like to thank all the study participants for their time and valuable contributions to this study. The authors received no specific funding for this work.

Generative artificial intelligence technologies, including Grammarly [58] and OpenAI's ChatGPT 3.5 [59], were used during manuscript preparation to improve grammatical accuracy and enhance clarity in the text. These tools were not used to refine or alter the original qualitative data or inform the thematic analysis. The qualitative data were analyzed in their original form, ensuring accuracy and consistency with participants' responses. The authors reviewed and edited all aspects related to grammatical accuracy and textual clarity to ensure alignment with the study's objectives and maintain content integrity, taking full responsibility for the content of the publication.

Authors' Contributions

Both SK and ND were involved in the study design. SK conducted the data collection and performed the initial data analysis, which was subsequently reviewed and refined by ND. Both the authors were actively involved in writing and reviewing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide.
[PDF File (Adobe PDF File), 72 KB - mededu_v11i1e67993_app1.pdf]

References

1. Gooren L. The biology of human psychosexual differentiation. Horm Behav 2006 Nov;50(4):589-601. [doi: 10.1016/j.yhbeh.2006.06.011] [Medline: 16870186]
2. Understanding transgender people, gender identity and gender expression. American Psychological Association. URL: https://www.apa.org/topics/lgbtq/transgender-people-gender-identity-gender-expression [accessed 2024-04-29]
3. Fiani CN, Han HJ. Navigating identity: experiences of binary and non-binary transgender and gender non-conforming (TGNC) adults. Int J Transgend 2019;20(2-3):181-194 [FREE Full text] [doi: 10.1080/15532739.2018.1426074] [Medline: 32999605]
4. Sumerau JE, Mathers LA. America through Transgender Eyes. Lanham, MD: Rowman & Littlefield; 2019.
5. Gates GJ, Distinguished W. How many people are lesbian, gay, bisexual, and transgender? Williams Institute. 2011. URL: https://williamsinstitute.law.ucla.edu/wp-content/uploads/How-Many-People-LGBT-Apr-2011.pdf [accessed 2024-04-29]
6. Conron KJ, Scott G, Stowell GS, Landers SJ. Transgender health in Massachusetts: results from a household probability sample of adults. Am J Public Health 2012 Jan;102(1):118-122. [doi: 10.2105/ajph.2011.300315]
7. Spizzirri G, Eufrásio R, Lima MC, de Carvalho Nunes HR, Kreukels BP, Steensma TD, et al. Proportion of people identified as transgender and non-binary gender in Brazil. Sci Rep 2021 Jan 26;11(1):2240 [FREE Full text] [doi: 10.1038/s41598-021-81411-4] [Medline: 33500432]

8. The struggle of trans and gender-diverse persons. Office of the United Nations High Commissioner for Human Rights (OHCHR). URL: https://www.ohchr.org/en/special-procedures/ie-sexual-orientation-and-gender-identity/struggle-trans-and-gender-diverse-persons [accessed 2024-05-12]
9. Kearns S, Hardie P, O'Shea D, Neff K. Instruments used to assess gender-affirming healthcare access: a scoping review. PLoS One 2024 Jun 3;19(6):e0298821 [FREE Full text] [doi: 10.1371/journal.pone.0298821] [Medline: 38829881]
10. Kearns S, Kroll T, O'Shea D, Neff K. Experiences of transgender and non-binary youth accessing gender-affirming care: a systematic review and meta-ethnography. PLoS One 2021 Sep 10;16(9):e0257194 [FREE Full text] [doi: 10.1371/journal.pone.0257194] [Medline: 34506559]
11. Jessani A, Berry-Moreau T, Parmar R, Athanasakos A, Prodger JL, Mujugira A. Healthcare access and barriers to utilization among transgender and gender diverse people in Africa: a systematic review. BMC Glob Public Health 2024 Jun 27;2(1):44 [FREE Full text] [doi: 10.1186/s44263-024-00073-2] [Medline: 38948028]
12. Renner J, Blaszcyk W, Täuber L, Dekker A, Briken P, Nieder TO. Barriers to accessing health care in rural regions by transgender, non-binary, and gender diverse people: a case-based scoping review. Front Endocrinol (Lausanne) 2021 Nov 18;12:717821 [FREE Full text] [doi: 10.3389/fendo.2021.717821] [Medline: 34867775]
13. Ahuja TK, Goel AD, Gupta MK, Joshi N, Choudhary A, Suman S, et al. Health care needs and barriers to care among the transgender population: a study from western Rajasthan. BMC Health Serv Res 2024 Aug 26;24(1):989 [FREE Full text] [doi: 10.1186/s12913-024-11010-2] [Medline: 39187822]
14. Gillespie C. 7 major health disparities affecting the LGBTQ+ community. Health. URL: https://www.health.com/mind-body/lgbtq-health-disparities [accessed 2025-01-08]
15. Health disparities and equitable access to health care persist with transgender adults. The American Heart Association. URL: https://newsroom.heart.org/news/health-disparities-and-equitable-access-to-health-care-persist-with-transgender-adults [accessed 2024-12-20]
16. Min L. Health disparities in the LGBTQ community. Int J Arts Humanit Soc Sci Stud 2023:1-10 [FREE Full text]
17. Hughes C. A review of: “Transgender Health and HIV Prevention: Needs Assessment Studies from Transgender Communities across the United States. Edited by Walter Bockting and Eric Avery". J HIV AIDS Soc Serv 2008 Oct 12;7(3):305-307. [doi: 10.1080/15381500802309712]
18. Melendez RM, Pinto R. 'It's really a hard life': love, gender and HIV risk among male-to-female transgender persons. Cult Health Sex 2007;9(3):233-245 [FREE Full text] [doi: 10.1080/13691050601065909] [Medline: 17457728]
19. Nemoto T, Sausa LA, Operario D, Keatley J. Need for HIV/AIDS education and intervention for MTF transgenders: responding to the challenge. J Homosex 2006;51(1):183-202. [doi: 10.1300/J082v51n01_09] [Medline: 16893831]
20. Lombardi EL, Wilchins RA, Priesing D, Malouf D. Gender violence: transgender experiences with violence and discrimination. J Homosex 2002 Mar 26;42(1):89-101. [doi: 10.1300/J082v42n01_05]
21. Bockting WO, Robinson BE, Forberg J, Scheltema K. Evaluation of a sexual health approach to reducing HIV/STD risk in the transgender community. AIDS Care 2005 Apr 27;17(3):289-303. [doi: 10.1080/09540120412331299825] [Medline: 15832877]
22. Grant JM, Mottet LA, Tanis J, Harrison J, Herman JL, Keisling M. Injustice at every turn: a report of the national transgender discrimination survey. National Center for Transgender Equality and National Gay and Lesbian Task Force. 2011. URL: https://transequality.org/sites/default/files/docs/resources/NTDS_Report.pdf [accessed 2024-04-29]
23. Soled KR, Dimant OE, Tanguay J, Mukerjee R, Poteat T. Interdisciplinary clinicians' attitudes, challenges, and success strategies in providing care to transgender people: a qualitative descriptive study. BMC Health Serv Res 2022 Sep 08;22(1):1134 [FREE Full text] [doi: 10.1186/s12913-022-08517-x] [Medline: 36076288]
24. Reisner SL, Bailey Z, Sevelius J. Racial/ethnic disparities in history of incarceration, experiences of victimization, and associated health indicators among transgender women in the U.S. Women Health 2014;54(8):750-767 [FREE Full text] [doi: 10.1080/03630242.2014.932891] [Medline: 25190135]
25. Canvin L, Twist J, Solomons W. How do mental health professionals describe their experiences of providing care for gender diverse adults? A systematic literature review. Psychol Sex 2021 Apr 23;13(3):717-741 [FREE Full text] [doi: 10.1080/19419899.2021.1916987]
26. Mikulak M, Ryan S, Ma R, Martin S, Stewart J, Davidson S, et al. Health professionals’ identified barriers to trans health care: a qualitative interview study. Br J Gen Pract 2021 Jun 15;71(713):e941-e947. [doi: 10.3399/bjgp.2021.0179]
27. Agapoff J. Supporting and understanding non-binary and gender diverse youth: a physician’s view. Child Adolesc Psychiatry Ment Health 2024 Aug 24;18(1):105 [FREE Full text] [doi: 10.1186/s13034-024-00798-w] [Medline: 39182103]
28. Ilango TS, Karthikeyan S, Sumithra Devi S, Arumuganathan S, Usaid S, Sethumadhavan V. An online survey of education, knowledge and attitude toward homosexuality in adults. Indian J Soc Psychiatry 2020;36(4):344. [doi: 10.4103/ijsp.ijsp_43_20]
29. Dubin SN, Nolan IT, Streed Jr CG, Greene RE, Radix AE, Morrison SD. Transgender health care: improving medical students' and residents' training and awareness. Adv Med Educ Pract 2018;9:377-391 [FREE Full text] [doi: 10.2147/AMEP.S147183] [Medline: 29849472]

30. Tollemache N, Shrewsbury D, Llewellyn C. Que(e) rying undergraduate medical curricula: a cross-sectional online survey of lesbian, gay, bisexual, transgender, and queer content inclusion in UK undergraduate medical education. BMC Med Educ 2021 Feb 12;21(1):100 [FREE Full text] [doi: 10.1186/s12909-021-02532-y] [Medline: 33579262]
31. Persson Tholin J, Broström L. Transgender and gender diverse people's experience of non-transition-related health care in Sweden. Int J Transgend 2018 May 23;19(4):424-435 [FREE Full text] [doi: 10.1080/15532739.2018.1465876]
32. Nahata L, Quinn GP, Caltabellotta NM, Tishelman AC. Mental health concerns and insurance denials among transgender adolescents. LGBT Health 2017 Jun;4(3):188-193. [doi: 10.1089/lgbt.2016.0151] [Medline: 28402749]
33. Hughto JM, Reisner SL, Pachankis JE. Transgender stigma and health: a critical review of stigma determinants, mechanisms, and interventions. Soc Sci Med 2015 Dec;147:222-231 [FREE Full text] [doi: 10.1016/j.socscimed.2015.11.010] [Medline: 26599625]
34. Bradford J, Reisner SL, Honnold JA, Xavier J. Experiences of transgender-related discrimination and implications for health: results from the Virginia Transgender Health Initiative Study. Am J Public Health 2013 Oct;103(10):1820-1829. [doi: 10.2105/AJPH.2012.300796] [Medline: 23153142]
35. Do TT, Nguyen AT. 'They know better than we doctors do': providers' preparedness for transgender healthcare in Vietnam. Health Sociol Rev 2020 Mar 29;29(1):92-107. [doi: 10.1080/14461242.2020.1715814] [Medline: 33411663]
36. Allen LR, Adams N, Ashley F, Dodd C, Ehrensaft D, Fraser L, et al. Principlism and contemporary ethical considerations for providers of transgender health care. Int J Transgend Health 2024 Jan 19:1-19 [FREE Full text] [doi: 10.1080/26895269.2024.2303462]
37. Tamim RM, Bernard RM, Borokhovski E, Abrami PC, Schmid RF. What forty years of research says about the impact of technology on learning. Rev Educ Res 2011 Mar 01;81(1):4-28. [doi: 10.3102/0034654310393361]
38. Davies RS, West RE. Technology integration in schools. In: Spector JM, Merrill MD, Elen J, Bishop MJ, editors. Handbook of Research on Educational Communications and Technology. Cham, Switzerland: Springer; 2014:841-853.
39. Proenca L, Mendes JJ, Botelho J, Machado V. E-learning and Digital Training in Healthcare Education: Current Trends and New Challenges. New York, NY: MDPI; 2023.
40. eHealth. World Health Organization. URL: http://www.emro.who.int/health-topics/ehealth/ [accessed 2024-04-29]
41. Young J, Gregory J, Rojas M, Justin G, Kalir T. Transgender healthcare: development of an illustrated elearning tool for medical education. MedEdPublish (2016) 2021;10(1):159 [FREE Full text] [doi: 10.15694/mep.2021.000159.1] [Medline: 38486569]
42. Sittig DF, Singh H. A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. Qual Saf Health Care 2010 Oct;19 Suppl 3(Suppl 3):i68-i74 [FREE Full text] [doi: 10.1136/qshc.2010.042085] [Medline: 20959322]
43. Flick U, von Kardoff E, Steinke I. A Companion to Qualitative Research. Thousand Oaks, CA: Sage Publications; 2004.
44. Saunders B, Sim J, Kingstone T, Baker S, Waterfield J, Bartlam B, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. Qual Quant 2018 Mar;52(4):1893-1907 [FREE Full text] [doi: 10.1007/s11135-017-0574-8] [Medline: 29937585]
45. Braun V, Clarke VZ. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
46. Saladana J. The Coding Manual for Qualitative Researchers. 2nd edition. Thousand Oaks, CA: Sage Publications; 2013.
47. The common rule. U.S. Department of Health & Human Services. URL: https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html [accessed 2024-04-29]
48. UK policy framework for health and social care research. Health Research Authority. URL: https://www.hra.nhs.uk/planning-and-improving-research/policies-standards-legislation/uk-policy-framework-health-social-care-research/ [accessed 2024-04-29]
49. National ethical guidelines for biomedical and health research involving human participants. Indian Council of Medical Research. URL: https://ethics.ncdirindia.org/asset/pdf/ICMR_National_Ethical_Guidelines.pdf [accessed 2024-04-29]
50. The NHS ends the "gender-affirmative care model" for youth in England. Society for Evidence-Based Gender Medicine. URL: https://segm.org/England-ends-gender-affirming-care [accessed 2024-04-29]
51. Bans on gender-affirming care for transgender youth: a 2023 legislative update. Williams Institute. URL: https://williamsinstitute.law.ucla.edu/wp-content/uploads/Trans-Youth-Health-Bans-Mar-2023.pdf [accessed 2024-04-29]
52. Ahmad T, Robinson L, Uleryk E, Yu C. Trans health training objectives: a scoping review. Clin Teach 2024 Feb 08;21(1):e13673. [doi: 10.1111/tct.13673] [Medline: 37806669]
53. Cooper RL, Ramesh A, Radix AE, Reuben JS, Juarez PD, Holder CL, et al. Affirming and inclusive care training for medical students and residents to reduce health disparities experienced by sexual and gender minorities: a systematic review. Transgend Health 2023 Aug 01;8(4):307-327 [FREE Full text] [doi: 10.1089/trgh.2021.0148] [Medline: 37525832]
54. Davidge-Pitts CJ, Nippoldt TB, Natt N. Endocrinology fellows' perception of their confidence and skill level in providing transgender healthcare. Endocr Pract 2018 Dec;24(12):1038-1042. [doi: 10.4158/ep-2018-0307]
55. Lacombe-Duncan A, Logie CH, Persad Y, Leblanc G, Nation K, Kia H, et al. Implementation and evaluation of the 'Transgender Education for Affirmative and Competent HIV and Healthcare (TEACHH)' provider education pilot. BMC Med Educ 2021 Nov 04;21(1):561 [FREE Full text] [doi: 10.1186/s12909-021-02991-3] [Medline: 34732178]

56. Light AD, Obedin-Maliver J, Sevelius JM, Kerns JL. Transgender men who experienced pregnancy after female-to-male gender transitioning. Obstet Gynecol 2014 Dec;124(6):1120-1127 [FREE Full text] [doi: 10.1097/AOG.0000000000000540] [Medline: 25415163]
57. Mansh M, White W, Gee-Tong L, Lunn MR, Obedin-Maliver J, Stewart L, et al. Sexual and gender minority identity disclosure during undergraduate medical education: "in the closet" in medical school. Acad Med 2015 May;90(5):634-644. [doi: 10.1097/ACM.0000000000000657] [Medline: 25692563]
58. Grammarly. URL: https://www.grammarly.com/ [accessed 2025-03-04]
59. ChatGPT 3.5. OpenAI. URL: https://chatgpt.com/g/g-F00faAwkE-open-a-i-gpt-3-5 [accessed 2025-03-04]

Abbreviations

ICT: information and communication technology
TNBI: transgender, nonbinary, and intersex

Edited by B Lesselroth; submitted 25.10.24; peer-reviewed by H Pilabré, J Wells; comments to author 03.12.24; revised version received 14.01.25; accepted 16.02.25; published 06.03.25.

Please cite as:
Katta S, Davoody N
Exploring Health Care Professionals' Perspectives on Education, Awareness, and Preferences for Digital Educational Resources to Support Transgender, Nonbinary, and Intersex Care: Interview Study
JMIR Med Educ 2025;11:e67993
URL: https://mededu.jmir.org/2025/1/e67993
doi:10.2196/67993
PMID:40053815

Original Paper

A Web-Based Training Intervention for Primary Care Providers on Preparing Patients for Cancer Treatment Decisions and Conversations About Clinical Trials: Evaluation of a Pilot Study Using Mixed Methods and Follow-Up

Naomi D Parker[1], PhD, MPA; Margo Michaels[2], MPH; Carla L Fisher[1], PhD; Alyssa Crowe[1], BS; Elisa S Weiss[3], PhD; Maria Sae-Hau[3], PhD; Jason Arnold[4], EdD; Andrea Cassells[5], MPH; Domenic Durante[4], MEd; Ji-Hyun Lee[6], DrPH; Raymond Mailhot Vega[7], MD, MPH; Ana Natale-Pereira[8], MD, MPH; Taylor S Vasquez[9], PhD; Zhongyue Zhang[10], MS; Carma L Bylund[1], PhD

[1]Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, FL, United States
[2]School of Public Health, Boston University, Boston, MA, United States
[3]The Leukemia & Lymphoma Society, Washington, DC, United States
[4]Department of E-Learning, Technology, and Communications, College of Education, University of Florida, Gainesville, FL, United States
[5]Clinical Directors Network, Inc, New York, NY, United States
[6]Department of Biostatistics, College of Public Health and Health Professions, University of Florida, Gainesville, FL, United States
[7]Department of Radiation Oncology, College of Medicine, University of Florida, Gainesville, FL, United States
[8]Department of Medicine, New Jersey Medical School, Rutgers University, Newark, NJ, United States
[9]Department of Public Relations, College of Journalism and Communications, University of Florida, Gainesville, FL, United States
[10]UF Health Cancer Center, University of Florida, Gainesville, FL, United States

Corresponding Author:
Naomi D Parker, PhD, MPA
Department of Health Outcomes and Biomedical Informatics
College of Medicine
University of Florida
1889 Museum Rd
Suite 7000
Gainesville, FL, 32611
United States
Phone: 1 352 627 9467
Email: naomiparker@ufl.edu

Abstract

Background: Recruitment to cancer clinical trials (CCTs) is low, particularly for underrepresented groups such as uninsured patients, those with low-income status, and racial and ethnic minoritized individuals. A significant barrier is that treating oncologists often fail to inform patients about the possibility of CCT participation as an option for quality cancer care. Therefore, patient inquiries about trials before starting treatment should be normalized and encouraged, particularly for underrepresented groups. Primary care providers (PCPs) are uniquely suited to do this because they interact with patients at the time of cancer diagnosis, provide ongoing care, and are trusted sources of information.

Objective: This study was designed to pilot an innovative web-based CCT training intervention for PCPs, including practicing clinicians and trainees, to increase their ability to prepare patients for cancer treatment decisions and conversations with oncologists about clinical trials.

Methods: We conducted an evaluation of a pilot study using a self-guided, 1-hour web-based training intervention for PCPs with survey assessments before the intervention, immediately after the intervention, and at the 3-month follow-up. We used a mixed methods approach, incorporating quantitative and qualitative data collection and analysis. The evaluation was guided by the Kirkpatrick evaluation model, focusing on levels 1 (reaction), 2 (learning), and 3 (behavior).

Results: A total of 29 PCPs completed the intervention and pre- and postintervention measures, with 28 (97%) PCPs completing the 3-month follow-up assessment. Of these 28 PCPs, 8 (29%) participated in a qualitative interview after the 3-month follow-up assessment. Participants reported high levels of satisfaction with the course. CCT knowledge, as well as attitudes and beliefs, improved after the course and were sustained at the 3-month follow-up. PCPs reported willingness to communicate with patients about cancer treatment options, including CCTs, and willingness to talk with their colleagues about potential changes in referral practices. However, fewer PCPs had actually engaged in these conversations by the 3-month follow-up. In the interviews, PCPs cited limited interprofessional knowledge sharing and organizational constraints as barriers. Notably, PCPs reported changes in their communication behavior with patients: a higher proportion reported communicating with patients at the time of referral about cancer treatment options and clinical trials at the 3-month follow-up than at baseline. In the interviews, PCPs reported that they felt more comfortable and empowered to have these conversations.

Conclusions: This pilot study found that a self-guided, 1-hour web-based training intervention for PCPs resulted in improved knowledge, attitudes, and beliefs, as well as improved communication with patients, to prepare them for discussions with oncologists about cancer treatment and CCTs. Future dissemination of this course has the potential to make an impact on CCT accrual.

(JMIR Med Educ 2025;11:e66892) doi:10.2196/66892

KEYWORDS

cancer clinical trials; continuing medical education; continuing education; primary care providers; provider-patient communication; cancer treatment; referral practices; online learning

Introduction

Background

Cancer clinical trials (CCTs) are essential for advancing cancer treatment and ensuring quality of care for patients [1]. However, for cancer research to benefit all patients, it is critical to include participants who represent the diversity of the American population. Unfortunately, participation in cancer treatment trials remains low, with an overall accrual rate of only 7.1% [2]. Moreover, only approximately 5% of trial participants are Black or Hispanic, although these groups represent 15% and 13%, respectively, of people with cancer [3]. The ongoing underrepresentation of minoritized populations in CCTs not only raises concerns about equitable access to care but also limits the generalizability of the findings across diverse patient groups.

Various barriers at the system, institutional, clinician and research team, and patient levels contribute to low CCT enrollment [4-11]. One significant barrier is that a substantial portion of eligible patients are not approached about the possibility of receiving treatment through clinical trials [12]. This issue is especially pronounced for patients from demographic groups who are typically underrepresented in trials (eg, uninsured individuals, members of racial and ethnic minoritized groups, and people living in rural areas) [12-19] and who experience higher cancer mortality rates [20]. Even when the option of clinical trials is discussed with eligible patients, the communication is frequently unclear or inequitable, often reflecting racial disparities in how comprehensively information about CCTs is conveyed [21].

Research has shown that educating patients before their first oncologist visit can improve their knowledge, attitudes, and readiness for treatment decision-making, as well as increase their willingness to ask about and consider cancer treatment trials [22]. Therefore, it is important to normalize receiving treatment through clinical trials and encourage patient inquiries about trials before starting treatment, particularly for underrepresented groups. To effectively inquire about clinical trials, patients need to understand that quality care can be provided through CCTs, that asking about CCTs is encouraged and supported, that their participation in treatment decisions is encouraged [23,24], and that they are capable of fulfilling this role [8-10]. Patient education can start with the patient's trusted primary care provider (PCP).

PCPs are a potential gateway facilitating access to CCTs because they interact with patients at the time of cancer diagnosis, provide ongoing care, and are trusted sources of information [25,26]. PCPs also report believing that they play an important role in patient care across the cancer continuum, including being actively involved during treatment [27]. Moreover, studies have shown that a trusted physician's recommendation is a primary factor influencing patients' decisions to enroll in clinical trials when offered [5,10,28-34] and that this influence may even be more acute for minoritized populations [35]. Therefore, improving PCPs' attitudes, beliefs, and behaviors around CCTs may significantly influence patients' attitudes and openness toward participation [36-39].

Although PCPs have expressed a desire for more of their patients with cancer to participate in CCTs, they often feel inadequately prepared to discuss trials due to their own limited understanding of CCTs as a high-quality treatment option [40,41]. In addition, they need guidance on effectively communicating with patients about potential trial participation. This gap in knowledge and skills presents a significant opportunity to educate PCPs about CCTs and their role in preparing patients for the oncology referral. These new skills can potentially improve trial accrual and patient access, especially for underrepresented groups.

Previous research, including our own, has demonstrated that conducting education with PCPs who serve underrepresented populations can contribute to breaking down key barriers to trial participation [42-44]. This can be achieved by improving PCPs' capacity to both (1) educate recently diagnosed patients about the possibility of trial participation at the time of treatment referral to cancer care and (2) provide decision-making support

to patients if advice is sought about trial participation after the patient has already met with the oncologist.

Objectives

The purpose of this study was to pilot-test an innovative web-based training intervention for PCPs, including both practicing clinicians and trainees, designed to help them prepare patients for discussions about cancer treatment and CCTs. There are several types of CCTs, including prevention, screening, behavioral, and supportive care trials. However, for the purpose of this training, we focused specifically on treatment trials (eg, those involving chemotherapy or immunotherapy). We aimed to evaluate (1) PCPs' general impressions of the training intervention and how the intervention impacted (2) their knowledge about CCTs, (3) their perceptions of their roles in caring for patients with cancer, (4) their approach toward and communication with patients, and (5) their willingness and behaviors in enacting practice-level changes.

Methods

Study Design

A 1-hour web-based training intervention for PCPs was developed and evaluated in a single-arm pilot study with assessments conducted before the intervention, immediately after the intervention, and at 3-month follow-up. We used a mixed methods approach, using quantitative and qualitative data collection and analysis to evaluate the training intervention [45]. We used the Kirkpatrick model to guide our evaluation [46]. The Kirkpatrick model proposes 4 sequential levels of training evaluation that become progressively more distal from the original training. In this study, we focused on levels 1 (reaction), 2 (learning), and 3 (behavior).

After conducting separate analyses, we integrated the results to allow the qualitative data to enhance and provide context for the quantitative data [47,48]. This approach integrates the data for triangulation, which strengthens the validity of the findings and allows for a more comprehensive understanding of the data [49-51]. This study followed the SQUIRE 2.0 (Standards for Quality Improvement Reporting Excellence) guidelines for assessing quality and reporting results [52].

Training Intervention

We collaborated with instructional designers (DD and JA) from the University of Florida's E-Learning, Technology, and Communications department to develop a 1-hour web-based training intervention for PCPs titled "Preparing Patients for Cancer Treatment Decisions: The Critical Role of Primary Care Providers in Facilitating Equitable Access to Care and Clinical Trials." The content of the intervention was based on the authors' collective research on the topic [40-43,53] as well as current literature on the role of PCPs in cancer treatment and referral, barriers to CCTs, patient activation, and best practices in medical education.

The asynchronous training intervention was hosted by a PCP (AN-P) and a radiation oncologist (RMV) who introduced content through engaging conversations and videos. Using a model of cognitive dissonance, the course introduced new information about CCTs, with the intent of helping providers realize inaccuracies in their knowledge, attitudes, and behaviors around CCTs and the referral to cancer treatment [54,55]. A clinical trials content expert (MM) additionally provided facts about CCTs, cancer care disparities, barriers to CCT participation, and the role of PCPs in cancer care. The 4 intervention modules covered content to enhance PCPs' knowledge and communication skills surrounding cancer treatment discussions, including clinical trials and strengthening preparation related to the oncology referral process. The communication skills taught in the curriculum centered on the 5 E's communication model (ie, explore, educate, encourage, engage in planning, and emphasize partnership), a framework designed to support patients in becoming more active participants in cancer treatment decision-making. Informed by principles of patient activation, treatment decision support, shared decision-making, and patient-centered communication, the model provides a simple mnemonic that clinicians can use to educate patients; prepare them for oncology referrals; and encourage inquiry about treatment options, including cancer treatment trials [56-59]. The learning objectives for each of the 4 modules are displayed in Table 1.

Table 1. Intervention modules and learning objectives.

Module	Learning focus	Learning objectives
1	Disparities in cancer care and clinical trial participation	• List the reasons why cancer treatment trials are defined as quality cancer care • Describe disparities in cancer care and in cancer treatment trial participation
2	Importance of strong PCP[a] role in referrals to cancer treatment	• Assess the critical role of the PCP in the referral process • Assess the PCP's role in preparing patients around treatment decision-making with the specialist
3	Communication skills to foster patient engagement in cancer treatment	• Apply a 5-step communication approach to prepare patients for the specialist referral and their engagement in treatment planning
4	Needed changes with colleagues in the referral process	• Evaluate the need to refine organizational procedures to improve one's own referral practices for patients recently diagnosed with cancer

[a]PCP: primary care provider.

Throughout the modules, video vignettes featuring actors as patients and physicians further illustrate the 5E's communication model. The modules also include case studies where participants can explore patients' disease histories, backgrounds, and cancer diagnoses. Participants complete each module at their own pace.

Evaluation of the Training Intervention: Quantitative Assessments

Participants and Recruitment

We recruited PCPs to complete a web-based training intervention, pre- and postintervention surveys, and a postintervention interview. Recruitment was based on the following inclusion criteria: (1) being a PCP, including a physician (with MD or DO degrees), nurse practitioner (NP), physician assistant, or trainee (specifically postgraduate year [PGY] 2 or above in a family or internal medicine residency program) with at least 3 months of outpatient primary care experience; (2) working in an outpatient setting; (3) having made patient referrals to specialists for cancer treatment in the past year; and (4) speaking English and residing in the United States or its territories. Between July and December 2023, we emailed invitations that included a study description and instructions for those interested to contact a research coordinator. The invitations were distributed through email lists provided by three partner organizations: (1) the Clinical Directors Network, a national not-for-profit practice-based research network and clinician training organization; (2) Penn Medicine, an academic health system; and (3) the University of Florida's Department of Community Health and Family Medicine and its family medicine and internal medicine residency programs, which are based in a large public academic health system. Once enrolled in the study, participants had up to 4 weeks to complete the training intervention.

Procedure

Participants completed a pretraining survey, an immediate posttraining survey, and a 3-month follow-up survey. Data were collected on the web via QualtricsXM (Qualtrics International Inc) [60]. The quantitative and qualitative phases of data collection and descriptions of each method are presented in Table 2.

Table 2. Descriptions of data collection measures and timeline.

Measures	Description	Before the training	After the training	3-mo follow-up
Participant characteristics	A 10-item survey that asks about participants' demographics, professional background, and current professional setting	✓		
Knowledge of CCTs[a]	A 7-item measure comprising true or false statements about CCTs	✓	✓	✓
CCT attitudes and beliefs	A 4-item Likert-type measure that assesses attitudes, beliefs, and behavioral intentions related to CCTs	✓	✓	✓
Patient communication and referral practices	A 14-item measure that assesses the communication and referral behaviors of PCPs[b] before (7 questions) and after (7 questions) returning patients' initial referral to a cancer specialist to discuss treatment options	✓		✓
Willingness to change	A 3-item Likert-type measure that focuses on PCPs' willingness to make changes in their practice	✓	✓	✓
Willingness to communicate	A 6-item Likert-type measure that focuses on PCPs' willingness to engage with patients who have a cancer diagnosis	✓	✓	✓
Intervention usability	A 5-item survey that asks about participants' device (eg, tablet computer), browser type, and whether they encountered any technical issues while taking the web-based course		✓	
Overall course satisfaction	A 10-item Likert-type measure where participants evaluate their training experience (eg, content and logistics) and intervention acceptability, along with 2 open-ended questions for additional feedback		✓	
Interview	A semistructured interview based on the quantitative survey measures, with an emphasis on exploring PCPs' experiences with the course			✓

[a]CCT: cancer clinical trial.

[b]PCP: primary care provider.

Before beginning the training intervention and after providing informed consent, participants were asked to complete a 44-item web-based survey. The preintervention survey included 5 measures—knowledge of CCTs, CCT attitudes and beliefs, patient communication and referral practices, willingness to communicate, and willingness to change—all of which were informed by our previous work [40,41,43,53]. The preintervention survey also collected participant characteristics, including demographics and professional background. Immediately after the intervention, participants were asked to complete a 35-item web-based survey that included the knowledge of CCTs, CCT attitudes and beliefs, willingness to communicate, and willingness to change measures. In addition, they were asked about intervention usability and overall course satisfaction. Completing the intervention and the pre- and postintervention surveys fulfilled participants' training requirements, at which point they were compensated US $150 for their participation.

Three months after the intervention, participants were asked to compete an additional 34-item web-based survey that included the knowledge of CCTs, CCT attitudes and beliefs, patient communication and referral practices, willingness to communicate, and willingness to change measures. After completing the 3-month follow-up survey, participants were compensated an additional US $45 for their participation. All surveys are shown in Multimedia Appendix 1.

Statistical Analysis

Descriptive statistics were used to analyze participants' demographic characteristics. The data were analyzed using R software (version 4.4.0; R Foundation for Statistical Computing) [61]. Cronbach α values were used to measure the internal consistency among the CCT attitudes and beliefs, willingness to communicate, and willingness to change measures, and we report both individual item means and aggregated means for these measures. Frequencies (%) and mean (SD) scores for all survey measures were reported. The McNemar chi-square test and the Wilcoxon signed rank test were used to compare the results of the preintervention and postintervention surveys, including the 3-month follow-up survey. Due to the exploratory nature of this pilot study, adjustments for multiple comparisons were not made.

Evaluation of the Training Intervention: Qualitative Phase

To gain further insights into the survey results, we conducted follow-up interviews with PCPs who completed the 3-month follow-up survey.

Participants and Recruitment

At the end of the 3-month follow-up survey, interested PCPs were invited to participate in a brief web-based interview to discuss their experience with the course. Between December 2023 and February 2024, members of the research team (A Crowe and NDP) contacted interested PCPs to confirm their participation and to schedule the interviews.

Interview Procedure

Between December 2023 and February 2024, two authors (NDP and A Crowe) interviewed PCPs via Zoom (Zoom Video Communications, Inc). The semistructured interview guide was designed based on the quantitative survey measures, with an emphasis on exploring PCPs' experiences with the course. Participants were asked to share their impressions of the training intervention, their attitudes and beliefs regarding the role of PCPs, and to describe any changes they had made to their practice since taking the course. The interviews were audio recorded and professionally transcribed verbatim before analysis. Participants who completed the interview received an additional US $50 in compensation.

Qualitative Analysis

The interview data and analysis were managed using ATLAS.ti software (version 24.1.1; ATLAS.ti Scientific Software Development GmbH) [62]. Data collection and analysis were conducted concurrently, enabling an initial rapid analysis to capture and categorize data into broad themes [63,64]. Subsequently, a comprehensive thematic analysis was conducted by the first author (NDP), using both deductive and inductive coding strategies guided by the constant comparative method [65,66]. Throughout the analysis, 2 authors (NDP and CLF) met to review and discuss codes and emerging themes and to develop a codebook that guided the full thematic analysis [65,67]. Similar codes identified across all transcripts were collapsed into overarching themes, followed by axial coding to further characterize thematic properties [67].

Data Integration

After conducting individual quantitative and qualitative analyses, we jointly interpreted the findings from both phases to enhance reporting. We report quantitative data first, which are then elaborated with the qualitative findings. When appropriate, we created joint displays to illustrate this data integration and to organize the results [48,68].

Ethical Considerations

The study procedures were approved by the University of Florida Institutional Review Board (202200894). Before the PCPs participated in the surveys and interviews, they were provided a statement of participant rights and responsibilities and given the option to withdraw from the study at any time. All survey participants checked a box agreeing to participate before taking the survey. All interview participants verbally confirmed their willingness to participate, had the opportunity to ask questions before the interview, and were informed that interview data would be transcribed for analysis but only available to members of the research team. All data were deidentified before reporting in this study. Participants who completed the intervention and the pre- and postintervention surveys were compensated US $150. Participants who completed the 3-month follow-up survey were compensated an additional US $45.

Results

Overview

A total of 29 PCPs completed the study and preintervention and postintervention surveys; the sample included 15 (52%) women and 14 (48%) men. Of the 29 PCPs, 28 (97%) completed the 3-month follow-up survey (Figure 1). Of these 28 participants, 8 (29%) took part in the interviews. Of the 29 participants, 26 (90%) had MD degrees, 2 (7%) were NPs, and 1 (3%) had a DO degree. More than half were physicians in training, with 17 (59%) describing themselves as residents or fellows in PGY2 (n=7, 41%), PGY3 (n=9, 53%), or PGY4 (n=1, 6%). Participants described their practice areas as internal medicine (16/29, 55%), family practice (11/29, 38%), or other (2/29, 7%). Most reported that their practice or training setting was an academic-affiliated hospital (17/29, 59%) or community-based and owned by an academic medical center (7/29, 24%). Complete demographic and professional characteristics of both survey and interview participants are shown in Table 3.

Figure 1. Flow diagram depicting study enrollment and participation.

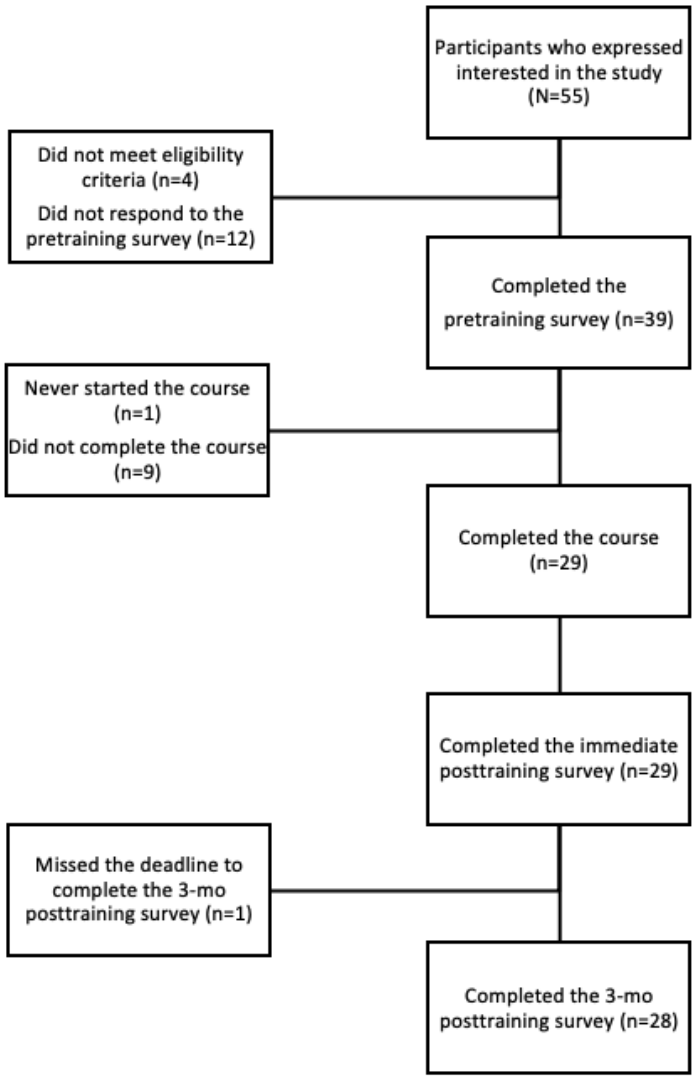


Table 3. Characteristics of surveyed and interviewed participants[a].

Characteristics	Survey (n=29), n (%)	Interview (n=8), n (%)
Sex		
Female	15 (52)	7 (88)
Male	14 (48)	1 (12)
Racial identity		
American Indian or Alaska Native	1 (3)	0 (0)
Asian	9 (31)	4 (50)
Black or African American	2 (7)	2 (25)
White	13 (45)	2 (25)
Multiracial	2 (7)	0 (0)
Other	1 (3)	0 (0)
Prefer not to respond	1 (3)	0 (0)
Ethnicity		
Hispanic or Latino	1 (3)	0 (0)
Not Hispanic or Latino	27 (93)	8 (100)
Prefer not to respond	1 (3)	0 (0)
Clinician type		
Doctor of medicine	26 (90)	7 (88)
Nurse practitioner	2 (7)	1 (12)
Doctor of osteopathic medicine	1 (3)	0 (0)
Stage of training		
Resident or fellow	17 (59)	4 (50)
Not resident or fellow	7 (24)	3 (38)
Prefer not to respond	5 (17)	1 (12)
Postgraduate year of training (if resident or fellow)		
2	7 (24)	1 (12)
3	9 (31)	2 (25)
4	1 (3)	1 (12)
N/A[b]	12 (41)	0 (0)
Missing	0 (0)	4 (50)
Practice area		
Internal medicine	16 (55)	5 (62)
Family practice	11 (38)	3 (38)
Other	2 (7)	0 (0)
Practice or training setting		
Academic-affiliated hospital	17 (59)	5 (62)
Community-based, owned by an academic medical center	7 (24)	1 (12)
Community-based, independent private	2 (7)	1 (12)
Community hospital	2 (7)	1 (12)
Community-based, owned by a large, nonhospital entity	1 (3)	0 (0)
Is practice a designated federally qualified health center?		
Yes	15 (52)	4 (50)

Characteristics	Survey (n=29), n (%)	Interview (n=8), n (%)
No	14 (48)	4 (50)
Is practice supported by the Indian Health Service?		
No	27 (93)	8 (100)
Yes	2 (7)	0 (0)

[a]Percentages may not add up to 100 because of rounding.

[b]N/A: not applicable.

The mixed methods findings were used to (1) characterize PCPs' general impressions of the training intervention; in addition, we used these evaluation findings to capture how the training intervention impacted their (2) knowledge, (3) perceptions of their roles in caring for patients with cancer, (4) communication with their patients, and (5) practice-level change willingness and behaviors. The quantitative results are reported first and then elaborated upon with the qualitative findings within each of the 5 areas of impact.

General Impressions of the Course

Both survey and interview findings illustrated PCPs' positive impressions of, and high satisfaction with, the course. Specifically, PCPs expressed satisfaction with the course content, design and functionality, course elements, and course length. Their general impressions are presented in a joint display (Table 4) to depict the mixed methods results comprehensively.

Table 4. Primary care providers' impressions of the course.

Course characteristics and quantitative measures	Survey question ratings[a], mean (SD)	Illustrative quotes from the interviews
Content		
"Course content was relevant to my work."	4.6 (0.5)	"The content was excellent, and the way that it was provided was also very good. I found it very useful to learn information that I didn't know before." [P1]
"The course provided an adequate evidence base to support the content."	4.5 (0.6)	"I thought it was kind of helpful...I'd never looked into discrepancies in cancer care. I probably could've suspected but didn't know that was a fact." [P8]
"The right amount of information was provided."	4.6 (0.5)	"I felt [the course was] kind of straight to the point, and the topic was well laid out." [P2]
Design and functionality		
"The course presented information in a clear and organized manner."	4.7 (0.5)	"From an educational perspective, I thought it was designed pretty well. It was logical in the way that the modules flowed from one to the next." [P3]
"The graphics (graphs, pictures, illustrations) added to the effectiveness of the presentation."	4.5 (0.6)	"I thought [the course] was really well organized and laid out, and it had a nice, I guess, sequence of training." [P4]
Course elements		
"The course scenarios facilitated my understanding."	4.6 (0.5)	"I really enjoyed the videos, the examples, and the quiz questions...[and the] summary at the end, which I felt was helpful." [P1]
"The interactive features in the course activities helped me learn."	4.4 (0.8)	"I liked the different scenarios, and the short videos that you could engage with, and then immediately afterwards, trying to apply that to some questions." [P3]
"The clinical interactions shown in the course felt authentic."	4.3 (0.8)	"I really liked the case scenarios where it gave you different options for how you would respond to a patient. I thought that gave really valuable feedback in how we phrase some of the topics that we bring up." [P4]
"The clinical interactions shown in the course were relatable to me."	4.4 (0.6)	"I think, even as a primary care doctor who sees a lot of patients and who does diagnose cancer sometimes...It was really good content to push me to be more hands-on in my patients' cancer care process." [P3]
Length of course		
"The length of the course was appropriate."	4.7 (0.5)	"I thought that [the course] was educational and informative, and it was fairly straightforward and quick." [P8]

[a]Survey questions were scored on a 5-point scale ranging from 1=strongly disagree to 5=strongly agree.

CCT Knowledge

PCPs' positive feedback about the course aligns with their significant improvement in CCT knowledge after participating in the course. The survey results (Table 5) indicated that PCPs' knowledge of CCTs was low before the training, with a mean score of 55% (SD 17.9%) across all 7 CCT knowledge items. Immediately after the training, PCPs' knowledge of CCTs improved significantly to a mean score of 81% (SD 12.1%; $P<.001$). This improvement from before the training was sustained at the 3-month follow-up, with a mean score of 73% (SD 18.4%; $P=.005$).

Table 5. Comparison of correct responses on the pretraining and posttraining knowledge of cancer clinical trials (CCTs) survey[a].

True or false CCT knowledge questions	Before the training (n=29), n (%)	After the training (n=29), n (%)	3-mo follow-up (n=28), n (%)	*P* value[b] (before the training vs after the training)	*P* value[b] (before the training vs 3-mo follow-up)
"In a cancer treatment trial, patients will receive a placebo alone or the new treatment being tested."	19 (66)	26 (90)	21 (75)	*.046* [c]	.70
"While about 25% of U.S. adults with cancer are eligible to participate in cancer treatment trials, only about 8% participate."	20 (69)	29 (100)	27 (96)	*.008*	*.046*
"Patients determined to be eligible for cancer treatment trials are usually offered the opportunity to participate."	14 (48)	23 (79)	22 (79)	*.04*	.06
"Patients from racial and ethnic minority groups eligible to participate in cancer treatment trials tend to be offered the opportunity to participate less frequently than are white patients."	19 (66)	25 (86)	21 (75)	.15	.60
"When offered the opportunity to participate in a cancer treatment trial, patients from racial and ethnic minority groups agree to participate at about the same rate as do white patients."	0 (0)	5 (17)	5 (18)	.07	.07
"Recommendation by a trusted physician is a primary factor influencing a patient's decision to enroll in a cancer treatment trial."	28 (97)	29 (100)	27 (96)	1	1
"Federal law now requires private insurers, Medicare and Medicaid to cover routine patient care costs in most cancer treatment trials."	11 (38)	28 (97)	20 (71)	*<.001*	*.03*

[a]Percentage of aggregated correct responses: before the training, n=55 (18%); after the training, n=81 (12%); 3-mo follow-up, n=73 (18%); before the training versus after the training *P*<.001*;* before the training versus 3-mo follow-up *P*=.005.

[b]*P* values were calculated using the McNemar chi-square test with continuity correction, including only the participants who completed both pre- and posttraining surveys.

[c]Italicization indicates a statistically significant *P* value.

PCPs' Perceptions of Their Roles in Caring for Patients With Cancer

Overview

The findings demonstrated that the training intervention was effective in changing PCPs' attitudes and beliefs about their role in caring for patients with cancer, which was sustained at the 3-month follow-up. As shown in Table 6, the scores across all 4 items increased significantly from before the training (mean 4.2, SD 0.6) to after the training (mean 4.6, SD 0.4; *P*<.001), as PCPs agreed that they play a role in their patients' cancer care. This change was sustained at the 3-month follow-up (mean 4.5, SD 0.5; *P*=.004). Most significantly, PCPs were more likely to agree that they "have an important role in educating patients about the possibility of receiving cancer treatment through a clinical trial, before the referral to a specialist" immediately after the training (mean 4.5, SD 0.6; *P*<.001) compared to before the training (mean 3.6, SD 1.0). This individual item improvement was also sustained at the 3-month follow-up (mean 4.4, SD 0.6; *P*<.001).

Table 6. Comparison of pretraining and posttraining scores on the cancer clinical trial attitudes and beliefs survey[a].

Questions[b]	Before the training (n=29), mean (SD)	After the training (n=29), mean (SD)	3-mo follow-up (n=28), mean (SD)	*P* value[c] (before the training vs after the training)	*P* value[c] (before the training vs 3-mo follow-up)
"Health care providers like me have an important role in educating patients that there is often more than one option for treatment, before the referral to a specialist."	4.2 (0.8)	4.6 (0.5)	4.5 (0.6)	*.008* [d]	.07
"Health care providers like me have an important role in educating patients about the possibility of receiving cancer treatment through a clinical trial, before the referral to a specialist."	3.6 (1.0)	4.5 (0.6)	4.4 (0.6)	*<.001*	*<.001*
"Health care providers like me have an important role in supporting patients' decision to participate in a cancer treatment trial."	4.4 (0.7)	4.7 (0.5)	4.6 (0.6)	.07	.30
"Health care providers like me can make a difference in the quality of cancer care our patients receive."	4.6 (0.6)	4.7 (0.5)	4.5 (0.6)	.11	.80

[a]Mean of aggregated CCT attitudes and beliefs questions: before the training, mean 4.2 (SD 0.6); after the training, mean 4.6 (SD 0.4); 3-mo follow-up, mean 4.5 (SD 0.5); before the training versus after the training *P*<.001*;* before the training versus 3-mo follow-up *P*=.004. Survey questions were scored on a 5-point scale ranging from 1=strongly disagree to 5=strongly agree.

[b]Cronbach α value for the 4 CCT attitudes and beliefs questions was 0.82.

[c]*P* values were calculated using the Wilcoxon signed rank test with continuity correction, including only the participants who completed both pre- and posttraining surveys.

[d]Italicization indicates a statistically significant *P* value.

The qualitative findings shed light on these results by illustrating PCPs' perceived role in caring for patients with cancer. After the 3-month follow-up, PCPs described fulfilling three key roles: (1) *partner*, (2) *educator*, and (3) *interpreter.* However, they also acknowledged struggling or feeling restricted in these roles due to *expertise boundaries*.

PCP as Partner

PCPs explained how they can fulfill roles beyond offering medical expertise by being partners. They emphasized the importance of *being a consistent source of support* by remaining involved in patients' cancer journeys:

> *As a primary doctor, you really want to make your patient feel that you're there. You're not leaving...You still want to be a part of whatever plan is going to occur.* [P2]

Another PCP highlighted the collaborative nature of this ongoing support:

> *You have to look forward and make a plan, let [patients] know that they take one step, I take one step forward, and we're kind of in tandem.* [P7]

PCPs also discussed *providing emotional support*, including helping patients manage their emotions and "the mental health issues that come along with [a] cancer diagnosis and cancer treatment, in general" (PCP1). Another PCP explained as follows:

> *When you initially tell the patient, they're kind of like flabbergasted, and they're in a state of shock...They're just thinking "I'm going to die," so I...try to help the patient get through the emotions, that initial shock.* [P6]

PCP as Educator

PCPs also discussed having roles as educators, where they focused on preparing patients for consultations with specialists by *discussing treatment options*, including introducing the possibility of receiving care through a CCT. PCPs emphasized being more proactive after the training intervention in ensuring that their patients were adequately informed to engage with specialists:

> *[The course] helped me to focus more on general treatment ideas, like radiation, surgery, chemotherapy, and sort of introduce them to those different treatment options so that then they could take that information and ask their oncologist and oncology surgeons about what is truly available to them, including clinical trials.* [P4]

A PCP additionally described *exploring patients' understanding* of their treatment options, including clarifying misconceptions:

> *Whether it's clinical trials or normal treatment, it's important to explore [the patient's] understanding of what's going on, then you can kind of course correct, educate them on what actually might be happening and maybe where there'd be misconceptions.* [P7]

PCP as Interpreter

PCPs described taking on an interpreter role that centered on enhancing patients' comprehension of medical information. PCPs described being a "go-between" (ie, a bridge) by acting as intermediaries who can interpret information provided to patients by oncologists and other cancer specialists:

> *[Patients] should come to me if they have questions, concerns, don't understand something, because I am great at being a go-between in terms of reading notes and/or contacting oncologists, surgeons, interpreting things for them...and analyzing their options with them in ways that they understand.* [P5]

PCPs also recognized their role in *translating medical information*, such as information related to medications and laboratory results, to "kind of frame [the information] in terms the patient may know, just because some of the language isn't as clear all the time" (PCP8). They recognized the value of making information more accessible for patients, although they were not cancer specialists themselves:

> *As a patient, it's very overwhelming...but that's kind of like why I'm here. I can try to interpret, maybe, better what's going on...Like medications. What is it? What does it do? I mean, obviously, I'm not a specialist. I don't know the ins and outs of everything, but I think kind of interpreting labs, a patient, a lot of times, they don't [understand]...Just maybe a more kind of hands-on understanding for the patient.* [P2]

PCPs Navigating Expertise Boundaries

Finally, despite completing the training intervention, PCPs emphasized that they were still struggling with expertise boundaries. They admitted that their *lack of specialized knowledge* in oncology made them cautious about discussing cancer-related topics in detail:

> *I think the challenges as just a PCP...is that I just don't have that knowledge of oncology, in specific. I don't know enough about that specific clinical trial or that specific diagnosis, the prognosis, and things like that...Due to the lack of expertise, it makes it hard to communicate.* [P1]

PCPs further highlighted feeling the constraints of *role limitations* in guiding patients through complex cancer treatment decisions:

> *Oftentimes the questions [that patients are] asking about the treatments are very nuanced...I think there's only so much that the primary care doctor can answer.* [P7]

Communication About Treatment Options

Overview

The survey findings demonstrated that the training intervention enhanced PCPs' *willingness to communicate* with their patients about cancer treatment options in general, including CCTs (Table 7). Mean scores across the 6 items increased significantly from before the training (mean 4.4, SD 0.5) to immediately after the training (mean 4.7, SD 0.5; P=.007). This improvement was not sustained at the 3-month follow-up. However, 1 item—PCPs' willingness to "educate these patients about the possibility of receiving treatment within a cancer clinical trial"—showed significant improvement at the 3-month follow-up (mean 4.4, SD 0.6; P=.04) compared to before the training (mean 4.0, SD 0.8).

Table 7. Comparison of pretraining and posttraining scores on the willingness to communicate survey[a].

Questions[b]	Before the training (n=29), mean (SD)	After the training (n=29), mean (SD)	3-mo follow-up (n=28), mean (SD)	*P* value[c] (before the training vs after the training)	*P* value[c] (before the training vs 3-mo follow-up)
"I am willing to explore patients' concerns about cancer treatment."	4.4 (0.5)	4.7 (0.5)	4.5 (0.5)	*.01* [d]	.60
"I am willing to educate these patients that there is often more than one option for treatment."	4.4 (0.7)	4.7 (0.6)	4.5 (0.6)	*.04*	.70
"I am willing to educate these patients about the possibility of receiving treatment within a cancer clinical trial."	4.0 (0.8)	4.6 (0.6)	4.4 (0.6)	*.003*	*.04*
"I am willing to encourage these patients to ask questions of the specialist about different treatment options."	4.6 (0.6)	4.7 (0.5)	4.4 (0.7)	.30	.30
"I am willing to encourage these patients to ask questions of the specialist about receiving treatment within a cancer clinical trial."	4.4 (0.6)	4.7 (0.5)	4.5 (0.5)	.14	.60
"I am willing to emphasize my role as a partner throughout these patients' cancer care."	4.6 (0.5)	4.7 (0.5)	4.6 (0.6)	.20	.80
Mean of aggregated willingness to communicate questions	4.4 (0.5)	4.7 (0.5)	4.5 (0.5)	*.007*	.50

[a]Mean of aggregated willingness to communicate questions: before the training, mean 4.4 (SD 0.5); after the training, mean 4.7 (SD 0.5); 3-mo follow-up, mean 4.5 (SD 0.5); before the training versus after the training *P*=.007*;* before the training versus 3-mo follow-up *P*=.50. Survey questions were scored on a 5-point scale ranging from 1=*strongly disagree* to 5=*strongly agree*.

[b]Cronbach α value for the 6 willingness to communicate questions was 0.91.

[c]*P* values were calculated using the Wilcoxon signed rank test with continuity correction, including only the participants who completed both pre- and posttraining surveys.

[d]Italicization indicates a statistically significant *P* value.

The survey results also showed that the training intervention influenced PCPs' self-reported *patient communication and referral practices* immediately after the training and at the 3-month follow-up. As shown in Table 8, seven items on the patient communication and referral practices measure focused on the 5Es model from the curriculum (ie, explore, educate, encourage, engage in planning, and emphasize partnership). We asked PCPs about their communication in two different contexts of seeing patients: (1) before making an initial patient referral and (2) after patients returned following their referral appointment. After the training, PCPs reported improvements in both contexts. Specifically, they reported discussing cancer treatment options with a higher percentage of patients in the "before making a referral" context (ie, context 1). Overall, 6 (86%) of the 7 items showed improvement, with the most significant changes seen on the 2 items focusing specifically on CCTs. PCPs reported educating a higher percentage of patients "about receiving treatment with a cancer clinical trial" 3 months after completing the training intervention (before the training: mean 15%, SD 20.2%; 3-mo follow-up: mean 48%, SD 30.5%; *P*<.001). They also reported educating a higher percentage of patients to "encourage inquiry about receiving treatment through a cancer clinical trial" (before the training: mean 22%, SD 29.4%; 3-mo follow-up: mean 62%, SD 34.9%; *P*<.001). When surveyed about their behaviors after patients returned following their referral appointment (ie, context 2), these same 2 items were the only ones that showed significant improvement.

Table 8. Comparison of pretraining and 3-month follow-up scores on the patient communication and referral practices survey.

Questions (regarding patients diagnosed with cancer in the past 3 mo)	Before the training (n=29), mean (SD)	3-mo follow-up (n=28), mean (SD)	*P* value[a]	Illustrative quotes
"**Prior to making a referral, approx. what percentage of patients did you:"**[b]				
"Explore concerns about cancer treatment in general"	59.8 (35.9)	75.4 (27.1)	*.03* [c]	"[The course] reaffirmed the importance of leveraging the primary care relationship...A lot of my patients circle back to ask about my opinions before starting a specific type of cancer treatment." [P3]
"Educate about cancer treatment in general"	54.8 (32.7)	73.4 (34.2)	*.006*	"[Patients are] meeting these oncologists for the first time and...these are people I've known for years, so it makes sense that I also still have a role in making sure that they understand clinical trials, but also the treatment options that have been made available to them by the oncologist." [P7]
"Educate about receiving treatment with a cancer clinical trial"	15.2 (20.2)	47.9 (30.5)	*<.001*	"[Before the course] I didn't realize that PCPs can also play a role and influence patient decisions [about clinical trials]." [P1]
"Encourage inquiry about treatment options"	68.8 (37.6)	83.8 (28.0)	*.04*	"The important role of the PCP, that was interesting to me. I thought...the oncologist probably had more impact. But I can see that if you have a good relationship with your patient, then you can also encourage them one way or the other, or maybe as a second opinion, to help them figure out what they want for their treatment plan." [P1]
"Encourage inquiry about receiving treatment through a cancer clinical trial"	21.7 (29.4)	61.6 (34.9)	*<.001*	"[I now remember] to emphasize and mention the possibility of clinical trials when they see their specialists, particularly when I'm working with minority patients to...remind them to consider [clinical trials] and to ask" [P5]
"Engage in helping patients plan these inquiries with a specialist"	48.3 (37.2)	70.2 (30.2)	*.007*	"Putting on people's radars...the availability of trials and asking patients to talk to specialists about the potential to participate in trial was valuable." [P5]
"Emphasize partnership as patients go through cancer care"	80.7 (32.6)	83.6 (29.4)	.60	"I kind of felt like I needed to do a better job on being more intentional about the follow-ups...Maybe having more check-ins to see how the treatment is going, how their mental health is doing and stuff like that is important. So that was one thing I've been trying to do is schedule more follow-ups during their active cancer treatment." [P3]
"**After patients returned following their initial referral, approx. what percentage did you:"**[b]				
"Explore concerns about cancer treatment in general"	62.3 (38.6)	53.4 (44.3)	.40	"What I've noticed is a patient will get diagnosed with something...then they'll come to us as a primary care doctor and sort of ask questions, like, they want us to help guide them through the decision-making process, which is great." [P7]
"Educate about cancer treatment in general"	56.9 (35.0)	58.1 (42.6)	.60	"In a couple of cases [my patient] point blank said, 'You've explained my diagnosis and my treatment options better than the other people that I talked to...I felt very prepared with asking [my oncologist] about specific treatment options.'" [P4]
"Educate about receiving treatment with a cancer clinical trial"	23.3 (32.4)	50.4 (40.7)	*.003*	"[A takeaway for me was] the involvement of a PCP in helping make some of those decisions and helping patients feel open to [cancer clinical trials]. I do get the sense of encouraging the patient or empowering them a little bit and supporting them in terms of whatever their [treatment] decision might be. It was helpful to see that there is a true role for that." [P8]
"Encourage inquiry about treatment options"	59.6 (37.8)	64.4 (41.5)	.70	"I think...just presenting the different treatment options for the patients really got them to feel more comfortable asking questions when they were ready to see their oncologist." [P4]
"Encourage inquiry about receiving treatment through a cancer clinical trial"	19.6 (25.9)	55.9 (43.1)	*<.001*	"I actually encourage [my patients] to ask more treatment discussion and decision-making questions with their oncologist. I just explain to them that, you should write [questions] down in ways that they'll remember, right?" [P7]
"Engage in helping patients plan these inquiries with a specialist"	47.8 (40.2)	61.0 (42.7)	.13	"I try to use that connection and that existing relationship to help patients feel comfortable in the decisions they're making and then also to support the specialists in our academic institution to try to help them also build confidence in the new specialists that they're seeing." [P3]

Questions (regarding patients diagnosed with cancer in the past 3 mo)	Before the training (n=29), mean (SD)	3-mo follow-up (n=28), mean (SD)	*P* value[a]	Illustrative quotes
"Emphasize partnership as patients go through cancer care"	73.2 (36.0)	69.5 (41.4)	.70	"I think what stood out to me is there's an emphasis on making sure patients understand their diagnosis...and the importance of following up...and [that] the oncologist is meeting them probably for the first or second time when you're talking about clinical trials and cancer treatments." [P7]

[a]*P* values were calculated using the Wilcoxon signed rank test with continuity correction, including only the participants who completed both pre- and posttraining surveys.

[b]Each question was answered as a percentage, ranging from 0 to 100.

[c]Italicization indicates a statistically significant *P* value.

The findings from the interviews expand on these results by illustrating how PCPs' approach to patient communication evolved after the course. Collectively, they reported an increased sense of self-efficacy, feeling more comfortable and empowered in discussing cancer treatment options with patients. PCPs emphasized the importance of three goals in their approach: (1) engage patients, (2) empower patients, and (3) promote equity. PCPs described using specific communication approaches to achieve each of these goals.

Engage Patients

First, PCPs described a shift in practice after taking the course, working to engage patients in informed discussions about treatment options, including CCTs, rather than deferring these conversations to oncologists. They emphasized taking a proactive approach by initiating discussions about treatment options:

> *[The course] helped me to be more aware [of treatment options], so I can at least have that conversation with the patient.* [P2]

> *[The course] certainly opened my eyes to the fact that [treatment options] should be mentioned, something I would not have previously done.* [P8]

PCPs also highlighted encouraging consideration of clinical trials when communicating with patients about their cancer treatment options:

> *After the course, I'm just more likely to bring up clinical trials, whether there's a clinical trial [available] or not.* [P7]

> *My approach, it changed a little bit, to the point that I'm more willing to talk to [patients] a little bit more about [CCTs], rather than just immediately just defer everything to oncology.* [P1]

Empower Patients

Next, PCPs highlighted their efforts in empowering patients to "take ownership of their own health care" (PCP4). PCPs discussed prompting patients to thoroughly *explore treatment options*, ensuring that they are aware of the possibilities available to them:

> *I...either encourage them to at least take a look at what that [treatment] option is for them or encourage them to deep-dive into that option.* [P1]

Moreover, PCPs emphasized encouraging their patients to *ask specialists questions about treatment options*:

> *[I tell patients] "I can answer as many questions as I can, but the ones that I can't...you can go back to the oncologist and ask them." So, I encourage them to ask more treatment discussion and decision-making questions with their oncologist.* [P7]

Promote Equity

Finally, PCPs discussed being more aware of the importance of promoting equity when discussing cancer treatment options. PCPs described *tailoring patient communication* to incorporate patients' diverse needs and to actively engage those who are part of groups who are underrepresented in clinical trials:

> *[Before the course] I didn't realize...that there is a discrepancy on, for example, race, and those that participate in clinical trials...When I see that patient, I now know that I should actively engage them and talk to them about clinical trials and prime them before they go in to see the oncologist, for example...If you know where...the lack of utilization is, you can kind of intervene and make it more of an active conversation with the patients.* [P7]

PCPs also highlighted being "far more cognizant" (PCP5) after the course about *facilitating patients' access* to treatment options to ensure "patients are getting offered the same [care] as everyone else" (PCP2). A PCP emphasized as follows:

> *I was appreciative that the [course] modules focused on [disparities in CCT participation], and I am trying to be more equitable in the care that I provide to my patients as well, by being sure that I offer all the treatment options...You never know how much that could change someone's life.* [P4]

PCPs' Practice-Level Changes

Overview

We also examined PCPs' willingness or ability to enact practice-level changes when referring patients to cancer specialists (Table 9). At the posttest assessment, PCPs reported high levels of willingness to enact these changes (mean 4.3, SD 0.6). However, at the 3-month follow-up, PCPs reported lower levels of agreement, although still above midpoint, with statements indicating that they had had these conversations (mean 3.7, SD 0.8). Notably, although 24 (83%) of the 29

participating PCPs completed the course's interactive action plan, none of those interviewed reported printing the plan for future reference, which was provided as a direction in the course. Of the 8 PCPs who were interviewed, 5 (63%) stated that they did not have access to a printer, 2 (25%) mentioned being too busy, and 1 (12%) did not provide an answer.

Table 9. Comparison of pretraining and posttraining and 3-month follow-up scores on the willingness to change survey[a].

Assessments and questions[b]	Scores, mean (SD)
Posttraining survey (n=29)[c]	
"I am willing to make needed changes to whom I refer my patients for cancer care, in order to improve their access to cancer treatment trials."	4.3 (0.7)
"I am willing to talk with colleagues in my practice about needed changes in how we educate our patients with cancer prior to referral."	4.3 (0.7)
"I am willing to talk with colleagues in my practice about needed changes to whom we refer our patients for cancer care, in order to improve their access to cancer treatment trials."	4.3 (0.6)
3-mo follow-up survey (n=28)[d]	
"I have made changes to whom I refer my patients for cancer care, in order to improve their access to cancer treatment trials."	3.9 (1.0)
"I have talked with colleagues in my practice about needed changes in how we educate our patients with cancer prior to referral."	3.7 (1.0)
"I have talked with colleagues in my practice about needed changes to whom we refer our patients for cancer care, in order to improve their access to cancer treatment trials."	3.5 (0.9)

[a]Survey questions were scored on a 5-point scale ranging from 1=strongly disagree to 5=strongly agree.

[b]Cronbach α values for the 3 willingness to change questions was 0.9.

[c]Mean of aggregated willingness to change questions=4.3 (SD 0.6).

[d]Mean of aggregated willingness to change questions=3.7 (SD 0.8).

The qualitative findings further illuminate the underlying factors contributing to PCPs' hesitancy or unwillingness to implement practice-level changes. Most PCPs reported facing systemic barriers, specifically highlighting (1) *limited interprofessional knowledge sharing* and (2) *organizational constraints*.

Limited Interprofessional Knowledge Sharing

First, PCPs cited how limited interprofessional knowledge sharing within their practice environments was a barrier to discussing and enacting change for improving cancer care and referral practices. Instead, PCPs shared, "we kind of just do our own thing" (PCP2). PCPs emphasized that a *lack of collaboration opportunities* made informal or formal discussions within their practice settings scarce:

> *Part of the course involved talking with colleagues about best ways to advocate for our patients and potentially system-change stuff for advising patients about clinical trials...Loved the idea, but...We typically work at our individual desks through lunch and rarely get together and talk...When we do have staff meetings, they are typically putting out the biggest fires then the smallest fires, rather than things that would address this particular topic.* [P5]

This challenge was compounded by *time constraints* as PCPs noted the difficulty in finding moments amid busy schedules to discuss referral practices or share insights from training:

> *In clinic we're just always so busy. There aren't really many opportunities for any major interprofessional collaboration. I mean, if you have a medical question or something, of course, yes, you can ask your colleague, but as far as just sharing information, usually the days are pretty jam-packed.* [P6]

Organizational Constraints

Second, PCPs described being limited by *organizational constraints*, particularly within large institutions where high specialization diminishes the perceived need to refine referral practices:

> *We're at a large academic institution, things are so specialized that we just assume our patients are going to get very high-quality cancer care...often our main objective is to get our patients to the oncologist as quickly as possible.* [P3]

Institutional policies such as set referral practices further reduce the flexibility needed to refine these processes:

> *We do get pretty stuck here with referring. We have oncologists at [my institution] that we refer to and then if they feel they don't have adequate options available, the oncologist usually refers the patient out [to another institution] which is right across the street from us. So, it is sort of a closed community.* [P4]

PCPs further expressed having *limited power* to make changes within their current roles, pointing to obstacles such as systemic barriers and hierarchical structures:

> *The health care system and the clinic schedule, in general, are very overwhelming, and there are a lot of burdens, as the PCP...So, I think it's the system structure, in general, makes it difficult to make the*

> *changes, and it's just not encouraging enough to really motivate me to make the changes.* [P1]

Another PCP cited hierarchical barriers as limiting their ability to effect change:

> *Because I'm a resident, this is a little bit trickier, honestly. I think when I'm attending, I'll have more structural influence on how things are practiced.* [P7]

Finally, *fiscal policies* impacted PCPs' ability to refine their referral practices:

> *Some of our referral practices are going to be based off what type of insurance [a patient has]. Or for us, like I mentioned at [my institution], our clinical decision-making is oftentimes within the realm of what [our institution] is willing to cover.* [P4]

Discussion

Principal Findings

This mixed methods pilot study provides a comprehensive evaluation of an innovative web-based training intervention aimed at enhancing PCPs' knowledge, attitudes, communication, and practices regarding referrals to cancer care and CCTs. The Kirkpatrick evaluation model [46] provided a helpful framework for these evaluation findings. It is widely used in medical education to assess the outcomes of training and learning programs.

Level 1 of the Kirkpatrick evaluation model focuses on user reaction to the training intervention. Participants had high levels of satisfaction with the training intervention and reported high scores and positive comments about various aspects. Although satisfaction is important for participant engagement, it is not necessarily sufficient to produce desired changes.

Kirkpatrick level 2 focuses on learning as the next important outcome. Before the intervention, PCPs demonstrated relatively low levels of CCT knowledge, with a mean score of 55% (SD 17.9%) across 7 knowledge items, consistent with our earlier assessments of PCPs' understanding of clinical trial concepts [40]. We found that PCPs' knowledge about CCTs improved significantly after the intervention and was sustained at the 3-month follow-up, replicating what our team found in a prior study of PCPs in the New York City area [40]. This improvement is particularly important given PCPs' previously self-identified limitations in cancer-specific knowledge [33,41]. Our evaluation also found an improvement in PCPs' self-reported attitudes and beliefs about their role in working with patients with cancer, which was sustained at the 3-month follow-up, as well as willingness to communicate with their patients about CCTs, although this willingness to communicate was not sustained at the 3-month follow-up. However, willingness may be less important than actual behaviors.

With a foundation of change in knowledge, attitudes, and behavioral intentions, Kirkpatrick level 3 examines the impact of the educational intervention on behavior. According to our survey data, PCPs reported using the communication skills taught in the course with a higher percentage of patients about cancer treatment and CCTs at the 3-month follow-up than they did before the training within the context of making an initial referral for cancer care. Of the 7 communication items on the PRCP, only 1 (14%)—*emphasize partnership as patients go through cancer care*—did not show an improvement in scores. There may have been a ceiling effect on this item because the baseline score was already high. With respect to communication skills with patients who had already met with an oncologist, only 2 (29%) of the 7 items on the PRCP saw significant increases, and these 2 items focused specifically on the discussion of CCTs as a potential treatment option. It may be that the other items are not as relevant for a patient who has already had their initial meeting with the oncologist.

Another type of behavior we assessed was participants' willingness to change their referral behavior after the training and the reports of these behavior changes at the 3-month follow-up. As these are slightly different concepts, we did not compare them; however, it was not unexpected to see lower scores on actual behavior change than intention.

Overall, the qualitative data underscored the positive impact that the training intervention had on the participating PCPs in terms of their evaluation of the intervention, their learning, and their behavior change. The qualitative data highlighted ways in which the intervention facilitated a more informed and patient-centered approach to care by increasing PCPs' self-efficacy and confidence.

In addition to these positive outcomes, PCPs expressed some barriers to changing their CCT communication behaviors. They voiced concerns about their perceived lack of cancer expertise despite the course addressing the boundary of what PCPs should be expected to discuss. In our prior work, PCPs expressed a similar reluctance to discuss cancer treatments with patients [40], further highlighting a need to expand education regarding the role of PCPs in encouraging patients to engage in decision-making about their cancer treatment, including the consideration of CCTs. Importantly, the training does not suggest that PCPs determine patient eligibility for specific trials but rather prepares them to initiate supportive, patient-centered conversations that help patients feel empowered to ask questions and engage in screening discussions with their oncology team. PCPs can also support patients by reviewing laboratory or imaging results, helping to translate complex findings into accessible language, and explaining how these may relate to cancer treatment decisions or clinical trial eligibility without making determinations about eligibility themselves.

PCPs also relayed broader systemic challenges in their work settings that hindered making behavioral changes. The interview data helped to explain the lower scores on the willingness to change referral practices, as PCPs discussed the systematic barriers hindering collaborative efforts among health care professionals; these barriers in turn hindered their ability to integrate new knowledge and practices into their clinical settings.

Implications for Practice

Our findings underscore the need for ongoing education to address gaps in the PCPs' knowledge and skills needed to effectively communicate with patients about the potential of

participating in CCTs as a high-quality treatment option. The evaluation of this educational intervention suggests that sustained change in knowledge, attitudes, and behavior can be achieved through an engaging, self-directed web-based intervention. As this training intervention is easily scalable, further dissemination may be able to positively impact quality of care and participation in clinical trials.

Limitations and Future Research

As this was a pilot study, the sample size was small, although not too small to have enough power to detect significant results. Notably, the majority of the participants (n=17, 59%) were trainees working in academic settings, which may influence the generalizability of findings to the broader population of practicing PCPs. Furthermore, their receptiveness to the training and self-reported outcomes may not fully reflect the attitudes and behaviors of more experienced PCPs. Future research using a larger and more diverse sample of practitioners, including both trainees and practicing PCPs, would allow us to look for differences in evaluation metrics among different types of PCPs (residents, practicing physicians, physician assistants, and NPs). This study is also limited by its reliance on self-reported measures. Future research could incorporate patient interviews to provide additional insights into the effectiveness of the intervention from the patients' perspectives. Using larger, more diverse samples to validate the intervention's effectiveness across different practice settings and demographics is also recommended. Furthermore, including more follow-up surveys (eg, at 6 mo and 12 mo) could evaluate the sustained impact of educational interventions on PCPs' knowledge retention and clinical practices over time. While this study identified several systemic barriers that may hinder CCT recruitment, addressing these challenges was beyond the scope of this pilot intervention. Future research should explore multilevel strategies to overcome these barriers and support sustained changes in clinical practice. This study, while focused on oncology referrals, is intended to be proof of concept for an intervention that can be adapted for other disease areas requiring specialist referrals, such as cardiology, rheumatology, and gastroenterology.

Conclusions

This pilot study found that a self-guided, 1-hour web-based training intervention for PCPs improved their knowledge about CCTs as well as their attitudes and beliefs regarding their role in discussing treatment options with patients. The intervention also enhanced PCPs' ability to communicate with patients about CCTs and prepare them for subsequent steps in the referral process, including discussions with specialists. Further testing of the training intervention in a larger sample can lead to future dissemination, with the potential to improve CCT accrual, especially among underrepresented groups. Furthermore, this study underscores the effectiveness of targeted educational interventions in equipping PCPs with the knowledge and confidence to communicate with patients about the potential of trial participation as a high-quality treatment option.

Acknowledgments

This research study was sponsored in part by the Leukemia & Lymphoma Society, AbbVie Inc, Amgen Inc, AstraZeneca Pharmaceuticals LP, and Bristol Myers Squibb.

Authors' Contributions

CLB, MM, ESW, and MS-H conceptualized the study. NDP and A Crowe curated the data. NDP, CLF, CLB, JHL, and ZZ were responsible for formal analysis. CLB, ESW, and MS-H were responsible for funding acquisition. A Crowe was responsible for investigation. CLB, MM, A Crowe, ESW, MS-H, JA, DD, RMV, AN-P, and TSV were responsible for methodology. CLB, A Crowe, and A Cassells were responsible for project administration. CLB was responsible for supervision. NDP and ZZ were responsible for visualization. NDP, CLF, CLB, and A Crowe wrote the original draft. All authors reviewed and edited the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full surveys, including demographic questions.
[DOCX File , 23 KB - mededu_v11i1e66892_app1.docx]

References

1. Institute of Medicine. A National Cancer Clinical Trials System for the 21st Century: Reinvigorating the NCI Cooperative Group Program. Washington, DC: The National Academies Press; 2010.
2. Unger JM, Shulman LN, Facktor MA, Nelson H, Fleury ME. National estimates of the participation of patients with cancer in clinical research studies based on Commission on Cancer Accreditation data. J Clin Oncol 2024 Jun 20;42(18):2139-2148 [FREE Full text] [doi: 10.1200/JCO.23.01030] [Medline: 38564681]

3. Oyer RA, Hurley P, Boehmer L, Bruinooge SS, Levit K, Barrett N, et al. Increasing racial and ethnic diversity in cancer clinical trials: an American Society of Clinical Oncology and Association of Community Cancer Centers joint research statement. J Clin Oncol 2022 Jul 01;40(19):2163-2171. [doi: 10.1200/JCO.22.00754] [Medline: 35588469]
4. Townsley CA, Selby R, Siu LL. Systematic review of barriers to the recruitment of older patients with cancer onto clinical trials. J Clin Oncol 2005 May 01;23(13):3112-3124. [doi: 10.1200/JCO.2005.00.141] [Medline: 15860871]
5. Avis NE, Smith KW, Link CL, Hortobagyi GN, Rivera E. Factors associated with participation in breast cancer treatment clinical trials. J Clin Oncol 2006 Apr 20;24(12):1860-1867. [doi: 10.1200/JCO.2005.03.8976] [Medline: 16622260]
6. Pinto HA, McCaskill-Stevens W, Wolfe P, Marcus AC. Physician perspectives on increasing minorities in cancer clinical trials: an Eastern Cooperative Oncology Group (ECOG) Initiative. Ann Epidemiol 2000 Nov;10(8 Suppl):S78-S84. [doi: 10.1016/s1047-2797(00)00191-5] [Medline: 11189096]
7. Jones JM, Nyhof-Young J, Moric J, Friedman A, Wells W, Catton P. Identifying motivations and barriers to patient participation in clinical trials. J Cancer Educ 2006 May 1;21(4):237-242. [doi: 10.1080/08858190701347838] [Medline: 17542716]
8. Melisko ME, Hassin F, Metzroth L, Moore DH, Brown B, Patel K, et al. Patient and physician attitudes toward breast cancer clinical trials: developing interventions based on understanding barriers. Clin Breast Cancer 2005 Apr;6(1):45-54. [doi: 10.3816/CBC.2005.n.008] [Medline: 15899072]
9. Meropol NJ, Buzaglo JS, Millard J, Damjanov N, Miller SM, Ridgway C, et al. Barriers to clinical trial participation as perceived by oncologists and patients. J Natl Compr Canc Netw 2007 Sep;5(8):655-664. [doi: 10.6004/jnccn.2007.0067] [Medline: 17927923]
10. Comis RL, Miller JD, Aldigé CR, Krebs L, Stoval E. Public attitudes toward participation in cancer clinical trials. J Clin Oncol 2003 Mar 01;21(5):830-835. [doi: 10.1200/JCO.2003.02.105] [Medline: 12610181]
11. Wong YN, Schluchter MD, Albrecht TL, Benson AB, Buzaglo J, Collins M, et al. Financial concerns about participation in clinical trials among patients with cancer. J Clin Oncol 2016 Feb 10;34(5):479-487 [FREE Full text] [doi: 10.1200/JCO.2015.63.2463] [Medline: 26700120]
12. Barriers to patient enrollment in therapeutic clinical trials for cancer. American Cancer Society Cancer Action Network. 2018 Apr 11. URL: https://www.fightcancer.org/policy-resources/barriers-patient-enrollment-therapeutic-clinical-trials-cancer [accessed 2025-06-10]
13. George S, Duran N, Norris K. A systematic review of barriers and facilitators to minority research participation among African Americans, Latinos, Asian Americans, and Pacific Islanders. Am J Public Health 2014 Feb;104(2):e16-e31. [doi: 10.2105/AJPH.2013.301706] [Medline: 24328648]
14. Senft N, Hamel LM, Manning MA, Kim S, Penner LA, Moore TF, et al. Willingness to discuss clinical trials among Black vs White men with prostate cancer. JAMA Oncol 2020 Nov 01;6(11):1773-1777 [FREE Full text] [doi: 10.1001/jamaoncol.2020.3697] [Medline: 32940630]
15. Simon MS, Du W, Flaherty L, Philip PA, Lorusso P, Miree C, et al. Factors associated with breast cancer clinical trials participation and enrollment at a large academic medical center. J Clin Oncol 2004 Jun 01;22(11):2046-2052. [doi: 10.1200/JCO.2004.03.005] [Medline: 15082724]
16. Joseph G, Dohan D. Diversity of participants in clinical trials in an academic medical center: the role of the 'Good Study Patient?'. Cancer 2009 Feb 01;115(3):608-615 [FREE Full text] [doi: 10.1002/cncr.24028] [Medline: 19127544]
17. Ford JG, Howerton MW, Lai GY, Gary TL, Bolen S, Gibbons MC, et al. Barriers to recruiting underrepresented populations to cancer clinical trials: a systematic review. Cancer 2008 Jan 15;112(2):228-242 [FREE Full text] [doi: 10.1002/cncr.23157] [Medline: 18008363]
18. Wendler D, Kington R, Madans J, Van Wye G, Christ-Schmidt H, Pratt LA, et al. Are racial and ethnic minorities less willing to participate in health research? PLoS Med 2006 Feb 6;3(2):e19 [FREE Full text] [doi: 10.1371/journal.pmed.0030019] [Medline: 16318411]
19. Murthy VH, Krumholz HM, Gross CP. Participation in cancer clinical trials: race-, sex-, and age-based disparities. JAMA 2004 Jun 09;291(22):2720-2726. [doi: 10.1001/jama.291.22.2720] [Medline: 15187053]
20. Cancer disparities. National Institutes of Health National Cancer Institute. URL: https://www.cancer.gov/about-cancer/understanding/disparities [accessed 2025-06-10]
21. Eggly S, Barton E, Winckles A, Penner LA, Albrecht TL. A disparity of words: racial differences in oncologist-patient communication about clinical trials. Health Expect 2015 Oct 02;18(5):1316-1326 [FREE Full text] [doi: 10.1111/hex.12108] [Medline: 23910630]
22. Meropol NJ, Wong Y, Albrecht T, Manne S, Miller SM, Flamm AL, et al. Randomized trial of a web-based intervention to address barriers to clinical trials. J Clin Oncol 2016 Feb 10;34(5):469-478 [FREE Full text] [doi: 10.1200/JCO.2015.63.2257] [Medline: 26700123]
23. O'Connor AM, Tugwell P, Wells GA, Elmslie T, Jolly E, Hollingworth G, et al. Randomized trial of a portable, self-administered decision aid for postmenopausal women considering long-term preventive hormone therapy. Med Decis Making 1998;18(3):295-303. [doi: 10.1177/0272989X9801800307] [Medline: 9679994]
24. Patient decision aids developed by or with our research team. Ottawa Hospital Research Institute. URL: https://decisionaid.ohri.ca/decaids.html [accessed 2025-06-10]

25. Hickner J, Kent S, Naragon P, Hunt L. Physicians' and patients' views of cancer care by family physicians: a report from the American Academy of Family Physicians National Research Network. Fam Med 2007 Feb;39(2):126-131. [Medline: 17273955]
26. Klabunde CN, Ambs A, Keating NL, He Y, Doucette WR, Tisnado D, et al. The role of primary care physicians in cancer care. J Gen Intern Med 2009 Sep 14;24(9):1029-1036 [FREE Full text] [doi: 10.1007/s11606-009-1058-x] [Medline: 19597893]
27. Dossett LA, Hudson JN, Morris AM, Lee MC, Roetzheim RG, Fetters MD, et al. The primary care provider (PCP)-cancer specialist relationship: a systematic review and mixed-methods meta-synthesis. CA Cancer J Clin 2017 Mar 11;67(2):156-169 [FREE Full text] [doi: 10.3322/caac.21385] [Medline: 27727446]
28. Howard JM, DeMets D. How informed is informed consent? The BHAT experience. Control Clin Trials 1981 Dec;2(4):287-303. [doi: 10.1016/0197-2456(81)90019-2] [Medline: 6120794]
29. Ellis PM, Butow PN, Tattersall MH, Dunn SM, Houssami N. Randomized clinical trials in oncology: understanding and attitudes predict willingness to participate. J Clin Oncol 2001 Aug 01;19(15):3554-3561. [doi: 10.1200/JCO.2001.19.15.3554] [Medline: 11481363]
30. Grunfeld E, Zitzelsberger L, Coristine M, Aspelund F. Barriers and facilitators to enrollment in cancer clinical trials: qualitative study of the perspectives of clinical research associates. Cancer 2002 Oct 01;95(7):1577-1583 [FREE Full text] [doi: 10.1002/cncr.10862] [Medline: 12237928]
31. Kinney AY, Richards C, Vernon SW, Vogel VG. The effect of physician recommendation on enrollment in the Breast Cancer Chemoprevention Trial. Prev Med 1998 Sep;27(5 Pt 1):713-719. [doi: 10.1006/pmed.1998.0349] [Medline: 9808803]
32. Umutyan A, Chiechi C, Beckett LA, Paterniti DA, Turrell C, Gandara DR, et al. Overcoming barriers to cancer clinical trial accrual: impact of a mass media campaign. Cancer 2008 Jan 01;112(1):212-219 [FREE Full text] [doi: 10.1002/cncr.23170] [Medline: 18008353]
33. Lawrence RA, McLoone JK, Wakefield CE, Cohn RJ. Primary care physicians' perspectives of their role in cancer care: a systematic review. J Gen Intern Med 2016 Oct 24;31(10):1222-1236 [FREE Full text] [doi: 10.1007/s11606-016-3746-7] [Medline: 27220499]
34. Klabunde CN, Han PK, Earle CC, Smith T, Ayanian JZ, Lee R, et al. Physician roles in the cancer-related follow-up care of cancer survivors. Fam Med 2013;45(7):463-474 [FREE Full text] [Medline: 23846965]
35. Sprague Martinez L, Freeman E, Winkfield KM. Perceptions of cancer care and clinical trials in the Black community: implications for care coordination between oncology and primary care teams. Oncologist 2017 Sep;22(9):1094-1101 [FREE Full text] [doi: 10.1634/theoncologist.2017-0122] [Medline: 28706009]
36. Sateren WB, Trimble EL, Abrams J, Brawley O, Breen N, Ford L, et al. How sociodemographics, presence of oncology specialists, and hospital cancer programs affect accrual to cancer treatment trials. J Clin Oncol 2002 Apr 15;20(8):2109-2117. [doi: 10.1200/JCO.2002.08.056] [Medline: 11956272]
37. Silberlust J, Suarez MM, Caban-Martinez AJ. Disparities in clinical trial participation and the influence of physician specialty. Clin Trials 2021 Feb 14;18(1):127-129. [doi: 10.1177/1740774520956578] [Medline: 32921166]
38. Barnes EA, Hanson J, Neumann CM, Nekolaichuk CL, Bruera E. Communication between primary care physicians and radiation oncologists regarding patients with cancer treated with palliative radiotherapy. J Clin Oncol 2000 Aug 15;18(15):2902-2907. [doi: 10.1200/JCO.2000.18.15.2902] [Medline: 10920139]
39. Baquet CR, Commiskey P, Daniel Mullins C, Mishra SI. Recruitment and participation in clinical trials: socio-demographic, rural/urban, and health care access predictors. Cancer Detect Prev 2006 Jan;30(1):24-33 [FREE Full text] [doi: 10.1016/j.cdp.2005.12.001] [Medline: 16495020]
40. Bylund CL, Weiss ES, Michaels M, Patel S, D'Agostino TA, Peterson EB, et al. Primary care physicians' attitudes and beliefs about cancer clinical trials. Clin Trials 2017 Oct 11;14(5):518-525 [FREE Full text] [doi: 10.1177/1740774517717722] [Medline: 28693389]
41. Michaels M, D'Agostino TA, Blakeney N, Weiss ES, Binz-Scharf MC, Golant M, et al. Impact of primary care provider knowledge, attitudes, and beliefs about cancer clinical trials: implications for referral, education and advocacy. J Cancer Educ 2015 Mar 9;30(1):152-157 [FREE Full text] [doi: 10.1007/s13187-014-0662-6] [Medline: 24805229]
42. Michaels M, Weiss ES, Guidry JA, Blakeney N, Swords L, Gibbs B, et al. "The promise of community-based advocacy and education efforts for increasing cancer clinical trials accrual". J Cancer Educ 2012 Mar 22;27(1):67-74. [doi: 10.1007/s13187-011-0271-6] [Medline: 21938600]
43. Bylund CL, Michaels M, Weiss ES, Patel S, D'Agostino TA, Binz-Scharf MC, et al. The impact of an online training program about cancer clinical trials on primary care physicians' knowledge, attitudes and beliefs, and behavior. J Cancer Educ 2021 Oct 10;36(5):1039-1044 [FREE Full text] [doi: 10.1007/s13187-020-01731-3] [Medline: 32157570]
44. Robinson MK, Tsark JU, Braun KL. Increasing primary care physician support for and promotion of cancer clinical trials. Hawaii J Med Public Health 2014 Mar;73(3):84-7; quiz 88 [FREE Full text] [Medline: 24660125]
45. Creswell JW, Creswell JD. Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Thousand Oaks, CA: SAGE Publications; 2018.
46. Hutchinson L. Evaluating and researching the effectiveness of educational interventions. BMJ 1999 May 08;318(7193):1267-1269 [FREE Full text] [doi: 10.1136/bmj.318.7193.1267] [Medline: 10231262]

47. Creswell JW, Plano Clark V. Designing and Conducting Mixed Methods Research Third Edition. Thousand Oaks, CA: SAGE Publications; 2017.
48. Guetterman TC, Fetters MD, Creswell JW. Integrating quantitative and qualitative results in health science mixed methods research through joint displays. Ann Fam Med 2015 Nov 09;13(6):554-561 [FREE Full text] [doi: 10.1370/afm.1865] [Medline: 26553895]
49. Plano Clark VL, Ivankova NV. Mixed Methods Research: A Guide to the Field. Thousand Oaks, CA: SAGE Publications; 2015.
50. Ivankova NV, Creswell JW, Stick SL. Using mixed-methods sequential explanatory design: from theory to practice. Field Methods 2006 Feb 01;18(1):3-20. [doi: 10.1177/1525822X05282260]
51. Bergman MM. Advances in Mixed Methods Research: Theories and Applications. Thousand Oaks, CA: SAGE Publications; 2008.
52. Ogrinc G, Davies L, Goodman D, Batalden P, Davidoff F, Stevens D. SQUIRE 2.0 (Standards for QUality Improvement Reporting Excellence): revised publication guidelines from a detailed consensus process. BMJ Qual Saf 2016 Dec 14;25(12):986-992 [FREE Full text] [doi: 10.1136/bmjqs-2015-004411] [Medline: 26369893]
53. Shen MJ, Binz-Scharf M, D'Agostino T, Blakeney N, Weiss E, Michaels M, et al. A mixed-methods examination of communication between oncologists and primary care providers among primary care physicians in underserved communities. Cancer 2015 Mar 15;121(6):908-915 [FREE Full text] [doi: 10.1002/cncr.29131] [Medline: 25377382]
54. Braun LT, Schmidmaier R. Dealing with cognitive dissonance: an approach. Med Educ 2019 Dec 18;53(12):1167-1168. [doi: 10.1111/medu.13955] [Medline: 31532838]
55. Klein J, McColl G. Cognitive dissonance: how self-protective distortions can undermine clinical judgement. Med Educ 2019 Dec;53(12):1178-1186. [doi: 10.1111/medu.13938] [Medline: 31397007]
56. Epstein RM, Street RLJ. Patient-Centered Communication in Cancer Care: Promoting Healing and Reducing Suffering. Bethesda, MD: National Cancer Institute; 2007.
57. Hibbard JH, Stockard J, Mahoney ER, Tusler M. Development of the Patient Activation Measure (PAM): conceptualizing and measuring activation in patients and consumers. Health Serv Res 2004 Aug;39(4 Pt 1):1005-1026 [FREE Full text] [doi: 10.1111/j.1475-6773.2004.00269.x] [Medline: 15230939]
58. Hibbard JH, Mahoney E, Sonet E. Does patient activation level affect the cancer patient journey? Patient Educ Couns 2017 Jul;100(7):1276-1279. [doi: 10.1016/j.pec.2017.03.019] [Medline: 28330715]
59. Elwyn G, O'Connor A, Stacey D, Volk R, Edwards A, Coulter A, et al. Developing a quality criteria framework for patient decision aids: online international Delphi consensus process. BMJ 2006 Aug 26;333(7565):417 [FREE Full text] [doi: 10.1136/bmj.38926.629329.AE] [Medline: 16908462]
60. Qualtrics XM: the leading experience management software. Qualtrics XM. URL: https://www.qualtrics.com [accessed 2025-06-10]
61. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2018. URL: https://www.R-project.org/ [accessed 2025-06-10]
62. ATLAS.ti scientific software development GmbH, version 9.1.7. ATLAS.ti. URL: https://atlasti.com [accessed 2025-06-10]
63. Qualitative methods in implementation science. Division of Cancer Control and Population Sciences, National Cancer Institute. 2020. URL: https://cancercontrol.cancer.gov/sites/default/files/2020-09/nci-dccps-implementationscience-whitepaper.pdf [accessed 2025-06-10]
64. Hamilton A. Rapid qualitative analysis: updates/developments. U.S. Department of Veterans Affairs. 2020 Sep 29. URL: https://www.hsrd.research.va.gov/for_researchers/cyber_seminars/archives/video_archive.cfm?SessionID=3846 [accessed 2025-06-10]
65. Glaser BG, Strauss AL. Discovery of Grounded Theory: Strategies for Qualitative Research. New York, NY: Routledge; 1999.
66. Strauss A, Corbin J. Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory. Thousand Oaks, CA: Sage Publications, Inc; 1998.
67. Corbin J, Strauss A. Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory. Thousand Oaks, CA: SAGE Publications; 2014.
68. Guetterman TC, Fàbregues S, Sakakibara R. Visuals in joint displays to represent integration in mixed methods research: a methodological review. Methods Psychol 2021 Dec;5:100080. [doi: 10.1016/j.metip.2021.100080]

Abbreviations

CCT: cancer clinical trial
NP: nurse practitioner
PCP: primary care provider
PGY: postgraduate year
SQUIRE 2.0: Standards for Quality Improvement Reporting Excellence

Edited by M Montagna; submitted 02.10.24; peer-reviewed by D Bracken-Clarke, M Rattray, A Arbabisarjou, S Nakhoda; comments to author 14.03.25; revised version received 18.04.25; accepted 30.04.25; published 17.07.25.

Please cite as:
Parker ND, Michaels M, Fisher CL, Crowe A, Weiss ES, Sae-Hau M, Arnold J, Cassells A, Durante D, Lee JH, Vega RM, Natale-Pereira A, Vasquez TS, Zhang Z, Bylund CL
A Web-Based Training Intervention for Primary Care Providers on Preparing Patients for Cancer Treatment Decisions and Conversations About Clinical Trials: Evaluation of a Pilot Study Using Mixed Methods and Follow-Up
JMIR Med Educ 2025;11:e66892
URL: https://mededu.jmir.org/2025/1/e66892
doi:10.2196/66892
PMID:

Pharmacists' Attitudes, Perceptions, and Preferences Regarding Continuing Education: Cross-Sectional Study in Vietnam

Trung Quang Vo[1*], PhD; Phuoc Duy Le[1*], BPharm; Hien Thi Bich Tran[1*], MSc; Hieu Thi Thanh Nguyen[2], BPharm; Thoai Dang Nguyen[1], PhD; Trang Nguyen Khanh Huynh[3], PhD; Bay Van Vo[1], PhD

[1]Department of Economic and Administrative Pharmacy, Faculty of Pharmacy, Pham Ngoc Thach University of Medicine, No 2. Duong Quang Trung, Hoa Hung Ward, Hochiminh City, Vietnam

[2]Social, Economic and Administrative Pharmacy (SEAP) Graduate Program, Faculty of Pharmacy, Mahidol University, Bangkok, Thailand

[3]Department of Obstetrics and Gynecology, Faculty of Medicine, Pham Ngoc Thach University of Medicine, Hochiminh City, Vietnam

[*]these authors contributed equally

Corresponding Author:
Bay Van Vo, PhD
Department of Economic and Administrative Pharmacy, Faculty of Pharmacy, Pham Ngoc Thach University of Medicine, No 2. Duong Quang Trung, Hoa Hung Ward, Hochiminh City, Vietnam

Abstract

Background: The evolution of the health care landscape necessitates expanding the roles of pharmacists in patient-centered care to encompass direct patient management, collaborative practice, and preventive service. These responsibilities can be fulfilled by pharmacists through ongoing professional development, in which continuing education (CE) is instrumental to career advancement and improved patient care.

Objective: This cross-sectional study aimed to assess Vietnamese pharmacists' attitudes, perceptions, and preferences regarding CE.

Methods: Participants were recruited via convenience and snowball sampling, after which a validated 42-item questionnaire was administered to them through online and offline channels from December 2024 to February 2025. The data were examined via descriptive statistical analysis using SPSS (version 26.0; IBM Corp). The associations between participant characteristics and attitudinal or perception scores (P<.05) were assessed using 1-way ANOVA with 1000 bootstrap samples.

Results: This study involved 508 pharmacists, most of whom were aged 25 to 30 years (n=197, 38.8%), and the majority held university degrees (n=360, 70.9%). Their mean attitudinal score was 44.4 (SD 5.5), reflecting generally positive attitudes toward CE. However, significant differences in mean attitudinal scores were found across groups categorized by education level, job position, and frequency of overtime (P<.05). More than half of the participants derived good scores on their perceptions of CE, with their preferred CE formats including computer- and internet-based learning, as well as the use of medical search engines. Finally, the pharmacists expressed a strong preference for CE topics focusing on skill development.

Conclusions: The Vietnamese pharmacists exhibited positive attitudes toward CE, favoring flexible learning formats and practical topics. These insights can inform the efforts of policymakers and educators to enhance CE accessibility, improve pharmacists' competencies, and, ultimately, advance patient care.

(JMIR Med Educ 2025;11:e77013) doi:10.2196/77013

KEYWORDS

attitudes; continuing education; pharmacist; perception; preferences; Vietnam

Introduction

Continuing education (CE) is defined by the World Health Organization (WHO) as the process of updating professional knowledge after the completion of formal education, enabling health care professionals to maintain and enhance their clinical practice [1]. In 2015, the Accreditation Council for Pharmacy Education expanded its CE framework by introducing the concept of continuing professional development (CPD), which is characterized as a self-directed, ongoing, systematic, and goal-oriented learning process embedded within professional practice [2]. CPD is to describe the broader, holistic process by which health professionals maintain and enhance their knowledge, skills, and performance throughout their careers [2]. This encompasses not only formal learning activities but also self-directed learning, quality improvement initiatives, and mentorship. CE, in contrast, refers specifically to structured, formal educational activities, such as courses, workshops, and conferences, which are a subset of CPD [2,3]. In Vietnam, the Ministry of Health has enacted CE policies since 2013 through

Circular No. 22/2013/TT-BYT, later amended through Circular No. 26/2020/TT-BYT [4]. The Vietnamese primary framework for lifelong learning is designated as CE, which includes various short-term training programs such as professional development courses, continuing medical education, and technology transfer training [5]. From a broader international perspective, this CE framework constitutes a central component of CPD, which encompasses a wider range of formal and informal learning activities. Circular No. 32/2023/TT-BYT requires a minimum of 120 CE credit hours over 5 years from the medical practitioners [6], while Decree No. 163/2025/NĐ-CP, which amends the Law on Pharmacy No. 105/2016/QH13 [7], mandates the completion of at least 8 credit hours every three years by pharmacists [8]. These regulatory distinctions reflect a progressive shift toward role-specific CE requirements for different health care professionals. However, the absence of a globally standardized definition of CE brings about variations in CE awareness, structural implementation, and regulatory frameworks across countries. While the policy framework is being constructed, there is a striking lack of empirical research into how Vietnamese pharmacists themselves perceive and experience this evolving CE system.

CE is pivotal in ensuring the continued competence of health care professionals, ultimately contributing to improved patient care and efficiency in health care systems. Among such professionals, pharmacists need to take on roles that go beyond traditional dispensing duties to include direct patient care, collaborative practice, comprehensive medication management, and preventive care service [9]. This means that CE is essential to updating clinical knowledge and skills. CE is critical for maintaining excellent pharmacy practice and optimizing health care outcomes. Its importance grows as technology advances, evidence-based medicine expands, and health care becomes increasingly globalized. CE is also a prerequisite for the renewal of pharmacist licenses, with specific credit requirements enforced in jurisdictions such as France and the United States. Similarly, the United Kingdom and Canada implement a mandatory CPD approach [10]; European countries, such as Belgium and Norway, link CE completion to salary increments; and the Netherlands, Austria, Switzerland, Spain, Hungary, and Italy treat CE as compulsory for health care professionals [11]. The same holds true for Vietnam, where pharmacists must fulfill CE requirements within 3 years of obtaining their practice certificates to maintain licensure [7].

Accordingly, assessing pharmacists' attitudes, perceptions, and preferences regarding CE is crucial for optimizing training programs and ensuring their relevance to professional practice. In this context, identifying cognitive gaps—defined as the discrepancies between pharmacists' current knowledge, skills, or perceptions and the expected competencies—serves a dual role. First, it enhances the evaluation of attitudes by revealing potential misconceptions or underrecognized needs that may influence learning motivation. Second, it informs how CE programs are designed and delivered. This ensures content and instructional strategies match real-world professional demands and learners' expectations. Several studies have examined these variables in relation to CE participation, particularly in Gulf and Middle Eastern countries [12-18]. These studies demonstrated that most pharmacists recognize CE as essential for professional development. Research in Kuwait indicated that over 60% of pharmacists exhibit positive attitudes toward CE, with workshops being the most preferred modality [16]. Enhancing pharmacists' engagement with CE requires the provision of educational content in preferred formats, such as interactive workshops or structured discussions, to improve knowledge retention and application [12]. Meanwhile, a study in Ethiopia found that 56.5% of pharmacists are unfamiliar with the concept of CE, highlighting regional disparities in CE awareness and accessibility [19].

Despite the growing importance of CE, limited evidence has been derived as to Vietnamese pharmacists' attitudes, perceptions, and preferences regarding CE. To address this deficiency, this study evaluated the aforementioned variables to contribute to essential endeavors to strengthen the national CE system in the country and improve pharmacists' access to CE opportunities. The findings will provide the first crucial evidence to inform the effective development of the national CE system, ensuring it is responsive to its end users. From a global perspective, this research offers a valuable case study of CE implementation in a developing country, contributing comparative insights into how regulatory evolution interacts with professional motivation in a unique Southeast Asian context.

Methods

Study Design

This cross-sectional study involved the distribution of a self-administered questionnaire online and offline from December 2024 to February 2025 to evaluate the attitudes, perceptions, and preferences regarding CE of pharmacists in southeastern Vietnam.

Eligibility and Sample Size

The inclusion criteria were (1) pharmacists employed in both the government and the private sector in the southeastern region of Vietnam, covering the provinces of Binh Phuoc, Binh Duong, Ba Ria-Vung Tau, Dong Nai, and Tay Ninh as well as Ho Chi Minh City, and (2) those proficient in reading and comprehending Vietnamese. Pharmacists who submitted incomplete questionnaires or provided the same answer (eg, only "A") throughout the questionnaire were excluded.

The minimum sample size was determined using the formula used by the WHO [20]:

$$N=(Z\alpha/2)2\times P(1-P)d2$$

where $Z_{\alpha/2}$ denotes normal distribution with a 95% CI ($Z_{\alpha/2}$=1.96), d is an error margin of 0.05, and P denotes the proportion of pharmacists with positive attitudes toward CE (P=.7, based on the pilot study). To account for potential exclusions, an additional 10% was incorporated into the sample size, resulting in a minimum requirement of 355 participants.

Data Collection

Questionnaire Development

A questionnaire was developed following a comprehensive review of the literature [12,13,15,16,19,21,22], after which it was translated on the basis of conceptual equivalence using a four-step process adapted from the WHO [23]: (1) two independent forward translations, of which the initial translation from the original English version into Vietnamese was performed by a professional translator, (2) a review of the first draft by an expert panel comprising five pharmacists operating in different fields relevant to the study, (3) cognitive interviews performed by 25 pharmacists who did not participate in the main research, and (4) revision based on feedback and suggestions. Face and content validity were determined in steps 2 and 3.

A pilot study was performed on a convenience sample of 30 pharmacists to identify ambiguities and ensure that the questionnaire items would yield reliable data. Revisions were made on the grounds of feedback derived during the pilot study. The data collected during this phase were excluded from the analyses carried out in the main research. The reliability of the questionnaire's subcomponents was assessed. The Cronbach α coefficients of the subscales on attitudes toward CE (14 items), perceptions regarding CE (4 items), and satisfaction with CE activities (8 items) were 0.841, 0.821, and 0.897, respectively. These values indicate good internal consistency.

Questionnaire

The final questionnaire comprised 4 sections (42 items): demographic characteristics (10 items), attitudes toward CE (14 items), perceptions regarding CE (4 items), and preferences for CE types and topics (14 items; Multimedia Appendix 1).

The demographics section was intended to assess sex, year of birth, marital status, education level, years of experience, ethnicity, organizational type, job position, frequency of overtime, and number of CE courses attended. The section on attitudes toward CE was adapted from the Jefferson Scale of Physician Lifelong Learning (JSPLL) [16]. The 14 items were ranked using a 4-point Likert scale ranging from 1=*strongly disagree* to 4=*strongly agree*. The sum of the item scores was used to compute the total attitudinal score, which was interpreted and categorized as follows: a total score ranging from 14 to 28 was regarded as indicative of poor attitudes, a score of 29 to 42 was a reflection of fair attitudes, and a score of 43 to 56 represented good attitudes.

In the section on perceptions regarding CE, statements were rated on a 5-point Likert scale ranging from 1=*strongly disagree* to 5=*strongly agree*. The sum of the item scores was used to compute the total perception score, which was interpreted and categorized into 3 levels: a total score ranging from 5 to 10 denoted poor perceptions, a score of 11 to 15 reflected fair perceptions, and a score of 16 to 20 represented good perceptions.

The section on CE preferences covered pharmacists' satisfaction with CE activities (8 items) and their interest in CE topics (6 items), as well as the perceived impact of such education on professional practice and knowledge. The items were rated on a 5-point Likert scale (*strongly disagree*, *disagree*, *neutral*, *agree*, and *strongly agree*).

Measures

Data were collected using both paper-based and online questionnaires. A combination of convenience and snowball sampling was used. Participant recruitment was conducted systematically through multiple channels to enhance sample diversity:

1. On-site recruitment: research team members directly approached pharmacists at a selected variety of community pharmacies and hospitals across the urban and suburban areas of Ho Chi Minh City and the neighboring provinces of Dong Nai and Binh Duong. At these sites, pharmacists were informed about the study and could immediately participate by scanning a QR code to access the online questionnaire or by completing a paper version.
2. Digital outreach: the survey was disseminated through closed professional social media groups and forums dedicated to Vietnamese pharmacists, accompanied by a post explaining the study's purpose.
3. Institutional distribution: the questionnaire was distributed via the official mailing lists of partner universities and professional pharmacist associations in the region. The online questionnaire was administered using Google Forms, with access provided via a QR code and a direct link. To maintain data integrity, the platform was configured to prevent multiple submissions from the same device. As part of the snowball sampling approach, all participants, regardless of recruitment channel, were explicitly encouraged to share the survey link with their pharmacist colleagues.

Data Processing and Statistical Analyses

The initial sample consisted of 550 pharmacists, among whom 42 were excluded because they provided the same response across the questionnaire. The final sample consisted of 508 pharmacists who satisfied the inclusion criteria. The data were entered directly into Microsoft Excel 2019 for cleaning and checking for missing answers.

The SPSS software (IBM Corp) was used to analyze the data. More specifically, descriptive analysis was conducted to determine frequencies and percentages for categorical variables, while means and SDs were calculated for continuous variables. A one-way ANOVA with 1000 bootstrap resamples was carried out to assess differences in the mean attitudinal and perception scores of two or more unrelated groups based on the same continuous dependent variable, with a $P<.05$ considered indicative of statistical significance.

Ethical Considerations

This study adhered to the ethical principles outlined in the Declaration of Helsinki [24] and received approval from the Scientific Research Ethics Committee at Pham Ngoc Thach University of Medicine (1237/TĐHYKPNT-HĐĐĐ). Informed consent was obtained from all participants prior to enrollment. Participants were informed of their right to withdraw at any

time without consequence. Participants did not receive any compensation for their involvement in the study. To ensure confidentiality in the online survey, the study was designed to be fully anonymous; no personally identifiable information (eg, names and IP addresses) was collected. Data were collected via a secure platform with encrypted transmission and stored on a password-protected, encrypted server accessible only to the principal investigators. In compliance with institutional data governance policy, the anonymized dataset will be retained for a period of 5 years following study completion, after which it will be permanently deleted.

Risk of Bias Assessment

A risk of bias assessment was conducted using the Appraisal Tool for Cross-Sectional Studies (AXIS; Checklist 1) [25] to determine whether bias mitigation strategies were implemented in this study rather than the provision of a numerical rating of bias risk. The assessment consisted of 20 questions directed toward the study's design, analysis, and reporting processes, with the response options available being "yes," "no," or "don't know." To enhance objectivity and reduce potential bias, 2 authors independently conducted the assessment.

Results

Sociodemographic Characteristics

Table 1 summarizes the sociodemographic characteristics of pharmacists. The majority of the 508 respondents were female (331/508, 65.2%). The most prevalent age group was 25 to 30 years, who accounted for 38.8% (197/508) of the sample. More than half of the pharmacists were employed in enterprises operating in the private sector (339/508, 66.7%). A substantial proportion of them were single (318/508, 62.6%) and had less than 5 years of professional experience (311/508, 61.2%). Most of the pharmacists attended at least 1 CE course (458/508, 90.2%).

Table . Demographic characteristics (N=508).

	Value, n (%)
Sex	
Male	177 (34.8)
Female	331 (65.2)
Age (y)	
21-24	109 (21.4)
25-29	197 (38.8)
30-34	76 (15)
≥35	126 (24.8)
Marital status	
Single	318 (62.6)
Married	190 (37.4)
Highest level of education	
Elementary or intermediate or college[a]	108 (21.2)
University	360 (70.9)
Postgraduate	40 (7.9)
Ethnic	
Kinh	486 (95.7)
Other	22 (4.3)
Organizational type	
State administrative agency system	166 (32.7)
Nongovernmental organizations	3 (0.6)
Enterprise-private system	339 (66.7)
Job position	
Staff	446 (87.8)
Manager	62 (12.2)
Years of experience	
<5	311 (61.2)
≥5	197 (38.8)
Frequency of overtime (times/wk)	
<4	457 (90)
≥4	51 (10)
Number of CE[b] courses attended	
Never	50 (9.8)
At least 1 course	458 (90.2)

[a]Educational levels in the Vietnamese context: elementary=vocational secondary training; intermediate=technical diploma; and college=3-year college degree.

[b]CE: continuing education.

Attitudes Toward CE

As shown in Table 2, participants generally held positive attitudes toward CE, with a majority (311/508, 61.2%) demonstrating good attitudes and a mean score of 44.4 (SD 5.5). There was a near-universal agreement (457/508, 89.9% combined) with the statement “I take every opportunity to gain new knowledge/skills.” A notable proportion of respondents reported disengaging from routine professional activities, with 25.2% (128/508) disagreeing that they “routinely attend meetings of pharmacy organizations” and 31.9% (162/508)

disagreeing that they “read professional journals at least once every week.”

Table . Attitudes toward continuing education (CE)[a].

Item	Pharmacists' attitude	Likert scale, n (%)				Value, mean (SD)
		1[b]	2[c]	3[d]	4[e]	
A1	Searching for the answer to a question is in and by itself rewarding.	2 (0.4)	33 (6.5)	271 (53.3)	202 (39.8)	3.3 (0.6)
A2	CE is a professional responsibility of all pharmacists.	4 (0.8)	46 (9.1)	263 (51.8)	195 (38.3)	3.3 (0.7)
A3	I enjoy reading articles in which issues of pharmacy are discussed.	2 (0.4)	56 (11.0)	310 (61.0)	140 (27.6)	3.2 (0.6)
A4	I routinely attend meetings of pharmacy organizations.	11 (2.2)	128 (25.2)	267 (52.6)	102 (20.0)	2.9 (0.7)
A5	I read professional journals at least once every week.	26 (5.1)	162 (31.9)	246 (48.4)	74 (14.6)	2.7 (0.8)
A6	I routinely search for computer databases to find out about new developments in my specialty.	11 (2.2)	60 (11.8)	297 (58.4)	140 (27.6)	3.1 (0.7)
A7	I believe that I would fall behind if I stopped learning about new developments in pharmacy.	2 (0.4)	16 (3.2)	251 (49.4)	239 (47.0)	3.4 (0.6)
A8	One of the important goals of Faculties of Pharmacy is to develop students' lifelong learning skills.	3 (0.6)	32 (6.3)	258 (50.8)	215 (42.3)	3.3 (0.6)
A9	Rapid changes in therapeutics require constant updating of knowledge and the development of new professional skills.	2 (0.4)	13 (2.6)	256 (50.4)	237 (46.6)	3.4 (0.6)
A10	I always make time for self-directed learning, even when I have a busy work schedule and other obligations.	7 (1.4)	89 (17.5)	306 (60.2)	106 (20.9)	3.0 (0.7)
A11	I recognize my need to constantly acquire new professional knowledge.	3 (0.6)	24 (4.7)	269 (53.0)	212 (41.7)	3.4 (0.6)
A12	I routinely attend CE courses to improve patient care.	14 (2.7)	121 (23.8)	263 (51.8)	110 (21.7)	2.9 (0.7)

Item	Pharmacists' attitude	Likert scale, n (%)				Value, mean (SD)
		1[b]	2[c]	3[d]	4[e]	
A13	I take every opportunity to gain new knowledge/skills that are important.	5 (1.0)	46 (9.1)	321 (63.1)	136 (26.8)	3.2 (0.6)
A14	My preferred approach in finding an answer to a question is to search for the appropriate computer databases.	6 (1.2)	47 (9.3)	285 (56.0)	170 (33.5)	3.2 (0.7)

[a]Poor attitude: 4/508 (0.8%); fair attitude: 193/508 (38%); and good attitude: 311/508 (61.2%). The average total score using bootstrap (1000 times) was 44.4 (SD 5.5).

[b]1: strongly disagree.

[c]2: disagree.

[d]3: agree.

[e]4: strongly agree.

Perceptions Regarding CE

Table 3 shows the findings related to the participants' perceptions regarding CE. More than half (288/508, 56.7%) held good perceptions, 39.4% (200/508) exhibited fair perceptions, and 3.9% (20/508) showed poor perceptions. Their mean perception score was 15.6 (SD 2.7). The majority of the pharmacists agreed and strongly agreed with the assertion that CE helps increase knowledge (434/508, 85.4%). Over 60% of the respondents agreed or strongly agreed with the remaining statements.

Table . Perceptions regarding continuing education (CE)[a].

Item	Pharmacist's perception	Likert scale					Value, mean (SD)
		1[b]	2[c]	3[d]	4[e]	5[f]	
P1	The value of the employer places on his participation in CE	8 (1.6)	41 (8.1)	136 (26.8)	235 (46.2)	88 (17.3)	3.7 (0.9)
P2	Your interest in/value of CE	6 (1.2)	12 (2.4)	110 (21.6)	283 (55.7)	97 (19.1)	3.9 (0.8)
P3	CE affects the way you practice	9 (1.8)	40 (7.9)	124 (24.4)	239 (47.0)	96 (18.9)	3.7 (0.9)
P4	CE helps increase your knowledge	5 (1.0)	3 (0.6)	66 (13.0)	230 (45.3)	204 (40.1)	4.2 (0.8)

[a]Poor perception: 20/508 (3.9%); fair perception: 200/508 (39.4%); and good perception: 288/508 (56.7%). The average total score using bootstrap (1000 times) was 15.6 (SD 2.7).

[b]1: strongly disagree.

[c]2: disagree.

[d]3: neutral.

[e]4: agree.

[f]5: strongly agree.

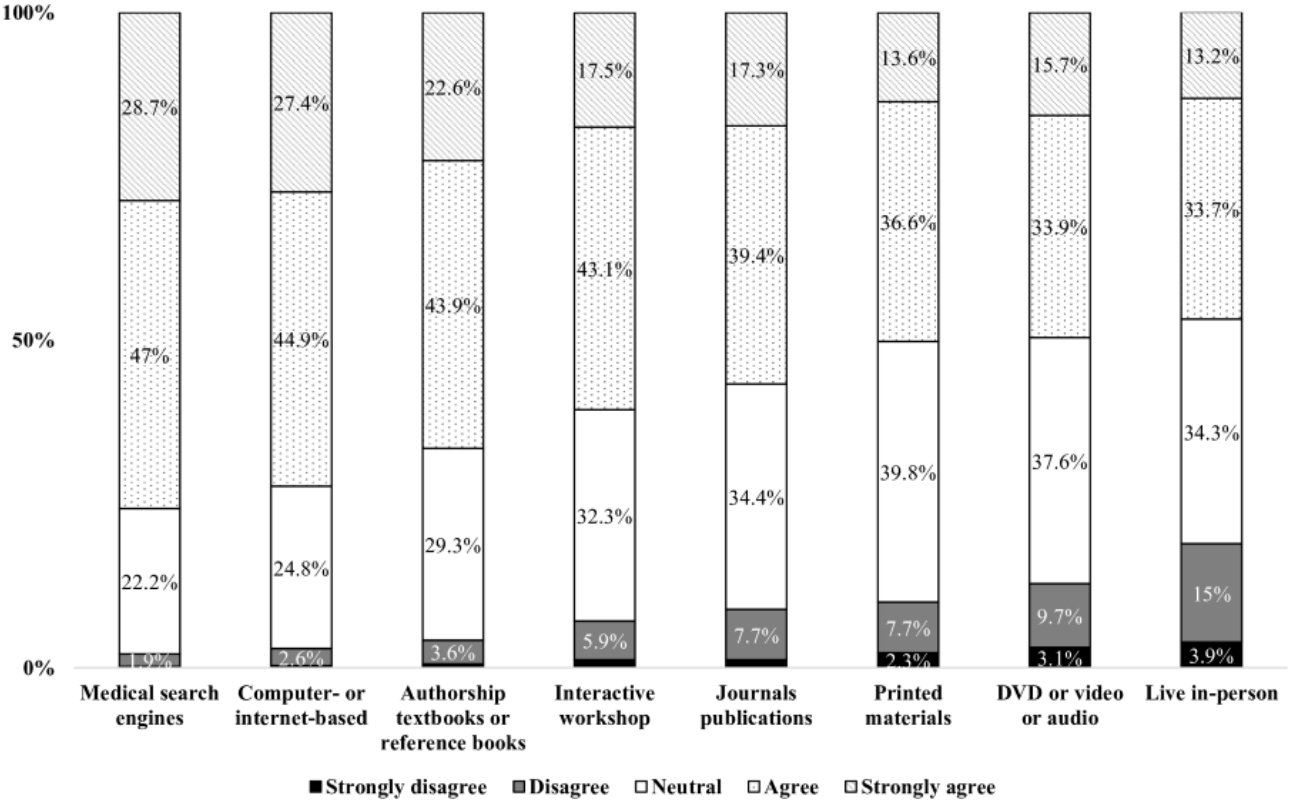
Preferences Regarding CE

Figure 1 illustrates the participants' levels of satisfaction with CE activities. The majority preferred computer- or internet-based CE (367/508, 72.2% agreed or strongly agreed with this option) and medical search engines (385/508, 75.7% agreed or strongly agreed with this option). The option of live in-person learning received mixed feedback from participants: 15% (76/508) of the respondents disagreed with the aforementioned options, 34.3% (174/508) exhibited neutrality, and 46.9% (238/508) expressed agreement. The pharmacists showed little preference for DVD- or video- or audio-based learning, with 9.7% (49/508) disagreeing and 37.6% (191/508) showing neutrality with

respect to this option. Textbooks or reference books were well received, with 43.9% (223/508) indicating preference and 22.6% (115/508) showing a strong preference for these materials.

Pharmacists' preferences for CE topics were assessed, with the most favored issue being results on skill development (415/508, 81.7% agreed or strongly agreed with this option), followed by innovations in pharmacy practice (399/508, 78.5% agreed or strongly agreed with this option), pharmacy management concepts (398/508, 78.3% agreed or strongly agreed with this option), and innovations in disease management (397/508, 78.1% agreed or strongly agreed with this option). The least preferred topic was innovations in pharmaceutical manufacturing (303/508, 59.6% agreed or strongly agreed with this option).

Figure 1. Pharmacists' level of satisfaction with continuing education activities.



Evaluation of Mean Scores for Perceptions, Attitudes, and Impact in Relation to Sociodemographic and Professional Characteristics

The mean attitudinal and perception scores of the pharmacists pointed to significant differences between groups with varying levels of education (43.3, SD 5.4 vs 44.6, SD 5.4 or 45.3, SD 6.0 for mean attitudinal score, *P*=.048 and 15.1, SD 2.6 vs 15.7, SD 2.8 or 15.6, SD 2.5 for mean perception score, *P*=.046; Table 4). Significant differences in mean attitudinal scores were found between respondents categorized by job positions (44.2, SD 5.3 vs 45.7, SD 6.2; *P*=.04) and frequency of overtime (44.2, SD 5.5 vs 45.9, SD 5.0; *P*=.03).

Table . Comparison of mean attitudinal and perception scores in relation to sociodemographic and professional characteristics.

	Attitude		Perception	
	Mean (SD)	*P* value[a]	Mean (SD)	*P* value[b]
Age (y)		.59		.46
21-24	44.9 (5.9)		15.9 (2.4)	
25-29	44.1 (5.1)		15.6 (2.7)	
30-34	44.7 (5.9)		15.3 (2.7)	
≥35	44.3 (5.3)		15.4 (2.9)	
Gender		.74		.14[c]
Male	44.5 (6.0)		15.3 (3.0)	
Female	44.3 (5.1)		15.7 (2.5)	
Ethnic		.27		.82
Kinh	44.4 (5.4)		15.5 (2.7)	
Other	43.1 (6.7)		15.7 (2.6)	
Marital status		.90		.38
Single	44.4 (5.5)		15.6 (2.6)	
Married	44.3 (5.4)		15.4 (2.9)	
Highest level of education		*.048*		*.046*
Elementary or intermediate or college[d]	43.3 (5.4)		15.1 (2.6)	
University	44.6 (5.4)		15.7 (2.8)	
Postgraduate	45.3 (6.0)		15.0 (2.5)	
Organizational type		.68[c]		.95
State administrative agency system	44.7 (6.2)		15.5 (3.0)	
Nongovernmental organizations	45.0 (4.4)		15.7 (0.6)	
Enterprise-private system	44.2 (5.1)		15.6 (2.6)	
Job position		*.04*		.61
Staff	44.2 (5.3)		15.6 (2.6)	
Manager	45.7 (6.2)		15.4 (3.2)	
Year of experience		.48		.59
<5	44.2 (5.4)		15.6 (2.6)	
≥5	44.6 (5.5)		15.5 (2.9)	
Frequency of overtime (times/wk)		*.03*		.85
<4	44.2 (5.5)		15.5 (2.7)	
≥4	45.9 (5.0)		15.6 (2.7)	
Number of CEs[e] attended		.52[c]		.55
Never	43.9 (5.0)		15.8 (2.7)	
At least 1 course	44.4 (5.5)		15.5 (2.7)	

[a]The italicized *P* values are used to denote statistical significance, following conventional academic practice. A *P* value <.05 indicates that the observed difference between groups is statistically significant.

[b]ANOVA test.

[c]Welch test.

[d]Educational levels in the Vietnamese context: elementary=vocational secondary training; intermediate=technical diploma; and college=3-year college degree.

[e]CE: continuing education.

Risk of Bias Assessment

The risk of bias assessment indicated a low risk across all evaluated domains (Checklist 1). The study had clearly defined objectives, an appropriate design and sampling strategy, and robust data analysis methods, although potential nonresponse bias was noted.

Discussion

Main Findings

This study evaluated the attitudes, perceptions, and preferences of 508 southern Vietnamese pharmacists regarding CE. Most of the respondents reported participation in at least 1 CE course. The majority demonstrated positive attitudes and considerably favorable perceptions of CE. Computer- or internet-based CE and the use of medical search engines were the preferred learning methods of the participants. The mean attitudinal scores varied significantly depending on education level, job position, and overtime frequency. Significant differences in the mean perception scores were found across groups with varying education levels.

Sociodemographic Characteristics

Most of the pharmacists were aged between 25 and 30 years (197/508, 38.8%), followed by those >35 years (126/508, 24.8%). The mixed methods approach to data collection, that is, the administration of both online and offline surveys, was intended to achieve an age-balanced sample. The online survey streamlined questionnaire distribution, facilitated data management, and enhanced engagement by allowing the respondents to conveniently complete the survey on their phones or computers from various locations. However, online surveys are often more accessible to younger individuals, potentially leading to an age imbalance due to sampling bias, low response rates, and data quality concerns [26,27]. To mitigate these limitations, we administered the survey offline, which also enabled direct interaction with the participants [26] and the provision of clarifications, thereby improving response accuracy and increasing participation among older pharmacists. Combining these methods enabled us to maximize data collection efficiency while ensuring a more representative age distribution.

The majority of the pharmacists hold university degrees. Similarly, Adhikari et al [28] reported that 60% of their pharmacist participants completed university education, whereas Poudel et al [29] indicated that 70% of their respondents acquired college diplomas or equivalents. According to Circular No. 22/2011/TT-BYT in Vietnam Law, most of the main divisions in the pharmacy departments of special-, first-, and second-grade hospitals require pharmacists to have a university degree or higher to practice [30]. Most of the pharmacists in the current work have attended at least 1 CE course, while the remaining individuals may have been newly graduating pharmacists who had not obtained a practicing certificate or taken CE classes. As the field of pharmacy continues to advance globally, pharmacy practitioners feel the need to keep in step with innovations to improve their skills in providing better care to patients [28]. Without CE engagement, pharmacists risk falling behind with regard to new drug developments, treatment protocols, and safety guidelines. This knowledge gap can directly impact the quality of patient care, leading to inaccurate medication counseling, increased potential for prescription errors, and compromised treatment outcomes. High-income countries require pharmacists to pursue CPD to maintain licensure and ensure exceptional patient care—this means frequent and regular participation in CE. As shown by Tjin et al [31], Dutch pharmacists participate in CE for an average of 27.0 hours over 11 months and prefer face-to-face learning (85.5%) over e-learning (13.8%).

Attitudes Toward CE

The questionnaire used in the present research included closed-ended items taken from the JSPLL [32], which has been used in similar studies conducted in Saudi Arabia [16] and Nigeria [21]. The JSPLL is the first instrument intended specifically to measure lifelong learning among health care providers with supporting psychometric evidence [32]. The majority demonstrated positive attitudes, aligning with the findings of Aldosari et al [16]. Specifically, 61.2% showed good attitudes regarding CE, 38% had fair attitudes, and only 0.8% indicated poor attitudes. These findings are consistent with those of Aldosari et al [16], who reported that 60%, 39%, and 1% of the pharmacists participating in their study showed good, fair, and poor attitudes, respectively [16]. The mean score of 44.4 in the present research pointed to overall good attitudes toward CE among the Vietnamese pharmacists. Specifically, most of them agreed that CE is an essential professional commitment for all pharmacists and that failing to keep up with pharmaceutical advancements could hinder their professional growth. Although the attitudinal distribution was similar to the findings of Aldosari et al [16], the underlying drivers for these attitudes are likely fundamentally different, highlighting the influence of the national context. The Vietnamese context lacks a fully institutionalized framework [3]. Therefore, the similarly favorable attitudes observed among pharmacists in this study are arguably more indicative of a strong intrinsic motivation for professional development. This suggests that, even in the absence of a coercive system, Vietnamese pharmacists personally value lifelong learning. This intrinsic drive is a response to the pressures and opportunities within an evolving national health care system, where rapid advancements and increasing patient expectations create a professional imperative to stay current. This finding underscores that the Vietnamese pharmacist workforce possesses a foundational readiness for CE, which could be powerfully leveraged as the formal system continues to develop. However, the greater proportion of them neither routinely participated in pharmacy organizational meetings nor engaged in weekly readings of professional journals. These results are consistent with the findings of Aldosari et al [16]. Acknowledging the responsibility to engage

in CE and lifelong learning is essential for pharmacists to maintain professional competence and deliver outstanding patient care as part of a patient-centered approach. The limited participation of pharmacists in professional organizational meetings and the low frequency of reading professional journals may be attributed to time and cost constraints. In addition, these types of CE may be less prevalent due to the increasing preference for internet-based resources. To solve these issues, professional meetings should be widely promoted at pharmacists' workplaces and scheduled with suitable time and cost considerations. The availability of professional journals should also be increased to enhance attitudes. In fact, increasing access to professional journals enhances attitudes by providing up-to-date research and expert insights that shape a deeper understanding of the field [33]. This exposure encourages critical thinking and helps professionals refine their practices, fostering more informed and positive attitudes.

Pharmacists with advanced degrees are often exposed to academic environments that promote evidence-based practice, critical thinking, and reflective learning, which enhance their appreciation of CE as an essential component of professional competence [34,35]. They also tend to possess stronger career motivation and clearer professional development goals, fostering a greater intrinsic drive for lifelong learning [36,37]. Their advanced level of education may also provide a stronger knowledge base, reducing uncertainty and fostering a more positive attitude of CE participation as a marker of professional identity and responsibility [38]. Consequently, pharmacists with high educational attainment tend to demonstrate stronger commitment to continuous learning than those with low qualifications [39]. The favorable attitudes among pharmacists in managerial positions align with the findings of Darwish et al [17], who reported greater awareness among pharmacist managers in Jordan. Managers often bear supervisory and decision-making responsibilities, which enhance their recognition of CE as essential for maintaining professional competence and ensuring the quality of pharmaceutical services [35,40]. Moreover, such positions may cultivate stronger intrinsic motivation toward continuous learning, as these individuals perceive CE as a pathway for leadership development and professional recognition [38]. Their access to institutional resources and professional networks could also facilitate participation in CE activities. These findings suggest that interventions to improve attitudes may need to be tailored, for instance, by emphasizing practical benefits for frontline staff and leveraging the advocacy of managerial and postgraduate champions.

Perceptions Regarding CE

The study found that Vietnamese pharmacists held considerably favorable perceptions of CE, with most agreeing that it increases their knowledge. This aligns with global trends observed in studies from Saudi Arabia [15] and Lebanon [22], suggesting a universal recognition of CE's core purpose among pharmacists. A notable point of divergence, however, lies in the perceived institutional support. While only 63.5% of the respondents agreed that their employers fully recognize the value of CE, this rate is substantially higher than the rates reported in the earlier international studies [15,22]. This divergence cannot be explained by regulatory mandate alone, as CE is compulsory in all three contexts. Instead, it likely reflects the unique and evolving state of Vietnam's CE landscape. Unlike the more established systems in other countries, Vietnam's framework has undergone significant recent reforms aimed at enhancing regulatory enforcement and accessibility [4,14]. This period of active development and the concurrent national focus on upgrading health care professional competency have created a distinctive environment. This has potentially fostered a stronger, more tangible sense among Vietnamese pharmacists that CE is becoming increasingly relevant and valued within their specific professional ecosystem, even if employer recognition is not yet universal.

In Vietnam, most pharmacists recognize CE activities as essential for enhancing their professional knowledge and skills [41]. Similarly, in their identification of pharmacists' professional learning needs in support of expanded roles in practice, Schindel et al [42] discovered that most of the respondents participate in CE activities out of personal interest or due to job requirements. Interestingly, we found that a considerable number of pharmacists disagreed with the idea that CE affects the way in which these professionals practice. This suggests that pharmacists pursue CE to maintain licensure rather than educate themselves on topics that can be applied to their current work. Therefore, understanding pharmacists' preferences and ensuring that CE topics align with their practice is a critical mission for policy makers and health care institutions.

A critical finding of this study is the apparent discrepancy between the broadly positive attitudes toward CE and the reported low engagement in specific activities like routine journal reading or organizational meeting attendance. This reflects the practical barriers within the Vietnamese context. The positive attitudes likely capture a genuine desire for knowledge and professional growth. However, the translation of this desire into consistent action may be hindered by significant systemic and personal obstacles. These could include overwhelming clinical workloads and limited paid leave, which restrict time for self-directed learning; financial constraints, where the cost of journal subscriptions or conference fees is prohibitive; a perceived lack of immediate relevance of available content to daily practice challenges; or limited access to resources, particularly for pharmacists in rural or community settings. This gap highlights that fostering positive attitudes is only the first step. For CE to be truly effective, systemic changes, such as providing protected time for learning, subsidizing costs, and ensuring the practical relevance of content, are necessary to enable pharmacists to convert their positive attitudes into consistent professional development behaviors.

Preferences Regarding CE

Computer- or internet-based and medical search engines were the preferred methods of learning by the Vietnamese pharmacists, as with other studies [16,21]. However, the mean scores for these resources were higher than those derived by Alharthi et al [15], suggesting that pharmacists in Vietnam prefer participating in online CE programs. This preference, in

turn, may be attributed to the flexibility of online offerings, allowing pharmacists to complete lessons at their convenience. They also reduce or eliminate program and travel costs, making them more accessible options [41,43]. Economic and financial factors may significantly influence the strong preference for online CE among pharmacists in this study. In resource-constrained settings like Vietnam, traditional in-person CE often entails direct costs such as travel, accommodation, registration fees, and indirect costs related to time away from work. For many pharmacists—particularly those working in rural or underfunded health facilities—these costs can be prohibitive. Online CE offers a cost-effective alternative, eliminating travel expenses and reducing opportunity costs, thereby making professional development more financially accessible. In addition, the increasing availability of free or low-cost online CE programs offered by hospitals, universities, and professional associations further reduces financial barriers. This shift in learning modality can also be interpreted as a response to broader economic pressures. Health care professionals, especially in low- and middle-income countries, often face modest salaries and must finance their own continuing education. These findings contribute to the growing body of literature on digital health education by emphasizing the role of economic factors in shaping learning preferences. From a policy perspective, this highlights the need for targeted investment in digital infrastructure and the subsidization of online CE initiatives. Such measures would not only improve access to education but also help build a more resilient and up-to-date health care workforce. In contrast, Al-Kubaisi et al[12] and other researchers [18,28,44] found that live in-person CE is favored over online learning. This preference stems from the fact that face-to-face classes allow for peer discussions, direct interactions with instructors, and the immediate clarification of questions, which enhance understanding and knowledge retention [28,45].

Most of the pharmacists in our study preferred topics related to skill development, innovations in pharmacy practice, and pharmacy management. These findings are consistent with Iskandar et al.'s study, which showed greater interest in pharmaceutical management than in any other matters [22]. In Dai et al's research [41], the majority of the participants were inclined to learn about regulations related to drug business operations (75.8%) and pharmaceutical practice (73.5%).

A larger proportion of pharmacists agreed that obtaining certification for a job and commitment to lifelong learning were important reasons for engaging with CE courses. Al-Kubaisi et al [12] found that the main motivation for CE participation among pharmacists is the relevance of a given topic. Therefore, incorporating more engaging and relevant subjects into CE programs can encourage greater participation. The development of future CE programs should focus on comprehensiveness and effectiveness in bridging the divide between theoretical learning and real-world application. CE topics must help pharmacists improve their competencies and abilities as they carry on with their practice. The applicability of what they learn from CE to their work would increase their interest in and motivation for such training. The organization of CE courses must also align with pharmacists' interests, especially those offered on the web. This goal can be achieved by enhancing engagement and support from employers and by advancing collaborative work between pharmacy regulators and CE providers, which can define the skills and competencies that need to be reinforced and offer CE programs intended for self-directed, lifelong learning.

Strengths of the Study

This study has several notable strengths. As the first study in Vietnam to explore pharmacists' attitudes, perceptions, and preferences regarding CE, it addresses a substantial gap in research and delivers novel insights into an underexamined aspect of pharmacy practice. By identifying the key attitudes, perceptions, and preferences that influence pharmacists' participation in CE, the study derives evidence that policymakers and health care institutions can use to design targeted CE programs that align with pharmacists' needs and preferences, ultimately enhancing participation rates and improving professional competencies. By offering evidence-based recommendations, this study contributes to the development of more effective CE initiatives, ensuring that they are accessible, relevant, and promotive of pharmacists' professional growth. Overall, the implication is support for improvements in pharmacy practice and health care quality in Vietnam.

Limitations

The limitations of this research are likewise worth discussing. First, the use of the JSPLL, which was originally designed to assess physicians' attitudes toward CE in the United States, may not have fully captured the perspectives of the Vietnamese pharmacists. Although the questionnaire has been validated and culturally adapted for Vietnamese people, the tool may not have reflected unique local professional development needs and contextual factors. The potential impact of this is a measurement bias that could have led to an underestimation of certain attitudes or perceptions highly specific to the Vietnamese pharmacy context. Consequently, some of the nuanced barriers or motivators for CE in Vietnam might remain unidentified. Second, the structure of the questionnaire may have limited the scope of responses, particularly in sections where predefined answer options were provided for CE preferences (eg, reading journal articles, and attending conferences or seminars). This limitation could have constrained the depth of insights into preferred learning methods and emerging CE trends. Third, since the study used a self-administered questionnaire, there was potential for response bias, wherein the participants may have tended to agree with the statements because this was expected or favorable rather than reflecting their true attitudes and behaviors. The likely impact is an inflation of positive scores on attitudinal items, and the prevalence of negative or ambivalent attitudes is likely underreported. Fourth, the study used a web-based questionnaire, which could not ensure that responses were exclusively from the target population, as anyone could access the QR code and complete the form. To address this limitation, a screening question was included to confirm whether the respondent was a pharmacist. The primary mode of data collection was a web-based questionnaire, which, even when combined with paper-based options and on-site recruitment, poses a potential risk of selection bias. The risk of selection bias arises from the fact that internet access is unevenly

distributed over the population [46]. Under-coverage occurs because people without internet access are excluded from the survey. It is possible that the sample may still be somewhat skewed toward younger, more technology-comfortable pharmacists who are inherently more receptive to digital surveys. While the mixed mode approach, particularly the on-site paper-based option, was implemented to counter this effect, it cannot be ruled out entirely. Therefore, the generalizability of the findings, especially regarding attitudes toward digital CE, should be interpreted with this potential bias in mind. Fifth, this study was based on quantitative survey methodology. While this approach was effective for identifying and statistically analyzing trends, attitudes, and perceptions across our sample, the closed-ended nature of the questions (including Likert scales and multiple-choice formats) inherently limits the depth of exploration into the underlying motivations, personal narratives, and nuanced contextual factors behind the responses. A qualitative research design would be necessary to provide that deeper layer of understanding. Finally, this study did not examine differences among pharmacists working in various practice settings (eg, community, hospital, and industry). By treating pharmacists as a homogeneous group, it may have obscured significant variations in learning needs, motivations, and perceived barriers. This limits the practical use of the findings for designing targeted CE programs, as the specific, setting-dependent drivers of engagement remain unexplored.

Future Research

Future research can expand this work by recruiting a nationally representative sample, including pharmacists from various regions of Vietnam, to better capture geographic and demographic differences in CE participation and preferences. Qualitative methods, such as in-depth interviews or focus group discussions, can help researchers more exhaustively illuminate the barriers to and motivators of engagement in CE. Studies should also focus on evaluating the effectiveness of different CE methods, identifying the most impactful learning formats so that they enhance pharmacists' knowledge, skills, and professional practice. Future studies should investigate the factors that influence these preferences to enable the design of more targeted and effective CE programs. Conducting subgroup analyses across different practice settings is necessary. Investigating the specific CE needs and preferences of pharmacists in community, hospital, industrial, and regulatory roles would provide invaluable data for designing highly targeted and effective CE programs that are responsive to the unique demands of each sector. Finally, longitudinal studies assessing the long-term impact of CE on pharmacists' career progression, clinical decision-making, and patient outcomes would be valuable in shaping future CE policies and programs.

The findings of this study offer clear, actionable guidance for optimizing CE in Vietnam. While the policies mandate CE, understanding the pharmacists' views and attitudes toward it is essential for improving both compliance and voluntary participation. For policymakers, the results underscore the need to move beyond a one-size-fits-all regulatory approach. The strong preference for digital and self-directed learning modalities supports the strategic expansion and accreditation of high-quality online CE platforms, which can improve accessibility for pharmacists across diverse geographic and practice settings. For pharmacy educators and training institutions, the high demand for topics in skill development, pharmacy management, and innovative practice indicates a critical need to shift content away from purely theoretical knowledge toward applied, practical competencies. Curricula should be co-designed with practicing pharmacists to ensure relevance. Furthermore, the significant intrinsic motivation observed suggests that promotional campaigns should frame CE not just as a mandatory requirement, but as a valuable tool for career advancement and professional excellence. By aligning program design with these evidence-based preferences, stakeholders can significantly enhance engagement and the overall impact of continuing professional development in Vietnam.

Conclusions

This study found that although most of the pharmacists exhibited positive attitudes and perceptions regarding CE, a substantial proportion had not participated in CE activities, indicating a gap between awareness and engagement. Variations in preferences for CE formats and topics underscore the need for tailored programs. Addressing participation barriers and aligning CE initiatives with pharmacists' professional needs may enhance engagement, support professional development, and improve patient care in Vietnam.

Acknowledgments

The authors would like to extend their appreciation to all the individuals who took part in this study.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The final questionnaires.
[DOCX File, 36 KB - mededu_v11i1e77013_app1.docx]

Checklist 1
AXIS checklist.
[DOCX File, 20 KB - mededu_v11i1e77013_app2.docx]

References

1. Global recognition of learning achievement framework. World Health Organization. 2006. URL: https://www.who.int/about/who-academy/quality-standards-credentialing/global-recognition-of-learning-achievement-framework [accessed 2025-11-12]
2. Continuing professional development (CPD) overview. Accreditation Council for Pharmacy Education. 2015. URL: https://www.acpe-accredit.org/continuing-professional-development [accessed 2025-11-12]
3. Institute of Medicine (US) Committee on Planning a Continuing Health Professional Education Institute. Redesigning Continuing Education in the Health Professions: National Academies Press; 2010.
4. The Ministry of Health. Circular No. 26/2020/TT-BYT Amendments to Circular No. 22/2013/TT-BYT dated August 9, 2013 of the Minister of Health guiding continuous training for medical officials. Vietnam Legal Documents. 2020. URL: https://vbpl.vn/TW/Pages/vbpq-toanvan.aspx?ItemID=147220 [accessed 2025-11-15]
5. The Ministry of Health. Circular No. 22/2013/TT-BYT guidelines on continuing education for healthcare personnel. Vietnam Legal Documents. 2013. URL: https://vbpl.vn/boyte/Pages/vbpq-van-ban-goc.aspx?ItemID=46966 [accessed 2025-11-15]
6. The Ministry of Health. Circular No. 32/2023/TT-BYT on elaboration of the law on medical examination and treatment. Vietnam Legal Documents. 2023. URL: https://vbpl.vn/bolaodong/Pages/ivbpq-thuoctinh.aspx?ItemID=167907&Keyword= [accessed 2025-11-15]
7. Viet Nam Congress. Law on pharmacy. Vietnam Government Portal. 2016. URL: https://vanban.chinhphu.vn/?pageid=27160&docid=184569 [accessed 2025-11-15]
8. Decree No. 163/2025/ND-CP detailed regulations and measures for the implementation and guidance of the law on pharmacy. Vietnam Government Portal. 2025. URL: https://vanban.chinhphu.vn/?pageid=27160&docid=214322 [accessed 2025-11-15]
9. Wheeler JS, Chisholm-Burns M. The benefit of continuing professional development for continuing pharmacy education. Am J Pharm Educ 2018 Apr;82(3):6461. [doi: 10.5688/ajpe6461] [Medline: 29692444]
10. Driesen A, Verbeke K, Simoens S, Laekeman G. International trends in lifelong learning for pharmacists. Am J Pharm Educ 2007 Jun 15;71(3):52. [doi: 10.5688/aj710352] [Medline: 17619652]
11. Braido F, Popov T, Ansotegui IJ, et al. Continuing medical education: an international reality. Allergy 2005 Jun;60(6):739-742. [doi: 10.1111/j.1398-9995.2005.00805.x] [Medline: 15876302]
12. Al-Kubaisi KA, Elnour AA, Sadeq A. Factors influencing pharmacists' participation in continuing education activities in the United Arab Emirates: insights and implications from a cross-sectional study. J of Pharm Policy and Pract 2023 Dec 31;16(1):112. [doi: 10.1186/s40545-023-00623-3]
13. Kandasamy G, Almaghaslah D, Almanasef M. An evaluation of continuing medical education among pharmacists in various pharmacy sectors in the Asir region of Saudi Arabia. Healthcare (Basel) 2023 Jul 19;11(14):2060. [doi: 10.3390/healthcare11142060] [Medline: 37510500]
14. Nguyen TH, Thai TT, Pham PTT, Bui TNM, Bui HHT, Nguyen BH. Continuing medical education in Vietnam: A weighted analysis from healthcare professionals' perception and evaluation. Adv Med Educ Pract 2021;12:1477-1486. [doi: 10.2147/AMEP.S342251] [Medline: 34938141]
15. Alharthi NM, Alsaeed MS, Alsharif MO, Almalki MG, Alshehri WS, Prabahar K. Assessment of pharmacists' perception toward continuing education. J Adv Pharm Technol Res 2021;12(4):368-372. [doi: 10.4103/2231-4040.329910] [Medline: 34820311]
16. Aldosari H, Alsairafi Z, Waheedi S. Continuing education in pharmacy: a cross-sectional study exploring pharmacists' attitudes and perceptions. Saudi Pharm J 2020 Jul;28(7):803-813. [doi: 10.1016/j.jsps.2020.05.008] [Medline: 32647481]
17. Darwish RM, Ammar K, Rumman A, Jaddoua SM. Perception of the importance of continuing professional development among pharmacists in a middle east country: a cross-sectional study. PLoS ONE 2023;18(4):e0283984. [doi: 10.1371/journal.pone.0283984] [Medline: 37058486]
18. Mohamed Ibrahim OH. Assessment of Egyptian pharmacists' attitude, behaviors, and preferences related to continuing education. Int J Clin Pharm 2012 Apr;34(2):358-363. [doi: 10.1007/s11096-012-9616-4] [Medline: 22354853]
19. Gelayee DA, Mekonnen GB, Birarra MK. Involvement of community pharmacists in continuing professional development (CPD): a baseline survey in Gondar, Northwest Ethiopia. Global Health 2018 Feb 1;14(1):15. [doi: 10.1186/s12992-018-0334-0] [Medline: 29391021]
20. de Bekker-Grob EW, Donkers B, Jonker MF, Stolk EA. Sample size requirements for discrete-choice experiments in healthcare: a practical guide. Patient 2015 Oct;8(5):373-384. [doi: 10.1007/s40271-015-0118-z] [Medline: 25726010]
21. Abdullahi AK, Salaudeen MA, Mosanya AU, et al. Exploring Nigerian pharmacists' attitudes and perceptions to continuing education and professional development. Pharm Educ 2023;23(1):89-99. [doi: 10.46542/pe.2023.231.8999]
22. Iskandar K, Raad EB, Hallit S, et al. Assessing the perceptions of pharmacists working in Lebanese hospitals on the continuing education preferences. Pharm Pract (Granada) 2018;16(2):1159. [doi: 10.18549/PharmPract.2018.02.1159] [Medline: 30023023]

23. Andersson S, Granat L, Brännström M, Sandgren A. Translation, cultural adaptation, and content validation of the Palliative Care Self-Efficacy scale for use in the Swedish context. Int J Environ Res Public Health 2022 Jan 20;19(3):1143. [doi: 10.3390/ijerph19031143] [Medline: 35162163]
24. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013;310(20):2191-2194. [doi: 10.1001/jama.2013.281053]
25. Downes MJ, Brennan ML, Williams HC, Dean RS. Development of a critical appraisal tool to assess the quality of cross-sectional studies (AXIS). BMJ Open 2016 Dec 8;6(12):e011458. [doi: 10.1136/bmjopen-2016-011458] [Medline: 27932337]
26. Nayak M, Narayan KA. Strengths and weakness of online surveys. IOSR J Humanit Soc Sci 2019;24:31-38. [doi: 10.9790/0837-2405053138]
27. Hargittai E, Hinnant A. Digital inequality: differences in young adults' use of the internet. Communic Res 2008;35(5):602-621. [doi: 10.1177/0093650208321782]
28. Adhikari B, Khatiwada AP, Shrestha R, Shrestha S. Assessing pharmacy practitioners' perceptions of continuing pharmacy education and professional development at an oncology service hospital in Nepal: a pilot study. Adv Med Educ Pract 2020;11:911-919. [doi: 10.2147/AMEP.S271129] [Medline: 33293884]
29. Poudel RS, Piryani RM, Shrestha S, Chaurasiya R, Niure BP. Opinion of hospital pharmacy practitioners toward the Continuing Pharmacy Education program: a study from a tertiary care hospital in central Nepal. Integr Pharm Res Pract 2017;6:157-161. [doi: 10.2147/IPRP.S145026] [Medline: 29354562]
30. Circular No 22/2011/TT-BYT on organization and operation of the department of pharmacy at the hospital. Vietnam Legal Documents. 2011. URL: https://vbpl.vn/boyte/Pages/vbpq-van-ban-goc.aspx?ItemID=26730 [accessed 2025-11-15]
31. Tjin A Tsoi S, de Boer A, Croiset G, Koster AS, Kusurkar RA. Factors influencing participation in continuing professional development: a focus on motivation among pharmacists. J Contin Educ Health Prof 2016;36(3):144-150. [doi: 10.1097/CEH.0000000000000081] [Medline: 27583989]
32. Hojat M, Nasca TJ, Erdmann JB, Frisby AJ, Veloski JJ, Gonnella JS. An operational measure of physician lifelong learning: its development, components and preliminary psychometric data. Med Teach 2003 Jul;25(4):433-437. [doi: 10.1080/0142159031000137463] [Medline: 12893557]
33. Tukhareli N. Bibliotherapy-based wellness program for healthcare providers: using books and reading to create a healthy workplace. J Can Health Libr Assoc 2017;38(2):44-50. [doi: 10.5596/c17-010]
34. Schön DA. The Reflective Practitioner: How Professionals Think in Action: Routledge; 2017.
35. Donyai P, Herbert RZ, Denicolo PM, Alexander AM. British pharmacy professionals' beliefs and participation in continuing professional development: a review of the literature. Int J Pharm Pract 2011 Oct;19(5):290-317. [doi: 10.1111/j.2042-7174.2011.00128.x] [Medline: 21899610]
36. Knowles MS. The Modern Practice of Adult Education: From Pedagogy to Andragogy: Cambridge; 1970.
37. Austin Z, Marini AE, Glover NM, Croteau D. Continuous professional development: a qualitative study of pharmacists' attitudes, behaviors, and preferences in Ontario, Canada. Am J Pharm Educ 2005 Sep;69(1):4. [doi: 10.5688/aj690104]
38. Cruess SR, Cruess RL, Steinert Y. Supporting the development of a professional identity: general principles. Med Teach 2019 Jun;41(6):641-649. [doi: 10.1080/0142159X.2018.1536260] [Medline: 30739517]
39. Brooks R, Everett G. The impact of higher education on lifelong learning. Int J Lifelong Educ 2008 May;27(3):239-254. [doi: 10.1080/02601370802047759]
40. Rouse MJ. Continuing professional development in pharmacy. J Pharm Technol 2004 Sep;20(5):303-306. [doi: 10.1177/875512250402000509]
41. Dai DX, Thuy NTX, Thao Đ, Nha PTT, Tram NTB. Online continuing education among pharmacists. J Nghien cuu Duoc Thong tin Thuoc 2023;14(1):47-54 [FREE Full text]
42. Schindel TJ, Yuksel N, Breault R, Daniels J, Varnhagen S, Hughes CA. Pharmacists' learning needs in the era of expanding scopes of practice: evolving practices and changing needs. Res Social Adm Pharm 2019 Apr;15(4):448-458. [doi: 10.1016/j.sapharm.2018.06.013] [Medline: 29941404]
43. Scott VG, Amonkar MM, Madhavan SS. Pharmacists' preferences for continuing education and certificate programs. Ann Pharmacother 2001 Mar;35(3):289-299. [doi: 10.1345/aph.10191] [Medline: 11261525]
44. Aziz Z, Jet CN, Abdul Rahman SS. Continuing professional development: views and barriers toward participation among Malaysian pharmacists. Eur J Soc Behav Sci 2013;4(1):45-55. [doi: 10.15405/ejsbs.2013.1.6]
45. Alsaaty FM, Carter E, Abrahams D, Alshameri F. Traditional versus online learning in institutions of higher education: minority business students' perceptions. Bus Manag Res 2016;5(2):31-41. [doi: 10.5430/bmr.v5n2p31]
46. Bethlehem J. Selection bias in web surveys. Int Statistical Rev 2010 Aug;78(2):161-188. [doi: 10.1111/j.1751-5823.2010.00112.x]

Abbreviations

AXIS: Appraisal Tool for Cross-Sectional Studies
CE: continuing education

CPD: continuing professional development
JSPLL: Jefferson Scale of Physician Lifelong Learning
WHO: World Health Organization

Edited by J Gentges; submitted 06.05.25; peer-reviewed by N Quang, N Sholihat, O Allela, SA Kristina; revised version received 15.10.25; accepted 17.10.25; published 16.12.25.

Please cite as:
Vo TQ, Le PD, Tran HTB, Nguyen HTT, Nguyen TD, Huynh TNK, Vo BV
Pharmacists' Attitudes, Perceptions, and Preferences Regarding Continuing Education: Cross-Sectional Study in Vietnam
JMIR Med Educ 2025;11:e77013
URL: https://mededu.jmir.org/2025/1/e77013
doi:10.2196/77013

Evaluation of an Interdisciplinary Educational Program to Foster Learning Health Systems: Education Evaluation

Sathana Dushyanthen[1], PhD; Nadia Izzati Zamri[2], MPharm; Wendy Chapman[1], PhD; Daniel Capurro[1,3], MD, PhD; Kayley Lyons[1], PharmD, PhD

[1]Centre for Digital Transformation of Health, University of Melbourne, Carlton, Australia

[2]Faculty of Pharmacy and Pharmaceutical Sciences, Monash University, Parkville, Australia

[3]School of Computing and Information Systems, University of Melbourne, Melbourne, Australia

Corresponding Author:
Kayley Lyons, PharmD, PhD
Centre for Digital Transformation of Health, University of Melbourne, Carlton, Australia

Abstract

Background: Learning health systems (LHS) have the potential to use health data in real time through rapid and continuous cycles of data interrogation, implementing insights to practice, feedback, and practice change. However, there is a lack of an appropriately skilled interprofessional informatics workforce that can leverage knowledge to design innovative solutions. Therefore, there is a need to develop tailored professional development training in digital health, to foster skilled interprofessional learning communities in the health care workforce in Australia.

Objective: This study aimed to explore participants' experiences and perspectives of participating in an interprofessional education program over 13 weeks. The evaluation also aimed to assess the benefits, barriers, and opportunities for improvements and identify future applications of the course materials.

Methods: We developed a wholly online short course open to interdisciplinary professionals working in digital health in the health care sector. In a flipped classroom model, participants (n=400) undertook 2 hours of preclass learning online and then attended 2.5 hours of live synchronous learning in interactive weekly Zoom workshops for 13 weeks. Throughout the course, they collaborated in small, simulated learning communities (n=5 to 8), engaging in various activities and problem-solving exercises, contributing their unique perspectives and diverse expertise. The course covered a number of topics including background on LHS, establishing learning communities, the design thinking process, data preparation and machine learning analysis, process modeling, clinical decision support, remote patient monitoring, evaluation, implementation, and digital transformation. To evaluate the purpose of the program, we undertook a mixed methods evaluation consisting of pre- and postsurveys rating scales for usefulness, engagement, value, and applicability for various aspects of the course. Participants also completed identical measures of self-efficacy before and after (n=200), with scales mapped to specific skills and tasks that should have been achievable following each of the topics covered. Further, they undertook voluntary weekly surveys to provide feedback on which aspects to continue and recommendations for improvements, via free-text responses.

Results: From the evaluation, it was evident that participants found the teaching model engaging, useful, valuable, and applicable to their work. In the self-efficacy component, we observed a significant increase (P<.001) in perceived confidence for all topics, when comparing pre- and postcourse ratings. Overall, it was evident that the program gave participants a framework to organize their knowledge and a common understanding and shared language to converse with other disciplines, changed the way they perceived their role and the possibilities of data and technologies, and provided a toolkit through the LHS framework that they could apply in their workplaces.

Conclusions: We present a program to educate the health workforce on integrating the LHS model into standard practice. Interprofessional collaborative learning was a major component of the value of the program. This evaluation shed light on the multifaceted challenges and expectations of individuals embarking on a digital health program. Understanding the barriers and facilitators of the audience is crucial for creating an inclusive and supportive learning environment. Addressing these challenges will not only enhance participant engagement but also contribute to the overall success of the program and, by extension, the broader integration of digital health solutions into health care practice and, ultimately, patient outcomes.

(JMIR Med Educ 2025;11:e54152) doi:10.2196/54152

KEYWORDS

continuing professional development; learning health system; flipped classroom; digital health informatics; data science; health professions education; interdisciplinary education; foster; foster learning; health data; design; innovative; innovative solution; health care workforce; Australia; real time; teaching model

Introduction

As health care delivery evolves in complexity and scope, the need for systems that promote continuous learning and adaptation is paramount. The learning health systems (LHS) concept has emerged as a transformative framework that bridges clinical practice with ongoing research, ensuring that health care institutions remain at the forefront of scientific and patient-centered care advancements [1,2]. Central to the LHS paradigm is the notion that data contribute to a broader system of knowledge and is used to refine care practices in real time [1,3]. Achieving this idea requires an interdisciplinary workforce adept in information systems, informatics, data interrogation, quality improvement and implementation methods, and system-based practice, to be able to use the existing data to inform future care [3]. Moreover, health care transformation such as this requires the skills of various professions working together towards solving these complex problems [4,5].

While there are previous studies that have described their LHS-focused programs, few have robustly evaluated the purpose of their implementations. Furthermore, other programs have focused on specific cohorts of participants such as PhD students [6], postdoctoral students [7,8], and clinical fellows [9-11] in the United States [6,9-11] and Canada [7]. Our study adds new insights to the literature given the interprofessional nature of the program, as well as its design (flipped classroom, working groups) and delivery (wholly online). To our knowledge, few programs have involved teaching a structured curriculum [8,12], while other programs have involved mainly project-based work and on-the job learning [7,10,13,14].

For such a dynamic and integrated approach to take root, educating the next generation of health care professionals about LHS principles is crucial. While the theoretical foundation of LHS has been well established, there has been a paucity of research evaluating the efficacy and impact of educational interventions centered on LHS. We developed a 13-week short course called Applied Learning Health Systems, which commenced in September 2021 and has now been running for 2 years [15]. The program is open to all professionals working in the health care setting—clinical and nonclinical—and focuses on interdisciplinary work; the LHS concept can be taught to both digital health and informatics generalists and specialists, clinicians and nonclinicians, front-line workers, and upper management [15].

As institutions increasingly incorporate LHS into their curricula, understanding the nuances of its educational translation becomes vital. This research aims to evaluate the motivations, experiences, and perceptions of participants learning in a collaborative learning environment, as well as the effectiveness, confidence, applicability, challenges, and outcomes of LHS education, providing insights that will shape pedagogical strategies and potentially influence the future of health care education.

The purpose of this paper is to explore participants' experiences and perspectives of participating in a wholly online interprofessional education program. This evaluation also aimed to assess the benefits, barriers, and opportunities for improvements, and identify future applications of the course materials to the participants' workplace endeavors. We will also discuss the implementation, feasibility, and outcomes of the program which aimed to foster LHS skills in the Australian health care workforce through didactic coursework, interactive workshops, and collaborative learning. By describing our program and its 2-year evaluation, we believe that current and future educators can learn from our experience when building their own programs. Additionally, our paper will contribute to the emerging education literature on how to foster LHS through workforce development and education. Compared to previous publications on LHS education programs, we are contributing novel insights to this literature through new perspectives based on our location (ie, Australia), the health system data infrastructure (ie, recent electronic medical record [EMR] implementations and digital immaturity), and our participants (ie, diverse interprofessionals). While we have had early successes, we also wish to highlight the obstacles we encountered and how we refined our approach in response. Our results will be valuable to other educators as they consider similar endeavors.

Methods

Study Design and Recruitment

We undertook a mixed methods study consisting of both quantitative and qualitative data collection methods. Surveys were conducted precourse, throughout teaching, and postcourse. The surveys consisted of metric scales, qualitative scales, and open free-text boxes. Participation in the research project was via opt-out. Therefore, all enrolled participants were eligible to participate in the project voluntarily, unless they chose not to. There were several modes of recruitment for the course itself. These included reaching out to existing precinct partners who undertook internal expression of interest recruitment processes to sponsor a number of places, social media advertising on X and LinkedIn, Google search search engine optimization, and university students undertaking electives or formal university-accredited certificates.

Ethical Considerations

This study was approved by the University of Melbourne ethics committee (project ID 22641). In certain parts of the study, participants had the option to opt out (eg, surveys) or provide consent to participate (eg, interviews). In terms of informed consent, participants were provided with a plain language statement describing the purpose and design of the study.

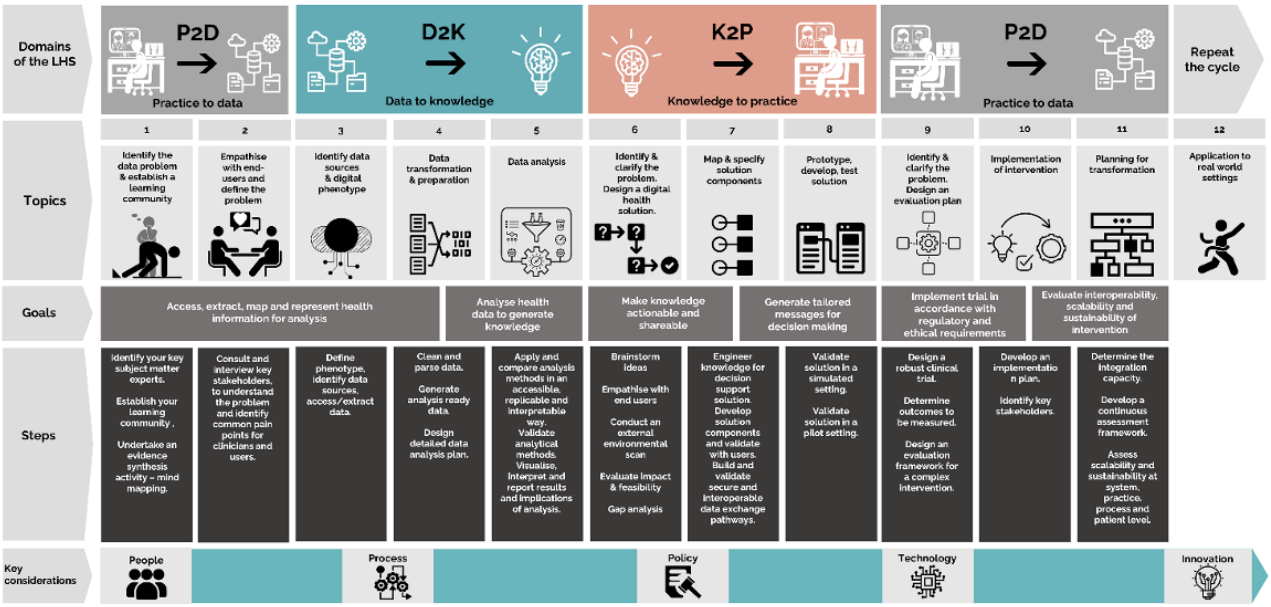
Participants were notified that participation was voluntary and were given the option to opt out. For privacy and confidentiality, data were completely deidentified and only aggregate data were analyzed and presented. Data were housed on secure University of Melbourne single sign-on Qualtrics servers and restricted access to OneDrive servers. As participation was completely voluntary, no compensation was provided to participants; however, participants in the pilot version of the course were given free scholarship admission in return for their feedback.

The Program

The LHS short course was created by the University of Melbourne Centre for Digital Transformation of Health, a high-research academic institution with existing partnerships with local and regional hospitals and primary care networks. The course has been delivered 5 times to 400 participants. Each iteration of the short course involved a 13-week online course revolving around LHS and was delivered wholly online, by diverse instructors, in a flipped classroom learning format. Participants were from a range of backgrounds, including working professionals in health care, PhD research students, masters-level university students, and consumers. The course structure involves 3 hours of weekly individual asynchronous prereadings, followed by 2.5 hours of weekly workshops. Each week participants work through activities associated with a threaded diabetes case scenario, in their assigned interprofessional working group [15].

We mapped the stages of the LHS system onto a swim lane diagram and created specific learning objectives for skills and knowledge at each stage, which were then operationalized into the diabetes scenario. Filling in this swim lane and competency map required knowledge from many disciplines, including data science and biostatistics, standards, user-centered design, change management, workflow mapping, app development, implementation science, and evaluation as well as expertise in the clinical domain and in how the Australian health system works. No single person could effectively design the course we developed, which posed challenges and opportunities for curriculum development. Using the LHS cycle enabled curriculum designers to join the varied subject matter expertise, by mapping it to an agreed framework. Details of the full course design, development, and curriculum outline are published elsewhere (Figure 1) [16].

Figure 1. The Applied Learning Health Systems short course curriculum [16].



Evaluation Framework

We used the Kirkpatrick model of evaluation [17] to map out our measurements (Table 1). This model is a widely used evaluation framework in education and is used to shift researchers away from simply measuring perceptions and satisfaction. We examined whether participants' attitudes, knowledge, behavior, and professional practice changed as a result. Additionally, we applied a mixed methods approach that included pre- and postsurveys, weekly surveys, and postinterviews.

Table . Application of the Kirkpatrick model of evaluation, adapted from Barr et al [18], to this project.

Level	Details	Evaluation measures and data sources in this project
1	Perception of training among subjects	Pre-, weekly, and postsurveys; postcourse participant interviews
2a	Change in the attitudes of subjects	Pre- and postchange in digital health interest and identity
2b	Change in the knowledge and/or skills of subjects	Pre- and post-self-efficacy changes in specific LHS[a] concepts (skills); pre- and postconcept maps (knowledge; out of scope for this paper)
3	Changes in the behavior of subjects	Postcourse participant interviews (will follow up in 1 year with participant interviews)
4a	Change in professional practice	Postcourse participant interviews (will follow up in 1 year with participant interviews)
4b	Changes in patients' condition	Not applicable

[a]LHS: learning health system.

Pre- and Postcourse Surveys

The pre- and postcourse surveys were developed by using a combination of psychological scales and open-ended questions. The pre- and postcourse surveys included the same self-efficacy scale (100 points; cannot do at all to highly certain can do) [19] which has significant evidence of reliability and validity. We choose to evaluate self-efficacy as it is one of the strongest proxy measures in education to predict actual and future performance, which are more difficult and take longer to measure [20]. The 10 items on the self-efficacy scale were adapted from the material taught in the LHS course and language from the LHS literature (eg, use machine learning algorithms to create a model for predicting a health outcome) [21,22]. The open-ended questions included demographic questions (eg, job title) and questions related to digital health identity development, course benefits, course barriers, what to keep, what to improve, and other suggestions or comments.

Surveys were designed and distributed via Qualtrics. Participants were invited to complete the surveys through emails and the learning management system. Responses to open-ended survey questions were also analyzed through qualitative content analysis. Two coders independently coded the text responses using NVivo (Lumivero) software. Coders met to resolve discrepancies and solidify themes and categories under each research question. The self-efficacy scales were analyzed using a 2-tailed, unpaired *t* test in GraphPad Prism to determine whether there was an improvement in self-efficacy across the 13 LHS concepts.

Weekly Surveys

Over the 12 weeks, participants had the opportunity to provide feedback on the level of engagement, usefulness, value, satisfaction, and areas for improvement in the course content, through participation in weekly surveys. These surveys contained scales (strongly disagree to strongly agree) and ask questions such as "how useful did you find this topic" and "how engaged did you feel" and open boxes for free-text responses. Descriptive statistics such as frequency, mean, and standard deviation will be used to summarize the data from these questions. Completion of these weekly surveys ranged from 2530 participants each week.

Qualitative Coding of Free-Text Responses

To analyze the text response according to our research questions, we first deidentify the transcripts for participant and institution names. The transcripts will be uploaded to NVivo software for qualitative content analysis [23]. A codebook was developed deductively from the literature and inductively from the research data. Two coders independently analyzed the transcripts according to the codebook. The 2 coders met to calculate an interrater agreement rate and resolve any discrepancies. The final codes were synthesized by creating summaries, narratives, and matrices. The final results included coding frequencies, themes, and categories according to the research questions.

Quantitative Statistical Analysis

For descriptive statistics, number of participants and proportion of participants are shown. For rating scales, frequency and proportion are shown. Pre- and postcourse self-efficacy comparisons were undertaken using a 2-tailed, unpaired *t* test. Incomplete or missing data were excluded from the analysis.

Results

Demographics

Thus far, the Applied Learning Health Systems program has had approximately 400 participants from various organizations (health care, government, research or university, industry) and job roles (clinician, researcher, data or information technology [IT], health services management, allied health, EMR implementation, health administration, consumer advocacy) (Table 2). Of the 400 participants, 343 (85.8%) completed the presurvey (week 0) and 200 (50%) completed the postsurvey (week 12). A few participants were lost to follow-up during the final week because they were ill, dropped out due to overcommitment, or did not respond to requests.

Table . Demographics shared by participants in the Applied Learning Health Systems program.

Characteristic		Participants, n (%)
Professional background (n=399)		
	Primary health care	44 (11)
	Tertiary health care	141 (35.3)
	Health services management	29 (7.3)
	Allied health	48 (12)
	Government	10 (2.5)
	Academia or research	73 (18.3)
	Business, IT[a], tech or data analytics	47 (11.8)
	Other	7 (1.8)
Role type (n=343)		
	Clinician (medical)	67 (17)
	Clinician (nursing)	25 (6.4)
	Clinical informatician	22 (5.6)
	Researcher (health services research or public health)	68 (17.3)
	Data analyst	28 (7.1)
	Allied health professional	58 (14.8)
	Health services manager	36 (9.2)
	Quality improvement lead	24 (6.1)
	Consultant or IT professional	19 (4.8)
	EMR[b] implementation team	18 (4.6)
	Health administration	8 (2)
	Consumer advocate	20 (5.1)

[a]IT: information technology.

[b]EMR: electronic medical record.

What Were Participants' Previous Encounters With the LHS Framework?

At the beginning of the course, participants were asked if they had any previous exposure to the LHS framework. Almost one-third of the participants had no previous experience with the LHS concept or any digital health concepts (121/343, 35.3%). Some participants stated that they had previous exposure to digital health and informatics concepts (50/343, 14.6%) through other courses and certifications (27/343, 7.8%), as well as through work-based activities, for example, EMR implementation and optimization (47/343, 13.1%), quality improvement, data interrogation (56/343, 16.3%), and various other health services projects (45/343, 13.1%). Others stated that they had no previous exposure to digital health or LHS concepts (49/343, 14.2%).

What Type of Teaching Approaches Did Participants Perceive as Effective?

Participants were asked to rate the usefulness and engagement of the topic's preclass learning and in-class sessions. In terms of usefulness, the majority found the preclass materials useful (880/956, 92.1%—"the preclass material was excellent and really helped to clarify many of the terms that I had heard people say but not truly understood") and in-class sessions useful (902/955, 94.5%—"analyzing the data during the class was useful and to see it connect with prelearning materials was good"). When asked to rate engagement, the majority found the preclass (881/954, 92.3%) and in-class activities engaging (881/955, 92.3%) (Tables 3-6).

Table . Ratings of usefulness and engagement with preclass learning materials and in-class Zoom sessions. Participants were asked to rate the agreement for usefulness (extremely useless to extremely useful) and engagement (extremely unengaged to engaged), weekly for each topic (1-13).

Questions	Rating							Total, n
	Extremely useless, n (%)	Moderately useless, n (%)	Slightly useless, n (%)	Neither useful nor useless, n (%)	Slightly useful, n (%)	Moderately useful, n (%)	Extremely useful, n (%)	
I found this topic's preclass learning useful (13 topics)	5 (0.5)	39 (4.1)	13 (1.4)	20 (2.1)	210 (22.0)	265 (27.7)	404 (42.3)	956
I found this topic's in-class session useful (13 topics)	2 (0.2)	13 (1.4)	9 (0.9)	28 (2.9)	103 (10.8)	360 (37.7)	440 (46.0)	955
I felt engaged when completing the preclass learning for this topic (13 topics)	5 (0.5)	9 (0.9)	29 (3.0)	30 (3.1)	127 (13.3)	428 (44.9)	326 (34.2)	954
I felt engaged when participating in the topic's in-class session (13 topics)	10 (1.0)	13 (1.4)	17 (1.8)	33 (3.5)	104 (10.9)	349 (36.5)	429 (44.9)	955

Table . Participants' ratings of value pertaining to overall value to personal career development for all topics.

Question	Rating					Total, n
	Highly unvaluable, n (%)	Unvaluable, n (%)	Neutral, n (%)	Valuable, n (%)	Highly valuable, n (%)	
Valuable to your personal career development (13 topics, n=189)	12 (0.6)	13 (0.6)	251 (12.3)	989 (48.5)	776 (38.0)	2041

Table . Participants' ratings of value pertaining to applicability to current workplace role for all topics.

Question	Rating					Total, n
	Highly not applicable, n (%)	Not applicable, n (%)	Neutral, n (%)	Applicable, n (%)	Highly applicable, n (%)	
Applicability to your current workplace role (13 topics, n=189)	64 (3.1)	154 (7.5)	325 (15.9)	837 (41.0)	661 (32.4)	2041

Table . Participants' ratings of value pertaining to overall satisfaction with the quality of the course, recommendation, instructors, and choice to revisit, as well as the value of educational activities (instructors, Zoom workshops, Canvas preclass activities, collaborative learning, the diabetes case scenario, Jupyter Notebooks, and discussion boards).

Questions	Rating							Total, n
	Extremely valueless, n (%)	Moderately valueless, n (%)	Slightly valueless, n (%)	Neither valuable nor valueless, n (%)	Slightly valuable, n (%)	Moderately valuable, n (%)	Extremely valuable, n (%)	
Collaborative learning in the working groups	1 (0.5)	6 (3.3)	3 (1.6)	5 (2.7)	29 (15.9)	63 (34.6)	75 (41.2)	182
Preclass learning activities on Canvas	0 (0)	3 (1.6)	2 (1.1)	2 (1.1)	21 (11.5)	77 (42.3)	77 (42.3)	182
In-class learning (Zoom) sessions	1 (0.5)	2 (1.1)	1 (0.5)	6 (3.3)	17 (9.3)	74 (40.7)	81 (44.5)	182
The diabetes case scenario	1 (0.5)	4 (2.2)	3 (1.6)	10 (5.5)	30 (16.5)	70 (38.5)	64 (35.2)	182
Jupyter Notebooks	0 (0)	11 (6.0)	11 (6.0)	20 (11.0)	53 (29.1)	55 (30.2)	32 (17.6)	182
Canvas learning management system	0 (0)	1 (0.5)	3 (1.6)	12 (6.6)	30 (16.5)	83 (45.6)	53 (29.1)	182
Discussion boards	5 (2.7)	10 (5.5)	10 (5.5)	43 (23.6)	62 (34.1)	39 (21.4)	13 (7.1)	182
The instructors	0 (0)	1 (0.5)	0 (0)	4 (2.2)	8 (4.4)	49 (26.9)	120 (65.9)	182

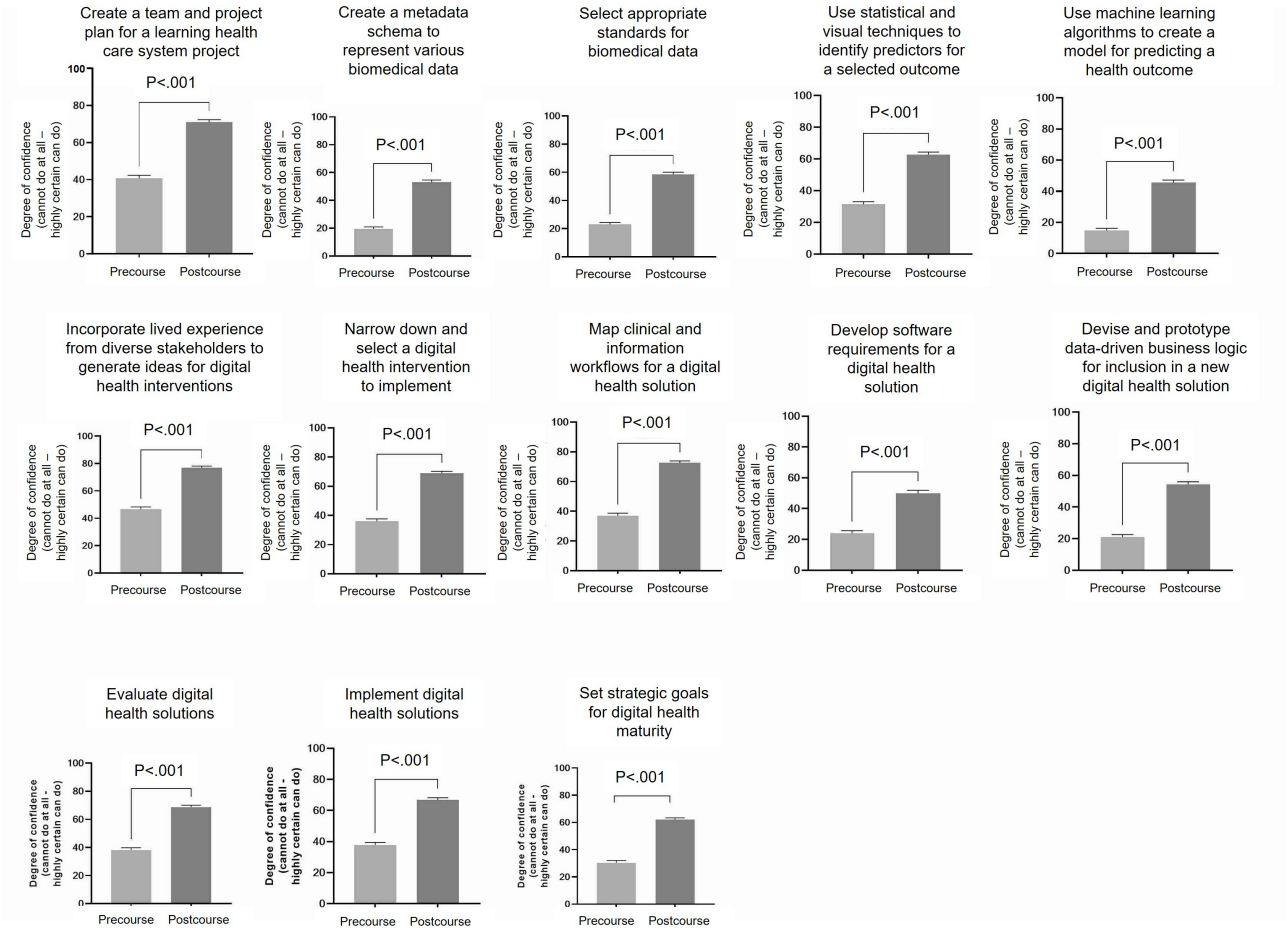
Responses to the question of satisfaction also yielded highly positive results. For the overall quality of the short course, most agreed that it was of a high standard (178/182, 97.8%), including the instructor quality (175 /182, 96.2%). When asked if they would recommend the short course to a colleague, 89.5% (163/182) said they would. In terms of revisiting the decision to complete it again, 85.1% (154/182) still said they would choose to take the course. When rating the value of the course to their personal career development, a majority found the course valuable (173/200, 86.5%). Participants were also asked to rate the applicability of the course to their day-to-day work, where 73.4% (134/182) found it applicable.

Given the number of facets implemented in the course, we asked participants to rate the value of these various elements. The most highly rated was the instructors: "the speakers were great, and the structure of having a short lecture and then doing an activity then coming back and having another lecture was good," with 92.8% (169/182) finding them moderately or extremely valuable. Next, in-class learning (155/182, 85.2%), preclass learning (154/182, 84.6%), collaborative learning (138/182, 75.8%), the diabetes case scenario (134/182, 73.7%), and the Canvas learning management system platform (136/182, 74.7%) rated similarly. The use of Jupyter Notebooks (87/182, 47.8%), and the discussion boards (52/182, 28.6%) rated lower (Table 6).

How Did Participants' Self-Efficacy for Digital Health Topics Change After the Course?

To explore the change in self-confidence levels pre- and postcourse, participants were surveyed on the key competencies for the 13 topics. Participants completed the same set of ratings at the beginning and at the end of the course, following completion of all the material. For all 13 learning outcomes, there was a statistically significant increase in self-efficacy (n=200, $P<.001$) (Figure 2).

Figure 2. Pre- and postcourse self-efficacy in LHS concepts. Participants rated confidence on a scale of 0 - 100 (0=cannot do at all to 100=highly certain can do). Two-tailed, unpaired *t* test was undertaken (n=296 precourse, n=200 postcourse). Changes from baseline to postcourse confidence are shown for each LHS concept. LHS: learning health system.



How Did Participants' Self-Perceived Role in Digital Health Change?

In the pre- and postsurvey, participants were asked to respond to the open-ended question of "What do you see as your role in digital health?" There were several types of roles that participants perceived themselves embodying postcourse. These included users of digital health or learners; champions, advocates, or change agents; researchers, innovators, or entrepreneurs; leaders, managers, strategic planners, or decision makers; educators or mentors; specialists or implementers; community builders, connectors, facilitators, collaborators, or translators (Table 7).

After the course, there was an increase in participants who viewed their role as an end user or learner and a community builder or facilitator, whereas there was a decrease in those who viewed their role as a champion or advocate and leader in digital health.

Table . Participants' perceived roles in digital health pre- and postcourse (qualitative themes).

	Precourse responses (n=274), n (%)	Postcourse responses (n=228), n (%)
End user of digital health or learner	41 (15.0)	54 (23.7)
Champion, advocate, or change agent	62 (22.6)	37 (16.2)
Researcher, innovator, or entrepreneur	32 (11.7)	30 (13.2)
Leader, manager, strategic planner, or decision maker	38 (13.9)	22 (9.6)
Educators or mentors	13 (4.7)	6 (2.6)
Specialist or implementer	57 (20.8)	41 (18.0)
Community builder, connector, facilitator, collaborator, or translator	31 (11.3)	38 (16.7)

What Did Participants Perceive as the Applications of the Learning in Their Workplace?

There were five main themes that arose for the types of applications that participants foresaw themselves using the course learnings: (1) learning and professional development: "upskilling in the current role, more understanding of the roles of my team members"; (2) using data and undertaking data analysis more effectively: "data mining and improving processes at work"; (3) implementing the LHS framework for digital health interventions: "we are embarking on establishing a data and analytics 3-year plan and we intend to incorporate LHS principals into this strategy"; (4) for undertaking research and quality improvement activities: "I now play a role in learning health networks for Safer Care Victoria, where I believe I could encourage digital health projects focused on quality improvement and patient safety"; and (5) collaborating and sharing knowledge and learnings with colleagues: "I intend to instill the LHS framework into my role, the work that I do and share it with my team" (Table 8).

Table . Participants' anticipated applications of learning in the workplace (qualitative themes).

	Precourse responses (n=338), n (%)	Postcourse responses (n=231), n (%)
Learning and professional development	74 (21.9)	47 (20.4)
Using data and undertaking data analysis	54 (16.0)	43 (18.6)
Implementing digital health solutions with the LHS[a] framework	63 (18.6)	52 (22.5)
Researching and quality improvement	82 (24.3)	62 (26.8)
Collaborating and knowledge sharing	65 (19.2)	27 (11.7)

[a]LHS: learning health system.

What Were the Perceived Benefits of the Program?

Participants were asked to state the benefits of the program. The major themes that arose were learning and knowledge acquisition: "the course material was presented well on Canvas and had a good mix of different learning resources to use," value of collaboration: "the course has been extremely eye-opening and has led me to begin collaborations on digital health projects through contacts made through the course," participant diversity and group work : "being in a group of people with all different work backgrounds and skills coming together with a common interest was really good for tackling the problems to solve in the class," beneficial course structure and content delivery (preclass: "the course material was presented well on Canvas and had a good mix of different learning resources to use" and in-class: "beneficial to be in a diverse group of other health care professionals - I learnt a lot from the robust and engaging discussions on Zoom)," and learning tools, importance of real-world applications: case study and personal work: "applying course concepts to this real-world scenario was instrumental in reinforcing their understanding," appreciation for instructors' diversity, expertise, engagement, and quality: "the instructors were very engaged and passionate about their topics," consumer focus, and focus on data analytics (Table 9).

Table . Beneficial elements of the course.

Theme	Responses (n=295), n (%)
Collaborative group work, diversity, or multidisciplinary approach	63 (21.4)
Course structure and content delivery or pre- and in-class material	54 (18.3)
Learning and knowledge acquisition	53 (18.0)
Real-world scenarios or real-world applicability	44 (14.9)
Exposure to tools and techniques	28 (9.5)
Appreciation for instructors	25 (8.5)
Exposure to complexity and challenges	14 (4.7)
Focus on consumers	14 (4.7)

What Were Participants' Barriers to Engaging With the Program?

When asked regarding barriers to participating in the course, participants' responses formed the following major categories: time constraints due to work, family, and other social commitments: "time constraints, balancing clinical work, other non-clinical work and home life," a lack of knowledge, terminology, and experience: "limited coal-face/frontline exposure and visibility of emerging frontline issues. I work at a more systems-based level and am not involved in interacting with patients day-to-day," technical challenges: "I found using so many new platforms eg Jupyter notebooks, BPMN so quickly challenging...," content complexity, and limited interactions online (Table 10).

Table . Barriers to effective participation.

Theme	Responses (n=259), n (%)
Time constraints or keeping up with materials	109 (42.1)
Lack of knowledge and experience	56 (21.6)
Family and personal commitments	37 (14.3)
Technical challenges	23 (8.9)
Health care terminology and clinical knowledge	14 (5.4)
Work commitments	14 (5.4)
Course structure and content	6 (2.3)

What Changes or Improvements Would Participants Suggest to the Short LHS Coursework?

While the majority of participants found beneficial elements to the course, there are always improvements that can be made. Areas in which changes were suggested were course structure, duration, and timing, suggesting concerns around the pace of the course and the amount of information and breadth covered: "it feels like a lot of materials are being cramped into 1 session and it was hard to appreciate the differences between the models" and the timing of delivery after a long work day; the usability of some learning tools, such as Jupyter Notebooks, difficulties with learning management platform navigation, more revision activities to reinforce learning and a desire for more printable or downloadable resources; questionable benefit of group work and collaborative work where students wanted more support and time to hear instructor expertise: "I feel there was too much reliance on group work and not enough input and guidance from the experts"; course delivery—online format, questioning whether networking opportunities were lost online; prerequisite skills required, given the difficulty of some content (Table 11).

Table . Participants suggested improvements to the course.

Theme	Responses (n=158), n (%)
Course content and structure—curriculum, quality, volume of material, level of complexity, clarity, usefulness, effectiveness, engagement, and applicability	64 (40.5)
Course logistics and administration—course duration, pace, delivery modality, pre-requisites, and learning platforms	37 (23.4)
Learning tools and materials—usability and accessibility	15 (9.5)
Group work and collaboration activities—diversity, effectiveness, and interaction	30 (19.0)
Instructor interactions in-class—interaction, engagement, and support	12 (7.6)

Discussion

Principal Findings

Despite the concept originating in 2007 [24], there is a lack of reports evaluating LHS education programs. In this evaluation, we discuss the findings of 2 years of implementation and iteration of an interdisciplinary Applied LHS professional development course (343/400, 85.8%, presurvey respondents; 200/400, 50%, postsurvey respondents), to a diverse range of professionals working and studying in health care, with an interest in digital health. Most of our participants were from Australia, where LHS was a novel but emerging concept [15,25-27]. The participants found the course engaging and relevant to their work. Participants highlighted specific benefits, barriers, and applications to this course and the LHS framework on their work.

Most health systems are actively seeking to increase the use of data and digital technology to drive improved health care delivery and health outcomes. A major ingredient needed to achieve that lofty goal is a workforce that knows how to not only thrive within the rapidly digitizing world but also how to innovate to improve value-driven care. Training a diverse workforce in the digital transformation of health poses an overwhelming number of choices about the most important learning objectives, competencies, and skills. The LHS framework [16] placed boundaries around the grand vision and enabled us to concretely tell a story that resonated with the goals of potential learners while lending itself to hands-on activities that invite learners to be part of that story.

In addition to the advantages of multidisciplinary curriculum development, the LHS framework was also a key part of the value of the course to interdisciplinary learners. We launched this course as a pilot and hand-selected 50 participants from a much larger pool of applicants with the aim of multidisciplinary involvement and of creating buzz around the course to encourage enrolment for a fee-paying version of the course. Medical directors, research leads, clinicians, and managers brought learnings from the course to hallway discussions and team meetings in their workplaces about how they could apply the LHS framework in specific projects. In addition to a better

understanding of how a project could go from idea to implementation and evaluation using the LHS principles, the framework provided a shared lexicon, a set of approaches like the creation of a learning community, and a toolkit of methods that learners could envision being used in their work. Their excitement was contagious, and a large proportion of our enrolees have come from organizations who continue to sponsor entire interdisciplinary teams of people to take the course together, because they see the value of the framework as a connector across disparate teams, such as clinicians, IT or EMR analysts, and health intelligence units, seeking to work toward a shared goal.

Overall, the course attracted a wide range of professionals at different levels (eg, medical students to directors of emergency departments), professions (eg, nursing and social work), consumers, researchers, and disciplines (eg, IT professionals). In this study, participants highly valued the interdisciplinary nature and collaborative learning activities in the course. Based on previous educational research, we purposefully sorted the groups for a diversity of professions and kept the participants within the same groups for the majority of the course to encourage relationship building. The interdisciplinary aspect of this course was a strength of our education model as it mimics the type of interdisciplinary practice required for complex LHS and digital health initiatives [28].

From several written comments and weekly surveys, we found that different disciplines struggled at different points within the course. For example, people without a research background found the data analysis topic and using Jupyter Notebooks the most challenging aspect of the course, whereas those with a nonclinical background struggled the most with mapping clinical workflows and implementation. Although we used these struggles as teaching moments to demonstrate the need for an interdisciplinary team in LHS, our experience indicates the need to improve our interdisciplinary education model. Previous education researchers and motivational theorists have established that optimal challenge is a key ingredient for engagement and learning [29]. If the material is too easy or too difficult, then learners disengage and, thus, do not learn the material. Many educators have described the challenge of designing a course for optimal challenge among a large cohort of uniprofessional courses [30]. However, our experience is that this challenge is even more dramatic in a one-size-fits-all model in an interdisciplinary course. The content we taught is still appropriate for all audiences, but each person may require more or less self-directed preparatory work as part of the flipped classroom model. Future researchers and educators should investigate how to continue serving an interdisciplinary audience while creating optimal challenges for all participants. For example, in future iterations, we will explore the use of generative artificial intelligence tools to personalize the self-directed online modules for participants' previous knowledge and professional context.

The participants' self-described digital health roles before and after the course only went through minor changes. There was a small conversion in participants who started out seeing themselves as leaders and then later described their roles as connectors. This phenomenon may have been due to instructors telling participants about the importance of connector roles within the LHS framework. Another reason for this effect may be the Dunning-Kruger effect [31]. The Dunning-Kruger effect is when individuals with low exposure to a topic often overestimate their abilities due to a lack of metacognitive awareness. As they gain more knowledge, they become more aware of the limitations. Despite the potential for the Dunning-Kruger effect, the lack of significant changes in participant digital health identity was in contrast to a similar evaluation of our parallel LHS education offering—a 1-year LHS fellowship program for clinicians [15]. In the fellowship program, half of the participants began the program by describing their role as champions and leaders, and then, by the middle of the program, all of the participants described their role as champions and leaders. This potential effect may be due to the benefits of the fellowship program; the fellowship is more experiential, project based, and explicitly focused on leadership development. Since self-identities are an important mediator of future performance [32], future educators and researchers should continue to investigate how LHS educational programs influence participants' self-described roles in the LHS framework and digital health.

Strengths, Limitations, and Future Directions

Overall, we achieved commendable survey response rates, suggesting a high level of engagement from participants. This study uniquely contributes to the existing literature by evaluating an interdisciplinary LHS education program—a domain previously underexplored. Our comprehensive approach encompassed both pre- and postcourse survey data, leveraging learning theories such as self-efficacy theory and the Kirkpatrick evaluation framework to inform our evaluation. Moreover, our qualitative analysis offers valuable insights into participants' perceptions, enriching our understanding of their experiences. However, a limitation is our current inability to capture the upper levels of the Kirkpatrick model, specifically how the LHS course may have influenced participants' workplace behaviors and the subsequent outcomes of those behaviors. In the long term, we aim to evaluate the impact this course and other LHS education offerings have had on individuals and their health organizations' journeys toward a learning health system and individual's career progression. We aim to do this by conducting follow-up, in-depth interviews with participants and organizational sponsors and thematically analyzing the changes that have occurred over time.

Achieving an LHS requires a symbiotic partnership between researchers and health services—by bridging theory and real-world application, future innovations emerging from an LHS will be evidence based and clinically relevant. To increase academic-practice collaboration, our LHS educational offerings aim to grow the understanding of LHS principles and skills in our health services partners and to provide insight into the enablers and barriers for their digital transformation. The shared LHS framework and increased mutual understanding from these programs are increasing trust and collaborative opportunities, leading toward joint translational LHS innovation programs within the health services. We hope that future educators and academic leaders see promise in our emerging LHS education evaluation work [15], other descriptions of LHS education

initiatives [6-11], and the success of LHS initiatives in health care practice [33-35].

By providing a professional development short course, we were able to serve a large market of health professionals who would not otherwise have participated in an expensive university degree. While some professionals like medical specialists receive a continuing medical education fund, most other disciplines are not provided with funding for professional development. Additionally, a major source of participants was partner organizations supporting and sending groups of staff through the program, to learn together as cohorts to develop communities of practice. In this scenario, enrollment was funded by their employers. This is crucial, as at the national and international level, we require a critical mass of appropriately skilled workforce to leverage LHS principles in improving the quality and value of health care delivery.

An interdisciplinary LHS short course has also provided a testbed for applying new technologies to learning. For instance, in the last iteration of the course, we experimented with generative AI feedback on the participants' learning. In their working groups, participants developed an evaluation plan. They fed their plans into ChatGPT, which we provided with structured, custom prompts to provide feedback and rate the quality of the plans. Although some students found the feedback to be generic, the depth of the feedback was dependent upon the richness of the data initially fed to the machine. In large group settings, where there are limited instructors and limited time to provide in-depth feedback to each interdisciplinary group or participant, ChatGPT may be a useful tool to assist with providing formative feedback. The use of this will be further explored in future iterations of the course.

Conclusions

Overall, the Applied Learning Health Systems course received significant positive feedback from interdisciplinary learners. They found the course to be well structured, engaging, and a valuable learning experience. The qualitative comments emphasized the importance of delivering courses that not only provide knowledge but also inspire and motivate learners, and provide concrete tools to apply in their workplaces. A significant number of participants expressed interest in future courses and opportunities for further learning, underscoring the potential for expanding and diversifying course offerings in the future. There is still a great deal of education that needs to be provided to upskill the workforce adequately enough to undertake digital health transformation, but it begins with a shared vision, a common language, and a mutual framework to follow.

Acknowledgments

We are grateful for the dedication of several instructors and contributors to the Applied Learning Health Systems program: Dr Mahima Kalla, Dr Kara Burns, Dr Chris McMaster, Associate Professor Brian Chapman, Dr Douglas Capurro, Reuben Daniels, Jennifer Morris, Dr Vlada Rozova, Anna Fedyukova, Dr Kit Huckvale, Professor Kathleen Gray, Associate Professor Graeme Hart, and Dr Debbie Passey.

Data Availability

All relevant data generated or analyzed during this study are included in this published article. The raw data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

SD played a key role in data curation, formal analysis, investigation, project administration, and visualization, and wrote the original draft, contributing to its review and editing. NZ focused on formal analysis. DC contributed by reviewing and editing the manuscript, ensuring its quality and coherence. WC conceptualized the project and also participated in the review and editing process. KL contributed to the investigation, methodology, supervision, writing the original draft, as well as reviewing and editing the final manuscript.

Conflicts of Interest

None declared.

References

1. Friedman CP. What is unique about learning health systems? Learn Health Syst 2022 Jul;6(3):e10328. [doi: 10.1002/lrh2.10328] [Medline: 35860320]
2. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. Sci Transl Med 2010 Nov 10;2(57):57cm29. [doi: 10.1126/scitranslmed.3001456] [Medline: 21068440]
3. Friedman CP, Allee NJ, Delaney BC, et al. The science of learning health systems: foundations for a new journal. Learn Health Syst 2017 Jan;1(1):e10020. [doi: 10.1002/lrh2.10020] [Medline: 31245555]
4. Detmer DE. Interprofessional clinical informatics education and practice: essentials for learning healthcare systems worldwide. J Interprof Care 2017 Mar;31(2):187-189. [doi: 10.1080/13561820.2016.1250554] [Medline: 28129005]
5. Gray K, Gilbert C, Butler-Henderson K, Day K, Pritchard S. Ghosts in the machine: identifying the digital health information workforce. Stud Health Technol Inform 2019;257:146-151. [Medline: 30741187]

6. McMahon M, Bornstein S, Brown A, Tamblyn R. Training for impact: PhD modernization as a key resource for learning health systems. Healthc Policy 2019 Oct;15(SP):10-15. [doi: 10.12927/hcpol.2019.25983] [Medline: 31755856]
7. Sim SM, Lai J, Aubrecht K, et al. CIHR health system impact fellows: reflections on“driving change” within the health system. Int J Health Policy Manag 2019 Jun 1;8(6):325-328. [doi: 10.15171/ijhpm.2018.124] [Medline: 31256564]
8. Grant RW, Schmittdiel JA, Liu VX, Estacio KR, Chen YFI, Lieu TA. Training the next generation of delivery science researchers: 10-year experience of a post-doctoral research fellowship program within an integrated care system. Learn Health Syst 2024 Jan;8(1):e10361. [doi: 10.1002/lrh2.10361] [Medline: 38249850]
9. Kalra A, Adusumalli S, Sinha SS. Cultivating skills for success in learning health systems: learning to lead. J Am Coll Cardiol 2017 Nov 7;70(19):2450-2454. [doi: 10.1016/j.jacc.2017.09.1086] [Medline: 29096813]
10. Wysham NG, Howie L, Patel K, et al. Development and refinement of a learning health systems training program. EGEMS (Wash DC) 2016;4(1):1236. [doi: 10.13063/2327-9214.1236] [Medline: 28154832]
11. Kohn MS, Topaloglu U, Kirkendall ES, Dharod A, Wells BJ, Gurcan M. Creating learning health systems and the emerging role of biomedical informatics. Learn Health Syst 2022 Jan;6(1):e10259. [doi: 10.1002/lrh2.10259] [Medline: 35036547]
12. Robinson CH, Thompto AJ, Lima EN, Damschroder LJ. Continuous quality improvement at the frontline: one interdisciplinary clinical team’s four-year journey after completing a virtual learning program. Learn Health Syst 2022 Oct;6(4):e10345. [doi: 10.1002/lrh2.10345] [Medline: 36263266]
13. Kasaai B, Thompson E, Glazier RH, McMahon M. Early career outcomes of embedded research fellows: an analysis of the health system impact fellowship program. Int J Health Policy Manag 2023;12:7333. [doi: 10.34172/ijhpm.2023.7333] [Medline: 37579439]
14. Lozano PM, Lane-Fall M, Franklin PD, et al. Training the next generation of learning health system scientists. Learn Health Syst 2022 Oct;6(4):e10342. [doi: 10.1002/lrh2.10342] [Medline: 36263260]
15. Dushyanthen S, Perrier M, Chapman W, Layton M, Lyons K. Fostering the use of learning health systems through a fellowship program for interprofessional clinicians. Learn Health Syst 2022 Oct;6(4):e10340. [doi: 10.1002/lrh2.10340] [Medline: 36263261]
16. Dushyanthen S, Choo D, Perrier M, et al. Designing an interprofessional online course to foster learning health systems. Stud Health Technol Inform 2024 Jan 25;310:1241-1245. [doi: 10.3233/SHTI231163] [Medline: 38270013]
17. Jones-Bonofiglio KD, Willett T, Ng S. An evaluation of flipped e-learning experiences. Med Teach 2018 Sep;40(9):953-961. [doi: 10.1080/0142159X.2017.1417577] [Medline: 29271281]
18. Barr H, Koppel I, Reeves S, Hammick M, Freeth D. Effective interprofessional education. In: Approaching Learning and Teaching: Blackwell Publishing Ltd; 2005:95-104. [doi: 10.1002/9780470776445]
19. Sherer M, Maddux JE, Mercandante B, Prentice-Dunn S, Jacobs B, Rogers RW. The self-efficacy scale: construction and validation. Psychol Rep 1982 Oct;51(2):663-671. [doi: 10.2466/pr0.1982.51.2.663]
20. Pajares F. Self-efficacy beliefs in academic settings. Rev Educ Res 1996 Dec;66(4):543-578. [doi: 10.2307/1170653]
21. Greenberg-Worisek AJ, Shippee ND, Schaffhausen C, et al. The learning health system competency appraisal inventory (LHS-CAI): a novel tool for assessing LHS-focused education needs. Learn Health Syst 2020 Apr;5(2):10.1002/lrh2.10218. [doi: 10.1002/lrh2.10218] [Medline: 33889729]
22. Forrest CB, Chesley FD, Tregear ML, Mistry KB. Development of the learning health system researcher core competencies. Health Serv Res 2018 Aug;53(4):2615-2632. [doi: 10.1111/1475-6773.12751] [Medline: 28777456]
23. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
24. Institute of Medicine (US) Roundtable on Evidence-Based Medicine; Olsen LA AD, Olsen LA, Aisner D, McGinnis JM. The Learning Healthcare System: Workshop Summary: National Academies Press (US); 2007. [doi: 10.17226/11903]
25. Enticott JC, Melder A, Johnson A, et al. A learning health system framework to operationalize health data to improve quality care: an Australian perspective. Front Med (Lausanne) 2021;8:730021. [doi: 10.3389/fmed.2021.730021] [Medline: 34778291]
26. Ellis LA, Sarkies M, Churruca K, et al. The science of learning health systems: scoping review of empirical research. JMIR Med Inform 2022 Feb 23;10(2):e34907. [doi: 10.2196/34907] [Medline: 35195529]
27. Dammery G, Ellis LA, Churruca K, et al. The journey to a learning health system in primary care: a qualitative case study utilising an embedded research approach. BMC Prim Care 2023 Jan 19;24(1):22. [doi: 10.1186/s12875-022-01955-w] [Medline: 36653772]
28. Friedman C, Rubin J, Brown J, et al. Toward a science of learning systems: a research agenda for the high-functioning learning health system. J Am Med Inform Assoc 2015 Jan;22(1):43-50. [doi: 10.1136/amiajnl-2014-002977] [Medline: 25342177]
29. Nakamura J, Csikszentmihalyi M. The concept of flow. In: Handbook of Positive Psychology 2002:89-105. [doi: 10.1093/oso/9780195135336.001.0001]
30. Cooper JL, Robinson P. The argument for making large classes seem small. New Drctns for Teach & Learn 2000 Mar;2000(81):5-16. [doi: 10.1002/tl.8101]
31. Kruger J, Dunning D. Unskilled and unaware of it: how difficulties in recognizing one’s own incompetence lead to inflated self-assessments. J Pers Soc Psychol 1999 Dec;77(6):1121-1134. [doi: 10.1037//0022-3514.77.6.1121] [Medline: 10626367]

32. Nolen SB, Horn IS, Ward CJ. Situating motivation. Educ Psychol 2015 Jul 3;50(3):234-247. [doi: 10.1080/00461520.2015.1075399]
33. Enticott J, Johnson A, Teede H. Learning health systems using data to drive healthcare improvement and impact: a systematic review. BMC Health Serv Res 2021 Mar 5;21(1):200. [doi: 10.1186/s12913-021-06215-8] [Medline: 33663508]
34. Casey JD, Courtright KR, Rice TW, Semler MW. What can a learning healthcare system teach us about improving outcomes? Curr Opin Crit Care 2021 Oct 1;27(5):527-536. [doi: 10.1097/MCC.0000000000000857] [Medline: 34232148]
35. Budrionis A, Bellika JG. The learning healthcare system: where are we now? A systematic review. J Biomed Inform 2016 Dec;64:87-92. [doi: 10.1016/j.jbi.2016.09.018] [Medline: 27693565]

Abbreviations

EMR: electronic medical record
IT: information technology
LHS: learning health system

Edited by TDA Cardoso; submitted 31.10.23; peer-reviewed by A Lakdawala, B Senst; revised version received 18.03.24; accepted 26.04.24; published 14.01.25.

Please cite as:
Dushyanthen S, Zamri NI, Chapman W, Capurro D, Lyons K
Evaluation of an Interdisciplinary Educational Program to Foster Learning Health Systems: Education Evaluation
JMIR Med Educ 2025;11:e54152
URL: https://mededu.jmir.org/2025/1/e54152
doi:10.2196/54152

Original Paper

Motivational Framing Strategies in Health Care Information Security Training: Randomized Controlled Trial

Thomas Keller[1], MSc; Julia Isabella Warwas[1], Prof Dr; Julia Klein[2], MSc; Richard Henkenjohann[3], MA; Manuel Trenz[3], Prof Dr; Simon Thanh-Nam Trang[4], Prof Dr

[1]Department of Business Education, University of Hohenheim, Stuttgart, Germany

[2]Research Group on Information Security and Compliance, University of Göttingen, Göttingen, Germany

[3]Chair of Interorganizational Information Systems, University of Göttingen, Göttingen, Germany

[4]Chair of Information Systems, in particular Sustainability, Paderborn University, Paderborn, Germany

Corresponding Author:
Thomas Keller, MSc
Department of Business Education
University of Hohenheim
Schloß Hohenheim 1
Stuttgart, 70599
Germany
Phone: 49 (0) 71145923430
Email: Thomas.Keller@uni-hohenheim.de

Abstract

Background: Information security is a critical challenge in the digital age, especially for hospitals, which are prime targets for cyberattacks due to the monetary worth of sensitive medical data. Given the distinctive security risks faced by health care professionals, tailored Security Education, Training, and Awareness (SETA) programs are needed to increase both their ability and willingness to integrate security practices into their workflows.

Objective: This study investigates the effectiveness of a video-based security training, which was customized for hospital settings and enriched with motivational framing strategies to build information security skills among health care professionals. The training stands out from conventional interventions in this context, particularly by incorporating a dual-motive model to differentiate between self- and other-oriented goals as stimuli for skill acquisition. The appeal to the professional values of responsible health care work, whether absent or present, facilitates a nuanced examination of differential framing effects on training outcomes.

Methods: A randomized controlled trial was conducted with 130 health care professionals from 3 German university hospitals. Participants within 2 intervention groups received either a self-oriented framing (focused on personal data protection) or an other-oriented framing (focused on patient data protection) at the beginning of a security training video. A control group watched the same video without any framing. Skill assessments using situational judgment tests before and after the training served to evaluate skill growth in all 3 groups.

Results: Members of the other-oriented intervention group, who were motivated to protect patients, exhibited the highest increase in security skills (ΔM=+1.13, 95% CI 0.82-1.45), outperforming both the self-oriented intervention group (ΔM=+0.55, 95% CI 0.24-0.86; *P*=.04) and the control group (ΔM=+0.40, 95% CI 0.10-0.70; *P*=.004). Conversely, the self-oriented framing of the training content, which placed emphasis on personal privacy, did not yield significantly greater improvements in security skills over the control group (mean difference=+0.15, 95% CI –0.69 to 0.38; *P*>.99). Further exploratory analyses suggest that the other-oriented framing was particularly impactful among participants who often interact with patients personally, indicating that a higher frequency of direct patient contact may increase receptiveness to this framing strategy.

Conclusions: This study underscores the importance of aligning SETA programs with the professional values of target groups, in addition to adapting these programs to specific contexts of professional action. In the investigated hospital setting, a motivational framing that resonates with health care professionals' sense of responsibility for patient safety has proven to be effective in promoting skill growth. The findings offer a pragmatic pathway with a theoretical foundation for implementing beneficial motivational framing strategies in SETA programs within the health care sector.

(JMIR Med Educ 2025;11:e73245) doi:10.2196/73245

KEYWORDS

cyber security; health care professionals; motivational framing; training programs; skill acquisition

Introduction

Background

In the digital era, information security has emerged as a pivotal challenge for a wide array of industries, most notably for critical infrastructure, where interruptions can exert detrimental impacts on public safety, the economy, and the daily lives of individuals [1]. Hospitals have become a primary target for cyberattacks due to the substantial financial value of sensitive medical data [2]. The *Ponemon Healthcare Cybersecurity Report 2023* indicates that 88% of health care organizations experienced an average of 40 cyberattacks in the past 12 months, significantly impacting patient care and causing financial losses amounting to millions of dollars [3]. The prevalence of incidents underscores the need to enhance measures of information security protection [4]. However, technological vulnerabilities are not the sole source of threat. The human factor frequently serves as a crucial gateway for criminals. Errors, negligence, and inadequate knowledge about countermeasures all contribute to successful cyberattacks [5]. In particular, a lack of security awareness encourages noncompliance with organizational security policies, thus leading to risky workplace behaviors such as using weak passwords and unsecured devices [6]. As a consequence, hospital staff unknowingly create vulnerabilities that can be exploited by attackers [7]. This problem is exacerbated by hospitals' increasing reliance on interconnected digital systems [2,8], including IoT-based health care solutions [9], and the high-pressure environment of health care work, which often prioritizes rapid patient care over rigorous security measures [10].

In response, health care organizations are expanding the implementation of Security Education, Training, and Awareness (SETA) programs to improve their employees' information security skills [11]. Despite these efforts, many SETA programs have proven ineffective in fostering secure working practices over the long term by focusing on formal regulatory requirements without addressing the work processes of health care professionals and, thus, the application contexts of users [12-15]. Adapting SETA programs to the tasks and inherent security threats that health care professionals face every day seems imperative [16]. However, this target group rarely perceives information security as part of their primary job responsibilities [17]. Despite the provision of domain- and context-specific knowledge through training, health care professionals may still rate the importance of safeguarding medical data in their daily workflows as low [18]. A substantial proportion of professionals even demonstrate a flawed understanding of the consequences of inadequate security practices, including ransomware attacks that access electronic health records or data manipulations that lead to treatment errors [5]. This inner devaluation of risky behaviors and their consequences threatens the confidentiality, availability, and integrity of sensitive patient data. In essence, motivational deficits on the part of the professionals can be a serious barrier for acquiring skills that are necessary to act in accordance with security policies at work—even if training content is adequately customized to match their daily professional activities.

Enhancing Training Engagement Through Motivational Framing Strategies

A recognized method of invigorating involvement in training and promoting compliance with information security policies at work applies *fear appeals*, which raise awareness of the *personal relevance* of counteracting security threats [19,20]. As Johnston et al [21] point out with reference to protection motivation theory (PMT), powerful fear appeals address both formal sanctions (eg, punishment and loss of valuable information) and informal consequences (eg, social disapproval and reputational damage) in future cases of noncompliance. Nevertheless, an exclusive emphasis on avoiding unfavorable outcomes may not be the most effective strategy to encourage engagement with training content, as evident in research on controlled versus autonomous motivation of learning [22]. Research based on self-determination theory [23] shows that controlled motivation, such as compliance driven by external pressure, often leads to superficial learning. By contrast, training approaches that cultivate autonomous motivation by connecting to individual goals or professional purpose are more likely to elicit deeper engagement and enduring behavioral change [24]. Moreover, where consequences for unnoticed violations are missing, there may be little long-term motivation to follow information security rules in daily work [19].

Motivational framing offers a promising alternative for establishing the personal relevance of secure workplace behaviors and encouraging their acceptance as an integral part of one's professional practice. It aligns the content and structure of a message or topic with the motivational states, pursued goals, or cherished values of the target audience, thereby increasing the topic's subjectively assessed importance [25]. Joyal-Desmarais et al [26] report that motivational framing has a strong impact when the message aligns specifically with the individual's intrinsic motivation or core values. Their meta-analysis of 702 experimental studies highlights how messages that are tailored to resonate with such personal reference points of the recipients can significantly increase the persuasiveness of provided information, thus contributing to marked changes in attitudes, intentions, and behaviors. This raises the question of which core values a security training in hospitals should best appeal to.

Since hospital staff provide assistance to others, their professional activities can most generally be termed prosocial [27]. According to the *dual-motives model*, however, prosocial behavior can be driven either by self-serving (self-oriented) or by altruistic (other-oriented) motives [28]. While the former focus on personal benefits, such as enhancing one's reputation, gaining recognition, or fulfilling job-specific duties as part of

one's paid employment, the latter prioritize the needs and well-being of others, such as ensuring the safety and dignity of patients [29]. A closer investigation of the *core professional values* endorsed by health care professionals strongly suggests that an other-oriented (ie, patient-oriented) framing of training content on information security could indeed function as the most effective prompt to encourage attention and engagement—an assumption that will be further elaborated in subsequent sections.

At this point, we can conclude that the "personal relevance" of averting security threats as a health care professional could become more of a voluntary activity in accordance with one's professional values than merely a pointless duty imposed by regulatory bodies if training participants are prompted to detect this accordance. The general idea of addressing a professional commitment as a motivational stimulus is also fully compatible with the Fogg Behavior Model (FBM). The FBM posits that the performance of a target behavior is contingent upon 3 factors: sufficient motivation, the capacity to execute this behavior, and an adept cue that prompts its execution. The occurrence of these factors must be simultaneous for the behavior to manifest [30]. A training program that incorporates appeals to those salient values that generally drive professional action and, thus, enriches practical applicability with targeted messages or cues, therefore, seems particularly adept for building individual capacities for secure workplace behaviors.

Against this background, this study seeks to determine whether a motivational framing of information security training content, which varies in its alignment with the professional values of the target audience, impacts the acquisition of information security skills to varying degrees.

Hypotheses Development

The training for health care professionals has been meticulously developed, integrating problem-solving approaches to learning and recommended features for video formats while delivering domain-specific, contextualized content (see section on Instructional and Assessment Design for details). Therefore, the investigated hypotheses focus on the "dual" strategies of motivational framing and their potential to establish a connection between the adoption of information security practices and the commitment to professional values.

Since core professional values form the foundation of health care practices and the inner ethical obligations of health care staff [31], appeals to these values should promote engagement with learning material in an effort to expand one's repertoire of security skills [32]. Among these values, patient-centered care stands out as a primary commitment, emphasizing the safety, dignity, and well-being of patients. Professional integrity, another critical value, entails adherence to ethical standards and accountability, particularly in safeguarding sensitive patient information. Altruism and compassion further underscore the motivation of health care professionals to prioritize patient needs, often going beyond contractual obligations to act in the best interest of those they serve [33]. Consequently, an *other-oriented* framing of training content, which directs the participants toward *protecting information of and about their patients*, should be most adept to foster the acquisition of information security skills. In support of this idea, White and Peloza [34] report that people are more likely to respond to other-oriented appeals in contexts where social responsibility is emphasized.

Nevertheless, self-oriented appeals, emphasizing the individual benefits of protecting personalized data, should not be without any effects. Support for this perspective comes from PMT [35], which argues that individuals are most likely to engage in protective behavior when they anticipate being affected individually by the consequences of successful threat aversion or their failure to do so, while concurrently believing that they possess the knowledge and tools to respond adequately. Self-oriented framing consequently stresses these personal risks and the importance of self-protection. It makes consequences, such as service disruption or privacy loss, more immediate. Research supports the notion that perceived personal vulnerability drives individuals to adhere to safety measures. AlSobeh and colleagues' [36] findings suggest that adolescents who perceive cybersecurity risks to impact their personal quality of life demonstrated a heightened level of security awareness. In accordance with this line of thinking, self-oriented framing is appealing to defenses of one's own (informational) integrity.

Therefore, we expect an other-oriented framing of information security training, accentuating patient protection, to promote greater skill acquisition than delivering the training without any framing. It should also be superior to self-protection appeals, which should again be more effective than no framing at all. This leads to the following hypotheses:

- Hypothesis 1: Participants who receive a self-oriented framing of training content will demonstrate higher levels of skill acquisition compared to a control group without any framing.
- Hypothesis 2: Participants who receive an other- (ie, patient-) oriented framing will demonstrate higher levels of skill acquisition compared to those with a self-oriented framing.

Methods

Instructional and Assessment Design

Enhancing Information Security Skills Through Job-Specific, Problem-Oriented, Video-Based Training

In determining the relevant skillset to be acquired, we drew on a framework that was developed specifically for information-secure workplace behaviors [37,38]. The model proposes professional activity to be targeted, knowledgeable, justifiable, and responsible [39,40]. It further aligns with models of (expert) problem-solving [41,42] by positing that a proficient handling of security threats spans a comprehensive cycle of reasoned action. This cycle starts with an early detection of risky situations within one's own working environment and extends up to the execution of follow-up procedures that are beneficial for the whole team or organization. Thus, the skillset comprises seven elements. (1) *Threat awareness* emphasizes the ability to recognize potential security threats and remain vigilant, distinguishing between threatening and nonthreatening work situations. (2) *Threat identification* focuses on accurately

identifying the presence of a security threat and understanding its specific nature. (3) *Threat impact assessment* ensures that individuals understand the consequences of not addressing current security threats. (4) *Tactic choice* enables individuals to select appropriate measures by aligning their actions with prevailing best practices and established rules of information security while tailoring those actions to the specific threat scenario they face. (5) *Tactic justification* emphasizes the importance of a structured, goal-oriented approach in which individuals can justify their actions based on relevance, effectiveness, and superiority. With (6) *tactic mastery,* individuals can effectively implement the security measure of choice. Finally, (7) *tactic check and follow-up* involves evaluating the effectiveness of implemented measures and, where possible, taking follow-up actions to further improve the organization's security.

A critical evaluation of the strengths and limitations of various training methods (for an overview, see [43]) revealed that video-based training offers a promising solution for the health care sector. First, it provides a largely visual presentation of content, thereby facilitating the acquisition of knowledge and skills regardless of the learners' literacy [44,45]. Second, health care professionals often have unpredictable schedules and limited time for instructor-led classroom training. In such circumstances, training videos provide a convenient and accessible learning route, ensuring the delivery of essential strategies and reasoning in a clear, concise, and engaging manner [46]. Third, the standardization and repeatability of video-based training material fosters the dissemination of consistent, high-quality information to all employees, enabling them to progress through the material at their own pace [47].

Several research-based recommendations were implemented in the video production. Given the condensed information typically conveyed in this format, empirical studies suggest limiting the duration of training videos to no more than 6 minutes to maintain attention and engagement [48,49]. Storytelling and sensory activation [50,51] appeal to different learning styles [52] and promote information retention. At the beginning of each training video, participants are presented with general orienting information, including the learning objectives to be achieved, the approximate completion time, and an explanation of the badges to be earned [53]. These badges provide transparency of individual learning progression (beginner, advanced, or expert), thereby contributing to learner motivation and self-directed learning [54]. Upon completion, participants have mastered all training content for analyzing and averting security threats. Figure 1 illustrates these design principles as a multipanel overview covering the key visual elements of the training: (A) introduction, (B) learning objectives and badges, (C) other-oriented nudge, and (D) the "tactic justification" dimension.

Figure 1. Design highlights from the training module "handling sensitive and critical information on social networks.".

In line with the general objective of SETA programs described in the introduction, the training aimed to develop the ability to select and implement safe behaviors that are compatible with the respective employee's daily work and that are particularly relevant to the risk factors present there. Based on a comprehensive analysis of information security behavior in the health care sector [16], different *threat vectors* have been identified for typical *job profiles*. These job profiles differ in their levels of interaction with patients and their reliance on IT systems, which directly influence their exposure to information security risks.

Physicians interact with patients frequently and use IT systems moderately to extensively, depending on their responsibilities for diagnosis, treatment planning, and documentation. Common

threats they face include the *inadvertent disclosure* of sensitive information, for example, by forwarding patient-related data via private messaging services, as well as *spear phishing*, where attackers send personalized emails designed to appear legitimate.

Nurses typically have the most frequent patient contact but rely on IT systems only to a limited extent, mainly for tasks such as documentation and medication administration. They, too, are particularly vulnerable to *inadvertent disclosure* of sensitive information, such as through social networking or the use of private messaging services, which could be used for expeditious communication.

Administrative staff rarely interact directly with patients but rely heavily on IT systems. Tasked with the critical function of managing both digital and physical (patient) data, scheduling, and billing, they are frequent targets of *phishing attacks*, such as fake websites. Additionally, a disorganized workspace increases the probability of sensitive information being exposed on desks or computer screens. Failure to comply with the hospital's *clean desk policy* thus increases their risk of granting unauthorized access and violating data confidentiality.

By assessing the criticality and phenomenology of these threat vectors through expert interviews [16], we designed customized training videos that specifically address these prevalent and consequential security threats of different job profiles. In line with the problem-solving skills model described above, the videos take the occurrence of threat vectors as the starting point of a problem-solving process in a prototypical workplace scene. Thus, each video begins with visual or auditory stimuli that represent variants of an authentic workplace situation that could pose an information security threat, encouraging employees to assess associated risks and participate in finding a solution [55].

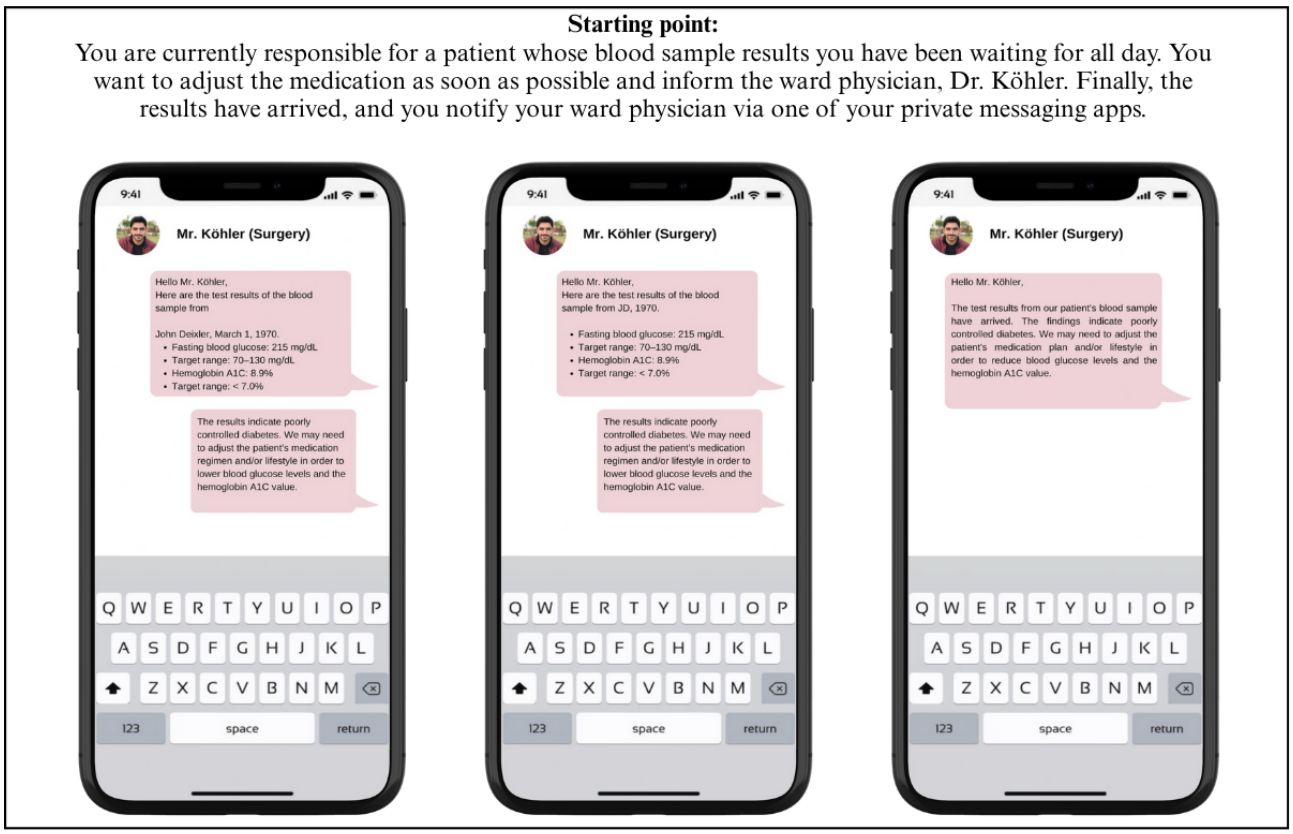
Creating Authentic Learning Assessments Through Situational Judgment Tests

Situational judgment tests (SJTs) provide a robust approach to assessing general and job-specific procedural knowledge in an authentic environment. They are applicable to a wide range of fields, including personnel selection and development, medical licensing and certification, education, and psychological assessment (eg, personality) [56-58]. Moreover, they hold significant potential for expanding into emerging applications, including health behavior and interpersonal skills [59]. SJTs are characterized by a strong intuitive prompt that asks test takers to place themselves in specific situations [60]. In this study, participants face a security threat scenario. Their task is then to select the most adept judgments and the most effective actions from several alternative responses, which are presented hereafter. In accord with the problem-solving cycle delineated above, which guided the participants through the training videos, the test items for evaluating their skill gains cover 7 dimensions, ranging from *threat awareness* to *tactic check and follow-up*. Each of these dimensions was represented by one test item within a thematically coherent testlet based on a realistic threat scenario.

Thus, the central tenets of the developed SJTs lie in their contextualization and authenticity. By immersing participants in a fictional university hospital and presenting prototypical threat scenarios through authentic image and video stimuli, these SJTs provide a cost-effective and valid method to assess employees' capabilities to manage these threats without the need to simulate factual security incidents. This approach is particularly salient in professional contexts where interruptions pose a significant threat to patient health and, in extreme cases, patient lives. The impracticality, cost, and ethical challenges associated with extensive simulations of, for example, data breaches render SJTs a compelling alternative. An example of such a scenario is shown in Figure 2, illustrating the *threat awareness* dimension of the SJT. Participants were asked to evaluate the relative severity of various messaging behaviors that could lead to the inadvertent disclosure of sensitive patient data via private messaging apps, and to rank them using the following scale: 1=most threatening, 2=less threatening, and 3=least threatening.

In several workshops, we not only developed problem-based threat scenarios but also professionally grounded response options for information-secure actions. The scenarios and response options were iteratively revised in close collaboration with information security experts and test developers, who paid particular attention to their alignment with the respective skill dimensions and security policies. To allow for a differentiated assessment of both high and low levels of the targeted construct, the number of response options per item was limited to 4 to 6. The tests used a forced response format with different response types, including single choice, multiple choice, and rating, to enable a nuanced assessment of decision-making and judgment in information security contexts. To examine the quality of the testlets, we conducted a pretest with 100 physicians (37 female, 63 male; mean age 34.73, SD 10.95 years), 101 nurses (41 female, 60 male; mean age 41.40, SD 9.85 years), and 102 administrative staff (24 female, 76 male, 2 diverse; mean age 28.26, SD 7.29 years). The results indicate an acceptable level of difficulty for the target population, as reflected in the range of the item difficulty index (0.71-0.78). In addition, key metrics such as variance and discrimination index confirmed that the test items effectively detect interindividual differences in test performance. Consistently positive feedback on the usability and authenticity of the test underscored its acceptance among the target groups [37].

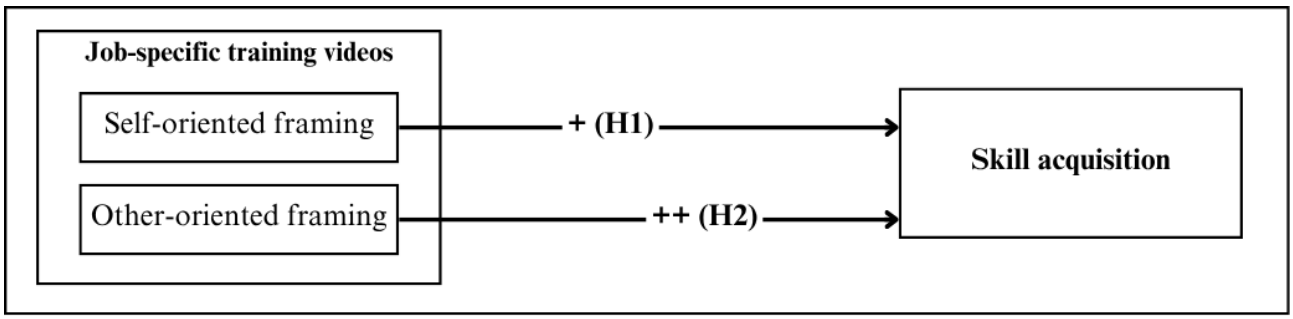
Figure 2. Example item from the situational judgment test targeting the dimension "threat awareness.".



Intervention Implementation

To test our hypotheses and proposed research model (Figure 3), we conducted a randomized controlled trial with hospital staff of 3 German university hospitals in late 2023. Data were collected via the online platform Qualtrics. The participating organizations were free to decide whether to share the link to this platform in their intranet or via newsletter, sent from either their marketing or their information security department. The training phase was open for 2 weeks, during which time participants could schedule the training at their convenience. At the beginning, participants were asked to indicate their job profile. Based on this information, participants were assigned a job-specific training video and corresponding pre- and posttests, each tailored to the demands and risks associated with their professional role. An attention check in the form of a single content-related question was included after the video, without revealing or overlapping with any later test content. All participants correctly answered the attention check question, indicating continuous attention to the video content. Additionally, the time participants spent on the training page was recorded as an indicator of intervention exposure. To ensure standardization and minimize external influences, all participants completed the baseline-skill assessment (pretest), the video intervention with or without motivational framing, and the achieved-skill assessment (posttest) in immediate succession on the platform.

Figure 3. Conceptual model illustrating the hypothesized effects of self-oriented and other-oriented motivational framing of job-specific training videos on skill acquisition.



Randomization and Blinding

Within each job profile, participants were randomly assigned to either the treatment or control condition. The allocation was conducted using simple randomization through the online platform at the start of the study. This automated process was embedded in the predefined succession of tests and training (see above) to ensure allocation concealment and prevent any influence from the researchers or participants on group assignment. Professionals in the control group were provided

only with their respective job-specific training videos. Those in the two intervention groups watched the same videos (for their profile) with an additional motivational framing at the beginning that was either self-oriented or other-oriented (Table 1). The total length of the training videos was approximately 5 minutes of pure information security content and an additional 30 seconds for a motivational framing. Participants were unaware that different motivational framings were used in the training materials and assumed that all participants received the same intervention. Skill assessment in both the pre- and the posttests was fully automated through a predefined scoring algorithm. No manual input or subjectively biased coding influenced the outcome data, ensuring objectivity and consistency.

Table 1. Overview of motivational framing strategies in the intervention groups.

Group	Motivational framing
Other-oriented framing (n=43)	The dangers posed by incorrect IT security behavior are very high and can lead to the interruption of services or the disclosure of sensitive data. Therefore, it is important that you are able to protect your patients. The content from this training will help you build the skills necessary to protect your patients and their privacy.
Self-oriented framing (n=43)	The dangers posed by incorrect IT security behavior are very high and can lead to the interruption of services or the disclosure of sensitive data. Therefore, it is important that you are able to protect yourself. The content from this video will help you build the skills necessary to protect yourself and your privacy.
Control (n=44)	No motivational framing.

Measures

The intervention's efficacy was evaluated using SJTs, which were developed to assess information security skills. To serve the purpose of pre- and posttraining assessments, these SJTs used 2 threat vectors from the same risk category for each job profile. For physicians and nurses, the risk category of inadvertently disclosing sensitive information was operationalized in two ways: (1) disclosing sensitive patient data via private messenger services and (2) publishing images via social networks. For administrative staff, two threat vectors were developed to reflect noncompliance with the clean desk policy: (1) leaving confidential documents openly visible on the desk and (2) leaving sensitive electronic data accessible by failing to lock the computer when unattended. Therefore, the pre- and posttest scenarios and their related items were not identical, although they followed the same structure and belonged to the same risk category. Nonetheless, the SJTs were presented in a randomized order to minimize potential biases arising from test familiarity. Each participant was randomly assigned 1 of the 2 SJTs corresponding to their job profile for the baseline assessment. The other test served to measure posttraining skill levels. Each test item, corresponding to 1 of the 7 information security skill dimensions, was scored from 0 to a maximum of 2 points, resulting in a total score of up to 14 points per testlet. Age and gender were included as control variables to account for potential demographic influences on skill gains.

Participants

A total of 130 participants were included in the study, with a balanced gender distribution (53.8% male and 46.2% female). Their age ranged from 18 to 71 years (mean 42.6, SD 12.21 years). Regarding job profiles within the health care sector, 50% of the participants primarily fulfilled administrative tasks in their respective hospitals, 30.8% were occupied as nurses, and another 19.2% as physicians. This diverse sample composition allows for a comprehensive analysis of the effects of the intervention across different job profiles.

Power Calculation

A power analysis was conducted using the R package pwr [61] to determine the necessary sample size for a 1-way ANOVA with 3 groups. Assuming a medium effect size (f=0.30 [62]), an alpha level of .05, and a desired statistical power of 0.80, at least 37 participants would be needed per group, which was exceeded in the present sample. The underlying assumption was supported by effect sizes reported in comparable studies on motivational framing and message matching, which demonstrated moderate effects on attitude and behavior change [26].

Data Analysis

Before running the main analysis, we compared the baseline characteristics of the experimental groups to ensure group equivalence at the beginning of the study. Categorical variables were analyzed using chi-square tests, while age was examined using 1-way ANOVA. Baseline skill levels (pretest scores) were compared using Kruskal-Wallis tests to account for potential violations of normality. No significant differences were found between groups with respect to age (P=.39), gender (P=.95), or job profile (P=.91). However, baseline skill levels differed significantly between groups, as indicated by the Kruskal-Wallis test (χ^2_2=9.4, P=.009). Therefore, the baseline skill level was included as a covariate in subsequent analytical steps. This allowed us to adjust for initial differences in participants' prior knowledge and to isolate the effect of motivational framing on skill gain with greater precision.

Before testing hypotheses of motivational framing, exploratory paired-samples t tests within each job profile were conducted. This step established the training intervention's baseline efficacy, allowing the interpretation of any estimated framing effects as modulators of learning processes. Its outcome is reported in the descriptive results below.

Building on this, we examined the expected differential extents of skill acquisition yielded by the applied types of motivational framing. An analysis of covariance (ANCOVA) was performed with skill acquisition (measured by the difference between posttest and pretest scores) serving as the dependent variable

and the experimental condition (control, self-oriented, or other-oriented) representing the independent variable. The requirements for applying ANCOVA are fulfilled. Observations were independent by design, as each participant was assigned to only one experimental condition. The assumption of homogeneity of variances was met, as indicated by the Levene test ($F_{2,127}$=0.37; P=.69). To detect potential outliers in skill acquisition within each experimental group, we calculated z scores separately for each group. No outliers were identified based on this criterion, suggesting that extreme values in skill acquisition were not present within any experimental condition. Although the Shapiro-Wilk test (W=0.967; P=.003) indicated a slight deviation from normality, given the sample size and the demonstrable absence of extreme outliers, the ANCOVA can be considered to deliver robust estimations [63]. The assumption of homogeneity of regression slopes was also fulfilled, as the interaction between group and pretest score was not significant ($F_{2,124}$=0.30; P=.74). All statistical analyses were conducted using R version 4.3.2 (October 31, 2023; R Core Team).

Ethical Considerations

The study was reviewed by the ethics committee of the Georg-August University of Göttingen, Germany (approval issued May 8, 2023). In an official statement issued in May 2023, the committee raised no ethical concerns about implementing the project as long as all applicable data protection regulations were followed. All participants were informed of the purpose of the training and how the tests would be used to evaluate its effectiveness. Participation was voluntary, and informed consent was obtained beforehand. Data handling complied with the General Data Protection Regulation and institutional data protection policies. No financial or material compensation was provided to participants.

Results

Descriptive Results

Building on the sample description above, Figure 4 depicts the flow of participants through recruitment, allocation, and analysis in accordance with the CONSORT (Consolidated Standards of Reporting Trials) guidelines [64] (checklist in Multimedia Appendix 1). Table 2 shows the descriptive study results for the 3 job profiles sorted by the experimental groups. Paired-samples t tests (2-tailed) indicate significant gains in information security skills in all job profiles following the training intervention. For administrative staff, the analysis reveals a significant improvement from pretest scores to posttest scores (t_{64}=3.30; P=.002) with a mean difference of 0.51 (95% CI 0.20-0.81) and an effect size of Cohen d=0.41. Similarly, a significant increase was observed for nurses (t_{39}=2.63; P=.01), with a mean difference of 0.65 (95% CI 0.15-1.15) and an effect size of d=0.48. The greatest improvement was found among physicians (t_{24}=4.11; P<.001), with a mean increase of 1.24 (95% CI 0.62-1.86) and an effect size of d=1.21. Taking an average across all job profiles, information security skills reached significantly higher levels after the training compared to before the training, with a mean difference of 0.69 (95% CI 0.45-0.94; t_{129}=5.58; P<.001), and an overall effect size of d=0.52, indicating a medium effect. These findings document that skill gains were not limited to an average value for the entire sample, but were consistently observed within each job profile.

Figure 4. CONSORT flow diagram illustrating the progression of participants through the randomized controlled trial. CONSORT: Consolidated Standards of Reporting Trials.

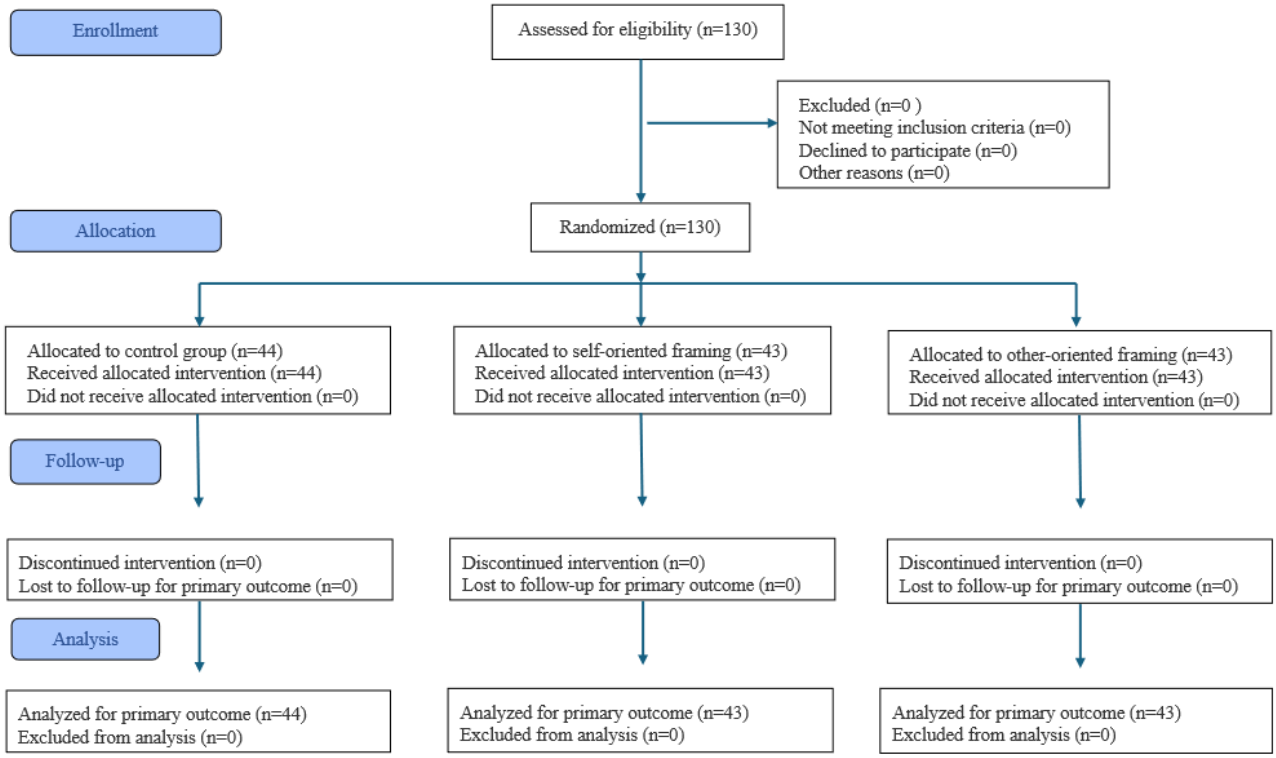


Table 2. Pretest, posttest, and skill acquisition scores by experimental group and job profile.

Intervention condition and job profile	Control group (n=44), mean (SD)	Self-oriented framing (n=43), mean (SD)	Other-oriented framing (n=43), mean (SD)
Administrative staff (n=65)			
Pretest	12.42 (1.25)	12.81 (1.33)	12.05 (1.28)
Posttest	12.67 (1.20)	13.00 (1.22)	13.20 (1.06)
Skill acquisition	+0.25 (1.33)	+0.19 (1.12)	+1.15 (1.04)
Nurses (n=40)			
Pretest	10.92 (1.66)	12.07 (1.27)	10.77 (1.54)
Posttest	11.54 (1.13)	12.07 (1.07)	12.15 (0.99)
Skill acquisition	+0.62 (1.66)	0.00 (1.36)	+1.38 (1.45)
Physicians (n=25)			
Pretest	12.86 (0.69)	12.13 (1.13)	11.20 (1.32)
Posttest	13.14 (0.69)	13.13 (0.83)	13.30 (0.67)
Skill acquisition	+0.29 (0.76)	+1.00 (1.31)	+2.10 (1.66)

Impact of Motivational Framing Strategies on Skill Acquisition

While overall improvements in information security skills are evident, the question remains whether the specific type of motivational framing influenced their magnitude. To address this, we tested hypotheses 1 and 2 using an ANCOVA procedure. The results reveal a significant main effect of motivational framing on skill acquisition, controlling for initial skill level ($F_{2,126}$=5.92; P=.004; partial η^2=0.09). Post hoc pairwise comparisons with Bonferroni adjustment across all job profiles indicate that the other-oriented framing group achieved significantly greater skill growth (mean +1.133, SD 0.159) compared to the control group (mean +0.400, SD 0.154; t_{126}=3.30; P=.004), as well as to the self-oriented framing group (mean +0.551, SD 0.158; t_{126}=2.54; P=.04). No significant differences were found between the control and self-oriented framing groups (t_{126}=0.687, P>.99). In addition to the effect of the framing strategy, a substantial portion of the variance in skill growth can be attributed to each participant's baseline skills ($F_{1,126}$=86.88; P<.001; partial η^2=0.41). Participants who started at a higher skill level demonstrated smaller gains from the training videos than those who started at lower or even rudimentary skill levels. These findings offer substantial evidence for hypothesis 2, which predicted the other-oriented framing of training content to be superior to a self-oriented framing in the studied target audience of health care professionals. Contradicting hypothesis 1, however, self-oriented framing did not yield greater enhancements of information security skills than no framing at all.

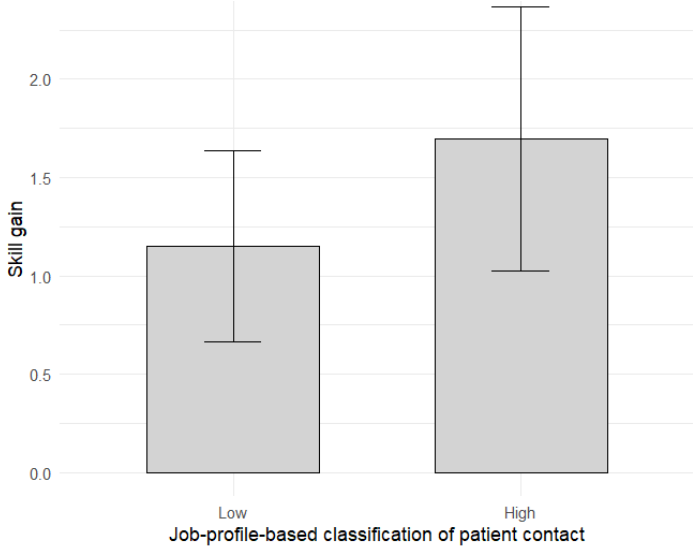
Furthermore, a closer inspection of the descriptive means presented in Table 2 suggests that the generally superior effectiveness of the other-oriented framing (see hypothesis 2 above, comparing intervention and control groups) appears to resonate more strongly with certain job profiles than with others. Physicians demonstrated the largest improvement (mean increase +2.10), followed by nurses (mean increase +1.38), and administrative staff (mean increase +1.15), whose gains, although statistically significant, were markedly smaller. This variability suggests that although all professional sectors within hospitals responded to an other-oriented motivational framing of training content, the extent of responsiveness may be influenced by individual or contextual characteristics. Drawing on models of social responsibility and professional identity [31,32], the degree of personal and direct patient contact may influence how strongly hospital staff respond to other-oriented appeals. Specifically, the frequency with which one engages personally and directly in medical and care work, as well as ethically charged decision-making, may further increase the salience of other-oriented appeals, thereby enhancing their persuasive potency.

A rigorous moderation analysis examining other-oriented framing effects contingent on contact frequency was not feasible, as a direct and isolated measure of individual patient contact was unavailable in this study. However, contact frequency is typically, although not deterministically, higher for certain job profiles than for others (see profile descriptions above). Consequently, we grouped nurses and physicians to compare them to the administrative staff. Nurses and physicians are embedded in the core clinical care process. They operate in close proximity to patients and often form interprofessional treatment teams within inpatient settings. Both focus on health-related goals and are most immediately responsible for the well-being of vulnerable individuals. Administrative staff primarily perform organizational and procedural tasks that serve patients' best interests, such as ensuring they receive swift treatment and proper accommodations. However, most of these tasks are structurally removed from the clinical care environment.

Therefore, we conducted a descriptive subgroup analysis restricted to participants who received the other-oriented framing. Specifically, we compared skill gains between roles involving high (n=23) and low (n=20) frequencies of direct patient contact. We applied the Welch t test, which yields robust

estimates. While the high-contact group showed greater skill gains (mean increase 1.70, SD 0.150) than the low-contact group (mean increase 1.15, SD 0.104), this difference did not reach statistical significance, $t_{38.67}$=1.37, *P*=.18, 95% CI –1.35 to 0.26. Figure 5 illustrates this descriptive pattern, showing that when receiving training content that is framed toward patient protection, participants whose professional roles typically involve frequent direct patient contact experience an even stronger increase in skill gain than those in lower contact-frequency roles. However, this observation is limited to the present sample.

Figure 5. Mean skill gain by the "profile-typical" frequency of patient contact within the other-oriented framing condition.



Discussion

Principal Findings

This study sheds new light on the effectiveness of motivational framing strategies in information security training for health care professionals. Participants of a video-based, job-specific training demonstrated significant improvements in information security skills. These improvements were observed across all job profiles, including administrative staff, nurses, and physicians. While empirical evidence that a contextualized and psychologically grounded content increases the effectiveness of security training has already been provided by previous studies [65], our randomized controlled trial with 130 participants demonstrates that the targeted activation of professional values can further enhance skill acquisition.

In particular, our results underscore the widely recognized limitations of traditional SETA programs, which focus on formal content and guidelines without addressing motivational aspects [66]. In this study, both the control group and the group that received a self-oriented motivational framing achieved comparably small skill gains. This indicates that appeals to individual protection or self-interest do not prompt heightened engagement with training content on information-secure workplace behaviors in professions with high levels of social responsibility, even if this content is well tailored to their daily tasks. Therefore, hypothesis 1 was not supported. This finding contrasts with studies from other professional sectors that used self-oriented *fear* appeals to promote secure behavior [67,68]. While self-oriented motivational framings have proven effective in private settings, such as protecting personal data, they appear to be less effective in clinical settings where the exertion of professional roles is strongly guided by professional-ethical values. As set out in the introduction, patient-centered care represents a primary commitment of health care professionals, which prioritizes the safety, dignity, and well-being of patients. Such concepts are integral to professional identity in the health care sector [33].

Consequently, and in accordance with hypothesis 2, framing security training content in an other-oriented way, thereby highlighting the protection of patients through information-secure workplace behaviors, proved to be the superior strategy. Participants in this intervention group demonstrated significantly greater improvements in information security skills compared to those in the control and self-oriented intervention groups. These results underscore the importance of aligning the messages that precede and accompany a training program with the core professional values of the target audience. The principle of value congruence [69], which posits that individuals respond more positively to messages that resonate with their core values, further substantiates this interpretation. Similar effects have been demonstrated in prosocial messaging research, where other-oriented appeals have a markedly stronger impact than self-oriented ones when applied in morally salient contexts [34]. Our findings extend this body of work to cybersecurity education, suggesting that value congruence can increase the effectiveness of SETA programs in socially responsible professions.

Supplementary exploratory examinations of subgroups indicate that certain variations in individual skill gains within the (generally superior) other-oriented framing condition might depend on the frequency of patient interaction. Participants whose job profiles typically require frequent direct and personal contact with patients, that is, physicians and nurses, displayed even greater increases in information security skills when the

training content was framed in an other-oriented manner than administrative staff participants did. This (merely descriptive) pattern is consistent with identity-based motivation theory [70], which posits that individuals are more likely to respond to messages that activate their specific role characteristics in a social or professional community. Thus, it provides initial support for the idea that the frequency of patient interactions (as an individual feature of participants) may further enhance the general responsiveness of health care professionals to other-oriented approaches.

Taken together, the findings highlight the importance of value-based design in SETA programs. Aligning training content with the core professional values of health care workers, such as patient protection and social responsibility, can help interventions achieve better learning outcomes. This interdisciplinary approach is a promising avenue for future research and practice on cybersecurity education.

Practical Implications

Not only the obtained empirical results but also the applied instructional and assessment design provide practical implications for the construction of SETA training and the implementation of motivational framing strategies. Structuring information security content around a substantiated skill model and tailoring it to employees' specific tasks and demands can arguably promote attention and engagement levels throughout the training. Put simply, professionals are more encouraged to extend their skill repertoire when the content is directly applicable to their daily work routines. Furthermore, tailored training programs are effective in reducing the duration of the program by eliminating irrelevant information. This approach prevents participants from feeling either overwhelmed by abstract input or overlooked as an agentic professional. Moreover, it facilitates an efficient use of training resources.

Explaining the repercussions and rationales for security-compliant conduct to participants is another important feature of effective training programs. Employees who comprehend the potential consequences of noncompliance, including data breaches and compromised patient safety, are more inclined to adopt secure practices (“threat impact assessment” according to the skill model in the present training design). The articulation of the reasons underlying particular security measures ensures that employees not only adhere to the prescribed procedures blindly (if at all) but also discern their significance (“tactic justification” according to the skill model in the present training design). In the long run, this can foster a culture of accountability and vigilance within the organization by promoting consistent and deliberate practice of safe behaviors.

Our empirical findings emphasize that a motivational framing of information security training should align with employees' professional values to maximize its impact. In the context of health care, patient protection was identified as a core value that can demonstrably be addressed through appropriate framing strategies. In other professional fields, formulating an adequate motivational framing necessitates a thorough understanding of the professional values advocated by the respective workforce.

Another recommendation is to introduce the topic of information security early in a professional’s development. This can be done through customized, easily accessible videos, such as those used in this study, or even through playful formats. Addressing this topic during the initial phases of formal vocational education can foster the understanding that information security is an integral part of responsible health care work. Initial vocational education provides more time for reflection on the compatibility of professional values and the protection of sensitive data than short-term pedagogical interventions that accompany daily professional routines and must therefore rely on short framing messages.

From an evaluative perspective, the study underscores the importance of accounting for baseline skills, particularly in light of ceiling effects, where participants with higher pre-training skill levels tend to show less growth. This underlines the need for adaptive and personalized training interventions that can optimize learning outcomes across various starting points [71]. By assessing skill levels across a range of skill sets prior to an intervention, organizations can more effectively target their training efforts to focus on areas where improvement is truly needed. Resources can be allocated to address individual skill gaps rather than providing redundant and one-size-fits-all training to the entire workforce. In the long run, this allows each employee to experience individual skill development throughout the training process, but also facilitates meaningful evaluation of the training program.

Limitations and Future Research

The study has several limitations that must be acknowledged. First, the investigation was conducted in German university hospitals with a sample of 130 health care professionals, which restricts the generalizability of the findings to other national, cultural, and organizational contexts. While patient protection served as a powerful motivator in the investigated health care context, other professional fields (eg, finance, public administration, and manufacturing) may be guided by entirely different core values, such as diligence, transparency, or efficiency. Further efforts are therefore required to develop problem-oriented, context-sensitive, and profession-specific training content for other occupational fields, as information security demands and role-related responsibilities differ substantially across sectors. In particular, the successful design of motivational framing strategies in these sectors requires a deep understanding of their respective professional ethos.

Second, the study examines short-term effects on skill acquisition, leaving open the question of long-term retention. Future research should focus on the long-term effects of information security interventions, ideally by using follow-up assessments weeks or months after the initial training. This would allow researchers to determine whether the training effects are stable, fade over time, or possibly even increase through continued application in practice.

Another limitation concerns potential self-selection bias. As participation in the study was voluntary and recruitment was conducted via internal channels, such as newsletters and hospital intranet systems, individuals with a stronger interest in information security, greater intrinsic motivation, or a higher

sense of professional responsibility were likely to be overrepresented. This may have led to an overestimation of intervention effects. It certainly helps to explain the relatively high baseline skill levels observed across the sample. Participants with advanced prior knowledge may be overrepresented compared to the general hospital workforce. Incorporating random sampling procedures or mandatory participation (in addition to the reported random assignment to framing conditions) could help to ensure a more representative selection of participants.

Results from exploratory subgroup comparisons for the generally beneficial other-oriented framing condition should encourage future investigations into moderator variables that capture individual features of health care professionals. In this study, a grouping approach to contrast job profiles that typically involve high versus low personal contact with patients (physicians and nurses versus administrative staff) provides preliminary support of the assumption that individual interaction frequency with patients might further enhance the impact of patient-oriented appeals. However, this pattern should be interpreted with caution, as the grouping by job profiles offers only a rough approximation of the psychological conditions that may shape responsiveness to value-based framings. Though commonly used to classify professional functions, job profiles often conceal variability within these profiles. For instance, not all physicians are equally involved in direct patient interaction, while certain administrative staff, such as case managers or receptionists, may routinely interact with patients. To specify features that further enhance or reduce the effectiveness of other-oriented framings in the health care sector, future research should therefore consider direct and controllable measures of individual-level moderators. This approach may also include individual assessments of empathy, moral sensitivity, or motivational regulation styles [72,73]. On a contextual level, variables such as organizational culture or leadership commitment to cybersecurity might also be considered as moderating the perceived relevance of information-secure behavior, engagement with training content, or responsiveness to different motivational framing strategies [74].

Finally, this study has put forth several didactical arguments to substantiate that demonstrable skill gains occur because instructional design elements and, in particular, motivational framing strategies encourage deeper engagement with the training content. Studies in other educational contexts have already established the mediating role of learning-process characteristics such as cognitive elaboration and engagement [75]. For information security training in health care settings, statistical evidence of the mediated learning path is still pending.

Conclusions

The increasing prevalence of cyberattacks and data breaches, particularly in the health care sector, underscores the need to enhance the information security skills of health care professionals through SETA programs. This experimental study highlights the effectiveness of customized security training videos with strategically framed motivational prompts in improving related skills, as assessed by elaborated SJTs. A key aspect of promoting skill acquisition of professionals is to elucidate that compliance with security policies can be fully compatible with their core professional values and that mastering information secure workplace behaviors contributes to enacting these values.

At least in the health care sector studied, where information security breaches have been documented to jeopardize patient safety, this alignment appears to be beneficial. By establishing a link between internalized professional responsibilities to protect patients and methods to ensure the confidentiality, availability, and integrity of sensitive patient data, employees develop a deeper commitment to security policies and to acquiring the necessary skills. In the long run, motivational framing strategies in information security training can become an important tenet in building organizational resilience to cyberattacks. They can not only raise awareness of security-related issues in day-to-day operations but also cultivate the deliberate and regular use of information-secure behaviors.

Acknowledgments

This research paper has been developed as part of the research project “KISK: Kompetenzbasierte Entwicklung von ITS-Trainingsangeboten in Krankenhäusern,” funded by the German Federal Ministry of Health. We gratefully acknowledge the support of the Federal Ministry of Health and all other project members, especially Dr. Kristin Masuch, Florian Schütz, and Florian Rampold. We would also like to thank the Chief Information Security Officers of the participating hospitals, whose cooperation made this research possible. Publishing fees are supported by the funding program Open Access Publishing of the University of Hohenheim.

Data Availability

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

Authors' Contributions

TK, JW, and JK wrote and revised the original manuscript draft. TK performed the investigation, designed the methodology, and performed the formal analysis. JW designed the methodology and performed the formal analysis as well as project administration. JK performed the investigation. RH conceptualized the study and performed the investigation. MT conceptualized the study,

reviewed and edited the manuscript draft, and performed project administration. ST reviewed and edited the manuscript draft and performed project administration.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT - eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 591 KB - mededu_v11i1e73245_app1.pdf]

References

1. Chowdhury N, Gkioulos V. Cyber security training for critical infrastructure protection: a literature review. Comput Sci Rev 2021;40:100361. [doi: 10.1016/j.cosrev.2021.100361]
2. Kruse CS, Frederick B, Jacobson T, Monticone DK. Cybersecurity in healthcare: a systematic review of modern threats and trends. Technol Health Care 2017;25(1):1-10 [FREE Full text] [doi: 10.3233/THC-161263] [Medline: 27689562]
3. 2023 healthcare cybersecurity report. Ponemon Institute. 2023. URL: https://www.proofpoint.com/uk/resources/threat-reports/ponemon-healthcare-cybersecurity-report [accessed 2025-02-25]
4. Die Lage der IT-Sicherheit in Deutschland 2023. Bundesamt für Sicherheit in der Informationstechnik (BSI). 2023 Nov 2. URL: https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Publikationen/Lageberichte/Lagebericht2023.pdf [accessed 2025-10-21]
5. Coventry L, Branley D. Cybersecurity in healthcare: a narrative review of trends, threats and ways forward. Maturitas 2018;113:48-52. [doi: 10.1016/j.maturitas.2018.04.008] [Medline: 29903648]
6. Gordon WJ, Fairhall A, Landman A. Threats to information security - public health implications. N Engl J Med 2017;377(8):707-709. [doi: 10.1056/NEJMp1707212] [Medline: 28700269]
7. Evans M, He Y, Maglaras L, Yevseyeva I, Janicke H. Evaluating information security core human error causes (IS-CHEC) technique in public sector and comparison with the private sector. Int J Med Inform 2019;127:109-119. [doi: 10.1016/j.ijmedinf.2019.04.019] [Medline: 31128822]
8. Naidoo R. A multi-level influence model of COVID-19 themed cybercrime. Eur J Inf Syst 2020;29(3):306-321. [doi: 10.1080/0960085x.2020.1771222]
9. Harasees A, Al-Ahmad B, Alsobeh A, Abuhussein A. A secure IoT framework for remote health monitoring using fog computing. In: 2024 International Conference on Intelligent Computing, Communication, Networking and Services (ICCNS). New York: IEEE; 2024:17-24.
10. Jalali MS, Kaiser JP. Cybersecurity in hospitals: a systematic, organizational perspective. J Med Internet Res 2018;20(5):e10059 [FREE Full text] [doi: 10.2196/10059] [Medline: 29807882]
11. Williams JA, Zafar H, Gupta S. Fortifying healthcare: an action research approach to developing an effective SETA program. Comput Secur 2024;138:103655. [doi: 10.1016/j.cose.2023.103655]
12. Reeves A, Delfabbro P, Calic D. Encouraging employee engagement with cybersecurity: how to tackle cyber fatigue. Sage Open 2021;11(1). [doi: 10.1177/21582440211000049]
13. Hu S, Hsu C, Zhou Z. Security education, training, and awareness programs: literature review. J Comput Inf Syst 2021;62(4):752-764. [doi: 10.1080/08874417.2021.1913671]
14. Posey C, Roberts TL, Lowry PB. The impact of organizational commitment on insiders' motivation to protect organizational information assets. J Manag Inf Syst 2016;32(4):179-214. [doi: 10.1080/07421222.2015.1138374]
15. Vinaykumar S, Zhang C, Shahriar H. Security and privacy of electronic medical records. 2019 Presented at: Proceedings of the Southern Association for Information Systems (SAIS) Conference; March 22-23, 2019; St. Simons Island, GA.
16. Rampold F, Heinsohn J, Schütz F, Klein J, Keller T, Masuch K, et al. Custom solutions for diverse needs: laying the foundation for tailored SETA programs in the healthcare domain. 2024 Presented at: Proceedings of the 57th Hawaii International Conference on System Sciences; January 3-6, 2004; Hilton Hawaiian p. 3719-3728. [doi: 10.24251/hicss.2024.449]
17. Kirlappos I, Beautement A, Sasse MA. "Comply or die" is dead: long live security-aware principal agents. In: Adams AA, Brenner M, Smith M, editors. Financial Cryptography and Data Security. FC 2013. Lecture Notes in Computer Science, vol 7862. Berlin: Springer; 2013:70-82.
18. Aslan, Aktuğ SS, Ozkan-Okay M, Yilmaz AA, Akin E. A comprehensive review of cyber security vulnerabilities, threats, attacks, and solutions. Electronics 2023;12(6):1333. [doi: 10.3390/electronics12061333]
19. Johnston, Warkentin. Fear appeals and information security behaviors: an empirical study. MIS Q 2010;34(3):549-566. [doi: 10.2307/25750691]
20. Boss S, Galletta D, Lowry P, Moody G, Polak P. What do systems users have to fear' using fear appeals to engender threatsfear that motivate protective security behaviors. MIS Q 2015;39(4):837-864. [doi: 10.25300/misq/2015/39.4.5]

21. Johnston AC, Warkentin M, Siponen M. An enhanced fear appeal rhetorical framework: leveraging threats to the human asset through sanctioning rhetoric. MIS Q 2015;39(1):113-134. [doi: 10.25300/misq/2015/39.1.06]
22. Deci EL, Ryan RM. The "what" and "why" of goal pursuits: human needs and the self-determination of behavior. Psychol Inq 2000;11(4):227-268. [doi: 10.1207/s15327965pli1104_01]
23. Kranz JJ, Haeussinger FJ. Why deterrence is not enough: the role of endogenous motivations on employees' information security behavior. 2014 Presented at: Proceedings of the International Conference on Information Systems (ICIS); December 14-17, 2014; Auckland, New Zealand.
24. Brunet J, Gunnell KE, Gaudreau P, Sabiston CM. An integrative analytical framework for understanding the effects of autonomous and controlled motivation. Pers Individ Dif 2015;84:2-15. [doi: 10.1016/j.paid.2015.02.034]
25. Carpenter C, Boster F, Andrews K. Functional attitude theory. In: The SAGE Handbook of Persuasion: Developments in Theory and Practice. 2nd ed. Thousand Oaks: Sage Publications; 2013:104-119.
26. Joyal-Desmarais K, Scharmer AK, Madzelan MK, See JV, Rothman AJ, Snyder M. Appealing to motivation to change attitudes, intentions, and behavior: a systematic review and meta-analysis of 702 experimental tests of the effects of motivational message matching on persuasion. Psychol Bull 2022;148(7-8):465-517. [doi: 10.1037/bul0000377]
27. Messineo L, Seta L, Allegra M. The relationship between empathy and altruistic motivations in nursing studies: a multi-method study. BMC Nurs 2021;20(1):124 [FREE Full text] [doi: 10.1186/s12912-021-00620-4] [Medline: 34233674]
28. Lynne GD. Toward a dual motive metaeconomic theory. J Socio-Econ 2006;35(4):634-651. [doi: 10.1016/j.socec.2005.12.019]
29. Cornelis I, Van Hiel A, De Cremer D. Volunteer work in youth organizations: predicting distinct aspects of volunteering behavior from self‐ and other‐oriented motives. J Appl Soc Psychol 2013;43(2):456-466. [doi: 10.1111/j.1559-1816.2013.01029.x]
30. Fogg B. A behavior model for persuasive design. In: Persuasive '09: Proceedings of the 4th International Conference on Persuasive Technology. New York: Association for Computing Machinery; 2009:1-7.
31. Brickner S, Fick K, Panice J, Bulthuis K, Mitchell R, Lancaster R. Professional values and nursing care quality: a descriptive study. Nurs Ethics 2024;31(5):699-713. [doi: 10.1177/09697330231200567] [Medline: 37739396]
32. Kanofsky S. Professionalism for physician assistants. Physician Assist Clin 2020;5(1):11-26. [doi: 10.1016/j.cpha.2019.08.002]
33. Moyo M, Goodyear-Smith FA, Weller J, Robb G, Shulruf B. Healthcare practitioners' personal and professional values. Adv Health Sci Educ 2016;21(2):257-286. [doi: 10.1007/s10459-015-9626-9] [Medline: 26215664]
34. White K, Peloza J. Self-benefit versus other-benefit marketing appeals: their effectiveness in generating charitable support. J Mark 2009;73(4):109-124. [doi: 10.1509/jmkg.73.4.109]
35. Rogers RW. A protection motivation theory of fear appeals and attitude change. J Psychol 1975;91(1):93-114. [doi: 10.1080/00223980.1975.9915803] [Medline: 28136248]
36. AlSobeh AMR, AlAzzam I, Shatnawi AMJ, Khasawneh I. Cybersecurity awareness factors among adolescents in Jordan: mediation effect of cyber scale and personal factors. Online J Commun Media Technol 2023;13(2):e202312. [doi: 10.30935/ojcmt/12942]
37. Keller T, Warwas J, Schütz F, Rampold F. Developing a situational judgement test for cybersecurity in healthcare: an important diagnostic prerequisite for fostering cybersecurity behavior among hospital staff. 2024 Presented at: 16th International Conference on Education and New Learning Technologies; July 1-3, 2024; Palma, Spain. [doi: 10.21125/edulearn.2024.1279]
38. Köpfer P, Warwas J, Schütz F, Rampold F, Masuch K, Trang S. A competence-based screening of instructional designs in trainings for IT-security at the workplace. 2023 Presented at: 2023 American Educational Research Association (AERA) Annual Meeting; April 13-16, 2023; Chicago. [doi: 10.3102/2019008]
39. Epstein RM, Hundert EM. Defining and assessing professional competence. JAMA 2002;287(2):226-235. [doi: 10.1001/jama.287.2.226] [Medline: 11779266]
40. Evetts J. The concept of professionalism: professional work, professional practice and learning. In: Billett S, Harteis C, Gruber H, editors. International Handbook of Research in Professional and Practice-Based Learning. Dordrecht, Netherlands: Springer; 2014:29-56.
41. Lohman MC. Cultivating problem‐solving skills through problem‐based approaches to professional development. Hum Resour Dev Q 2002;13(3):243-261. [doi: 10.1002/hrdq.1029]
42. Chacon JA, Janssen H. Teaching critical thinking and problem-solving skills to healthcare professionals. Med Sci Educ 2021;31(1):235-239 [FREE Full text] [doi: 10.1007/s40670-020-01128-3] [Medline: 34457878]
43. Abawajy J. User preference of cyber security awareness delivery methods. Behav Inf Technol 2012;33(3):237-248. [doi: 10.1080/0144929x.2012.708787]
44. Jahn D, Tress D, Attenberger C, Chmel L. Lernvideos können mehr als nur Erklären: Eine Studie zum Einsatz von narrativen Film-Ankern in einer hochschuldidaktischen Online-Weiterbildung. In: Buchner J, Freisleben-Teutscher CF, Haag J, Rauscher E, editors. Inverted Classroom – Vielfältiges Lernen. St. Pölten: Fachhochschule St. Pölten GmbH; 2018:149-164.
45. Zander S, Behrens A, Mehlhorn S. Erklärvideos als Format des E-Learnings. In: Niegemann N, Weinberger A, editors. Handbuch Bildungstechnologie: Konzeption und Einsatz digitaler Lernumgebungen. Berlin: Springer; 2020:247-258.

46. Kropp M. Studie zur digitalen Transformation: 90% der DAX Unternehmen nutzen Erklärvideos. Connectar. 2015 Feb 9. URL: https://www.connektar.de/informationen-medien/studie-zur-digitalen-transformation-90-der-dax-unternehmen-nutzen-erklaervideos-30442 [accessed 2025-10-21]
47. Wolf KD. Bildungspotenziale von Erklärvideos und Tutorials auf YouTube: Audiovisuelle Enzyklopädie, adressatengerechtes Bildungsfernsehen, Lehr-Lern-Strategie oder partizipative Peer Education? merz 2015;59(1):30-36. [doi: 10.21240/merz/2015.1.11]
48. Guo PJ, Kim J, Rubin R. How video production affects student engagement: an empirical study of MOOC videos. In: L@S '14: Proceedings of the First ACM Conference on Learning @ Scale Conference. New York: Association for Computing Machinery; 2014:41-50.
49. Findeisen S, Horn S, Seifried J. Lernen durch Videos – Empirische Befunde zur Gestaltung von Erklärvideos. Medien Pädagogik 2019;19(1):16-36. [doi: 10.21240/mpaed/00/2019.10.01.x]
50. Masemann S, Messer B. Improvisation und Storytelling in Training und Unterricht. Weinheim: Beltz Verlag; 2009.
51. Sadik A. Digital storytelling: a meaningful technology-integrated approach for engaged student learning. Educ Technol Res Dev 2008;56(4):487-506. [doi: 10.1007/s11423-008-9091-8]
52. Vester F, Denken L. Denken, Lernen, Vergessen: Was geht in unserem Kopf vor, wie lernt das Gehirn und wann läßt es uns im Stich?. München: Deutscher Taschenbuch Verlag; 1998.
53. Kerres M. Mediendidaktik: Konzeption und Entwicklung mediengestützter Lernangebote. München: Oldenbourg Wissenschaftsverlag; 2013. URL: https://www.degruyter.com/document/doi/10.1524/9783486736038/html [accessed 2025-02-25]
54. Mah D. Learning analytics and digital badges: potential impact on student retention in higher education. Tech Know Learn 2016;21(3):285-305. [doi: 10.1007/s10758-016-9286-8]
55. Herrington J, Oliver R. An instructional design framework for authentic learning environments. Educ Technol Res Dev 2000;48(3):23-48. [doi: 10.1007/BF02319856]
56. Christian M, Edwards B, Bradley J. Situational judgment tests: constructs assessed and a meta-analysis of their criterion-related validities. Pers Psychol 2010;63(1):83-117. [doi: 10.1111/j.1744-6570.2009.01163.x]
57. McDaniel MA, Morgeson FP, Finnegan EB, Campion MA, Braverman EP. Use of situational judgment tests to predict job performance: a clarification of the literature. J Appl Psychol 2001;86(4):730-740. [doi: 10.1037/0021-9010.86.4.730] [Medline: 11519656]
58. Polyhart R, MacKenzie W. Situational judgment tests: a critical review and agenda for the future. In: APA Handbook of Industrial and Organizational Psychology. Washington, DC: American Psychological Association; 2011:237-252.
59. Kepes S, Keener SK, Lievens F, McDaniel MA. An integrative, systematic review of the situational judgment test literature. J Manag 2024;51(6):2278-2319. [doi: 10.1177/01492063241288545]
60. Stemler SE, Sternberg RJ. Using situational judgment tests to measure practical intelligence. In: Situational Judgment Tests: Theory, Measurement, and Application. Mahwah, NJ: Lawrence Erlbaum; 2006:107-131.
61. Champely S. pwr: Basic functions for power analysis. R package version 1.3-0. 2020. URL: https://CRAN.R-project.org/package=pwr [accessed 2025-10-21]
62. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd edition. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
63. Schmider E, Ziegler M, Danay E, Beyer L, Bühner M. Is it really robust? Methodology 2010;6(4):147-151. [doi: 10.1027/1614-2241/a000016]
64. Hopewell S, Chan A, Collins GS, Hróbjartsson A, Moher D, Schulz KF, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. BMJ 2025;389:e081123. [doi: 10.1136/bmj-2024-081123] [Medline: 40228833]
65. Puhakainen P, Siponen M. Improving employees' compliance through information systems security training: an action research study. MIS Q 2010;34(4):757-778. [doi: 10.2307/25750704]
66. Alshaikh M, Maynard S, Ahmad A, Chang S. An exploratory study of current information security training and awareness practices in organizations. 2018 Presented at: Proceedings of the 51st Hawaii International Conference on System Sciences (HICSS); January 3-6, 2018; Hawaii p. 5085-5094. [doi: 10.24251/hicss.2018.635]
67. Ng B, Kankanhalli A, Xu Y. Studying users' computer security behavior: a health belief perspective. Decis Support Syst 2009;46(4):815-825. [doi: 10.1016/j.dss.2008.11.010]
68. Johnston AC, Warkentin M, McBride M, Carter L. Dispositional and situational factors: influences on information security policy violations. Eur J Inf Syst 2017;25(3):231-251. [doi: 10.1057/ejis.2015.15]
69. Edwards JR, Cable DM. The value of value congruence. J Appl Psychol 2009;94(3):654-677. [doi: 10.1037/a0014891] [Medline: 19450005]
70. Oyserman D. Identity‐based motivation: implications for action‐readiness, procedural‐readiness, and consumer behavior. J Consum Psychol 2009;19(3):250-260. [doi: 10.1016/j.jcps.2009.05.008]
71. Kärner T, Keller T, Schneider A, Albaner D, Schumann S. Ein Rahmenmodell zur Gestaltung technologisch unterstützter adaptiver Lehr- und Lernprozesse. Z Für Berufs Wirtschaftspädagogik 2021;117(3):351-371. [doi: 10.25162/zbw-2021-0016]
72. Deci E, Ryan R. Intrinsic Motivation and Self-Determination in Human Behavior. New York: Plenum; 1985.

73. Gagné M, Forest J, Vansteenkiste M, Crevier-Braud L, van den Broeck A, Aspeli AK, et al. The multidimensional work motivation scale: validation evidence in seven languages and nine countries. Eur J Work Organ Psychol 2014;24(2):178-196. [doi: 10.1080/1359432x.2013.877892]
74. D?Arcy J, Greene G. Security culture and the employment relationship as drivers of employees' security compliance. Inf Manag Comput Secur 2014;22(5):474-489. [doi: 10.1108/imcs-08-2013-0057]
75. Alp Christ A, Capon-Sieber V, Grob U, Praetorius AK. Learning processes and their mediating role between teaching quality and student achievement: a systematic review. Stud Educ Eval 2022;75:101209. [doi: 10.1016/j.stueduc.2022.101209]

Abbreviations

ANCOVA: analysis of covariance
CONSORT: Consolidated Standards of Reporting Trials
FBM: Fogg Behavior Model
PMT: protection motivation theory
SETA: Security Education, Training, and Awareness
SJT: situational judgment test

Edited by B Lesselroth; submitted 04.03.25; peer-reviewed by B Erci, A AlSobeh, C Ebermann, S Anselmann; comments to author 11.07.25; revised version received 08.08.25; accepted 30.09.25; published 07.11.25.

Please cite as:
Keller T, Warwas JI, Klein J, Henkenjohann R, Trenz M, Trang STN
Motivational Framing Strategies in Health Care Information Security Training: Randomized Controlled Trial
JMIR Med Educ 2025;11:e73245
URL: https://mededu.jmir.org/2025/1/e73245
doi:10.2196/73245
PMID:

Original Paper

Using AI-Based Virtual Simulated Patients for Training in Psychopathological Interviewing: Cross-Sectional Observational Study

Daniel García-Torres[1], MSc; César Fernández[1], PhD; José Joaquín Mira[1,2], PhD; Alexandra Morales[1], PhD; María Asunción Vicente[1], PhD

[1]Departamento de Psicología de la Salud, Universidad Miguel Hernández, Elche, Spain

[2]Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunitat Valenciana, Alicante, Spain

Corresponding Author:
César Fernández, PhD
Departamento de Psicología de la Salud
Universidad Miguel Hernández
Avenida de la Universidad s/n
Elche, 03202
Spain
Phone: 34 966658423
Email: c.fernandez@umh.es

Abstract

Background: Virtual simulated patients (VSPs) powered by generative artificial intelligence (GAI) offer a promising tool for training clinical interviewing skills; yet, little is known about how different system- and user-level variables shape students' perceptions of these interactions.

Objective: We aim to study psychology students' perceptions of GAI-driven VSPs and examine how demographic factors, system parameters, and interaction characteristics influence such perceptions.

Methods: We conducted a total of 1832 recorded interactions involving 156 psychology students with 13 GAI-generated VSPs configured with varying temperature settings (0.1, 0.5, 0.9). For each student, we collected age and sex; for each interview, we recorded interview length (total number of question–answer turns), number of connectivity failures, the specific VSP consulted, and the model temperature. After every interview, students provided a 1-10 global rating and open-ended comments regarding strengths and areas for improvement. At the end of the training sequence, they also reported perceived improvement in diagnostic ability. Statistical analyses assessed the influence of different variables on global ratings: demographics, interaction-level data, and GAI temperature setting. Sentiment analysis was conducted to evaluate the VSPs' clinical realism.

Results: Statistical analysis showed that female students rated the tool significantly higher (mean rating 9.25/10) than male students (mean rating 8.94/10; Kruskal-Wallis test, H=8.7; P=.003). On the other side, no significant correlation was found between global rating and age (r=0.02, 95% CI –0.03 to 0.06; P=.42), interview length (r=0.04, 95% CI –0.2 to 0.10; P=.18), or frequency of participation (Kruskal-Wallis test, H=4.62; P=.20). A moderate negative correlation emerged between connectivity failures and ratings (r=–0.26, 95% CI –0.41 to –0.10; P=.002). Temperature settings significantly influenced ratings (Kruskal-Wallis test, H=6.93; P=.03; η^2=0.02), with higher scores at temperature 0.9 compared with 0.1 (Dunn's test, P=.04). Concerning learning outcomes, self-perceived improvement in diagnostic ability was reported by 94% (94/100) of students; however, final practical examination scores (mean 6.67, SD 1.42) did not differ significantly from those of the previous cohort without VSP training (mean 6.42, SD 1.56). Sentiment analysis indicated predominantly negative sentiment in GAI responses (median negativity 0.8903, IQR 0.306-0.961), consistent with clinical realism.

Conclusions: GAI-driven VSPs were well-received by psychology students, with student gender and system-level variables (particularly temperature settings and connection stability) shaping user evaluations. Although participants perceived the training as beneficial for their diagnostic skills, objective examination performance did not significantly differ from the previous cohort. However, lack of randomization limits the generalization of the results obtained, and further experiments are required.

(JMIR Med Educ 2025;11:e78857) doi:10.2196/78857

KEYWORDS

artificial intelligence; clinical reasoning; educational technology; natural language processing; patient simulation; psychopathology; simulation training

Introduction

In health education, the development of clinical reasoning is fundamental for preparing competent professionals capable of making accurate diagnostic and therapeutic decisions. However, formal instruction in clinical reasoning remains limited within many curricula, often due to time constraints and the lack of targeted pedagogical approaches. As a result, recent graduates frequently report feeling inadequately prepared to manage the ambiguity and complexity inherent in real-world clinical practice, particularly in clinical psychology, where effective diagnostic formulation requires integrating diverse, nuanced patient information [1,2].

Clinical skill development in psychology education, particularly in subjects such as psychopathology, presents a significant challenge for university programs. Successful clinical training necessitates the integration of theoretical knowledge—such as diagnostic criteria—and practical skills, such as conducting clinical interviews. Acquiring competencies such as symptom identification, differential diagnosis, clinical reasoning, and empathic communication extends beyond theoretical understanding. These competencies are deeply intertwined with practical experience, decision-making in uncertain contexts, and sustained exposure to complex clinical situations. Unfortunately, traditional teaching methods, such as paper-based clinical cases, offer limited opportunities for students to actively and progressively develop these skills, negatively affecting their confidence and preparedness.

To address these limitations, the use of virtual patients has increasingly emerged as an effective pedagogical strategy [3,4], offering simulations of realistic clinical encounters in a risk-free environment. These simulations allow students to practice crucial skills such as history taking, hypothesis formulation, and diagnostic reasoning without risking patient safety [5,6]. Virtual patient technologies have evolved considerably—from initial static textual cases to sophisticated interactive simulations powered by generative artificial intelligence (GAI) and natural language processing (NLP) technologies [3].

The integration of GAIs based on large language models (LLMs), such as ChatGPT into virtual patient platforms represents a significant advancement in educational simulations. These models facilitate realistic, responsive interactions that closely resemble genuine clinical dialogues, thereby increasing learner engagement and immersion [7]. Recent studies, including a systematic review, have shown that GAI-powered conversational virtual patients (virtual simulated patients [VSPs]) significantly enhance clinical reasoning skills and student satisfaction, especially when the interactions are perceived as authentic and dynamic [8].

Concerning authenticity, LLMs are parameterizable in different ways to adjust their behavior. In particular, the temperature parameter controls how random or deterministic LLMs' choices are: low temperature values produce more predictable and less spontaneous answers, whereas high temperature values produce more creative and natural-sounding answers (although less consistent). This effect is discussed in detail in the report presented by Peeperkorn et al [9]. Temperature control is thus relevant in a VSP, where natural-sounding answers are preferable, but consistency is also a requirement.

Despite the promising literature on VSPs, existing research has predominantly focused on medical education (eg, Peralta Ramírez et al [10] or Borg et al [11]) and nursing education (eg, Padilha et al [12] or Hu et al [13]). There remains a gap regarding their effectiveness in psychology education, particularly in the field of psychopathology. A complete review of VSP applications in psychology can be found in Imam Hossain et al [14]. Among the few previous studies in this field, the work by Lan et al [15] proposes an alternative to objective structured clinical examinations in psychology based on VSPs, which, however, are not powered by GAI. Another study from Walkiewicz et al [16] compares actors or standardized patients with VSPs, the main conclusion being that standardized patients were more effective for interview skills and VSPs were most effective for clinical reasoning skills. Also in this case, the VSP platform used was not powered by a GAI.

This study evaluates the students' perceptions of GAI-based VSPs for practical psychopathology training in an undergraduate psychology course of a public Spanish University.

Methods

Experimental Design

This study used a cross-sectional observational design to evaluate the effectiveness of GAI-based VSPs in training psychological diagnostic skills.

Every student-VSP session followed a similar schedule: the student started with no prior knowledge about the case, except from the name and age of the patient (eg, a session may start with a heading like "Simon, a 12-year-old boy, is your new patient"). With only this limited information, the student had to start the interview with the patient and ask all questions she or he found necessary to reach a conclusion about a diagnosis for the patient. When the student had gathered all information needed, she or he ended the interview and filled out a report specifying the diagnose and, depending on the patient, answering a set of extra questions related to the case.

Apart from that, the student also rated the tool after each session and evaluated self-perceived learning improvement. All sessions ended through 2 web-based questionnaires. Both questionnaires adhere to the Checklist for Reporting Results of Internet E-Surveys (CHERRIES) guidelines [17] (Multimedia Appendix 1).

The first questionnaire (student satisfaction, completed after each interview) consisted of 3 items, distributed across a single screen (page). The second questionnaire (learning improvement,

completed only once after all practice sessions) had 14 questions distributed across 5 screens (pages), although only one of these items is included in this study. The project team was multidisciplinary: the psychologists designed both questionnaires, and the engineers designed the responsive web application following this design and assuring correct behavior on different screen sizes.

The study was conducted as a "closed survey," requiring participants to log in via the university's virtual campus with their unique student credentials. Once the questionnaire had been submitted, the students could check their answers and the conversation with the VSP, but the submit button was disabled to prevent duplicate entries. Furthermore, the application only allowed the submission of fully completed questionnaires. To remove nonmeaningful interactions from the dataset, sessions with fewer than 3 questions in the conversation between student and VSP were excluded from analysis.

In selected sessions, the GAI model's temperature parameter was fixed randomly at one of 3 different levels: 0.1, 0.5, and 0.9. This setting was unknown to the students in all cases. As outlined in the Introduction section, temperature controls the degree of randomness in the model's responses: lower values (eg, 0.1) produce more deterministic and structured replies, while higher values (eg, 0.9) allow for more varied and unpredictable outputs. The study explored whether this parameter influenced students' perceptions of the tool (tool rating), as well as the length of the interviews (number of questions asked by the student).

The platform recorded the complete interaction history, including both student inputs and GAI responses. The length of each interview was measured in terms of the number of questions asked by the student and answered by the VSP. We also explored whether this parameter influenced students' ratings of the tool.

Due to internet connectivity issues, the GAI model was occasionally unreachable, and certain student questions were not answered by the VSP. In these cases, the message received by the student was "Connection error, please repeat your question." Interview length did not account such failed interactions. We recorded separately the number of these connectivity failures in every interview to evaluate their possible influence on student ratings.

Platform Development

The starting point for platform development was 13 cases of different psychopathologies described in terms of (1) symptoms, clinical history, and familial or social context; and (2) questions to be answered by the students, including a proposal of the correct diagnosis for the patient.

The desired final result was 13 GAI-based VSPs behaving accordingly to each of the 13 cases. The VSPs did not offer any initial information about their diseases, and the students were responsible for gathering all information by interviewing them. An important requirement was to allow interaction using unlimited natural language (ie, free text instead of selection from predefined questions). After the interviews, the software had to ask the students the questions related to the case, including the proposal of a correct diagnosis. The complete interview (student questions and VSP answers) had to be registered for further analysis.

Other goals to be fulfilled by the VSP platform included:

- It should enable health care educators without programming expertise to modify and adjust the VSPs.
- The reliability of the GAI responses had to be assured, to avoid hallucinations or incorrect VSP answers to student questions.
- It should allow an easy customization of key GAI parameters—such as temperature (controlling response randomness) and top_p (influencing response diversity).
- It should facilitate user satisfaction assessment by collecting qualitative feedback and improvement suggestions.

The tools selected for VSP development were the PHP programming language and 2 different GAI models (OpenAI and Mistral AI) accessed through their public APIs.

The platform was designed by a multidisciplinary team involving software engineers, psychologists, and docents. We followed a collaborative approach similar to that presented in Fernández et al [18], under an incremental and iterative software development life cycle [19], in which, for each added functionality, we carried out successive steps of development, revision by the complete team, redesign if needed, and validation. This incremental scheme aimed at 6 different development steps:

- Step 1: Working VSP for the first clinical case: must answer all student questions correctly, according to the patient symptoms and expected behavior in terms of expressiveness and feelings.
- Step 2: Working VSP for the first clinical case with adjustable temperature and top_p parameters for answer randomness control.
- Step 3: Working VSP for the first clinical case with closed-loop supervision by a secondary GAI model and temperature or top_p automatic adjustment.
- Step 4: Working VSP for the first clinical case integrated in a teaching and evaluation environment with access control, final questionnaire for students, and practice registration in the database.
- Step 5: Docent tool for creation and edition of VSPs. This tool will further be used to create the 13 required VSPs for each of the 13 cases.
- Step 6: VSPs created for all 13 cases.

A final validation step was carried out, with exhaustive tests performed by the psychologists and docents for each of the 13 VSPs developed, prior to the start of training sessions with the students.

Recruitment of Participants and Demographic Data Registered

Participants were recruited from second-year undergraduate psychology students enrolled in the psychopathology course at Miguel Hernández University (UMH), Elche, Spain. This mandatory course, part of the second year of the psychology degree program, was delivered during the first semester (October

2024 to January 2025) of the 2024-2025 academic year and carried a workload of 7.5 credits, according to the European Credit Transfer and Accumulation System. All enrolled students were invited to participate in the study, with no exclusion criteria applied. Participation in the study required attendance at least one of the 6 training sessions scheduled, each one involving interaction with 1-3 different VSPs (globally, 13 VSPs distributed across 6 training sessions; more details can be found on the website [20]).

The only demographic data registered for participants were age and gender.

Student Satisfaction

Upon completion of each session, participants rated their experience on a 1-10 scale. Ratings of exactly 5 were excluded from the analysis, as this value appeared as the default option on the evaluation form. Because it could not be determined whether these responses were selected intentionally or by omission, their inclusion was considered potentially biased. Therefore, they were removed to preserve the validity of the statistical analysis.

Each student was also encouraged to write 2 open-ended comments: the first detailing the positive aspects found in the tool and the second providing improvement suggestions. Multimedia Appendix 2 shows the structure of the questionnaire.

Student satisfaction was analyzed for relationships with frequency of participation (number of interviews carried out by each student), age and gender of the student, length of interviews, VSP interviewed, number of connectivity failures, GAI temperature parameter, and gender pairing. Gender pairing refers to the possible influence on the tool rating of the VSP and the student having the same or different genders. In other words, the goal is to check whether male or female students rated male or female VSPs differently.

Learning Improvement

Learning improvement was measured both in terms of perceived improvement and in terms of marks obtained by the students, compared to previous years.

For perceived learning improvement, a final questionnaire was completed (optionally) by the students after all VSP sessions had ended. The only item related to learning improvement was: “Do you consider that interacting with virtual patients helped you improve your ability to identify relevant symptoms during the clinical interview?”

The final questionnaire included other items that are out of scope of this study; more details can be found in Morales et al [21]. Multimedia Appendix 3 shows the structure of the questionnaire.

For mark comparison, the marks obtained by the students in courses 2023/2024 and 2024/2025 were compared. Two items were analyzed: the marks obtained by the students in the practice sessions (reflecting how challenging the practices were) and the marks obtained by the students in the final practice examination (reflecting the competencies they acquired). The final practical examination was a paper-based examination in both courses. The training was also similar in both courses, covering the same 13 clinical cases; however, this training was paper-based in course 2023/2024 and VSP-based in course 2024/2025. For the analysis of average session grades, students with zero attendance were excluded, and the mean was calculated using only attended practices, ensuring that absences did not function as “zero” scores and skew the results.

Sentiment Analysis

A sentiment analysis was performed on both student questions and GAI-generated responses using a Python script [22] and an NLP library, *Pysentimiento* [23]. This analysis classified the emotional tone of the interactions as positive, neutral, or negative, both at the individual exchange level and for the entire conversation.

Content Analysis

Regarding open-ended comments, an automated content analysis was carried out to extract the most repeated topics from all user comments, both in the set of positive comments (ie, positive aspects found in the tool) and in the set of critical comments (ie, improvement suggestions). The analysis was automated through GAI to extract the most repeated topics and their repetition counts. Similar automations have been tested in Prescott et al [24], with results comparable to those obtained by human coders, particularly in inductive analyses like the one carried out in this study.

Statistical Details

Excel (version 16.101.3 for MacOS; Microsoft Corp) was used for data storage. Data processing and analysis were conducted using R (version 4.4.2; R Core Team).

Measures of central tendency and dispersion were calculated for quantitative variables, while frequency distributions were computed for categorical variables. Group comparisons were performed using parametric tests when the assumptions of normality were met and nonparametric alternatives when those assumptions could not be satisfied.

To examine the relationship between students’ ratings of the tool and other quantitative variables, Pearson correlation analyses were conducted.

Ethical Considerations

This study was approved by the Research Ethics Office of UMH (code DPS.CFP.250116). According to the limited personal data registered (only age and gender), the Research Ethics Office considered the study anonymous, that is, it is not possible to identify a participant from these data. Multimedia Appendix 4 shows the ethical approval record.

Results were stored in a password-protected database whose access was restricted to the researchers taking part in the project.

All students accepted an informed consent prior to every VSP session. The conversation with a VSP did not start unless the student read and accepted the terms. The text of the informed consent was made intentionally clear and concise: “The conversation held with the virtual patient, as well as the answers given in the further questionnaire, will be analyzed in aggregated terms, ensuring privacy and anonymity, as part of a research

study whose goal is to improve the use of virtual patients for psychology education. Please confirm that you accept the treatment of your conversation and answers under these conditions."

After all practice sessions ended, a final, global questionnaire was also presented to the students, who were also required to accept a similar informed consent, with the text: "The results obtained in this questionnaire will be analyzed in aggregated terms, ensuring privacy and anonymity, as part of a research study whose goal is to improve the use of virtual patients for psychology education. By sending the questionnaire you accept the treatment of your answers under these conditions."

Students received no financial compensation for their participation in the study.

Results

Platform Developed

According to the incremental and iterative software development life cycle described in the Methods section, different versions of the application were developed, tested, and validated before proceeding to the next development step. Table 1 shows the development process followed, including development and validation dates.

Table 1. Incremental development steps for the virtual simulated patient (VSP) platform.

Step	Developed	Validated
Step 1: Working VSP for first case	April 15, 2024	May 9, 2024
Step 2: VSP with temperature and top_p control	May 15, 2024	May 21, 2024
Step 3: VSP with closed loop supervision	June 19, 2024	June 25, 2024
Step 4: VSP integrated in learning environment	July 3, 2024	July 9, 2014
Step 5: Tool for creating and editing VSPs	July 12, 2024	July 24, 2024
Step 6: VSPs created for each of the 13 cases	September 12, 2024	September 25, 2024

The platform was developed as a responsive web application, optimized for seamless use across desktops, tablets, and smartphones, and programmed using PHP [25].

Figure 1 shows the flowchart of a practice session, which required initial informed consent. The main screen of the application is the dialogue or interview with the VSP, which can be as complete as the students require (in terms of number of questions asked to the VSP). The students can also check extra information during the practice session, specifically a manual with information on how to diagnose a patient. Once the students access the practice questionnaire, it is allowed to return to the interview screen (to revise the conversation), but it is not allowed to ask new questions to the VSP. After sending the questionnaire with all items fulfilled, the practice session ends.

Figure 2 provides example screenshots of a generated VSP interaction: the left-hand image shows the ongoing text-based patient dialogue (ie, interview screen), while the right-hand image presents sample assessment questions provided to the student postinteraction (ie, questionnaire screen).

Figure 1. Flowchart of a practice session with a virtual simulated patient (VSP).

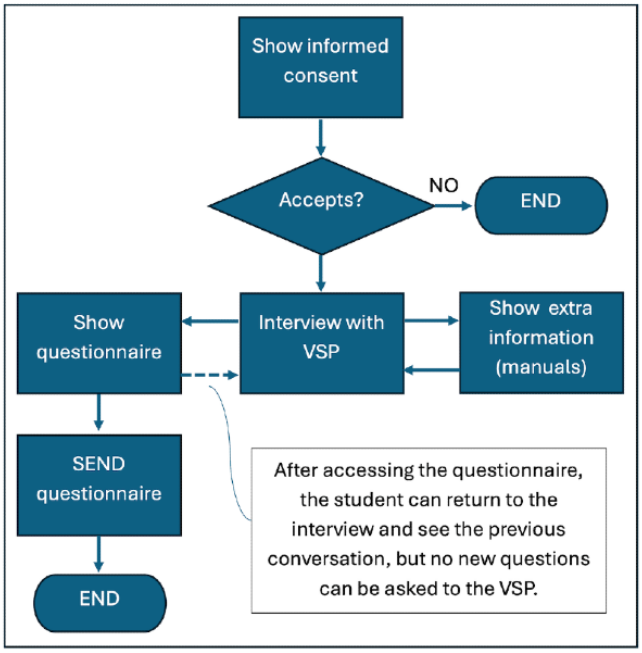
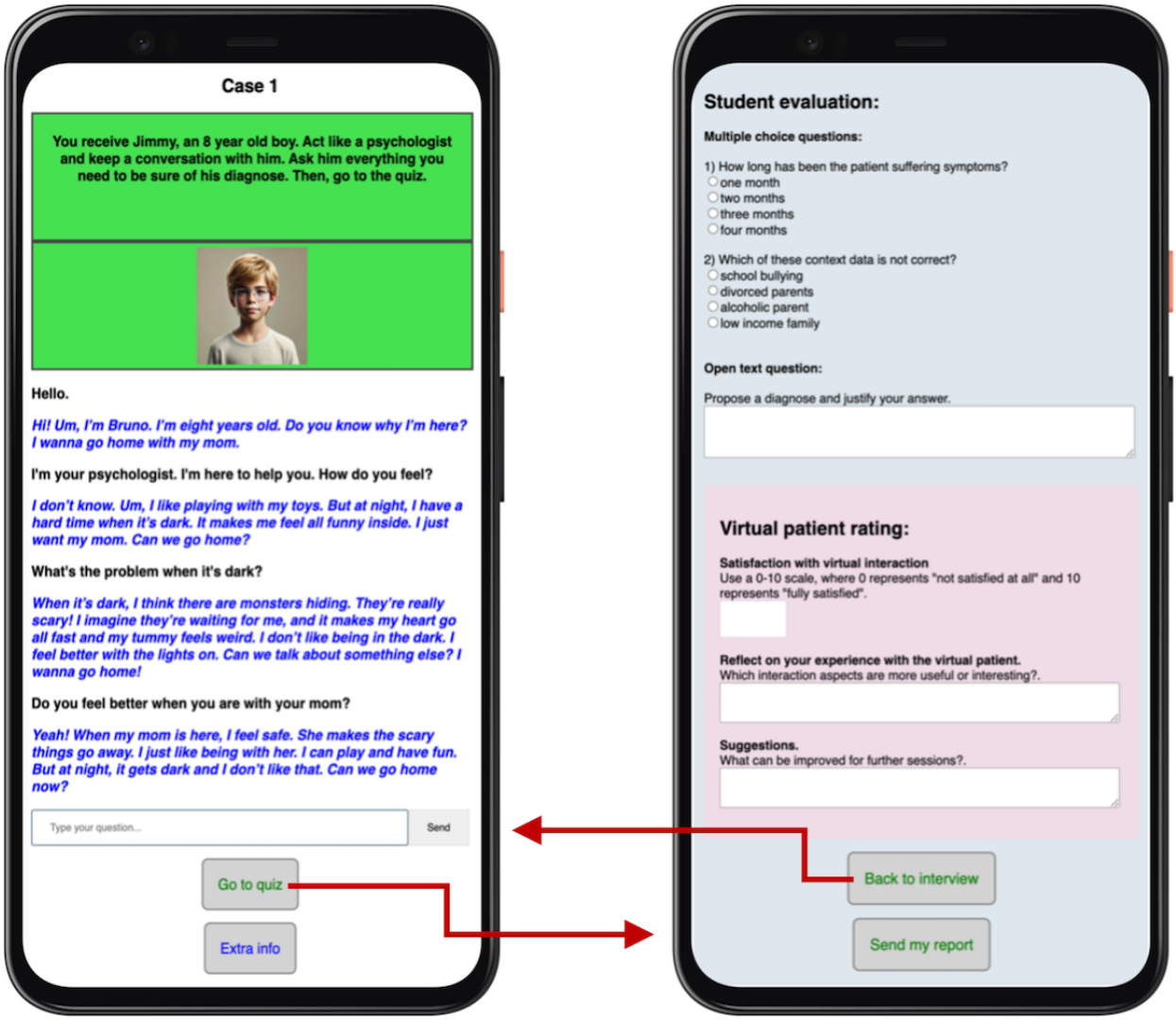


Figure 2. Example screenshots from the virtual simulated patient (VSP) application.



More details about the VSP platform developed, namely software architecture and the VSP generator for docents, are available in Multimedia Appendix 5.

Participant Demographics

A total of 156 unique participants took part in the study, carrying out 1832 interviews with VSPs (13 different VSPs). Most of the participants were aged 18-22 years, with a limited number of older participants (3 participants did not provide their ages). Table 2 shows the number of interviews carried out (frequency) per participant age range, among the 153 participants who provided age data.

The sample showed a marked gender imbalance, consisting mostly of female students (127/153, 83%), with male students representing 17% (26/153) of the total sample. Table 2 shows the complete age and gender distribution, which reflects the current trend in Spain, where the number of women enrolled in psychology degree programs significantly exceed that of men, a pattern observed in higher education statistics nationwide (77.2% of female psychology students as of the course 2022/2023 [26], and 79.9% of female psychology graduates [27], preliminary report for the course 2024/2025).

Table 2. Frequency and percentage distribution of participants by age range and gender.

Age range (years)	Men, n (%)	Women, n (%)	Total n (%)
18	0 (0)	8 (5.2)	8 (5.2)
19	14 (9.2)	86 (56.2)	100 (65.4)
20	4 (2.6)	10 (6.5)	14 (9.2)
21-25	5 (3.3)	10 (6.5)	15 (9.8)
26-30	1 (0.7)	5 (3.3)	6 (3.9)
31-35	1 (0.7)	4 (2.6)	5 (3.3)
36-40	0 (0)	1 (0.7)	1 (0.7)
41-45	0 (0)	2 (1.3)	2 (1.3)
46-50	1 (0.7)	0 (0)	1 (0.7)
51-55	0 (0)	1 (0.7)	1 (0.7)
Total	26 (17)	127 (83)	153 (100)

Student Satisfaction

Student Satisfaction Versus Demographics and Interview Length

Overall, high ratings (medians close to 10) remained consistent across different demographic groups and interaction levels.

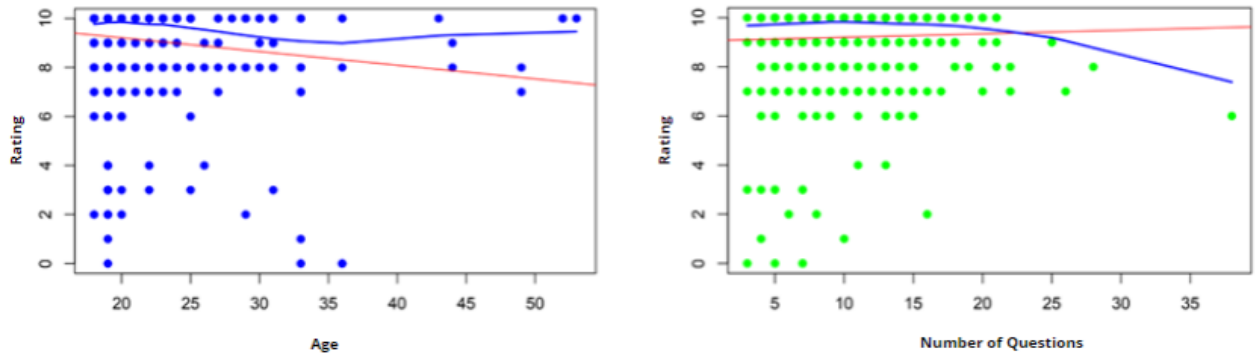
Female students rated the tool significantly higher (mean rating 9.25/10) than male students (mean rating 8.94/10; Kruskal-Wallis test, H=8.7; *P*=.003).

Concerning age, no significant correlation was found between participants' age and their overall rating of the tool (*r*=0.02, 95% CI –0.03 to 0.06; *P*=.42).

Similar results were obtained for interview length (number of questions posed by participants), with no significant correlation against the overall rating of the tool (*r*=0.04, 95% CI –0.2 to –0.10; *P*=.18). This indicates that the quantity of interaction did not notably influence students' evaluation of the platform.

Participants age and interview length are plotted against overall ratings in Figure 3. In interpreting these trends, no meaningful association emerged between participants' age and their rating of the tool: students of different ages consistently evaluated the tool positively, with only minimal variation across the age range. Likewise, although interviews involving a higher number of questions tended to show slightly lower ratings, this pattern was weak and did not indicate a substantial change in students' perceptions of the tool.

Figure 3. Relationship between participants' age and their rating of the tool (left panel) and between the number of questions posed and the rating provided (right panel).



Student Satisfaction Versus Frequency of Participation

On average, students rated the tool highly, with minor variations related to their frequency of participation. However, a modest positive trend in average ratings was observed, suggesting that increased exposure might slightly enhance perceptions of the platform's effectiveness (Table 3). To analyze this relationship between students' frequency of participation and their average ratings, Shapiro-Wilk tests indicated that ratings did not follow a normal distribution in any of the participation groups (*P*<.001 in all cases). To evaluate whether parametric methods could nevertheless be applied, several common transformations were tested (logarithmic, square root, Box-Cox, and Yeo-Johnson). Although the Yeo-Johnson transformation provided some improvement (eg, W=0.775; *P*<.001 for "Participated once"; W=0.911; *P*=.005 for "6-10 times"), none of the groups achieved normality. Consequently, a nonparametric Kruskal-Wallis test was used as the most appropriate analytic strategy. The results of this test showed that the effect of frequency of participation was not statistically significant (H=4.62, *P*=.20; Table 3).

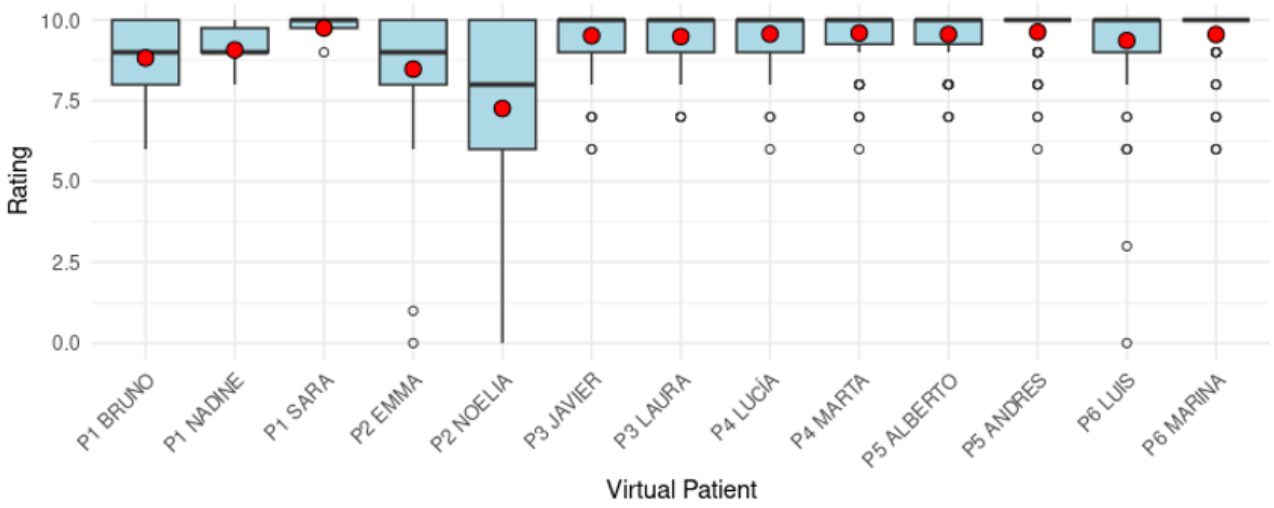
Table 3. Mean ratings of the platform based on student participation frequency (Kruskal-Wallis test, *P*=.20).

Participation frequency	Participants, n	Mean rating (95% CI)
Participated once	16	8.8 (7.3-9.6)
Participated 2-5 times	21	8.9 (8.3-9.4)
Participated 6-10 times	48	9.0 (8.6-9.4)
Participated >10 times	75	9.3 (9.1-9.6)

Student Satisfaction Versus VSP Interviewed and Connectivity Failures

When analyzing ratings by VSP, overall scores remained high, with most VSPs receiving median values near 10. However, some variation was observed, with median ratings ranging from approximately 8 to 10 across the 13 VSPs (Figure 4). Notably, Emma and Noelia received comparatively lower ratings. These 2 VSPs were involved in a session affected by a higher incidence of internet connectivity issues, which likely contributed to the reduced participant evaluations.

Figure 4. Distribution of participant ratings for each virtual simulated patient (VSP). The red dots represent the mean rating for each VSP. The label "P" indicates the practice session in which each VSP was used (eg, P1=Practice 1).



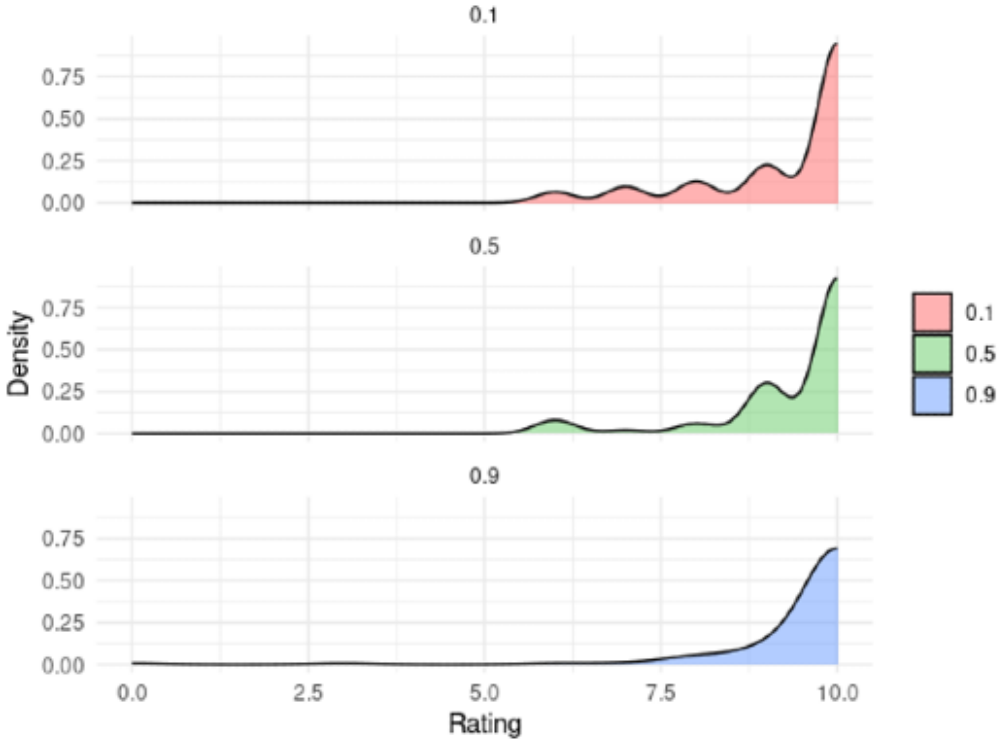
This finding aligns with a moderate negative correlation between the number of internet connectivity issues and participant ratings (r=–0.26, 95% CI –0.41 to –0.10; P=.002). This suggests that a higher number of connectivity failures was associated with lower ratings from students.

Student Satisfaction Versus GAI Temperature Parameter

Shapiro-Wilk tests conducted for each temperature level (0.1, 0.5, and 0.9) indicated strong departures from normality (P<.001 in all cases). Attempts to normalize the data through logarithmic and square root transformations were unsuccessful. The Box-Cox procedure suggested a transformation parameter far from 1, while the Yeo-Johnson approach estimated an extreme λ value ($\lambda \approx 11.2$), confirming severe nonnormality. Given these results, nonparametric Kruskal-Wallis tests were again retained as the most suitable analytic approach, revealing a statistically significant difference between them (H=6.93; P=.03). The effect size was small (η^2=0.02, 95% CI –0.00 to 0.07), suggesting that temperature explained only about 2% of the variance in ratings.

Post hoc comparisons using Dunn's test with Holm correction showed no significant difference between temperature levels 0.1 and 0.5 (P=0.62) nor between 0.5 and 0.9 (P=.14). However, a significant difference was found between 0.1 and 0.9 (P=.04), suggesting that higher ratings were associated with the 0.9 temperature condition (Figure 5).

Figure 5. Density plot showing the distribution of tool ratings across different temperature settings (0.1, 0.5, and 0.9).

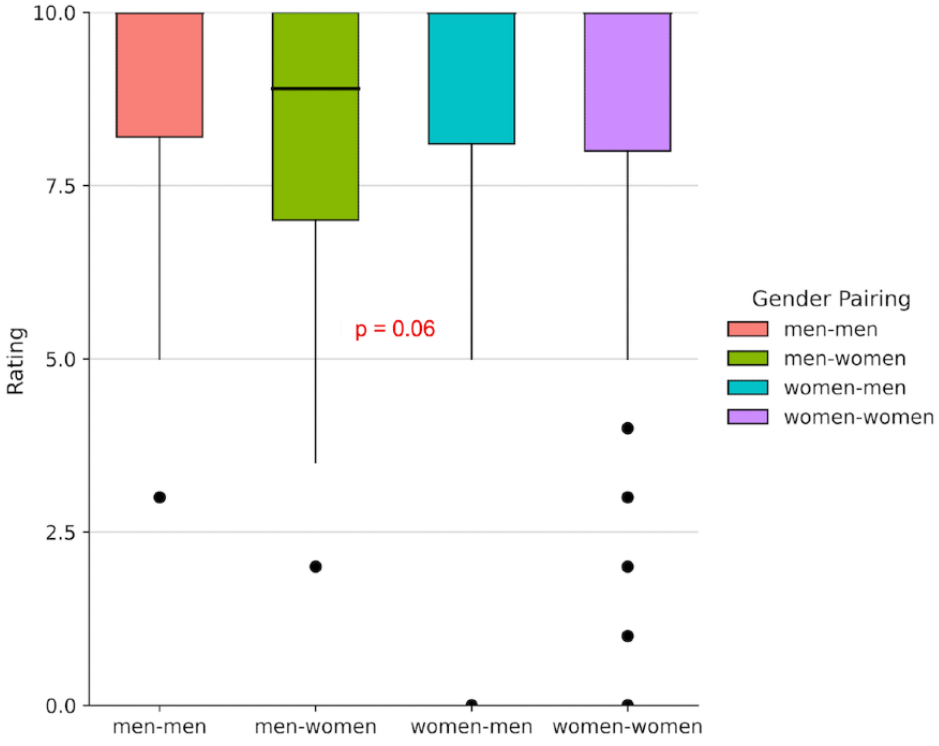


Student Satisfaction Versus Gender Pairing

Regarding the gender pairing between students and VSPs (Figure 6), the Kruskal-Wallis test revealed no statistically significant differences between groups (H=7.41, *P*=.06). Post hoc comparisons using Dunn's test with Bonferroni correction also showed no significant differences across any of the gender combinations evaluated. These results suggest that neither the participant's gender nor that of the VSP had a meaningful impact on how the tool was rated.

However, given that the *P* value was close to the conventional threshold for significance, it would be advisable to include a larger sample in future studies to more accurately assess whether gender pairing influences students' evaluations of the tool.

Figure 6. Boxplot showing the distribution of tool ratings by gender pairing between the participant and the virtual simulated patient (VSP).



Learning Improvement

A total of 100 students completed the optional final questionnaire. Table 4 shows the results obtained the question related to learning improvement.

According to the results obtained, the ability to identify relevant symptoms was mostly agreed (94/100, 94% of students found their ability had increased “a great deal” or “quite a lot”).

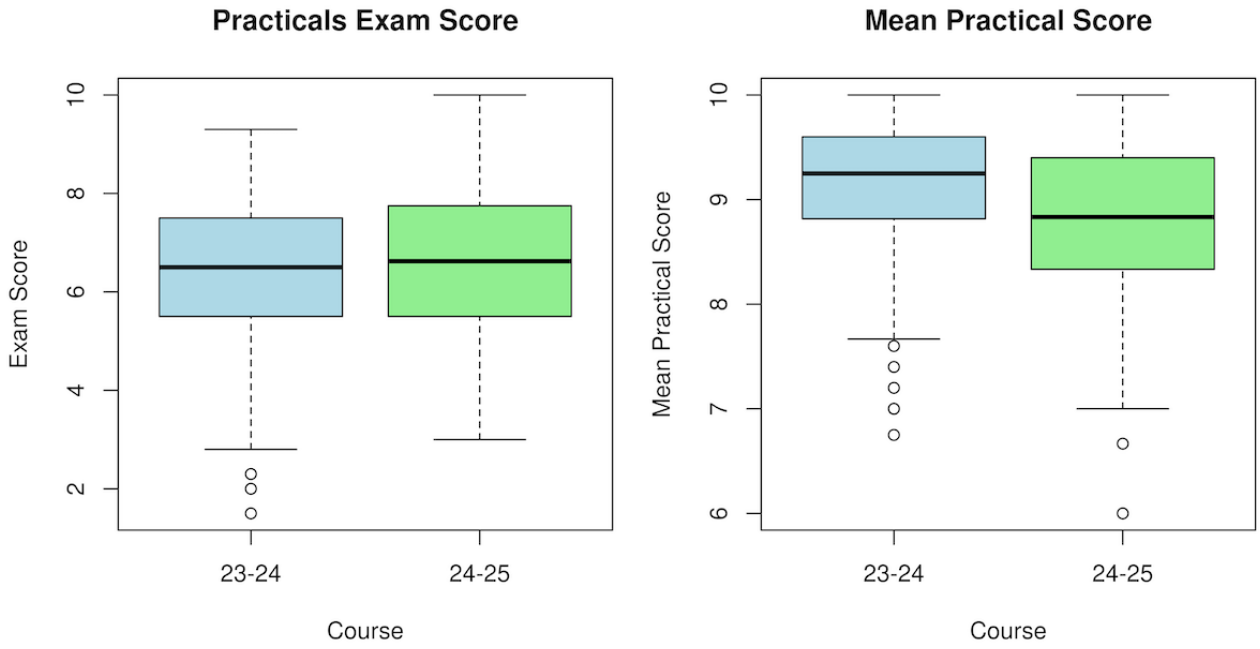
Table 4. Final questionnaire, item related to learning improvement.

Question: Do you consider that interacting with virtual patients helped you improve your ability to identify relevant symptoms during the clinical interview?	Answers, n (%)
A great deal	43 (43)
Quite a lot	51 (51)
Somewhat	5 (5)
A little	1 (1)
Not at all	0 (0)

The analysis of the final practical examination (Figure 7) showed that the mean score obtained in course 2024/2025 (mean 6.67, SD 1.42) was slightly higher than that of course 2023/2024 (mean 6.42, SD 1.56). However, this difference was not statistically significant (W=9297; *P*=.46).

Conversely, the analysis of the average practical session grades revealed that the scores from the 2024/2025 course (VSP-based; mean 8.8, SD 0.77) were significantly lower than those from the 2023/2024 course (paper-based; mean 9.14, SD 0.74; W=12,428; *P*<.001).

Figure 7. Mark comparison against previous course. Exam: examination.



Sentiment Analysis

All interactions with the platform (either student questions or GAI answers) were recorded and further processed using NLP, with the help of the *pysentimiento* library[23]. The output of the library rates the positive, neutral, and negative sentiments of each sentence, normalized so that positiveness + negativeness + neutralness = 1.

The first analysis carried out tried to explore whether the emotional tone of the GAI responses was influenced by the temperature parameter of the GAI model. We only show positiveness and negativeness results, since neutralness can be obtained from them. Figure 8 displays the total positive sentiment in responses (median 0.008, IQR 0.003-0.079). The results show a striking concentration of low positive sentiment across all temperature levels, especially at 0.1 and 0.5. Interestingly, temperature 0.9 shows slightly more dispersion, possibly due to more expressive or varied GAI outputs under higher randomness. Despite this, positivity in responses remains generally low, consistent with the structured, clinical nature of the interactions.

Figure 9 presents the total negative sentiment in responses, where a clear concentration of high negativity scores was observed across all temperature levels (median 0.890, IQR 0.306-0.961). This was particularly noticeable at temperatures 0.1 and 0.5. These findings may reflect the emotional content

inherent in the psychological case scenarios, in which patients often express distressing or symptomatic narratives.

Figure 8. Density plot of total positive sentiment in generative artificial intelligence (GAI) responses, grouped by model temperature (0.1, 0.5, and 0.9).

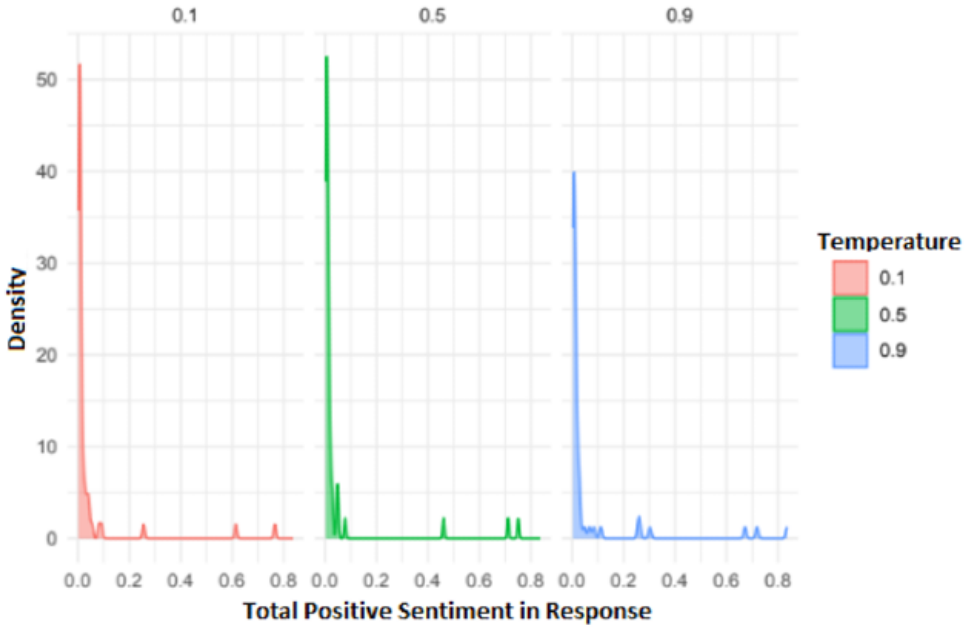
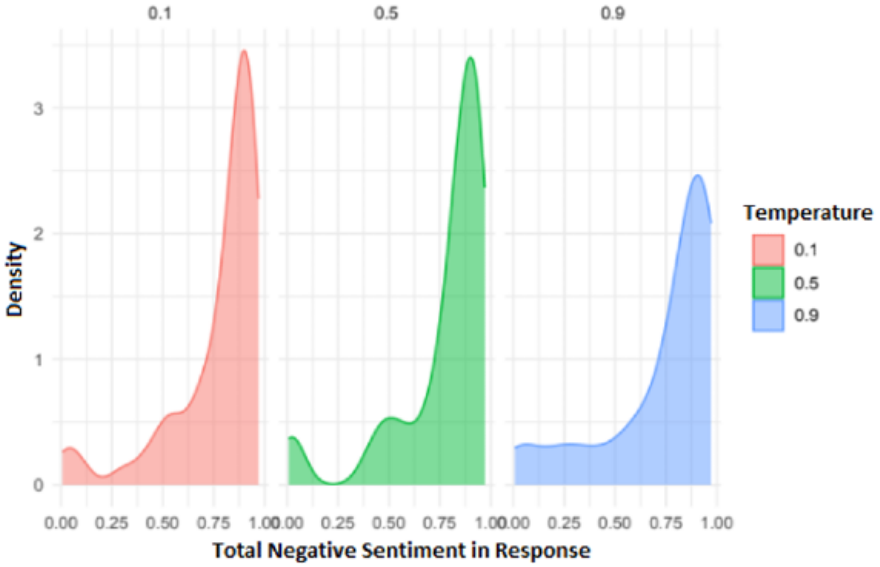


Figure 9. Density plot of total negative sentiment in generative artificial intelligence (GAI) responses, grouped by model temperature (0.1, 0.5, and 0.9).



This sentiment configuration shown in Figures 8 and 9 could be partially attributed to the design of the VSPs themselves, as they were intentionally modeled to represent clinical profiles commonly seen in mental health settings. These profiles often contain emotionally charged content, which likely contributes to the predominance of negative sentiment over positive sentiment in the GAI responses. Consequently, higher temperature values may lead the model to deviate from the expected clinical behavior, producing more creative and expressive responses that go beyond the original configuration of the VSPs [28]. This creative drift may result in a more positive tone in the interaction, as the model becomes less constrained by the simulated symptoms or emotional distress typically expected from a psychological patient.

Additional results concerning sentiment analysis are available in Multimedia Appendix 6, together with other statistical results not included in the main document.

Content Analysis

A total of 1708 valid answers (excluding empty answers, nonalphabetic answers, or answers without meaning) were registered as positive comments (positive aspects found in the tool). The automated content analysis of those comments is detailed in Table 5.

Concerning negative comments (or improvement suggestions), a total of 1604 valid answers were registered (using the same exclusion criteria as for positive comments). Automated content analysis of negative comments is summarized in Table 6.

Table 5. Content analysis of positive comments.

Positive aspect	Repetitions, n	Details
Educational usefulness and practical application	~570	Most valued aspect. Users report that the tool helps apply clinical knowledge, practice interviews, and develop professional skills in a safe environment. Universally described as useful, effective, and enriching.
Quality and clarity of the VSP's[a] responses	~470	Responses are accurate, clear, and coherent. Relevant for diagnosis, allowing interview progress. Includes completeness, correctness, and clinical utility.
Clarity and fluency in the interaction	~320	Emphasis on conversational naturalness, ease of use, and absence of glitches. Enhances the interview experience and realism.
Engagement, motivation, and dynamic experience	~240	Tool is engaging, maintains interest, and motivates learners. Nonmonotonous interaction supports student engagement and active learning.
Accurate symptom description and diagnostic support	~250	VSP provides rich and detailed symptom descriptions. Aids clinical reasoning and realistic hypothesis formulation.
Perceived improvement and positive comparison	~140	Perceived positive evolution in tool functionality and response quality. Increases satisfaction and perceived quality.
Perceived realism and immersiveness	~120	Interaction closely resembles real interviews. Realism improves pedagogical value and clinical preparation.

[a]VSP: virtual simulated patient.

Table 6. Content analysis of negative comments

Suggestion	Repetitions, n	Details
Realism and content of the VSP's[a] responses	~110	Suggestions focus on enhancing coherence, depth, and appropriateness of the VSP's clinical language. Proposals include: avoiding repetition, tailoring responses to age (eg, young children), adding relevant details, and ensuring internal consistency.
Diagnostic clarity and symptom presentation	~82	Many comments highlight difficulties in interpreting symptoms due to the similarity between disorders. Some users report that the patient directly reveals the diagnosis, undermining the clinical exercise. There is a request for more subtle clinical clues and better-differentiated scenarios.
Technical functionality and system errors	~74	Recurrent technical issues are reported: connection failures, GAI[b] model not being available, automatic deletion of student answers, and the need to reload the activity. In some cases, users are forced to repeat the task.
User interface and navigation	~49	Recommendations include improving navigation, enhancing the visibility of return buttons, enabling users to go back without losing information, and simplifying transitions between patients or tasks.
Linguistic clarity and textual formulation	~37	There is a call to improve the phrasing of both questions and responses. Suggestions include using clearer, more precise language appropriate to students' comprehension level.

[a]VSP: virtual simulated patient.

[b]GAI: generative artificial intelligence.

Discussion

Principal Findings

Concerning temperature influence on results, although not all observed effects reached statistical significance, clear trends emerged, particularly when comparing the lowest temperature level tested (0.1) with the highest one (0.9). The results in terms of user satisfaction were significantly higher for the 0.9 setting. This suggests that the temperature parameter may play a meaningful role in shaping students' perceptions of the interaction.

Contrary to expectations, no significant relationship was found between the number of questions asked during the simulation or the participants' age and the rating they assigned. However, as one might anticipate, a negative correlation was observed between the number of connectivity failures and the students' evaluation of the experience.

This suggests that students' perception of usefulness or satisfaction may not depend on the quantity of interaction, but rather on qualitative aspects, such as the fluidity of the dialogue or the perceived realism of the conversation.

A notable finding of this study is the apparent paradox in academic performance: while GAI-powered VSP implementation (course 2024/2025) led to significantly lower average grades in practical sessions compared to the traditional paper-based method (course 2023/2024), grades obtained in the final practical examination were slightly higher, although not statistically significant. Far from suggesting lower efficacy, we interpret this as evidence that the VSP simulations provide a more demanding and clinically realistic learning challenge. Traditional static paper-based cases reward methodical information retrieval [8], whereas the dynamic VSP tool required students to actively engage in real-time clinical interviewing and hypothesis formulation [6], better mirroring real-world clinical ambiguity [4]. Further randomized experiments are required to draw more reliable conclusions.

Comparison to Prior Work

Our findings on the influence of the temperature parameter are consistent with those found in previous literature. For instance, the experiments carried out by Davis et al [29] in different clinical research scenarios emphasize the compromise between creativity and consistency of the GAI answers and suggest specific temperature levels depending on the task. Other recent studies warn about the impact of inconsistencies and errors in ChatGPT's responses on user satisfaction when higher temperature settings are used [30], but in our case, the highest temperature tested (0.9) offered the best results in terms of user satisfaction.

The general evaluation of the VSP platform was highly positive, indicating strong acceptance of this type of simulation in clinical training contexts. In general, this result aligns with previous studies that have highlighted the potential of VSPs to create immersive learning environments that foster the development of clinical reasoning from the early stages of professional training [8].

Compatible with our results, the work presented by Peralta et al [10], based on an experiment with 32 medicine students, found highly valued student perceptions for both realism and consistency of the VSP responses. In particular, the students answered "agree" or "strongly agree" in 91% of the cases for the question "the scenario was realistic and similar to an authentic clinical situation," and in 94% of the cases for the question "the virtual patient responded appropriately to my actions and questions."

Focusing on specific aspects, the previous work on VSPs presented by Kamath et al [31] (pharmacology students, n=19) showed strongly positive user satisfaction for most aspects, particularly for "authenticity of patient encounter and consultation" (92.11% of positive responses), but low values for "learning effect of consultation" (47.37% of positive responses). In comparison, our experiments with psychology students agree on high user satisfaction for authenticity (Table 5, row 3: "conversational naturalness," "realism") and also offer strongly positive values for learning improvement, with 94% of students answering "a great deal" or "quite a lot" to the question "Do you consider that interacting with virtual patients helped you improve your ability to identify relevant symptoms during the clinical interview?" (Table 4). The difference in this particular result may be related with the specificities of pharmacology and psychology studies.

Another previous study, with medicine students (n=9) is presented by Cross et al [32]. Contrarily to our results, their students found verisimilitude issues and lack of empathy in the VSPs' answers. Such result may be related to the use of standard values for the temperature parameter (since the experiments were carried out directly from the web interface of ChatGPT) or a too strict definition of the clinical cases.

Strengths and Limitations

According to the results shown in Table 6, students find the tool helpful, relevant, and motivating. In addition, they particularly valued the realism of the interactions. The most common suggestions, as shown in Table 5, refer to improvements in the clinical language used by the VSPs, increasing the difficulty of the cases, avoiding connection failures, and improving the user interface.

The findings of this study provide preliminary evidence for the feasibility of using LLMs such as GPT-4o to simulate virtual patients in educational settings. The tool was rated positively by most participants, suggesting it can serve as an effective strategy for training fundamental clinical skills—such as conducting psychological interviews or gathering relevant case information—in a safe and controlled environment [6,31].

Moreover, the ability to adjust the model's temperature setting allows educators to tailor the GAI's behavior to specific learning objectives, making it possible to design adaptable training experiences that align with the learner's level of competence and the complexity of the scenario.

Concerning content analysis results, one of the most repeated positive comments was "responses are accurate, clear, and coherent. Relevant for diagnosis, allowing interview progress. Includes completeness, correctness, and clinical utility" (Table 5). On the other hand, the most repeated improvement suggestion was focused on "enhancing coherence, depth, and appropriateness of the virtual patient's clinical language" (Table 6). Surprisingly, the coherence of the VSPs' responses was considered both as a strength of the platform and as a topic requiring improvement. That suggests that, according to the students, coherence is a key point in a VSP.

This study has several limitations.

First, this was a cross-sectional, observational study, which limits the ability to draw causal conclusions from the findings. In addition, a potential source of bias was identified in the rating scale: the value "5" appeared as the default option in the evaluation form, making it unclear whether selections of this score were made intentionally or by oversight.

Second, another limitation involves the uneven usage of different VSP profiles and GAI models, which may restrict the generalizability of the results. Future research would benefit from a more balanced distribution of exposure to each virtual character and system configuration.

Third, the study's design lacked randomization. The comparison of academic performance was quasi-experimental, contrasting

the 2024/2025 cohort (which used the VSP tool) against the previous 2023/2024 cohort (which used paper-based cases) rather than using a randomized controlled trial. This nonrandomized approach means we cannot definitively attribute observed differences, or the lack thereof, in academic performance solely to the VSP intervention, as other unmeasured confounding variables between the two academic years may have influenced the findings.

Fourth, sentiment analysis was only focused on 2 topics: first, checking the predominant sentiment in VSP responses (which should be negative to reflect the clinical case situations), and second, determining whether sentiment in student questions influenced sentiment on VSP answers or vice versa (details of results are available in Multimedia Appendix 6). However, deeper analysis is needed to measure how closely the VSP reflects the correct sentiment for each case, following, for example, the guidelines that can be extracted from the study of Cero et al [33].

Finally, special attention should be given to the gender imbalance in the sample, which was composed predominantly of female students. Although no significant differences were found between male and female participants across the main variables, this disparity raises questions about potential gender-related biases in perception or interaction with the system. Future studies should aim to recruit more gender-balanced samples to assess these effects more thoroughly.

Future Directions

One promising line of inquiry is the integration of multimodal features into virtual patient simulations, including speech recognition, nonverbal communication (ie, gesture recognition), or even animated avatars, to increase realism and bring the experience closer to real clinical encounters. These enhancements would allow researchers and educators to assess not only the verbal content of the interaction but also paraverbal and behavioral cues, which are crucial in clinical practice. Nevertheless, in our experience, the VSPs have mostly been used in classroom settings during in-person practical sessions, where keyboard interaction remains the most reliable and least susceptible to disruption from peer interactions.

Another important direction involves carrying out randomized experiments for direct comparisons between GAI-based training and traditional educational methods, such as working with standardized patients or in-person role-play sessions. This would provide clearer insights into the relative effectiveness of each approach in developing specific clinical competencies, as well as students' perceived realism, usefulness, and transferability to real-world contexts.

Other future studies may explore the implementation of automated feedback systems or peer-based assessments using the transcripts generated during the interactions. These additions could further enhance the educational potential of GAI-powered simulations in hybrid or fully virtual learning environments.

Finally, this study has shown that the VSP generation tool we have developed offers enough flexibility to be adapted across various specialties within psychology, as well as in medicine and nursing. Currently, the tool is also being used in nursing and pediatrics, and we have received requests to implement it in other fields. Given this positive reception, our future goal is to create a complete hospital metaverse—a shared virtual environment that enables practical training across multiple specialties.

Acknowledgments

This study was supported by several research projects that contributed both to the development of the virtual patient platform and to the analyses presented in this study. We gratefully acknowledge the following funded initiatives: "Investigación sobre sesgos de género en la generación de imágenes por Inteligencia Artificial" (Reference: 19-3-ID24), funded by the Instituto de las Mujeres under the 2024 Call for Feminist, Gender, and Women's Studies Research (14/05/2024); "Pacientes virtuales basados en Inteligencia Artificial y con características específicas de género para la formación en medicina y enfermería sobre entornos gamificados" (Reference: PID2024-157772OB-I00), funded within the Spanish National R+D+i Program in the area of Information and Communication Technologies; "Nuevas metodologías basadas en Inteligencia Artificial para la educación en medicina y enfermería: pacientes virtuales en entornos gamificados con especificidades de género" (Expediente: CIAICO/2024/283), funded under the 2025 Call in the area of Information and Communication Technologies; and the internal research grant from Miguel Hernández University of Elche (UMH), awarded through the Resolución Rectoral 00686/2025 as part of the 2025 UMH Research Grants Program for Science, Technology, and Innovation Dissemination Projects (Code: 04-541-4-2025-0033-N). The authors extend their sincere gratitude to all the students who participated in the psychopathology course activities.

The authors declare the use of generative artificial intelligence (AI) in the research and writing process. According to the Generative AI in Data and Text (GAIDeT) taxonomy (2025), the following tasks were delegated to generative AI tools under full human supervision: code generation, code optimization, data analysis, summarizing text, translation, and reformatting. The generative AI tool used for writing and editing was ChatGPT-4o. Responsibility for the final manuscript lies entirely with the authors. Generative AI tools are not listed as authors and do not bear responsibility for the final outcomes. Additionally, OpenAI GPT-4o models and Mistral large-latest models were used in this study for the creation and operation of conversational virtual patients. AI tools were not used for drafting the article's scientific interpretations.

This declaration was submitted by DG-T, CF, JJM, AM, and MAV.

Data Availability

The datasets generated or analyzed during this study are available in the Open Science Framework (OSF) [34]. Documents in this repository are password protected; please contact the corresponding author for more information on how to access the data.

Authors' Contributions

CF and MAV conceived the study design, supervised the development of the virtual patient platform, and coordinated data collection. DGT, CF, and JJM contributed to data processing, statistical analysis, and methodological review. AM and MAV supported the implementation of training sessions, student coordination, and qualitative data extraction. AM and JJM also contributed to designing the behavioral profiles and conversational characteristics of the virtual patients used in the study. All authors reviewed and approved the final manuscript and contributed to the interpretation of results.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CHERRIES checklist.
[XLSX File (Microsoft Excel File), 16 KB - mededu_v11i1e78857_app1.xlsx]

Multimedia Appendix 2
First questionnaire.
[DOCX File , 15 KB - mededu_v11i1e78857_app2.docx]

Multimedia Appendix 3
Second questionnaire.
[DOCX File , 21 KB - mededu_v11i1e78857_app3.docx]

Multimedia Appendix 4
Ethics review board approval.
[PDF File (Adobe PDF File), 292 KB - mededu_v11i1e78857_app4.pdf]

Multimedia Appendix 5
Additional details about the VSP platform developed.
[DOCX File , 78 KB - mededu_v11i1e78857_app5.docx]

Multimedia Appendix 6
Additional statistical results.
[DOCX File , 729 KB - mededu_v11i1e78857_app6.docx]

References

1. Epstein RM, Hundert EM. Defining and assessing professional competence. JAMA 2002;287(2):226-235. [doi: 10.1001/jama.287.2.226] [Medline: 11779266]
2. Monrouxe LV, Grundy L, Mann M, John Z, Panagoulas E, Bullock A, et al. How prepared are UK medical graduates for practice? A rapid review of the literature 2009-2014. BMJ Open 2017;7(1):e013656 [FREE Full text] [doi: 10.1136/bmjopen-2016-013656] [Medline: 28087554]
3. Kononowicz AA, Woodham LA, Edelbring S, Stathakarou N, Davies D, Saxena N, et al. Virtual patient simulations in health professions education: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019;21(7):e14676 [FREE Full text] [doi: 10.2196/14676] [Medline: 31267981]
4. Plackett R, Kassianos AP, Mylan S, Kambouri M, Raine R, Sheringham J. The effectiveness of using virtual patient educational tools to improve medical students' clinical reasoning skills: a systematic review. BMC Med Educ 2022;22(1):365 [FREE Full text] [doi: 10.1186/s12909-022-03410-x] [Medline: 35550085]
5. Cook DA, Triola MM. Virtual patients: a critical literature review and proposed next steps. Med Educ 2009;43(4):303-311. [doi: 10.1111/j.1365-2923.2008.03286.x] [Medline: 19335571]
6. Isaza-Restrepo A, Gómez MT, Cifuentes G, Argüello A. The virtual patient as a learning tool: a mixed quantitative qualitative study. BMC Med Educ 2018;18(1):297 [FREE Full text] [doi: 10.1186/s12909-018-1395-8] [Medline: 30522478]

7. Dolianiti F, Tsoupouroglou I, Antoniou P, Konstantinidis S, Anastasiades S, Bamidis P. Chatbots in healthcare curricula: the case of a conversational virtual patient. In: Frasson C, Bamidis P, Vlamos P, editors. Brain Function Assessment in Learning. BFAL 2020. Lecture Notes in Computer Science. Cham: Springer; 2020.
8. García-Torres D, Vicente Ripoll MA, Fernández Peris C, Mira Solves JJ. Enhancing clinical reasoning with virtual patients: a hybrid systematic review combining human reviewers and ChatGPT. Healthcare (Basel) 2024;12(22):2241 [FREE Full text] [doi: 10.3390/healthcare12222241] [Medline: 39595439]
9. Peeperkorn M, Kouwenhoven T, Brown D, Jordanous A. Is temperature the creativity parameter of large language models? ArXiv. Preprint posted online on May 1, 2024 2024. [doi: 10.48550/arXiv.2405.00492]
10. Peralta Ramirez AA, Trujillo López S, Navarro Armendariz GA, De la Torre Othón SA, Sierra Cervantes MR, Medina Aguirre JA. Clinical simulation with ChatGpt: A revolution in medical education? J CME 2025;14(1):2525615 [FREE Full text] [doi: 10.1080/28338073.2025.2525615] [Medline: 40589612]
11. Borg A, Georg C, Jobs B, Huss V, Waldenlind K, Ruiz M, et al. Virtual patient simulations using social robotics combined with large language models for clinical reasoning training in medical education: mixed methods study. J Med Internet Res 2025;27:e63312 [FREE Full text] [doi: 10.2196/63312] [Medline: 40053778]
12. Padilha JM, Costa P, Sousa P, Ferreira A. The integration of virtual patients into nursing education. Simulation & Gaming 2024;56(2):178-191. [doi: 10.1177/10468781241300237]
13. Hu Y, Xiong Q, Yi L, Yoon I. Nurse town: An LLM-powered simulation game for nursing education. 2025 Presented at: IEEE Conference on Artificial Intelligence (CAI); May 5-7, 2025; Santa Clara p. 215-222. [doi: 10.1109/cai64502.2025.00041]
14. Imam Hossain S, Kelson J, Morrison B. The use of virtual patient simulations in psychology: a scoping review. AJET 2024;40(6):76-91. [doi: 10.14742/ajet.9559]
15. Lan Y, Chen W, Wang Y, Chang Y. Development and preliminary testing of a virtual reality measurement for assessing intake assessment skills. Int J Psychol 2023;58(3):237-246. [doi: 10.1002/ijop.12898] [Medline: 36720650]
16. Walkiewicz M, Zalewski B, Guziak M. Affect and cognitive closure in students-a step to personalised education of clinical assessment in psychology with the use of simulated and virtual patients. Healthcare (Basel) 2022;10(6):1076. [doi: 10.3390/healthcare10061076] [Medline: 35742127]
17. Eysenbach G. Improving the quality of Web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
18. Fernández C, Vicente MA, Guilabert M, Carrillo I, Mira JJ. Developing a mobile health app for chronic illness management: insights from focus groups. Digit Health 2023;9:20552076231210662 [FREE Full text] [doi: 10.1177/20552076231210662] [Medline: 37928329]
19. Alshamrani A, Bahattab A. A comparison between three SDLC models waterfall model, spiral model, and Incremental/Iterative model. International Journal of Computer Science Issues (IJCSI) 2015;12(1):106 [FREE Full text]
20. PsicoSimGPT virtual patient profiles. Miguel Hernández University. 2024. URL: https://lcsi.umh.es/psicosimgpt/ [accessed 2025-06-10]
21. Morales A, Hervás D, Fernández C, Fernández-Martínez I, Gonzálvez MT, Vicente MA. Pacientes virtuales con inteligencia artificial en psicopatología: una propuesta innovadora para la formación clínica universitaria. In: Satorre R, editor. Metodologías activas y tecnologías emergentes aplicadas a la docencia universitaria. Barcelona, Spain: Ediciones Octaedro; 2025.
22. 2025 python software foundation. Python software. URL: https://www.python.org/ [accessed 2025-06-10]
23. Pysentimiento: a python toolkit for sentiment analysis and social NLP tasks. GitHub, Inc. 2025. URL: https://github.com/pysentimiento/pysentimiento [accessed 2025-06-10]
24. Prescott MR, Yeager S, Ham L, Rivera Saldana CD, Serrano V, Narez J, et al. Comparing the efficacy and efficiency of human and generative AI: qualitative thematic analyses. JMIR AI 2024;3:e54482 [FREE Full text] [doi: 10.2196/54482] [Medline: 39094113]
25. Bakken SS, Suraski Z, Schmid E. PHP Manual: Volume 1. 2000. URL: http://citebay.com/how-to-cite/php/ [accessed 2025-12-12]
26. University students statistic. Spanish Ministry of Science, Innovation and Universities. URL: https://www.ciencia.gob.es/Ministerio/Estadisticas/SIIU/Estudiantes.html [accessed 2025-12-03]
27. Integrated university information system. Spanish Ministry of Science, Innovation and Universities. 2025. URL: https://www.ciencia.gob.es/dam/jcr:08f45793-116d-4df2-8ddd-207662c3c6ee/PrincipalesResultadosEstudiantes2025.pdf [accessed 2025-12-03]
28. Patel D, Timsina P, Raut G, Freeman R, Levin MA, Nadkarni GN, et al. Exploring temperature effects on large language models across various clinical tasks. medRxiv 2024 [FREE Full text] [doi: 10.1101/2024.07.22.24310824]
29. Davis J, Van Bulck L, Durieux BN, Lindvall C. The temperature feature of ChatGPT: modifying creativity for clinical research. JMIR Hum Factors 2024;11:e53559 [FREE Full text] [doi: 10.2196/53559] [Medline: 38457221]
30. Akamine A. Effects of temperature settings on information quality of ChatGPT-3.5. medRxiv 2024 [FREE Full text]

31. Kamath A, Ullal SD. Learning and clinical reasoning experience of second-year medical pharmacology students and teachers with virtual patients developed using openLabyrinth. Electronic Journal of General Medicine 2023;20(5):em509. [doi: 10.29333/ejgm/13289]
32. Cross J, Kayalackakom T, Robinson R, Vaughans A, Sebastian R, Hood R, et al. Assessing ChatGPT's capability as a new age standardized patient: qualitative study. JMIR Med Educ 2025;11:e63353 [FREE Full text] [doi: 10.2196/63353] [Medline: 40393017]
33. Cero I, Luo J, Falligant JM. Lexicon-based sentiment analysis in behavioral research. Perspect Behav Sci 2024;47(1):283-310. [doi: 10.1007/s40614-023-00394-x] [Medline: 38660506]
34. Using AI-based virtual simulated patients for training in psychopathological interviewing: cross-sectional observational study. Open Science Framework. URL: https://osf.io/cqvdx [accessed 2025-12-17]

Abbreviations

CHERRIES: Checklist for Reporting Results of Internet E-Surveys
GAI: generative artificial intelligence
LLM: large language model
NLP: natural language processing
UMH: Miguel Hernández University
VSP: virtual simulated patient

Edited by T Leung; submitted 10.06.25; peer-reviewed by K Yauy, L Jantschi; comments to author 20.08.25; accepted 09.12.25; published 23.12.25.

Please cite as:
García-Torres D, Fernández C, Mira JJ, Morales A, Vicente MA
Using AI-Based Virtual Simulated Patients for Training in Psychopathological Interviewing: Cross-Sectional Observational Study
JMIR Med Educ 2025;11:e78857
URL: https://mededu.jmir.org/2025/1/e78857
doi:10.2196/78857
PMID:41433050

Assessing and Improving Study Skills Support in Medical Education Through a Student-Staff Partnership: Mixed Methods Approach

Nicole Tay, BSc, MBBS; Anaïs Deere, BSc; Dhivya Ilangovan, BSc, MBBS; Carys F E Phillips, BMBS, BMedSci, MSc; Emma Kelley, MSc, MUDr

UCL Medical School, University College London, 74 Huntley Street, London, United Kingdom

Corresponding Author:
Nicole Tay, BSc, MBBS
UCL Medical School, University College London, 74 Huntley Street, London, United Kingdom

Abstract

Background: The necessity for self-regulated, lifelong learners in the rapidly evolving field of medicine underscores the importance of effective study skills. Efforts to support students with these skills have had positive outcomes but are often limited in scope and accessibility, with a tendency to target groups facing immediate challenges.

Objective: This study aimed to explore the student perspective on study skills support at University College London Medical School through a student-staff partnership, with the goal of guiding future improvements.

Methods: A mixed methods approach was adopted using an anonymous questionnaire and focus groups. After analyzing questionnaire responses using descriptive statistics to refine focus group questions, focus groups were conducted to delve deeper into identified issues. Transcripts were analyzed thematically using inductive coding.

Results: In total, 116 students completed the questionnaire in full and 6 students participated in 2 focus groups. The questionnaire revealed that 68% (68/100) of respondents felt that they never received study skills support at University College London Medical School. Preferred methods of support included small group sessions (56/100, 56%) and topics like examination preparation (83/100, 83%) and study skills specific to medicine (72/100, 72%). Focus group themes were the lack of current study skills support, delivery of study skills support, specific study skills for medical school, personalized approach to support needed, and accessing support. Findings informed the co-creation of study skills resources.

Conclusions: Overall, the findings highlight the need for strategically incorporating study skills support at medical school, emphasizing early and consistent promotion and tailored delivery methods.

(JMIR Med Educ 2025;11:e65053) doi:10.2196/65053

KEYWORDS

study skills; medical education; E-learning; student-staff partnership; student perspectives; students; studying; mixed methods; questionnaire; thematic analysis; school; learning; exams; medical school; teaching; education

Introduction

In the dynamic realm of modern medicine, the need for practitioners to be self-regulated, lifelong learners is imperative if they are to keep up with ever-evolving medical knowledge [1]. In White et al's evidence-based model [2], self-regulated learning (SRL) is composed of a set of learnable skills: "planning, learning, assessment, and adjustment," acquisition of which can be facilitated by educators. Study skills, or "learning strategies," have been identified as an important component of the learning phase of SRL, and they can be defined as an "integrated repertoire of tactics and strategies, which facilitate acquisition, organization, retention, and application of such knowledge" [3].

SRL and study skills are relevant across higher education, and their importance is especially pronounced in medical education which incorporates theoretical and practical knowledge to develop graduates with the ability to learn independently and adaptively throughout their future careers [4,5]. In the clinical environment, SRL has been positively associated with students' academic achievement, clinical skills performance, and mental health outcomes [6]. SRL strategies have also shown a particularly strong association with affective outcomes such as attitude, motivation, and confidence [7], suggesting that SRL may support not only learning performance but also how students feel about their learning. The association between SRL strategies and learning outcomes is more pronounced in clinical clerkship than in preclerkship phases [7], and literature suggests that support for SRL should be more explicit and structured earlier in training [8,9]. This highlights the importance of

preparing students to self-regulate before they enter the clinical environment. This may be best supported by approaching SRL as a shared endeavor from the outset, with medical schools playing an active role in its early development [8,10].

The broader literature highlights the value of educational interventions to support SRL in medical students, such as tailored workshops [11] and mentor guidance [10]. Classroom-based SRL interventions and one-on-one academic coaching have also been positively received by first-year students and linked to greater anticipated use of SRL strategies [12]. In clinical settings, factors such as access to full-time clinical teachers, peer collaboration, and mentor support have been identified as significant predictors of SRL and are associated with improved clinical performance [13]. However, study skills interventions in medical education often follow a reactive-deficit model that targets students who have already encountered difficulties, such as those at risk of academic failure [14] or academically low-achieving [15], or a proactive-deficit model that focuses on students deemed "at risk," such as new entrants to medical school [16]. While identifying and supporting students in need is essential, this approach may inadvertently exclude students across the wider academic spectrum. Evidence indicates that study strategies influence academic achievement across the performance spectrum [17-19], supporting the case for more universal and inclusive approaches to study skill development.

At University College London Medical School (UCLMS), the MBBS program is a 6-year integrated degree, including an intercalated BSc in the third year [20]. The curriculum comprises themed modules, with clinical and professional practice running vertically throughout the program. Year 1 and 2 modules focus on the fundamentals of clinical science, and Years 4 to 6 modules emphasize clinical practice across specialties [21]. At the time of the study, UCLMS offered an array of study skills resources which supported SRL, such as "study skills clinics" (one-to-one or small group sessions with clinical staff members), support from personal tutors, and online learning resources. University College London (UCL) also provided study skill support accessible to all students such as online academic resources [22], alongside broader institutional support for students with specific learning differences or long-term health conditions, including tailored accommodations for assessments and teaching settings [23].

Despite the range of available support, a 2020 internal survey at UCLMS revealed that many students felt undersupported in developing effective study skills. This perceived gap between resource availability and student experience signaled a need to evaluate and improve current provision. In response to this, a student-staff partnership (the authors) was formed to explore and address the perceived gap in study skills support. Although student input into medical curriculum design is often limited or restricted to a needs analysis [24], literature suggests that engaging learners in the design process through student-staff partnership can enhance learning engagement, teaching effectiveness, and assessment outcomes [25]. This project adopted a student–staff partnership model to facilitate collaborative educational research that would not only develop the research skills of the team but also support a more authentic, peer-informed exploration of the student voice. By positioning students as coresearchers, supported by faculty mentorship and institutional resources, the partnership aimed to generate deeper insights into the student experience [26].

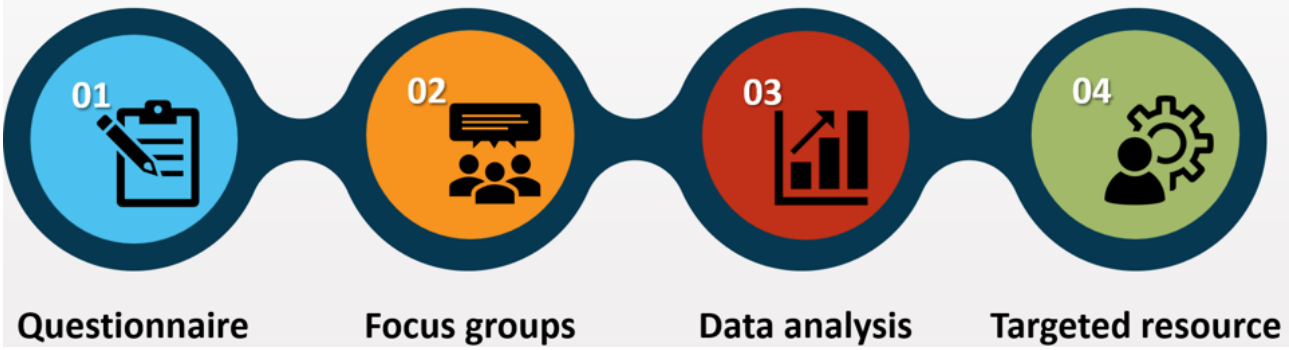
As interest in active student involvement in medical education grows [27], establishing models for successful collaboration is increasingly important. As a result, this research project aimed to use the student-staff partnership to explore the student perspective on study skills support at medical school, using findings to cocreate relevant resources and guide future improvements.

Methods

Overview

This is a mixed methods study consisting of a questionnaire and focus group, as depicted in Figure 1. A mixed methods design enabled collection of both qualitative and quantitative data, which allowed for greater confidence in the validity of any findings by means of triangulation [28], which was felt to be important given the relative lack of research in this specific area. This factor is also what led to the decision of conducting focus groups, due to their usefulness in exploratory research [29].

Figure 1. Summary of the methods.



An interpretivist paradigm was adopted for this study, recognizing that reality is subjective and socially constructed [30]. Aligned with social constructivism, this approach emphasizes the importance of understanding participants' experiences within their educational environment. It informed the use of focus groups and inductive thematic analysis to explore how students construct meaning based on their individual experiences. The collaborative student–staff research

model further reflects the interpretivist emphasis on coconstructed knowledge, positioning students as active contributors to the research process [26].

Questionnaire

Sampling Procedure and Recruiting Study Participants

The methodology of the web-based questionnaire is described in line with the CHERRIES checklist [31]. A web-based questionnaire was designed and distributed electronically, by purposive sampling of UCLMS medical students. Due to the diversity of intercalated BSc courses completed by students in their third year and the varying study skill requirements they entail, only students in Years 1, 2, 4, 5, and 6 of the MBBS program were recruited for this study.

Quality Assurance

A preliminary version of the questionnaire was developed by 2 students in the partnership and subsequently reviewed by the rest of the team. It was then further evaluated by the UCL Changemakers team, revised accordingly, and piloted by the other 2 students and 2 staff members within the partnership.

Recruitment Process

Students were initially sent a link to the questionnaire by email from the medical school, with a reminder email sent 1 month later. To further promote participation and increase response rates, online platforms such as WhatsApp [32] were also used. These channels were chosen due to their popularity among students and the observation by the medical student authors that their peers were particularly responsive through these platforms [33].

Survey Administration

The questionnaire was held on Opinio, accessible via link in the invitation email. As the study was voluntary, financial incentives were used to try and increase engagement [34]; students could enter a draw to win vouchers, which were funded by UCL Changemakers. The survey was anonymized with a separate link provided to enter the prize draw upon completion of the questionnaire. Data collection took place from March to May 2022.

Study Design

The questionnaire was structured to allow students to reflect on their current study skills, gather information on which existing UCL resources they accessed and their perceived effectiveness, and explore students' preferences for the delivery of future study skills support. It consisted of a total of 12 questions with various styles (multiple choice, multiselect multiple choice, Likert scale, and free text). The majority of questions used a 5-point Likert scale to enable participants' opinions to be quantitatively analyzed.

To facilitate targeted questioning on specific study strategies, seven study skill domains related to self-regulated learning were identified. This framework was developed through a scoping review of relevant literature in study skills and medical education and further refined through collaborative discussions within the research team, as shown in Textbox 1:

Textbox 1. Summary of the 7 study skill domains related to self-regulated learning.

- Time management
- Organization
- Exam preparation and technique
- Obtaining reliable sources of information
- Retaining information
- Study skills specific to a medical degree (eg, practical exams and navigating clinical years)
- Study skills and mental health (work or life balance and feeling overwhelmed)

A free space text box was provided near the end of the questionnaire for additional comments or suggestions. The full questionnaire can be found in "Multimedia Appendix 1: Questionnaire".

It was not mandatory for participants to complete all questions, and they could review and amend their answers before submission. We did not count view rates or participation rates, nor implement techniques which would prevent multiple entries. Partially completed questionnaires were analyzed.

Focus Groups

Before conducting the focus groups, the research team undertook a collaborative reflexive exercise to acknowledge potential biases during the research process and explore the diverse perspectives within the team [35]. This involved a semistructured discussion led by one of the staff researchers exploring individual assumptions and expectations, recognizing how the medical student researchers' unique position meant they had their own perceptions of study skills support and desirable outcomes of the project. Also discussed was how this position allowed them to better understand the student body and contextualize their opinions. The staff members also engaged in a separate reflexive discussion to consider their own positionality as educators and contributors to the support systems being evaluated.

Participants were recruited for the focus groups through the questionnaire, with interested individuals emailing the research team directly to maintain questionnaire anonymity. To encourage participation, each focus group attendee received a voucher funded by UCL Changemakers as a financial incentive.

The questionnaire results guided the development of the focus group questions. The focus groups were semistructured to allow students to freely elaborate on their views of study skills support

while maintaining enough structure to explore key questionnaire findings in greater depth. To ensure consistency, the interviewers followed an interview schedule with predetermined questions.

Two online focus group interviews were conducted in May 2022 using Zoom [36], a platform widely used for synchronous lectures and small-group teaching at UCLMS, to help participants feel comfortable engaging and turning on their cameras. Both focus groups were facilitated by 2 medical students from the research team to foster open, honest discussion and encourage participants to freely critique this study skills support without staff members present [37].

The focus groups began by inviting students to share their understanding of the term "study skills," with clarification provided as needed to ensure a clear understanding of the discussion topic. Respondents were asked a series of open-ended questions to establish their experience of study skills support at UCLMS, share their views on preferred methods of study skills support delivery, and explore further results of the questionnaire. Although questionnaire results were not shared with participants beforehand, they were introduced and explored during the discussions. The semi-structured focus group questions can be found in "Multimedia Appendix 2: Focus Group Questions".

Focus groups were video recorded and subsequently deleted after verbatim transcription by 2 of the authors, with all identifying information removed in accordance with the consent form the participants had signed before.

Analysis

Quantitative data from the survey were analyzed using descriptive statistics to compute percentages, means, and frequencies. Qualitative data from the focus groups were analyzed using inductive coding guided by Braun and Clark's phases of thematic analysis [38]. This process was informed by social constructivist principles, acknowledging that themes and insights emerge through the interactive process of data collection and analysis. Following familiarization with the transcripts through repeated readings, the questions were evenly distributed among the medical student researchers and coded independently. Initial codes were then reviewed collaboratively in meetings to standardize their interpretation across all questions. Each student researcher also reviewed and analyzed the coding completed by the other researchers to further ensure consistency. Through subsequent discussions, the codes were grouped into broad themes based on similarity, with subthemes developed where further categorization was needed. The codes and themes were then reviewed by the entire research team to refine them into more clearly defined categories. This process was supported by the program NVivo.

Ethical Considerations

Ethics approval for the study was granted by the Changemakers team—an initiative that provides funding and support for collaborative projects between students and staff aimed at enhancing the student learning experience at UCL (ethics approval number 12385/001). Informed consent was obtained from all questionnaire participants. No personally identifiable information was collected, aside from participants' year of study. All data were securely stored on a password-protected platform, accessible only to the research team. All data collected was anonymized in both the questionnaire and the focus group interviews. Students participating in the online questionnaire were given an option to partake in a USD $26.80 prize draw after completion; which was awarded to 1 student. Students participating in the focus group interviews all received compensation of USD $13.40.

Results

Questionnaire Results

In total, 1682 students across 5 years of medical school at UCL were sent the questionnaire. In total, 116 students out of 1682 (6.9% response rate) completed the online questionnaire in full; there was a 13.8% (16/116) drop-out rate from the start to the end of the survey. There was a UCL Central Study Skills Page available to all students for support, and 80% (80/100) of respondents indicated they had not accessed it. Among the 20% (20/100) of respondents who had accessed the page, its effectiveness was rated with a median score of 3/5.

A total of 68% (68/100) of students reported they never received study skills support during their time at UCLMS. Among the 32% (32/100) who reported they had received study skills support, the majority (81%, 26/32) indicated they received support 1 to 2 times a year.

When participants were asked to evaluate the effectiveness of various study skills delivery methods, all approaches—including lectures, self-directed learning, one-to-one support from staff or peers, personal tutors, and transition mentors (who provide weekly teaching sessions to support students' adjustment to university education in Years 1 and 2)—received a median effectiveness rating of 3 out of 5 across all year groups. In contrast, the study skills clinic received a lower median rating of 2.

The majority (56%, 56/100) of respondents indicated a preference for small group study skill sessions when asked about their preferred methods of delivery. Lectures (43%, 43/100), one-to-one support from staff (41%, 41/100), one-to-one support from peers (18%, 18/100), peer-to-peer support (34%, 34/100) and the study skills clinics (33%, 33/100) were also popular options. The 4 most popular topics preferred by students to be included within study skills support were: exam preparation (83%, 83/100), study skills specific to a medical degree (72%, 72/100), retaining information (69%, 69/100), and taking in new information (58%, 58/100). The demographics and summary of the questionnaire results can be found in the "Multimedia Appendix 3: Questionnaire results".

Focus Group Results

A total of 6 students participated in 2 focus groups: separated by those in Years 1 to 2 of study and those in Years 4 to 6 of study. This reflects that changes in SRL occur in a clinical learning environment [6], so the approach to study skills may differ in these groups as the modules in Years 4 to 6 are more focused on clinical practice.

Three individuals participated from Years 1 and 2, and three individuals participated from Years 4 to 6. Each focus group interview lasted around 40 minutes. Analysis identified 5 themes that explored the student perspective on study skills support: lack of current study skills support, delivery of study skills support, specific study skills for medical school, personalized approach to support needed, and accessing support.

Theme 1: Lack of Current Study Skills Support

Subtheme 1.1: - Lack of Support

In general, students felt there was a lack of study skills support provided by the University. This was more prominent in the Year 4 to 6 group, where students struggled to give any examples of formal teaching and acknowledged that the majority of their study support came from fellow students.

> *... I don't recall any, at least things I remember, any effective study skills teaching…* [(Participant 2, Year 4 to 6 group)]
>
> *... I feel like we haven't really had that much guidance, it's mainly from other students.* [(Participant 3, Year 4 to 6 group)]

In the Year 1 to 2 group, the problem identified was less to do with the availability of study skills support from the University, as students could provide several examples. Instead, their concern related to the effectiveness of that support.

> *... I don't think it [lecture on learning at medical school] was necessarily something that was particularly well supported or particularly well taught…* [(Participant 1, Year 1 to 2 group)]
>
> *... I signed up for it [study skills clinic], but I didn't end up going because it seemed like it was too generic, like it didn't seem medicine specific…* [(Participant 2, Year 1 to 2 group)]

Theme 2: Delivery of Study Skills Support

The way in which study skills support should be delivered was discussed. Although no unanimous consensus on the best method of delivery emerged, the advantages and disadvantages of various methods were acknowledged.

Subtheme 2.1 - One-to-One Staff Support

One-to-one staff support was a popular method of delivery among the students due to its focus on students' individual needs. Access to official guidance from medical school personnel, particularly those familiar with the curriculum and examination format, was considered especially beneficial. However, students emphasized the importance of staff fostering an open environment for students to voice their concerns without fear of judgment.

> *I think, as long as [the] staff member was kind of open [and] non judgmental and they make it a comfortable atmosphere for you to kind of open up and talk about your issues, I think that's what we need.* [(Participant 3, Year 1 to 2 group)]

Subtheme 2.2 - Peer Support

The concept of seeking study skills guidance from peers was positively received, with a preference for support from senior students who had completed the same style of assessments, allowing them to provide tailored, experience-based advice. In addition, while some students acknowledged the existence of established peer support systems in student societies, concerns were raised about the unfairness of the disparity in access to this support for students without such contacts.

> *...I feel like that's kind of unfair and some students also have like seniors who helped them out and stuff and not everyone has that, especially because of covid you know it's hard to make those contacts.* [(Participant 1, Year 1 to 2 group)]

Subtheme 2.3 - Small Group Teaching

Small group teaching was generally supported, as students valued the opportunity for interactive discussions and peer exchange of study habits and learning strategies. However, some expressed concern about feeling intimidated by the perceived workload and study strategies of their peers.

> *... sometimes I feel like other people are doing so much more than me or studying in so many like more efficient ways and I'm just like really behind and so yeah, I guess, in some ways it helps but in other ways, it can be quite intimidating as well.* [(Participant 3, Year 1 to 2 group)]

Subtheme 2.4 - Asynchronous vs Synchronous

Students preferred synchronous study skills lectures over asynchronous ones, as the live format was seen to promote better attendance.

> *... if it's asynchronous I don't think many people would go out of their way to do it because the workload is so much it wouldn't be anyone's priority* [(Participant 3, Year 4 to 6 group)]

Theme 3: Specific Study Skills for Medical School

Subtheme 3.1 - Medical School Is Unique

Students felt that studying medicine required a different learning approach than both school and other university degrees, particularly due to the unique nature of its assessment methods.

> *UCL medicine and medicine in general is really, really different to the way we're being assessed.* [(Participant 1, clinical group)]
>
> *... you can't just do it [learn] like A-levels or GCSE* [(Participant 2, Year 1 to 2 group)]

Subtheme 3.2 - Medical School Has a Lot of Content

Students often raised the fact that they felt overwhelmed with the volume of content they were expected to memorize in the course. They also found it challenging to discern the scope and depth of knowledge required, especially as expectations increased with each academic year.

> *... we still need to memorize the whole bunch of stuff to apply it and I struggle with that quite a bit just the sheer amount.* [(Participant 2, Year 4 to 6 group)]
>
> *I think it's really difficult to kind of ascertain what exactly we're expected to know and the level we're expected to know.* [(Participant 1, Year 4 to 6 group)]

Theme 4: Personalized Approach to Support Needed

Subtheme 4.1 - Study Skills Are Personal

Students recognized that not every approach to studying works for everyone as learning styles are very individual.

> *... everyone has like different circumstances and different areas that they struggle with; I struggle with time management.* [(Participant 3, Year 1 to 2 group)]

Subtheme 4.2 - Blind to Own Study Skill Weakness

Students expressed the challenge of identifying weaknesses in their study methods, noting that this awareness is an essential first step toward improvement.

> *"I don't know where I'm going wrong in order to fix it."* [(Participant 1, Year 4 to 6 group)]

Theme 5: Accessing Support

Subtheme 5.1 - Advertising Support

Students felt that study skills services available were not advertised well enough, suggesting promotion through The Royal Free, University College and Middlesex Medical Students' Association (RUMS), WhatsApp or Instagram (Meta) [20,28]. They also recommended the use of student testimonies so that students could gain a clearer insight into the benefits of the support.

> *If you made it a RUMS thing, because if you send it from the med school … I feel like people were kind of hesitant just because you know…I'm gonna be gone after or something. But if you make it like a RUMS thing it's like a community vibe…* [(Participant 3, Year 1 to 2 group)]

Subtheme 5.2 - Establishing Support Early On

Students believed that information about study skills support should be offered early in the year as opposed to just prior to exam season, allowing those struggling more time to access and benefit from the assistance.

> *... because they can connect earlier on in the year and kind of work on that, for the rest of the year rather than it being something that you do like a month before the exam and just panicking…* [(Participant 3, Year 1 to 2 group)]

Discussion

Principal Findings

Overall, the results from both the questionnaire and focus groups highlighted a perceived lack of study skills support. There was no unanimous consensus about students' preferences on how study skills support should be delivered. Despite the availability of established study skills services outlined in the introduction, 68% of questionnaire respondents (68/100) reported never receiving such support. This sentiment was echoed in the focus groups, where students indicated they primarily relied on peer advice for study skills guidance. This reinforces the importance of more personalized study support and the need for review of the promotion and accessibility of existing services.

Lectures, one-to-one support, peer group teaching, and study skills clinics all demonstrated comparable popularity as delivery methods; however, small group study skills sessions were the only format preferred by a majority of students (56%; 56/100) in the questionnaire. Focus group subthemes—"one-to-one staff support" and "small group teaching"—further illuminated mixed perceptions of small group sessions. While small groups facilitated peer sharing of study skills, some students found them intimidating due to comparisons with others. Conversely, whilst one-to-one support from staff was perceived as personalized to students' individual needs, especially if by a staff member familiar with the curricula and assessment process, its use was contingent on occurring in a nonjudgmental environment. These findings reflect the literature where both classroom-based learning and one-to-one academic coaching are shown to encourage the use of SRL techniques [12].

The focus group themes and subthemes highlighted that each method of study skills support delivery offers distinct advantages and disadvantages, catering to different student priorities. For example, some students valued direct support from staff for its perceived credibility, while others expressed concern about potential judgment when admitting difficulties. This stigma has been previously identified as a barrier towards medical students seeking help, particularly in the context of non-compulsory activities which offer additional support [39]. Moreover, regardless of delivery preferences, students acknowledged that study skills are highly individual, with no "one size fits all" solution—an insight consistent with existing literature [40]. This also aligns with the literature demonstrating the need for individualized support in the clinical environment to accommodate students' unique SRL strategies [10]. From a faculty perspective, these diverse needs and preferences must therefore be carefully considered when designing study skills support to maximize their effectiveness.

It is clear that social influences shape how students seek study skills support, with many highlighting that most of their support came from fellow peers. Help-seeking from classmates and senior students has been linked to higher levels of self-regulated learning [13], while friendships may also influence exam performance, likely through informal sharing of study strategies [41]. However, this dynamic can also result in unequal access to valuable resources, as demonstrated in the subtheme of "peer support," where senior students were reported to provide study skills guidance exclusively within their own societies. Similar patterns have been documented regarding the selective sharing of assessment materials, indicating that this issue extends beyond study skills alone [42]. Evidence from clinical skills education highlights that peer learning fosters a safe learning environment and aids SRL [43], whilst guidance from senior students has been identified as a positive predictor of SRL [13] . Given that future doctors must collaborate effectively, cultivating a culture

of inclusivity and cooperation around study practices is crucial, and medical schools are well positioned to facilitate this.

The four most popular topics identified in the questionnaire, for which students sought study skills support, were those most specific to medicine, primarily focusing on exam preparation and content management. This finding aligns with the theme "specific study skills for medical school" and underscores the distinctive nature of medical education compared to other disciplines, highlighting that students prefer support tailored to the unique demands of their degree rather than generic university-wide resources. This preference may partly reflect the distinctive assessment methods in medicine, which extend beyond cognitive knowledge to include behavioral competencies, necessitating learning methods that may not have been part of students' earlier educational experiences [44]. Furthermore, the vast volume of content in medical training—highlighted by the theme "medical school has a lot of content"—can be overwhelming and necessitates specific strategies for effective organization and retention.

Findings from both the questionnaire and focus group themes revealed that students highly value having access to a variety of study methods, including many which were already available at UCLMS or wider UCL. At the outset of the project, students in the research team recommended the creation of a centralized study skills page for the medical school on Moodle [22], the online platform used across UCL to deliver course material, and the data corroborated this suggestion from the wider student body. The research team, therefore, developed the page to provide UCLMS medical students with early and easy-to-access study skills support, directly addressing the need for improved accessibility highlighted in the theme "accessing support." The page signposts and advertises existing resources, including links to book onto one-to-one or group "study skills clinics" with staff. In addition, informed by study findings, new resources were made by the students in the partnership such as anonymous group forums for peer-to-peer interaction, evidence-based articles on effective study techniques, and videos with study skills tips from other students. Overall, the platform provides diverse support, allowing students to access guidance that meets their individual learning needs.

Strengths

This project offers several strengths. It adds to the existing literature on study skills support in medical education, with the advantage of a distinct emphasis on student perspectives. This was facilitated through multiple strategies, including the involvement of UCLMS students as members of the research team and as facilitators of the focus groups, which may have encouraged participants to speak more openly and candidly about the current provision of study skills support. Furthermore, while including the student authors' names in recruitment emails may have introduced some response bias, it likely also enhanced engagement by motivating students with personal connections to the authors to participate.

Staff involvement in the project enabled a well-rounded, collaborative approach to understanding and addressing the study skills needs of UCLMS students, increasing the likelihood of successful implementation of proposed changes. For students, this collaboration with staff was particularly beneficial as it fostered a sense of being heard and valued. It allowed them to communicate their needs more effectively and see their suggestions translated into action, which, in turn, enhanced their engagement and motivation. Ultimately, this student–staff partnership established a mutually inclusive and supportive environment that strengthened communication and promoted a more meaningful dialogue around study skills support.

The development of the online resource itself also has several key strengths. It provides a centralized, easily accessible platform that supports a diverse student cohort by consolidating a wide range of study skills resources in one place. While primarily designed for students, the resource also offers UCLMS staff valuable insights into student needs and perspectives, meaning that in the future, changes can be made on a wider scale if necessary. Its interactive design and capacity for continuous updates also ensure the resource remains responsive and relevant to evolving student requirements, preventing it from becoming outdated over time.

Weaknesses

The questionnaire had a low response rate, with just 6.9% (116/1682) of the medical student population participating, and 13.8% (16/116) of those respondents not completing all questions. This limits the generalizability of the findings and suggests results may disproportionately reflect students with a particular interest in study skills or those personally connected to the student researchers, whose names were included in the initial recruitment email. In addition, focus group recruitment relied on participants completing the questionnaire and then registering their interest via email at its conclusion, meaning the questionnaire dropout rate likely also contributed to the low focus group turnout. Finally, due to the small focus group sample, data saturation was not achieved, as novel themes emerged that could not be fully explored.

Future Directions

Future study proposals would like to evaluate student and staff opinion on the study skills resources created and implemented. Overall, this study provides insight into the medical student perspective on study skills at UCLMS, identifying key areas of improvement and methods to target this. Similar initiatives could be effectively undertaken at other medical schools.

Acknowledgments

We would like to extend our gratitude to Philip Marshall-Lockyer for assisting in data collection in this research. We also thank participants for completing the questionnaire and taking part in the focus groups and are grateful to the reviewers for their

constructive comments. We presented the findings at the ASME conference 2023, and we are thankful to the audience for their questions.

This study was funded by UCL Changemakers.

Data Availability

Data protection registration: Z6364106/2018/04/35 social research

All data generated or analyzed during this study are included in this published article [and its supplementary information files].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Table of contents outlining the questions within the questionnaire and answer options.
[PDF File, 136 KB - mededu_v11i1e65053_app1.pdf]

Multimedia Appendix 2
Table of contents of focus group questions and probing questions.
[PDF File, 107 KB - mededu_v11i1e65053_app2.pdf]

Multimedia Appendix 3
Table of contents of the results of the questions within the questionnaire. Given in mean, median, and percentage.
[PDF File, 133 KB - mededu_v11i1e65053_app3.pdf]

References

1. Marzo RR, Faculty of Medicine, Asia Metropolitan University, Johor, MALAYSIA. Role of medical education in cultivating lifelong learning skills for future doctors. EIMJ 2018;10(3):63-66. [doi: 10.21315/eimj2018.10.3.7]
2. White CB, Gruppen LD, Fantone JC. Self-regulated learning in medical education. In: In Understanding Medical Education Summary: John Wiley & Sons, Ltd; 2013:201-211. [doi: 10.1002/9781118472361.ch15]
3. Ball CR. Study skills. In: Goldstein S, Naglier JA, editors. Encyclopedia of Child Behavior and Development: Springer US; 2011. [doi: 10.1007/978-0-387-79061-9]
4. Siddaiah-Subramanya M, Nyandowe M, Zubair O. Self-regulated learning: why is it important compared to traditional learning in medical education? Adv Med Educ Pract 2017;8:243-246. [doi: 10.2147/AMEP.S131780] [Medline: 28360542]
5. Sandars J. When I say … self-regulated learning. Med Educ 2013 Dec;47(12):1162-1163. [doi: 10.1111/medu.12244] [Medline: 24206149]
6. Cho KK, Marjadi B, Langendyk V, Hu W. The self-regulated learning of medical students in the clinical environment - a scoping review. BMC Med Educ 2017 Jul 10;17(1):112. [doi: 10.1186/s12909-017-0956-6] [Medline: 28693468]
7. Zheng B, Sun T. Self-regulated learning and learning outcomes in undergraduate and graduate medical education: a meta-analysis. Eval Health Prof 2024 Oct 3:1632787241288849. [doi: 10.1177/01632787241288849] [Medline: 39361881]
8. Brydges R, Manzone J, Shanks D, et al. Self-regulated learning in simulation-based training: a systematic review and meta-analysis. Med Educ 2015 Apr;49(4):368-378. [doi: 10.1111/medu.12649] [Medline: 25800297]
9. van der Gulden R, Veen M, Thoonen BPA. A philosophical discussion of the support of self-regulated learning in medical education: the treasure hunt approach versus the (Dutch) “dropping” approach. Teach Learn Med 2023;35(5):623-629. [doi: 10.1080/10401334.2023.2187810] [Medline: 36939190]
10. van Houten-Schat MA, Berkhout JJ, van Dijk N, Endedijk MD, Jaarsma ADC, Diemers AD. Self-regulated learning in the clinical context: a systematic review. Med Educ 2018 Oct;52(10):1008-1015. [doi: 10.1111/medu.13615] [Medline: 29943415]
11. Sisa I, Garcés MS, Crespo-Andrade C, Tobar C. Improving learning and study strategies in undergraduate medical students: a pre-post study. Healthcare (Basel) 2023 Jan 28;11(3):375. [doi: 10.3390/healthcare11030375] [Medline: 36766950]
12. Boyd T, Besche H, Goldhammer R, Alblooshi A, Coleman BI. First-year medical students’ perceptions of a self-regulated learning-informed intervention: an exploratory study. BMC Med Educ 2022 Nov 29;22(1):821. [doi: 10.1186/s12909-022-03908-4] [Medline: 36447223]
13. Zhang JY, Liu YJ, Shu T, Xiang M, Feng ZC. Factors associated with medical students’ self-regulated learning and its relationship with clinical performance: a cross-sectional study. BMC Med Educ 2022 Feb 25;22(1):128. [doi: 10.1186/s12909-022-03186-0] [Medline: 35216585]

14. Stegers-Jager KM, Cohen-Schotanus J, Themmen APN. The effect of a short integrated study skills programme for first-year medical students at risk of failure: a randomised controlled trial. Med Teach 2013;35(2):120-126. [doi: 10.3109/0142159X.2012.733836] [Medline: 23110355]
15. Zarei Hajiabadi Z, Sandars J, Norcini J, Gandomkar R. The potential of structured learning diaries for combining the development and assessment of self-regulated learning. Adv Health Sci Educ Theory Pract 2024 Mar;29(1):27-43. [doi: 10.1007/s10459-023-10239-6] [Medline: 37273028]
16. Miller CJ. Implementation of a study skills program for entering at-risk medical students. Adv Physiol Educ 2014 Sep;38(3):229-234. [doi: 10.1152/advan.00022.2014] [Medline: 25179612]
17. Proctor BE, Prevatt FF, Adams K, Reaser A, Petscher Y. Study skills profiles of normal-achieving and academically-struggling college students. csd 2006 Jan;47(1):37-51. [doi: 10.1353/csd.2006.0011]
18. Ward PJ. Influence of study approaches on academic outcomes during pre-clinical medical education. Med Teach 2011;33(12):e651-e662. [doi: 10.3109/0142159X.2011.610843] [Medline: 22225447]
19. May W, Chung EK, Elliott D, Fisher D. The relationship between medical students' learning approaches and performance on a summative high-stakes clinical performance examination. Med Teach 2012;34(4):e236-e241. [doi: 10.3109/0142159X.2012.652995] [Medline: 22455715]
20. MBBS programme of study, Faculty of Medical Sciences. URL: https://www.ucl.ac.uk/medical-sciences/divisions/medical-school/information-current-mbbs-students-and-staff/mbbs-programme-study [accessed 2025-05-21]
21. Course structure, Faculty of Medical Sciences. URL: https://www.ucl.ac.uk/medical-sciences/divisions/medical-school/study/undergraduate/mbbs-programme/course-structure [accessed 2025-05-21]
22. UCL. Study skills, Students. 2025. URL: https://www.ucl.ac.uk/students/skills [accessed 2025-05-21]
23. UCL. The support we provide, Students. 2024. URL: https://www.ucl.ac.uk/students/support-and-wellbeing-services/disability-support/support-we-provide [accessed 2025-05-21]
24. Beckert L, Wilkinson TJ, Sainsbury R. A needs-based study and examination skills course improves students' performance. Med Educ 2003 May;37(5):424-428. [doi: 10.1046/j.1365-2923.2003.01499.x] [Medline: 12709183]
25. Bovill C. Student–staff partnerships in learning and teaching: an overview of current practice and discourse. Journal of Geography in Higher Education 2019 Oct 2;43(4):385-398. [doi: 10.1080/03098265.2019.1660628]
26. Han SP, Chua E, Rahadian RE, Mogali SR. How to … co-create research with medical students. Clin Teach 2025 Apr;22(2):e70066. [doi: 10.1111/tct.70066] [Medline: 40078108]
27. Martens SE, Wolfhagen I, Whittingham JRD, Dolmans D. Mind the gap: Teachers' conceptions of student-staff partnership and its potential to enhance educational quality. Med Teach 2020 May;42(5):529-535. [doi: 10.1080/0142159X.2019.1708874] [Medline: 31961749]
28. Mertens DM, Hesse-Biber S. Triangulation and mixed methods research. J Mix Methods Res 2012 Apr;6(2):75-79. [doi: 10.1177/1558689812437100]
29. Stalmeijer RE, Mcnaughton N, Van Mook W. Using focus groups in medical education research: AMEE guide no. 91. Med Teach 2014 Nov;36(11):923-939. [doi: 10.3109/0142159X.2014.917165] [Medline: 25072306]
30. Crotty M. The Foundations of Social Research: Meaning and Perspective in the Research Process: London: Routledge; 2021:978-971. [doi: 10.4324/9781003115700]
31. Eysenbach G. Improving the quality of web surveys: the checklist for reporting results of internet e-surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
32. WhatsApp: secure and reliable free private messaging and calling. WhatsApp.com. URL: https://www.whatsapp.com/ [accessed 2025-05-23]
33. Guckian J, Utukuri M, Asif A, et al. Social media in undergraduate medical education: a systematic review. Med Educ 2021 Nov;55(11):1227-1241. [doi: 10.1111/medu.14567] [Medline: 33988867]
34. Royal KD, Flammer K. Survey incentives in medical education: what do students say will entice them to participate in surveys? MedSciEduc 2017 Jun;27(2):339-344. [doi: 10.1007/s40670-017-0407-3]
35. Denniston C. Sharpening reflexive practice in health professional education research. FoHPE 2023:85-94. [doi: 10.11157/fohpe.v24i1.734]
36. One platform to connect. Zoom. URL: https://zoom.us/ [accessed 2025-05-23]
37. Djohari N, Higham R. Peer-led focus groups as 'dialogic spaces' for exploring young people's evolving values. Cambridge Journal of Education 2020 Sep 2;50(5):657-672. [doi: 10.1080/0305764X.2020.1754763]
38. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
39. Kelley E. You can lead a horse to water: exploring student engagement in non-compulsory activities to support learning [version 1; not peer reviewed]. Presented at: AMEE Conference, 2023; Aug 26-30, 2023; Scottish Event Campus (SEC) in Glasgow, Scotland, UK. [doi: 10.21955/mep.1115217.1]
40. Swanwick T, editor. Chapter 15: self-regulated learning in medical education. In: Understanding Medical Education: Evidence, Theory and Practice: UK: John Wiley & Sons, Ltd; 2013. [doi: 10.1002/9781118472361.ch15]

41. Woolf K, Potts HWW, Patel S, McManus IC. The hidden medical school: a longitudinal study of how social networks form, and how they relate to academic performance. Med Teach 2012;34(7):577-586. [doi: 10.3109/0142159X.2012.669082] [Medline: 22746963]
42. Gondhalekar AR, Rees EL, Ntuiabane D, et al. Levelling the playing field: students' motivations to contribute to an amnesty of assessment materials. BMC Med Educ 2020 Nov 23;20(1):450. [doi: 10.1186/s12909-020-02320-0] [Medline: 33225940]
43. Alzaabi S, Nasaif M, Khamis AH, Otaki F, Zary N, Mascarenhas S. Medical students' perception and perceived value of peer learning in undergraduate clinical skill development and assessment: mixed methods study. JMIR Med Educ 2021 Jul 13;7(3):e25875. [doi: 10.2196/25875] [Medline: 34021539]
44. Grant J. Learning needs assessment: assessing the need. BMJ 2002 Jan 19;324(7330):156-159. [doi: 10.1136/bmj.324.7330.156] [Medline: 11799035]

Abbreviations

RUMS: Royal Free, University College and Middlesex Medical Students' Association
SRL: self-regulated learning
UCL: University College London
UCLMS: University College London Medical School

Edited by D Chartash; submitted 04.08.24; peer-reviewed by S Otero, S Ganesh; revised version received 23.05.25; accepted 23.06.25; published 03.09.25.

Please cite as:
Tay N, Deere A, Ilangovan D, Phillips CFE, Kelley E
Assessing and Improving Study Skills Support in Medical Education Through a Student-Staff Partnership: Mixed Methods Approach
JMIR Med Educ 2025;11:e65053
URL: https://mededu.jmir.org/2025/1/e65053
doi:10.2196/65053

Recruiting Medical, Dental, and Biomedical Students as First Responders in the Immediate Aftermath of the COVID-19 Pandemic: Prospective Follow-Up Study

Nicolas Schnetzler[1,2], MSc; Victor Taramarcaz[1,2], MSc; Tara Herren[1,2], MSc; Eric Golay[2], MAS; Simon Regard[2,3], MD; François Mach[4], MD; Amanta Nasution[1,2], BSc; Robert Larribau[1,2], MD; Melanie Suppan[1,5], MD, MSc; Eduardo Schiffer[1,5,6], MD; Laurent Suppan[1,2], MD

[1]Department of Anaesthesiology, Pharmacology, Intensive Care and Emergency Medicine, Faculty of Medicine, University of Geneva, Geneva, Switzerland

[2]Division of Emergency Medicine, Department of Acute Care Medicine, Geneva University Hospitals, Rue Gabrielle-Perret-Gentil 2, Geneva, Switzerland

[3]Cantonal Physician Division, Cantonal Health Office, State of Geneva, Geneva, Switzerland

[4]Cardiology Department, University of Geneva Hospitals and Faculty of Medicine, Geneva, Switzerland

[5]Division of Anaesthesiology, Department of Acute Care Medicine, Geneva University Hospitals, Geneva, Switzerland

[6]Unit of Development and Research in Medical Education (UDREM), Faculty of Medicine, Geneva, Switzerland

Corresponding Author:
Laurent Suppan, MD
Department of Anaesthesiology, Pharmacology, Intensive Care and Emergency Medicine, Faculty of Medicine, University of Geneva, Geneva, Switzerland

Abstract

Background: Basic life support improves survival prognosis after out-of-hospital cardiac arrest, but is too rarely provided before the arrival of professional rescue services. First responder networks have been developed in many regions of the world to decrease the delay between collapse and initiation of resuscitation maneuvers. Their efficiency depends on the number of first responders available and many networks lack potential rescuers. Medical, dental, and biomedical students represent an almost untapped source of potential first responders, and a first study, carried out during the COVID-19 pandemic, led to the recruitment of many of these future professionals even though many restrictions were still in effect.

Objective: The objective of this study was to determine the impact of an enhanced strategy on the recruitment of medical, dental, and biomedical students as first responders in the immediate aftermath of the COVID-19 pandemic.

Methods: This was a prospective follow-up study, conducted between November 2021 and March 2022 at the University of Geneva Faculty of Medicine, Geneva, Switzerland. A web-based study platform was used to manage consent, registrations, and certificates. A first motivational intervention was held early in the academic year and targeted all first-year medical, dental, and biomedical students. Participants first answered a questionnaire designed to assess their initial basic life support knowledge before following an e-learning module. Those who completed the module were able to register for a face-to-face training session held by senior medical students. A course certificate was awarded to those who completed these sessions, enabling them to register as first responders on the Save a Life first responder network. Since the number of students who had enlisted as first responders 2 months after the motivational intervention was markedly lower than expected, a second, unplanned motivational intervention was held in an attempt to recruit more students.

Results: Out of a total of 674 first-year students, 19 (2.5%) students had registered as first responders after the first motivational intervention. This was significantly less than the proportion achieved through the initial study (48/529, 9.1%; $P<.001$). The second motivational intervention led to the enrollment of 7 more students (26/674, 3.9%), a figure still significantly lower than that of the original study ($P<.001$). At the end of the study, 76 (11.3%) students had been awarded a certificate of competence.

Conclusions: Contrary to expectations, an earlier presentation during the academic year outside the COVID restriction period did not increase the recruitment of medical, dental, and biomedical students as first responders in the immediate aftermath of the COVID-19 pandemic. The reasons underlying this drop in motivation should be explored to enable the design of focused motivational interventions.

(JMIR Med Educ 2025;11:e63018) doi:10.2196/63018

KEYWORDS

basic life support; out-of-hospital cardiac arrest; cardiopulmonary resuscitation; e-learning; blended learning; first responder; undergraduate medical education; student motivation; motivational strategies; medical student; COVID-19; pandemic; life support; survival prognosis; biomedical students; dental students; motivational interventions

Introduction

Background

Basic life support (BLS) improves survival prognosis after out-of-hospital cardiac arrest (OHCA) but is too rarely provided before the arrival of professional rescue services [1-6]. Without BLS, the probability of survival decreases by 10% for each minute that passes [7]. Thus, professional rescue is of limited worth if BLS has not been provided either by bystanders or by first responders [8-10]. Indeed, several studies have demonstrated that initiation of BLS maneuvers by nonprofessionals improves survival and neurological outcomes [1,11,12].

Increasing global awareness regarding the importance of quickly initiating BLS maneuvers after OHCA will take time, and barriers to action often prevent bystanders from initiating cardiopulmonary resuscitation [1,13-15]. To overcome this limitation, first responder systems have been developed in many regions of the world. These systems rely on BLS-certified professional or nonprofessional rescuers who accept a call to respond to OHCA alarms if they happen to be nearby.

In Geneva, Switzerland, the Save a Life project was initiated in October 2019 by the Swiss Emergency Responder Association, with the objective of developing a regional network of first responders [7]. When an OHCA is identified by the emergency medical call center, an alert is displayed on the Save a Life first responder app. If a first responder is near enough and agrees to intervene, the position of the nearest automatic external defibrillator (AED) is displayed along with the exact location of the intervention. The main limitations of this system are the limited number of first responders, their availability, and their geographical distribution.

To improve the number of first responders, Taramarcaz et al [16] designed a process to recruit first-year medical students. Their study took place while the COVID-19 pandemic was still ongoing, and many restrictions were still in effect. In addition, the motivational intervention designed to catch the students' interest was held online rather than in an auditorium and took place rather late after the beginning of the academic year and close to a critical exam session. Thus, the authors hypothesized that an intervention taking place earlier in the academic year, and without the constraints imposed by the COVID-19 pandemic, could lead to higher participation rates and the recruitment of a higher proportion of medical students as first responders [16].

Objective

The objective of this study was to determine the impact of the modifications proposed by Taramarcaz et al [16] on first responder recruitment.

Methods

Study Design

This prospective follow-up study was conducted between November 2021 and March 2022 and followed a structure and sequence similar to that described by Taramarcaz et al [16].

The study platform used for the initial study was reset and reused for this follow-up study. Given the use of a web-based platform, methods and results are reported according to the Checklist for Reporting Results of Internet E-Surveys (CHERRIES) guidelines when appropriate [17]. Data were managed in accordance with the European General Data Protection Regulation [18]. A more detailed description of the tools used can be found in Multimedia Appendix 1.

The learning path was identical to that described in Taramarcaz et al's [16] study and followed a flipped classroom design: after ensuring that no exclusion criteria were present, the first-year medical, dental, and biomedical students of the University of Geneva Faculty of Medicine (UGFM) answered a questionnaire designed to assess their initial BLS knowledge before following an e-learning module. After completing this module, they were able to register for a face-to-face training session held by senior medical students. The estimated time required to complete the e-learning and practice session was about an hour and a half. This duration was chosen as being long enough for learning and skill retention while avoiding an overt demand on their busy schedule. The participants who completed the entire learning path were awarded a BLS-AED course certificate enabling them to register as first responders on the Save a Life first responder network. The whole process, including the certification, was entirely free of charge, and there was no obligation for students to participate. The only incentive was to obtain a BLS-AED certificate.

Ethical Considerations

Since the regional ethics committee (Commission cantonale d'éthique de la recherche, Geneva, Switzerland) had already acknowledged that this design did not fall within the scope of the Swiss federal law on research involving human beings (Req-2020 - 01143), no further ethical assessment was required or requested.

Recruitment

A motivational intervention was performed live on November 29, 2021. This intervention was animated by 2 senior medical students and took place at the end of a basic medical science course. The presentation contained a short, humorous introductory video, a description of the project, and an overview of the Save a Life network. The last slide included a QR code and a URL link to the study platform as indicated in the research protocol [19]. On the same day, all potential participants received the same information by email through the class's

mailing list. To promote participation, a second motivational email was sent to the entire class on December 10.

Since participation was markedly lower than expected by the end of December, a second intervention was planned, this time at the start of a course on atherosclerosis given by the head of the cardiology department at Geneva University Hospital. This intervention took place on January 10, 2022, and a final reminder email was sent on January 13, 2022. For the second intervention, different support material was used, and various real-life scenarios were included, showcasing how BLS knowledge could enable them to act in the case of OHCA.

Enrollment

The QR code and URL provided during the motivational interventions and through the invitation emails redirected the students to an introductory page detailing the study's objectives and procedures. Those willing to participate were asked to answer 2 questions designed to detect the presence of either of 2 exclusion criteria: being registered as first responder and not being a UGFM student. If neither exclusion criterion was met, the students were redirected to a consent form (Table 1) including a disclaimer about data handling and security. Those who agreed were asked to create an account and to provide minimal personal information (first name, last name, and email address) for contact purposes and to allow for the creation of nominative BLS-AED certificates. The students who refused to participate and those who met either exclusion criteria were also given the possibility to follow the learning path and to receive a certificate allowing them to join a first responder system.

After completing the registration process, participants were asked to fill out a precourse questionnaire designed to gather demographic data and determine their precourse BLS knowledge (Table 2).

Table . Screening questionnaire and consent form (reused from Taramarcaz et al [16]).

Survey page, field, and question	Type of question
Page 1	
Already filled the questionnaire or exclusion criteria	
Already a first responder	Yes or no
Demographics	
Student at UGFM[a]	Yes or no
If no: current professional status	Open
Consent	
Agree to participate	Yes or no
If no: reasons for refusal	MAQ[b]
If no: access to the e-learning module	Yes or no

[a]UGFM: University of Geneva Faculty of Medicine.

[b]MAQ: multiple answer question.

Table . Precourse questionnaire (reused from Taramarcaz et al [16]).

Survey page, field, and question	Type of question
1: Demographics	
Year of birth	Open (Regex[a])
Gender	MCQ[b]
Medical, biomedical or dental medicine student	MCQ
Former student or graduate of another health care profession	MCQ
Target Specialty	MCQ
2: General BLS[c] knowledge	
Ever heard of BLS or ACLS[d] before	Yes/no
Meaning of AED[e,f]	Open
Year of the last BLS guidelines update	Open (Regex)
Phone number of the emergency medical communication center[f]	Open
3: Prior BLS experience	
Prior BLS training	MAQ[g]
Wish for additional BLS training	Yes/no
4: Specific BLS knowledge	
Criteria used to recognize OHCA[f,h]	MAQ
BLS-sequence[f]	Ordering
Artery for pulse assessment[f]	MCQ
Compression depth[f]	MCQ
Compressions: ventilation ratio[f]	MCQ
Compression rate[f]	MCQ
Compression-only CPR[fi]	Yes/no
Foreign body airway obstruction [f]	MCQ
5: Confidence	
Precourse confidence to act in an OHCA situation	Likert scale (1-5)

[a]A Regex validation rule was used to avoid invalid entries.
[b]MCQ: multiple choice question (only one answer accepted).
[c]BLS: basic life support.
[d]ACLS: advanced cardiovascular life support.
[e]AED: automatic external defibrillator.
[f]Items used to calculate the 10-point score (initial BLS knowledge).
[g]MAQ: multiple answer question (more than one answer accepted).
[h]OHCA: out-of-hospital cardiac arrest.
[i]CPR: cardiopulmonary resuscitation.

E-Learning and Practice Sessions

The interactive e-learning module used in Taramarcaz et al's [16] study was reused without any changes since it still matched the objectives, respected the Swiss Resuscitation Council's guidelines, and had not received any negative feedback from the students. This module was designed to last about 30 minutes, but no time limit was set and students were able to resume at will. A screen enabling participants to register for near-peer animated practice sessions was displayed upon completion of this e-learning module.

Practice sessions lasted 1 hour and were limited to 4 participants. A total of 32 sessions (128 slots) were planned between December 6, 2021 and March 11, 2022. The instructor-to-participant ratio (1:4) was kept unchanged to maintain high-quality training even though the COVID-19 restrictions had been lifted. The senior medical students who

animated these near-peer-led practice sessions were all certified as BLS-AED instructors according to the Swiss Resuscitation Council's guidelines. Most of the students who had already participated as instructors in Taramarcaz et al's [16] study (15/17, 88%) agreed to resume their involvement and 5 new instructors were trained. While all instructors were to ensure that the objectives had been met by using a standardized checklist, they were free to adapt the structure of their training sessions according to the participants' profiles.

Final Questionnaire and Certification

An email embedding a link to a postcourse questionnaire was sent to the students who successfully completed the practice sessions (Table 3). Participation in this questionnaire was mandatory to obtain a nominative BLS-AED certificate. These certificates, which had a 1-year validity, enabled participants to enroll as first responders on the Save a Life platform.

Table . Postcourse questionnaire (reused from Taramarcaz et al [16]).

Survey page, field, and question	Type of question
1: Opinion	
Appreciation	Yes/no
If yes: positive thoughts	MAQ[a]
If no: negative thoughts	MAQ
General comments	Free text
2: Confidence	
Postcourse confidence for OHCA[b] management	Likert scale (1-5)
Factors contributing to confidence	Likert scale (1-5)
Factors contributing to lack of confidence	Likert scale (1-5)
Other comments on confidence	Free text
3: First responders	
Intention to register as first responder	Yes/no
If yes: contributing factors	Likert scale (1-5)
If no: impeding factors	Likert scale (1-5)
Other factors	Free text
4: Improvement	
Suggestion for improvement	Free text

[a]MAQ: multiple answer question.

[b]OHCA: out-of-hospital cardiac arrest.

Adaptations From the Implementation Study

In line with this study's objectives, the main changes from the implementation study were that the initial presentation to first-year students and the training sessions were held earlier in the academic year [16], with the hypothesis that this would increase the number of registrations as first-year students would be further away from their final exams. Thus, the project was presented on November 29, 2 months earlier than the original study.

Despite this adaptation, and contrarily to our hypothesis, the number of participants was markedly lower than that in the original study. A second, initially unplanned intervention was therefore carried out in early January 2022, and constitutes the second major adaptation from the original implementation study.

Another difference was that biomedical students were also invited to participate in this study. Finally, the practice sessions were held between December 2021 and March 2022 in this study while they had taken place between January and April 2021 in the implementation study. The number of slots remained unchanged.

Outcomes

The primary outcome was the proportion of students who had registered as first responders before the second intervention took place, that is, by January 9, 2022. Secondary outcomes were the proportion of students who had registered following the second intervention, the overall proportion of students who had registered as first responders by May 1, 2022, and attrition at each step of the study [20]. The difference in self-reported confidence in performing BLS maneuvers was also assessed.

Statistical Analysis

Data curation and analysis were carried out using STATA/BE (version 17.0; StataCorp LLC). Descriptive statistics were used to describe the evolution of the number of students at each step of the learning path. Given the sample size, parametric tests were used when appropriate. A *P* value of less than .05 was considered statistically significant.

The chi-square test was used to assess the difference in student recruitment distribution between this study and Taramarcaz et al's [16]. This was carried out by reusing the original data file, which is freely available online as a Multimedia Appendix 1 of the original study. Since biomedical students had not been invited to participate in the original study, a sensitivity analysis was carried out by excluding them.

Potential differences between students who registered after following the first motivational intervention and those who registered after following the second one were looked for by applying a *t* test on the 10-point BLS score and by comparing attrition at each step. No weighting was used to compute the 10-point BLS knowledge score. A *t* test was performed to look for a difference between this score and interest in following BLS training.

A *t* test was also used to investigate whether there was a statistically significant difference between postcourse confidence and enrollment in the first responder network.

Given the presence of cells with very limited numbers (<5), Fischer tests were applied to analyze the factors influencing self-confidence and the desire to join the Save a Life first responder network.

Results

The 2021 - 2022 academic year included a total of 674 first-year students at UGFM in human and dental medicine and in biomedical sciences. The proportion of students who had registered as first responders after the first motivational intervention was 2.5% (19/674), significantly less than after Taramarcaz et al's [16] implementation study (48/529, 9.1%; *P*<.001). The second motivational intervention led to the enrollment of 7 more students (26/674, 3.9%). This figure is still significantly lower than that observed in Taramarcaz et al's study (*P*<.001) [16]. Even after excluding biomedical students from the analysis, the figure remained significantly lower (25/600, 4.2%) than in Taramarcaz et al's [16] study (*P*=.001).

A total of 502 (74.5%) students followed the link directing them to the study platform, of whom 447 (66.3%) students completed the screening questionnaire. Only 133 (19.7%) students registered on the platform and 76 (11.3%) students received a BLS-AED certificate at the end of the learning program. In the postcourse questionnaire, 68.4% (52/76) of students who obtained the certificate indicated a desire to join the network of first responders, but only 34.2% (26/76) of students followed through. Figure 1 shows participation at each step of the study.

There was a statistically significant relationship between prior BLS knowledge and e-learning completion (*P*=.007), practical session attendance (*P*<.001), and obtention of a BLS certificate (*P*=.003). Conversely, there was no statistically significant relationship between prior BLS-AED knowledge and enrollment in the Save a Life network (*P*=.05) or interest in the program (*P*=.94).

Students' confidence in their ability to initiate BLS maneuvers was significantly increased after following the learning path (*P*<.001, Multimedia Appendix 2). There was no statistical link between postcourse confidence and registration on the Save a Life platform (*P*=.09).

Postcourse satisfaction was 100% (76/76), as was the probability that students who had completed the learning path would recommend it to other students.

Figure 2 shows that a better understanding of health issues, a feeling of mastery of the subject, and an improvement in knowledge regarding resuscitation all contributed to promoting participant confidence.

Stress and fear of doing wrong were the 2 main factors reported as limiting one's confidence in performing BLS maneuvers (Figure 3).

Four factors promoting student willingness to register on the Save a Life platform were identified: feeling able to perform cardiopulmonary resuscitation, the possibility of making a difference, the stakes, and the desire to help (Figure 4).

Stress was the main factor preventing participants from registering as first responders (Figure 5).

Figure 1. Study flowchart. AED: automatic external defibrillator; BLS: basic life support.

Figure 2. Factors promoting student confidence in their ability to perform basic life support (BLS) maneuvers.

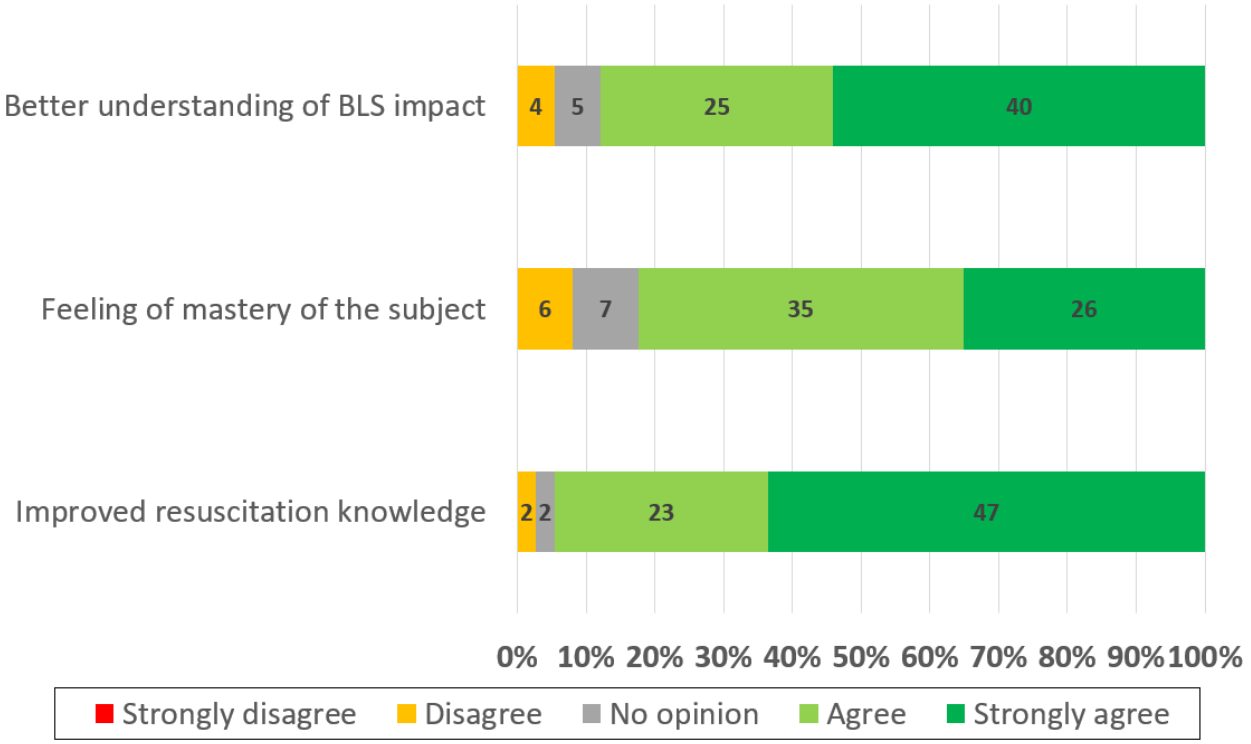


Figure 3. Factors limiting student confidence in performing basic life support maneuvers.

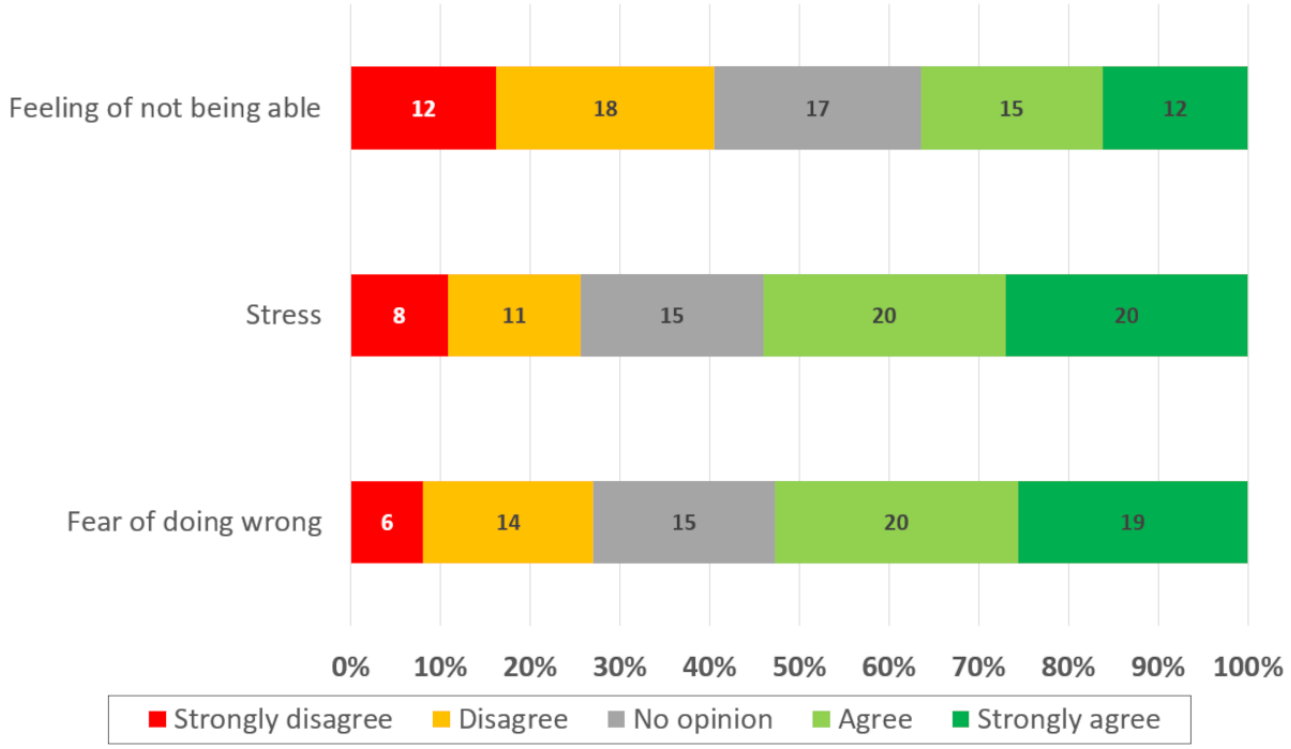


Figure 4. Factors promoting student willingness to register on the Save a Life platform. CPR: cardiopulmonary resuscitation.

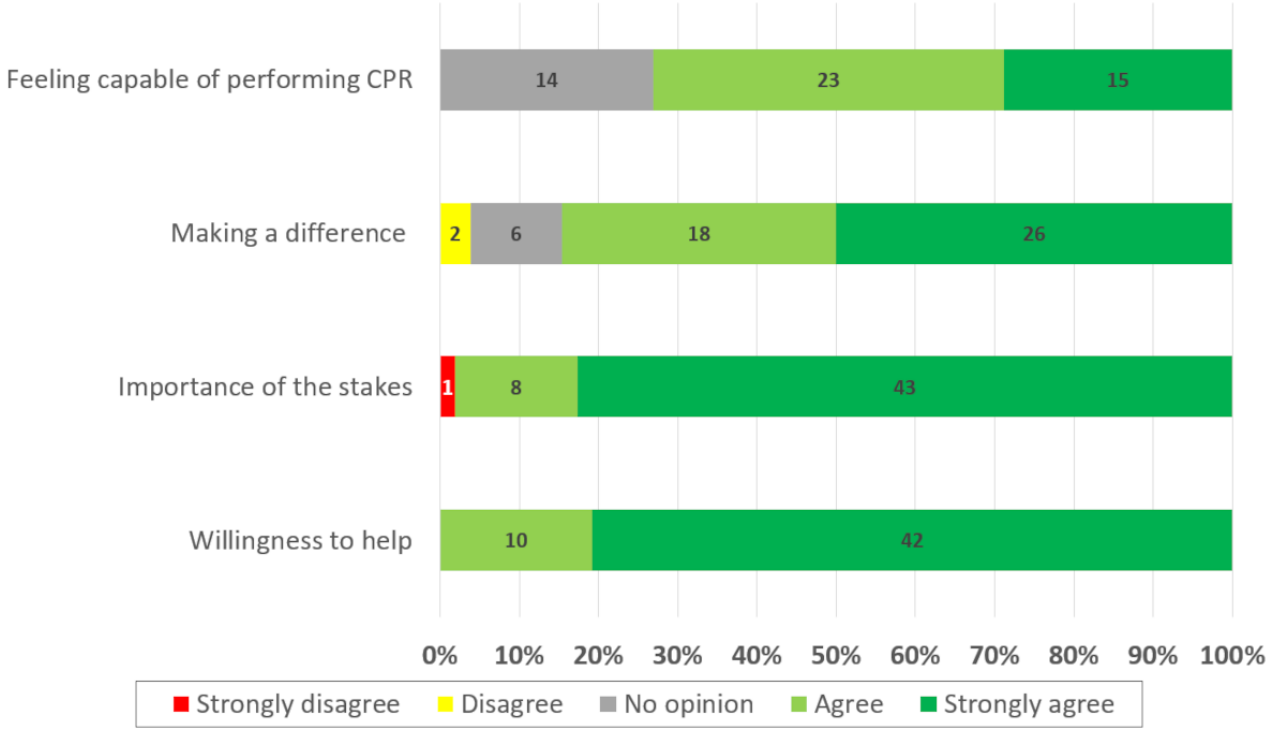
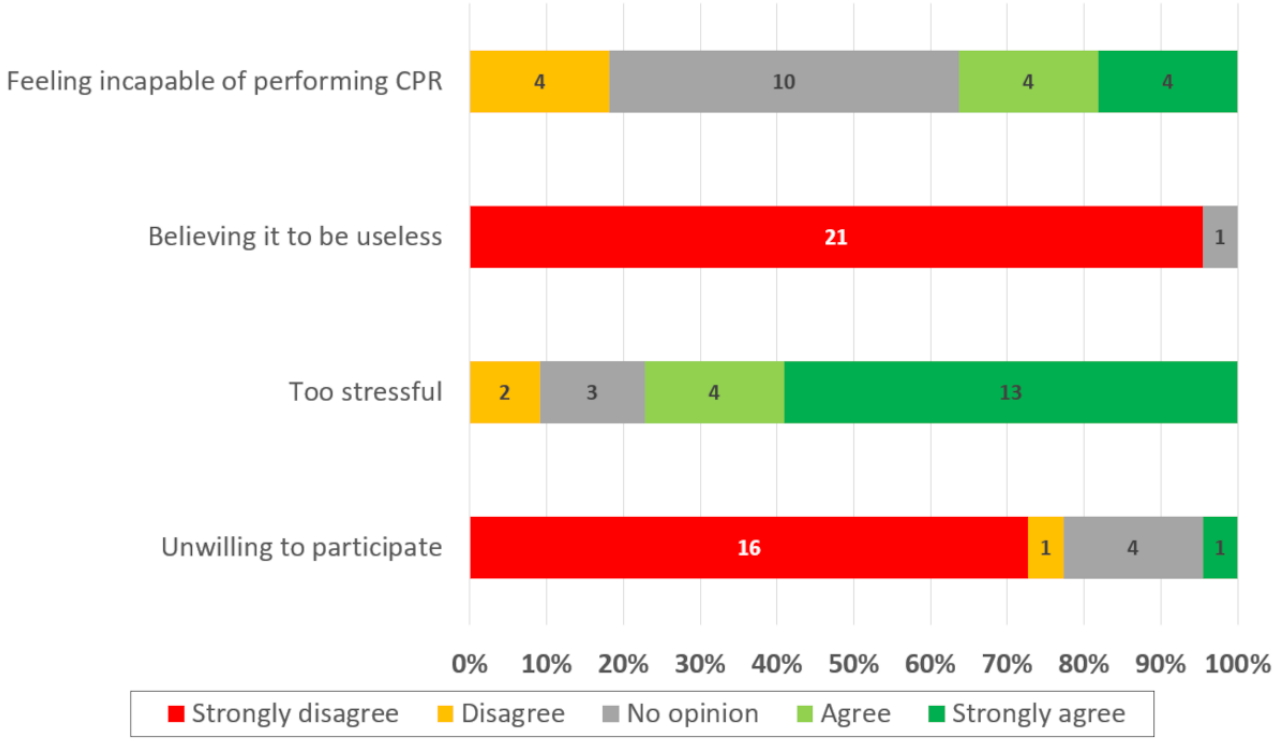


Figure 5. Factors limiting student willingness to register on the Save a Life platform. CPR: cardiopulmonary resuscitation.



Discussion

Main Considerations

Despite the modifications carried out according to the hypotheses outlined in the original study [16], and even after a second motivational intervention, recruitment was markedly lower than expected: indeed, in the original implementation study [16], the proportion of people who had registered on the platform at the end of the project was more than 2 times higher. The target set in the initial protocol was to recruit 10% of first-year students, a goal that has not yet been achieved.

These results deserve to be analyzed from a motivational point of view. The original study took place during the COVID-19 pandemic period, when student dynamics were probably

different, and students may have been more inclined to participate in a presential activity, due to the fact that they had no choice but to spend their first year remotely. The COVID period was a major awareness and health involvement on the part of medical students. Their involvement in medical tasks having a perceptible impact on the future of patients affected by the pandemic undoubtedly positively influenced their motivational reinforcement. This paradox probably partially explains the low recruitment observed in this study, carried out outside the pandemic context. Other factors may also have influenced student motivation during the COVID pandemic: during this period, the health care system was highly regarded by the population, and the project may have given the students a sense of belonging [21]. The feeling of being useless in the face of what was happening and the desire to help may also have been stronger at this time [21]. Conversely, the end of the restrictions may have decreased their motivation to take part in such a project, and students may have been keen to resume many of the activities they had been deprived of [22].

The profile of the teacher endorsing the motivational intervention may also have played a role since teachers can have a significant influence on their students [23]. Since most medical students are interested in the clinical field, any advice, opinion, or encouragement given by a clinician could have a particularly important influence on students [24]. Clinicians can also share their interest and experience in a subject [25], and their support can foster student interest in a particular field [24]. Indeed, human beings strive to feel connected to those they admire, and the sense of belonging that a prestigious clinician radiates can influence student motivation [24]. In addition, first-year students are more motivated by success, prestige, and money, compared with the upper years, who are more focused on the personal gratification of their activity [24,26]. Moreover, student motivation fluctuates over the years, both qualitatively and quantitatively. Understanding its evolution can help encourage students to enjoy their learning and possibly improve their performance [27].

According to the theory of self-determination, there are several types of motivation, depending on what influences it and what goals it aims to achieve: intrinsic motivation, extrinsic motivation, and amotivation [26]. Intrinsic motivation is linked to personal interest in or pleasure inherent to the activity. Extrinsic motivation aims at a goal, a consequence separable from the subject, such as a reward or the absence of inconvenience [26]. Extrinsic motivation can be described as a continuum through which a process of internalization takes place, finally resulting in integrating action towards self-determination [26]. In the educational environment, motivation can be seen as having 3 determinants: the perception of the value of an activity, its skill, and its controllability [28].

A clinician's valorization of the abilities and importance that each student can have in the health care system at their own level can influence the perception of their abilities, and their involvement and motivation [23]. A clinician's speech on public health issues can have a greater impact and radiate a positive perception of the values involved [24,28].

Despite these different aspects motivating first-year students to participate in an optional learning program can still be difficult since it does not bring them any short-term benefits, in this case passing their exams. According to Dweck, students pursue learning goals as well as performance goals [29]. In the short term, when the risk of success is low, students would restrict themselves to the performance goal and neglect activities they consider ineffective for success [28,30].

A considerable proportion of students did not continue with the learning path after completing the questionnaire assessing their BLS knowledge. These students' scores were lower than average, and their lack of knowledge may have had an impact on their perception of their skills for future activities, decreasing their self-confidence and motivation to continue their learning program [28]. This could be addressed by introducing BLS courses at school since, in Geneva, most schoolchildren receive only little, if any, first aid training before attending courses mandatory to obtain a driving license. Furthermore, training schoolchildren has been shown to improve OHCA outcomes [31]. Another option could be to remove this questionnaire from future studies to avoid any attrition linked to its administration.

Once enrolled in the learning program, students follow a certifying course, but registration on the Save a Life platform remains optional, and students therefore need further motivation to enlist as first responders. Participants agreed that perfecting their knowledge, mastering the subject, and understanding the health issues linked to early resuscitation all improved their self-confidence. This is in line with Viau and Louis's [28] opinion, that is, that the perception of the value of an activity and of one's own skills influence motivation.

Understanding the social impact of a first responder network could enable potential participants to internalize the values involved, and, according to the self-determination theory, increase motivation [26]. Thus, the societal impact of the Save a Life project could be further highlighted in future motivational interventions. This could improve recruitment since respondents unanimously agreed that the desire to help influenced their probability of registering as first responders.

Students reported that the main factors limiting their willingness to register as first responders were stress and the fear of making a mistake. Stress goes against the feeling of controllability of a situation, which is essential to self-confidence [28]. Addressing this issue will require further exploration, but a first step could be to point out the low risk of harm to the patient when practicing BLS maneuvers [14,15] and the clear benefits of early resuscitation early in the presentation [8-10].

Aspects modulating self-confidence need to be highlighted in future presentations to students, but also during the training program, to encourage these students as much as possible to join the Save a Life network. Their abilities and knowledge should be encouraged, and the efforts and gains they can make in the management of OHCA should be recognized. Any fears or doubts they may have must also be addressed during the learning path, and the effect of these motivational enhancements will need to be assessed in the next few years.

Limitations

Since the very low participation rate could not be anticipated, the design of this study had to be adapted. Even though the second motivational intervention was endorsed by a clinician while the first was endorsed by a specialist in basic medical science, the effect of each specific intervention could not be assessed given the design of this study. A randomized controlled trial could be considered to explore the effects of endorsement by either type of specialist. In addition, the motivational interventions themselves were also different, and the effect of specifically designed and theory-based motivational interventions would also deserve to be determined. Finally, the impact of specific factors on motivation was only assessed among the students who had followed the learning path, thereby leading to a selection bias. Therefore, participatory research should be considered to help identify better recruitment strategies, and focus groups held to gather a more thorough and less biased understanding of students' motivation and barriers to participation.

Conclusions

Contrary to expectations, an earlier presentation during the academic year outside the COVID restriction period did not increase the recruitment of medical students as first responders, which was more than 2 times lower than in the implementation study even after further motivational interventions. A thorough quantitative and qualitative exploration of motivational factors should be carried out to determine potential ways of improving the recruitment of first-year medical students as first responders.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Web-based platform.
[DOCX File, 15 KB - mededu_v11i1e63018_app1.docx]

Multimedia Appendix 2
Evolution of self-confidence in practice basic life support maneuver.
[PNG File, 25 KB - mededu_v11i1e63018_app2.png]

References

1. Panchal AR, Bartos JA, Cabañas JG, et al. Part 3: adult basic and advanced life support: 2020 American heart association guidelines for cardiopulmonary resuscitation and emergency cardiovascular care. Circulation 2020 Oct 20;142(16_suppl_2):S366-S468. [doi: 10.1161/CIR.0000000000000916] [Medline: 33081529]
2. Semeraro F, Greif R, Böttiger BW, et al. European resuscitation council guidelines 2021: systems saving lives. Resuscitation 2021 Apr;161:80-97. [doi: 10.1016/j.resuscitation.2021.02.008] [Medline: 33773834]
3. Olasveengen TM, Mancini ME, Perkins GD, et al. Adult basic life support: 2020 international consensus on cardiopulmonary resuscitation and emergency cardiovascular care science with treatment recommendations. Circulation 2020 Oct 20;142(16_suppl_1):S41-S91. [doi: 10.1161/CIR.0000000000000892] [Medline: 33084391]
4. Weisfeldt ML, Sitlani CM, Ornato JP, et al. Survival after application of automatic external defibrillators before arrival of the emergency medical system. J Am Coll Cardiol 2010 Apr;55(16):1713-1720. [doi: 10.1016/j.jacc.2009.11.077]
5. Arrêts cardio-respiratoires préhospitaliers: revue de la littérature et évolution du pronostic à genève entre 2009 et 2012 [Pre-hospital cardiopulmonary arrest: literature review and prognosis trends in Geneva between 2009 and 2012]. : Université de Genève; 2018 URL: https://access.archive-ouverte.unige.ch/access/metadata/9b59894e-ebb0-46cb-9c99-69897632b06b/download [accessed 2025-05-05]
6. Perkins GD, Graesner JT, Semeraro F, et al. European resuscitation council guidelines 2021: executive summary. Resuscitation 2021 Apr;161:1-60. [doi: 10.1016/j.resuscitation.2021.02.003] [Medline: 33773824]
7. Rapport d'activité [annual report]. : Save-a-Life; 2021 URL: https://www.save-a-life.ch/wp-content/uploads/2022/04/Save-a-Life-Rapport-dactivites-2021.pdf [accessed 2023-03-21]
8. Stiell IG, Wells GA, DeMaio VJ, et al. Modifiable factors associated with improved cardiac arrest survival in a multicenter basic life support/defibrillation system: OPALS study phase I results. Ann Emerg Med 1999 Jan;33(1):44-50. [doi: 10.1016/s0196-0644(99)70415-4] [Medline: 9867885]
9. Fordyce CB, Hansen CM, Kragholm K, et al. Association of public health initiatives with outcomes for out-of-hospital cardiac arrest at home and in public locations. JAMA Cardiol 2017 Nov 1;2(11):1226-1235. [doi: 10.1001/jamacardio.2017.3471] [Medline: 28979980]
10. Dami F, Fuchs V, Berthoz V, Carron PN. Régulation médicale: mise au point 2018 et développements futurs [medical dispatch: 2018 update and future developments]. Ann Fr Med Urgence 2018 Nov;8(6):376-382. [doi: 10.3166/afmu-2018-0089]

11. Iwami T, Kitamura T, Kiyohara K, Kawamura T. Dissemination of chest compression-only cardiopulmonary resuscitation and survival after out-of-hospital cardiac arrest. Circulation 2015 Aug 4;132(5):415-422. [doi: 10.1161/CIRCULATIONAHA.114.014905] [Medline: 26048093]
12. Yasunaga H, Horiguchi H, Tanabe S, et al. Collaborative effects of bystander-initiated cardiopulmonary resuscitation and prehospital advanced cardiac life support by physicians on survival of out-of-hospital cardiac arrest: a nationwide population-based observational study. Crit Care 2010;14(6):R199. [doi: 10.1186/cc9319] [Medline: 21050434]
13. Regard S, Rosa D, Suppan M, et al. Evolution of bystander intention to perform resuscitation since last training: web-based survey. JMIR Form Res 2020 Nov 30;4(11):e24798. [doi: 10.2196/24798] [Medline: 33252342]
14. Haley KB, Lerner EB, Pirrallo RG, Croft H, Johnson A, Uihlein M. The frequency and consequences of cardiopulmonary resuscitation performed by bystanders on patients who are not in cardiac arrest. Prehosp Emerg Care 2011;15(2):282-287. [doi: 10.3109/10903127.2010.541981] [Medline: 21250928]
15. Cassan P. Recommendations on basic cardiopulmonary resuscitation: main points. J Eur Urgences Reanim 2021;33(2):65-72. [doi: 10.1016/j.jeurea.2021.05.003]
16. Taramarcaz V, Herren T, Golay E, et al. A short intervention and an interactive e-learning module to motivate medical and dental students to enlist as first responders: implementation study. J Med Internet Res 2022 May 18;24(5):e38508. [doi: 10.2196/38508] [Medline: 35583927]
17. Eysenbach G. Improving the quality of web surveys: the checklist for reporting results of internet e-surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
18. Complete guide to GDPR compliance. GDPR.eu. URL: https://gdpr.eu/ [accessed 2022-03-21]
19. Suppan L, Herren T, Taramarcaz V, et al. A short intervention followed by an interactive e-learning module to motivate medical students to enlist as first responders: protocol for A prospective implementation study. JMIR Res Protoc 2020 Nov 6;9(11):e24664 [FREE Full text] [doi: 10.2196/24664] [Medline: 33155574]
20. Eysenbach G. The law of attrition. J Med Internet Res 2005 Mar 31;7(1):e11. [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
21. Zhang R, Pei J, Wang Y, et al. COVID-19 outbreak improves attractiveness of medical careers in Chinese senior high school students. BMC Med Educ 2022 Apr 4;22(1):241. [doi: 10.1186/s12909-022-03309-7] [Medline: 35379234]
22. Barrense-Dias Y, Urben S, Chok L, Schechter D, Suris JC. Exploration du vécu de la pandémie et du confinement dus à la COVID-19 des adolescent·es et des parents [exploration the COVID-19 pandemic and lockdown experience among teenagers and parents]. : Centre universitaire de médecine générale et santé publique (Unisanté); 2021. [doi: 10.16908/ISSN.1660-7104/320]
23. Guillemette F, Leblanc C. Favoriser l'expression de la motivation chez les étudiants [encouraging motivation among students]. Université du Québec. 2013. URL: https://pedagogie.uquebec.ca/le-tableau/favoriser-lexpression-de-la-motivation-chez-les-etudiants [accessed 2025-04-16]
24. Kusurkar RA, Ten Cate TJ, van Asperen M, Croiset G. Motivation as an independent and a dependent variable in medical education: a review of the literature. Med Teach 2011;33(5):e242-e262. [doi: 10.3109/0142159X.2011.558539] [Medline: 21517676]
25. Viau R. Des conditions à respecter pour susciter la motivation des élèves [conditions to be respected to motivate pupils]. Correspondance 2000 [FREE Full text]
26. Ryan RM, Deci EL. Intrinsic and extrinsic motivations: classic definitions and new directions. Contemp Educ Psychol 2000 Jan;25(1):54-67. [doi: 10.1006/ceps.1999.1020] [Medline: 10620381]
27. Orsini C, Binnie VI, Wilson SL. Determinants and outcomes of motivation in health professions education: a systematic review based on self-determination theory. J Educ Eval Health Prof 2016 May;13:19. [doi: 10.3352/jeehp.2016.13.19]
28. Viau R, Louis R. Vers une meilleure compréhension de la dynamique motivationnelle des étudiants en contexte scolaire [towards a better understanding of the motivational dynamics of students in a school context]. Can J Educ 1997;22(2):144. [doi: 10.2307/1585904]
29. Dweck CS. Motivational processes affecting learning. American Psychologist 1986;41(10):1040-1048. [doi: 10.1037//0003-066X.41.10.1040]
30. Fundations for a Psychology of Education: Routledge; 1989.
31. Böttiger BW, Semeraro F, Wingen S. "Kids Save Lives": educating schoolchildren in cardiopulmonary resuscitation is a civic duty that needs support for implementation. J Am Heart Assoc 2017 Mar 14;6(3):e005738. [doi: 10.1161/JAHA.117.005738] [Medline: 28292747]

Abbreviations

AED: automatic external defibrillator
BLS: basic life support
OHCA: out-of-hospital cardiac arrest
UGFM: University of Geneva Faculty of Medicine

Edited by A Bahattab; submitted 07.06.24; peer-reviewed by GJ Noordergraaf, J Llego; revised version received 14.03.25; accepted 15.03.25; published 24.04.25.

Please cite as:
Schnetzler N, Taramarcaz V, Herren T, Golay E, Regard S, Mach F, Nasution A, Larribau R, Suppan M, Schiffer E, Suppan L
Recruiting Medical, Dental, and Biomedical Students as First Responders in the Immediate Aftermath of the COVID-19 Pandemic: Prospective Follow-Up Study
JMIR Med Educ 2025;11:e63018
URL: https://mededu.jmir.org/2025/1/e63018
doi:10.2196/63018

Comparison of Learning Outcomes Among Medical Students in Thailand to Determine the Right Time to Teach Forensic Medicine: Retrospective Study

Ubon Chudoung, BSc; Wilaipon Saengon, BBA; Vichan Peonim, MD; Wisarn Worasuwannarak, LLB, MSc, MD

Department of Pathology, Faculty of Medicine Ramathibodi Hospital, Mahidol University, 270 Rama VI Road, Thung Phaya Thai, Bangkok, Thailand

Corresponding Author:
Wisarn Worasuwannarak, LLB, MSc, MD
Department of Pathology, Faculty of Medicine Ramathibodi Hospital, Mahidol University, 270 Rama VI Road, Thung Phaya Thai, Bangkok, Thailand

Abstract

Background: Forensic medicine requires background medical knowledge and the ability to apply it to legal cases. Medical students have different levels of medical knowledge and are therefore likely to perform differently when learning forensic medicine. However, different medical curricula in Thailand deliver forensic medicine courses at different stages of medical study; most curricula deliver these courses in the clinical years, while others offer them in the preclinical years. This raises questions about the differences in learning effectiveness.

Objective: We aimed to compare the learning outcomes of medical students in curricula that either teach forensic medicine at the clinical level or teach it at the preclinical level.

Methods: This was a 5-year retrospective study that compared multiple-choice question (MCQ) scores in a forensic medicine course for fifth- and third-year medical students. The fifth-year students' program was different from that of the third-year students, but both programs were offered by Mahidol University. The students were taught forensic medicine by the same instructors, used similar content, and were evaluated via examinations of similar difficulty. Of the 1063 medical students included in this study, 782 were fifth-year clinical students, and 281 were third-year preclinical students.

Results: The average scores of the fifth- and third-year medical students were 76.09% (SD 6.75%) and 62.94% (SD 8.33%), respectively. The difference was statistically significant (Kruskal-Wallis test: $P<.001$). Additionally, the average score of fifth-year medical students was significantly higher than that of third-year students in every academic year (all P values were <.001).

Conclusions: Teaching forensic medicine during the preclinical years may be too early, and preclinical students may not understand the clinical content sufficiently. Attention should be paid to ensuring that students have the adequate clinical background before teaching subjects that require clinical applications, especially in forensic medicine.

(JMIR Med Educ 2025;11:e57634) doi:10.2196/57634

KEYWORDS

multiple-choice question; MCQ; forensic medicine; preclinic; clinic; medical student

Introduction

Forensic medicine is a crucial field that intersects with the legal system. It involves the collection, analysis, interpretation, and presentation of evidence in legal cases [1]. Forensic medicine plays an essential role in assisting courts with making correct decisions by providing reliable and timely information. It also plays a critical role in protecting peoples' rights by ensuring that their legal, civil, and human rights are upheld throughout the legal process [2]. Furthermore, studying forensic medicine is important for medical students in different countries, as they are equipped with the necessary knowledge and skills to accurately assess and document injuries and provide expert opinions on causes of death and other relevant medical information that may have legal implications [3-6].

This subject is included among the professional subjects that every Thai medical student must study to comply with the Criminal Procedure Code of Thailand, which requires physicians working in public hospitals to be able to perform postmortem inquests with police in cases where no forensic physician is available [7]. The Medical Council of Thailand has included forensic medicine as a mandatory subject in every doctor of medicine program.

The doctor of medicine programs in Thailand are 6-year programs conducted after graduating from high school. They are generally divided into 3 years at the preclinical level (first through third year) and another 3 years at the clinical level (fourth through sixth year). The teaching of each university's curriculum differs in detail depending on various factors, such as the number of students, number of teachers, location, and

service characteristics. Forensic medicine is subject to these differences.

Studying forensic medicine involves dealing with dead bodies, crime scenes, and traumatic injuries that can be emotionally and mentally stressful for some students [8]. A study from Saudi Arabia revealed that medical students have poor attitudes toward and awareness of the importance of forensic medicine [9]. Additionally, forensic medicine courses cover a wide range of topics, such as anatomy, physiology, pathology, toxicology, psychology, and jurisprudence, which can be difficult to master and integrate [10,11].

Students with different levels of medical knowledge may experience different forensic medicine course outcomes. In Thailand, most medical curricula are currently designed to teach forensic medicine to medical students at the clinical level (fifth year) [12-14]. However, some curricula have been designed to teach forensic medicine to medical students at the preclinical or early clinical level (third or fourth year) [15]. There are no clear guidelines regarding the level of students who should be taught forensic medicine.

This study aims to compare the learning outcomes of medical students in a curriculum that teaches forensic medicine at the clinical level and those of medical students in a curriculum that teaches forensic medicine at the preclinical level.

Methods

Study Design

This retrospective study was conducted to compare multiple-choice question (MCQ) scores of fifth- and third-year medical students from two medical curricula that teach forensic medicine. Both groups of students studied forensic medicine with the same instructors, used similar content, and were assessed via MCQ examinations with similar difficulty levels. The scores indicated the participants' learning outcomes.

Setting and Participants

Samples

Our samples included (1) medical students in a curriculum that teaches forensic medicine at the clinical level (fifth year) through the Doctor of Medicine Program at Ramathibodi Hospital, Mahidol University (782 students), and (2) medical students in a curriculum that teaches forensic medicine as the last subject at the preclinical level (third year) through the Joint Program for Producing More Doctors for Rural Areas, Mahidol University (281 students).

Sample Size Calculation

The sample size was designed to compare 2-sided differences in the MCQ percentage scores between third- and fifth-year medical students studying forensic medicine. The null hypothesis (H_0) was that the MCQ percentage scores between third- and fifth-year medical students would not be significantly different. The alternative hypothesis (H_1) was that the MCQ percentage scores between third- and fifth-year medical students would be significantly different.

We calculated the sample size according to a 5% type 1 error (α) and an 80% study power ($1 - \beta$). The significant difference ($\mu 1 - \mu 2$) and SD (σ) were set at 10 and 11, respectively, based on MCQ score data for medical students who studied forensic medicine from 2010 to 2014. The required sample size was 38 (19 participants in each group; Multimedia Appendix 1) [16]. However, this study included more participants than the calculated sample size.

Intervention

Teaching Method

Both groups of medical students received on-site theoretical lectures before completing the MCQs. The content included basic knowledge of forensic pathology (including postmortem inquest, identification, time of death estimation, crime scene investigation, unnatural death, and sudden unexpected death), clinical forensic medicine (including patients who are wounded, child abuse, sexual assault, and forensic psychiatry), forensic evidence, forensic genetics, forensic toxicology, and medical law and ethics. Third-year medical students studied for 30 hours. Fifth-year medical students studied for 15 hours, using similar content that was more concise, and had the opportunity to visit a court for 3 hours. Neither group had the opportunity to attend crime scene investigations or autopsies (which they would attend later). This teaching method was performed regularly, and the authors did not intervene with any of the participants.

MCQ Examinations

For examinations, all teaching staff (4 staff members) created 5-option MCQs with a single best answer according to the topics they taught, including basic knowledge of forensic pathology (40% of questions), clinical forensic medicine (30% of questions), forensic evidence (5% of questions), forensic genetics (5% of questions), forensic toxicology (5% of questions), and medical law and ethics (15% of questions). The tests were designed to ensure that medical students are able to perform basic postmortem inquests, examine various types of forensic patients, produce accurate medicolegal reports, have basic knowledge of law and ethics, and understand the process of testifying in court. The MCQ examinations were structured via a balanced approach for cognitive function, allocating approximately 25% of the examination to knowledge, 30% to comprehension, 25% to application, and 20% to analysis level, according to the Bloom taxonomy. This distribution is maintained consistently from year to year. The examination was intended to have a moderate level of difficulty. Third-year medical students completed a 100-question examination in 2 hours, and fifth-year medical students completed an 80-question examination in 1.5 hours. Based on an analysis of the examination, most of the items had a difficulty level (p) in the range of 0.4 to 0.7 and a discriminatory power (r) in the range of 0.1 to 0.5. Internal consistency reliability (Kuder-Richardson Formula 20) was in the range of 0.6 to 0.7.

Data Collection

In this study, the data were collected retrospectively for 5 years, from academic years 2010 through 2014.

Statistical Analysis

For the comparison between the two groups, we used the means and SDs of the MCQ scores to test this study's hypothesis that the learning outcome is different between third- and fifth-year students. Kruskal-Wallis and Mann-Whitney *U* tests were used for continuous variables with normal and nonnormal distributions, respectively [17]. The significance level was set at 5% ($P<.05$). The program used for data analysis was SPSS software (version 26; IBM Corp).

Ethical Considerations

This study was approved by the Ethical Clearance Committee on Human Rights Related to Research Involving Human Subjects, Faculty of Medicine Ramathibodi Hospital, Mahidol University (MURA 2015/213). The need for informed consent was waived by the Ethical Clearance Committee on Human Rights Related to Research Involving Human Subjects, Faculty of Medicine Ramathibodi Hospital, Mahidol University. Data were collected by using an anonymous method—assigning numbers to all participants instead of names. No compensation was provided to participants.

Results

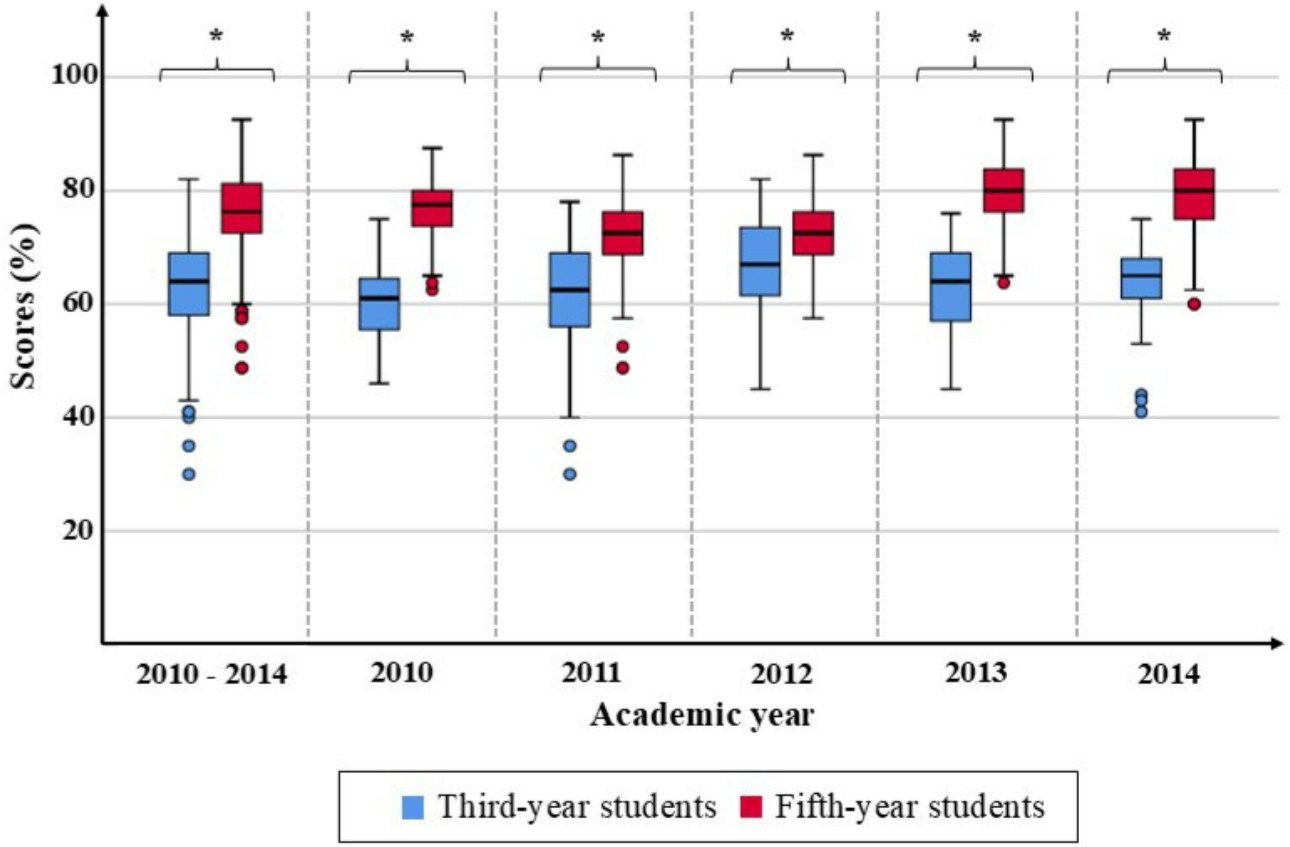
From the collection of MCQ scores of medical students from academic years 2010 to 2014 who were taught forensic medicine, the scores of 1063 students were used in this study. The scores were divided into scores of third-year medical students (n=281) and scores of fifth-year medical students (n=728), as shown in Table 1.

Table . Number of students in each academic year (N=1063).

Students	Academic year					Total
	2010	2011	2012	2013	2014	
Third-year students, n (%)						
Male	30 (2.8)	35 (3.3)	34 (3.2)	33 (3.1)	33 (3.1)	165 (15.5)
Female	21 (2)	23 (2.2)	21 (2)	23 (2.2)	28 (2.6)	116 (10.9)
Fifth-year students, n (%)						
Male	81 (7.6)	94 (8.8)	87 (8.2)	94 (8.8)	101 (9.5)	457 (43)
Female	53 (5)	64 (6)	71 (6.7)	64 (6)	73 (6.9)	325 (30.6)
Total, n (%)	185 (17.4)	216 (20.3)	213 (20)	214 (20.1)	235 (22.1)	1063 (100)

When comparing students' scores, it was found that fifth-year medical students had an average score of 76.09% (SD 6.75%), which was higher than that of third-year medical students (mean 62.94%, SD 8.33%). The difference was statistically significant (Kruskal-Wallis test: $P<.001$). In addition, when comparing the average scores in each academic year, it was found that the average score of fifth-year medical students was significantly higher than that of third-year students in every academic year (Mann-Whitney *U* test: all *P* values were <.001), as shown in Figure 1.

Figure 1. Comparing scores of third-year and fifth-year students. *Statistically significant (Mann-Whitney *U* test: *P*<.001).



Discussion

Principal Findings

According to this study's findings, fifth-year medical students achieved significantly higher marks on MCQs than those achieved by third-year medical students, despite the latter having more opportunities to prepare and take examinations due to their longer duration of study. The fact that the two groups of medical students had different scores may be due to their different levels of basic knowledge of medicine. Fifth-year medical students study basic clinical subjects. Therefore, they may have more comprehensive and complete basic medical knowledge and may be able to apply it to prove facts about legal cases better than third-year medical students who have not completed their basic clinical subjects. These results are consistent with a study in Italy, which showed that students' awareness of forensic medicine improved in the fifth or sixth year of a forensic medicine course [18].

When analyzing the data by academic year, fifth-year medical students still had higher MCQ scores than those of third-year medical students, with statistical significance for each academic year. These data show that the difference in MCQ scores was unlikely due to different medical students from year to year.

In forensic medicine, students should have the opportunity to learn about real cases, including examinations of legal patients, autopsies, and crime scene examinations. This would improve students' understanding of applying and ability to apply medical knowledge to legal applications. A study in India revealed that a court visit in a real scenario was the method that generated the most interest, and student-led objective tutorials comprised the method that best facilitated enhanced learning; the "model answer" method was also found to be an effective method for teaching forensic medicine [19]. Furthermore, a study in Mexico showed that crime scene investigation laboratory visits are an innovative method of learning that may help broaden medical students' perspectives on forensic sciences and help them understand the multidisciplinary processes of crime investigation [20].

By integrating forensic medicine into the medical curriculum, students also gain a deeper awareness of the complexities surrounding child abuse. Training on this topic not only enhances students' diagnostic skills but also instills a sense of responsibility to act in the best interests of the child, ensuring that they are better prepared to contribute to the early detection, intervention, and prevention of child abuse in their future careers [21].

This study used only MCQ scores from theoretical teaching, which may not measure all of the knowledge and skills of students. Although MCQs can test higher-order thinking, they are typically limited to the "application" and "analysis" levels of the Bloom taxonomy [22]. The use of MCQs is often driven by practical concerns, such as large class sizes, rather than pedagogical reasons. Although MCQs have their place, they may restrict the scope of teaching and require careful consideration to align with higher-order learning objectives [23]. Thus, a combination of test methods can be used. A study from Nepal found that objective structured practical examination is an acceptable and well-received method for medical students [24].

Integrating some content of clinical subjects via vertical integration for preclinical medical students may help to enhance their knowledge and understanding of forensic medicine. A previous study on learning environments found that undergraduate medical students from Egypt who received integrated curriculum teaching experienced a more positive learning environment [6]. Further, a similar study from Malaysia showed that integrated teaching positively affects medical students' learning environment [25]. These studies are also consistent with guidelines from the Medical Council of Thailand for developing medical curricula in Thailand, which support horizontal and vertical integration teaching [26]; that is, clinical teachers should teach about clinical experiences from the beginning and integrate basic medical science knowledge into the clinical years.

Limitations

A limitation of this study was its retrospective design; that is, past MCQ scores were analyzed to evaluate the medical curricula at the time of writing. No systematic interventions were conducted to test the hypothesis. In addition, this study used only MCQ scores; therefore, it may not include every learning outcome of the forensic medicine course.

Recommendations

Students' basic medical knowledge should be considered when teaching and learning subjects that require clinical application, especially in forensic medicine, which applies medical knowledge to law. Teaching such subjects to preclinical-level students, whose medical knowledge remains incomplete, may be too ambitious. It may be appropriate to integrate introductory content from clinical subjects to increase knowledge and understanding. In comparison, clinical-level students with complete basic knowledge may be more suitable for such clinical subjects.

Conclusion

Forensic medicine requires basic medical knowledge and the ability to apply this knowledge in legal cases. Students' basic medical knowledge should be considered when planning the teaching and learning of this subject. Teaching forensic medicine in the preclinical years may be too early, and doing so may result in students being unable to sufficiently understand the clinical content.

Acknowledgments

We would like to thank our colleagues, including the staff and officers of the Division of Forensic Medicine, for the support they have given us.

Data Availability

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

Authors' Contributions

Conceptualization: UC
Data curation: UC
Formal analysis: UC
Investigation: UC
Methodology: WW
Project administration: WW
Supervision: VP, WW
Validation: WW
Visualization: WW
Writing – original draft: UC
Writing – review & editing: UC, WS, VP, WW

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sample size calculation.
[DOCX File, 15 KB - mededu_v11i1e57634_app1.docx]

References

1. Shepherd R. Simpson's Forensic Medicine, 12th edition: Arnold; 2003.
2. Payne-James J. History and development of clinical forensic medicine. In: Clinical Forensic Medicine: A Physician's Guide: Humana Press; 2005:1-36. [doi: 10.1385/1-59259-913-3:001]

3. Tóth D, Petrus K, Heckmann V, Simon G, Poór VS. Application of photogrammetry in forensic pathology education of medical students in response to COVID-19. J Forensic Sci 2021 Jul;66(4):1533-1537. [doi: 10.1111/1556-4029.14709] [Medline: 33764562]
4. Sosa-Reyes AM, Villavicencio-Queijeiro A, Suzuri-Hernández LJ. Interdisciplinary approaches to the teaching of forensic science in the Forensic Science Undergraduate Program of the National Autonomous University of Mexico, before and after COVID-19. Sci Justice 2022 Nov;62(6):676-690. [doi: 10.1016/j.scijus.2022.08.006] [Medline: 36400489]
5. Marambe KN, Edussuriya DH, Somaratne PDIS, Piyaratne C. Do medical students who claim to be using deep learning strategies perform better at the Forensic Medicine examination? South-East Asian Journal of Medical Education 2009 Jun 30;3(1):25-30. [doi: 10.4038/seajme.v3i1.464]
6. Fayed MM, Abdo SA, Sharif AF. Preclinical and clinical medical students' perception of the learning environment: a reference to the Forensic Medicine and Clinical Toxicology course. Adv Med Educ Pract 2022 Apr 23;13:369-406. [doi: 10.2147/AMEP.S354446] [Medline: 35494484]
7. Thailand Criminal Procedure Code sections 148-156. Thai Law Forum. 2024 Jul 3. URL: http://www.thailawforum.com/thailand-criminal-procedure-code-sections-148-156/ [accessed 2025-01-28]
8. Papadodima SA, Sergentanis TN, Iliakis RG, Sotiropoulos KC, Spiliopoulou CA. Students who wish to specialize in forensic medicine vs. their fellow students: motivations, attitudes and reactions during autopsy practice. Adv Health Sci Educ Theory Pract 2008 Nov;13(4):535-546. [doi: 10.1007/s10459-007-9065-3] [Medline: 17486420]
9. Madadin MS. Assessment of knowledge about, attitudes toward, and awareness of a forensic medicine course among medical students at the University of Dammam. J Forensic Leg Med 2013 Nov;20(8):1108-1111. [doi: 10.1016/j.jflm.2013.10.003] [Medline: 24237831]
10. Wyatt JP, Squires T, Norfolk G, Payne-James J. Oxford Handbook of Forensic Medicine: Oxford University Press; 2011. URL: https://academic.oup.com/book/29998 [accessed 2026-01-17] [doi: 10.1093/med/9780199229949.001.0001]
11. Levinson SA, Muehlberger CW. An introductory course in legal medicine for medical students. Acad Med 1934 Sep;9(5):293-301. [doi: 10.1097/00001888-193409000-00006]
12. Faculty of Medicine Ramathibodi Hospital, Mahidol University. Doctor of Medicine Program, Revised Curriculum 2020 [Article in Thai]. Mahidol University. 2020. URL: https://www.rama.mahidol.ac.th/meded/sites/default/files/public/img2024/course/%E0%B8%A1%E0%B8%84%E0%B8%AD.2%20%E0%B8%AB%E0%B8%A5%E0%B8%B1%E0%B8%81%E0%B8%AA%E0%B8%B9%E0%B8%95%E0%B8%A3%20%E0%B8%9E%E0%B8%9A-63-pdf%20%E0%B8%A3%E0%B8%A7%E0%B8%A1-110863.pdf [accessed 2024-08-24]
13. Doctor of Medicine Program. Chiang Mai University. 2023. URL: https://www.cmu.ac.th/en/Faculty/course_detail/939ca112-78cc-444c-bfa4-5328c0484b7e [accessed 2024-08-12]
14. Chulalongkorn University. Doctor of Medicine Program (Revised Curriculum 2017) [Article in Thai]. Chulalongkorn University. 2017. URL: https://bhumibol-med.com/Media/media-2017-11-19-04-35-32.pdf [accessed 2024-08-12]
15. Clinical Medical Education Center, Sawanpracharak Hospital. Course Specification: Forensic Medicine 1 (NVEF411). Revised January 13, 2014. [Article in Thai]. Mahidol University. 2014. URL: https://www.rama.mahidol.ac.th/patho/sites/default/files/public/file/NVEF411.pdf [accessed 2024-08-12]
16. Noordzij M, Tripepi G, Dekker FW, Zoccali C, Tanck MW, Jager KJ. Sample size calculations: basic principles and common pitfalls. Nephrol Dial Transplant 2010 May;25(5):1388-1393. [doi: 10.1093/ndt/gfp732] [Medline: 20067907]
17. De Muth JE. Overview of biostatistics used in clinical research. Am J Health Syst Pharm 2009 Jan 1;66(1):70-81. [doi: 10.2146/ajhp070006] [Medline: 19106347]
18. Aulino G, Beccia F, Siodambro C, et al. An evaluation of Italian medical students attitudes and knowledge regarding forensic medicine. J Forensic Leg Med 2023 Feb;94:102484. [doi: 10.1016/j.jflm.2023.102484] [Medline: 36640545]
19. Gupta S, Parekh UN, Ganjiwale JD. Student's perception about innovative teaching learning practices in forensic medicine. J Forensic Leg Med 2017 Nov;52:137-142. [doi: 10.1016/j.jflm.2017.09.007] [Medline: 28922654]
20. Eraña-Rojas IE, López Cabrera MV, Ríos Barrientos E, Membrillo-Hernández J. A challenge based learning experience in forensic medicine. J Forensic Leg Med 2019 Nov;68:101873. [doi: 10.1016/j.jflm.2019.101873] [Medline: 31627125]
21. Aulino G, Beccia F, Rega M, et al. Child maltreatment and management of pediatric patients during COVID-19 pandemic: knowledge, awareness, and attitudes among students of medicine and surgery. A survey-based analysis. Front Public Health 2022 Sep 20;10:968286. [doi: 10.3389/fpubh.2022.968286] [Medline: 36203705]
22. Ehsan SB. Effectiveness of MCQs in assessing higher order cognition. Biomedica 2017;33(4):269-272 [FREE Full text]
23. Liu Q, Wald N, Daskon C, Harland T. Multiple-choice questions (MCQs) for higher-order cognition: perspectives of university teachers. Innovations in Education and Teaching International 2023 Jun 8;61(4):802-814. [doi: 10.1080/14703297.2023.2222715]
24. Menezes RG, Nayak VC, Binu VS, et al. Objective structured practical examination (OSPE) in forensic medicine: students' point of view. J Forensic Leg Med 2011 Nov;18(8):347-349. [doi: 10.1016/j.jflm.2011.06.011] [Medline: 22018165]
25. Yusoff MSB, Jaa'far R, Arzuman H, Arifin WN, Pa MNM. Perceptions of medical students regarding educational climate at different phases of medical training in a Malaysian medical school. Education in Medicine Journal 2013 Sep 1;5(3):e30-e41. [doi: 10.5959/eimj.v5i3.146]

26. Medical Council Announcement No. 76/2023 on the criteria for requesting to open/improve the Doctor of Medicine program and certify medical school institutions in 2023 [Article in Thai]. The Medical Council of Thailand. 2023. URL: https://tmc.or.th/index.php/News/Announcement/1089 [accessed 2024-08-12]

Abbreviations

MCQ: multiple-choice question

Edited by B Lesselroth; submitted 21.02.24; peer-reviewed by D Reham, G Aulino; revised version received 12.08.24; accepted 17.12.24; published 10.02.25.

Please cite as:
Chudoung U, Saengon W, Peonim V, Worasuwannarak W
Comparison of Learning Outcomes Among Medical Students in Thailand to Determine the Right Time to Teach Forensic Medicine: Retrospective Study
JMIR Med Educ 2025;11:e57634
URL: https://mededu.jmir.org/2025/1/e57634
doi:10.2196/57634

Original Paper

Understanding Community Health Care Through Problem-Based Learning With Real-Patient Videos: Single-Arm Pre-Post Mixed Methods Study

Kiyoshi Shikino[1,2,3], MD, MHPE, PhD; Kazuyo Yamauchi[1,2], MD, MHPE, PhD; Nobuyuki Araki[1,2], MD, PhD; Ikuo Shimizu[2,4], MD, MHPE, PhD; Hajime Kasai[2,4], MD, PhD; Tomoko Tsukamoto[2,3], MD, PhD; Hiroshi Tajima[2,4], MD, PhD; Yu Li[2,3], MD, PhD; Misaki Onodera[4], PhD; Shoichi Ito[1,2,4], MD, PhD

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

[2]Health Professional Development Center, Chiba University Hospital, Chiba University, Chiba, Japan

[3]Department of General Medicine, Chiba University Hospital, Chiba University, Chiba, Japan

[4]Department of Medical Education, Chiba University Graduate School of Medicine, Chiba University, Chiba, Japan

Corresponding Author:
Kiyoshi Shikino, MD, MHPE, PhD
Chiba University Graduate School of Medicine
Community-Oriented Medical Education
1-8-1, Inohana, Chu-ou-ku
Chiba, 2608670
Japan
Phone: 81 43 222 7171
Email: kshikino@gmail.com

Abstract

Background: Japan faces a health care delivery challenge due to physician maldistribution, with insufficient physicians practicing in rural areas. This issue impacts health care access in remote areas and affects patient outcomes. Educational interventions targeting students' career decision-making can potentially address this problem by promoting interest in rural medicine. We hypothesized that community-based problem-based learning (PBL) using real-patient videos could foster students' understanding of community health care and encourage positive attitudes toward rural health care.

Objective: This study investigated the impact of community-based PBL on medical students' understanding and engagement with rural health care, focusing on their knowledge, skills, and career orientation.

Methods: Participants were 113 fourth-year medical students from Chiba University, engaged in a transition course between preclinical and clinical clerkships from October 24 to November 2, 2023. The students were randomly divided into 16 groups (7-8 participants per group). Each group participated in two 3-hour PBL sessions per week over 2 consecutive weeks. Quantitative data were collected using pre- and postintervention questionnaires, comprehension tests, and tutor-assessed rubrics. Self-assessment questionnaires evaluated the students' interest in community health care and their ability to envision community health care settings before and after the intervention. Qualitative data from the students' semistructured interviews after the PBL sessions assessed the influence of PBL experience on clinical clerkship in community hospitals. Statistical analysis included median (IQR), effect sizes, and P values for quantitative outcomes. Thematic analysis was used for qualitative data.

Results: Of the 113 participants, 71 (62.8%) were male and 42 (37.2%) female. The total comprehension test scores improved significantly (pretest: median 4.0, IQR 2.5-5.0; posttest: median 5, IQR 4-5; P<.001; effect size r=0.528). Rubric-based assessments showed increased knowledge application (pretest: median 8, IQR 7-9; posttest: median 8, IQR 8-8; P<.001; r=0.494) and self-directed learning (pretest: median 8, IQR 7-9; posttest: median 8, IQR 8-8; P<.001; r=0.553). Self-assessment questionnaires revealed significant improvements in the students' interest in community health care (median 3, IQR 3-4 to median 4, IQR 3-4; P<.001) and their ability to envision community health care settings (median 3, IQR 3-4 to median 4, IQR 3-4; P<.001). Thematic analysis revealed key themes, such as "empathy in patient care," "challenges in home health care," and "professional identity formation."

Conclusions: Community-based PBL with real-patient videos effectively enhances medical students' understanding of rural health care settings, clinician roles, and the social needs of rural patients. This approach holds potential as an educational strategy

to address physician maldistribution. Although this study suggests potential for fostering positive attitudes toward rural health care, further research is needed to assess its long-term impact on students' career trajectories.

(JMIR Med Educ 2025;11:e68743) doi:10.2196/68743

KEYWORDS

community health care; community-oriented medical education; mixed method; problem-based learning; real-patient video

Introduction

Japan faces a significant health care delivery challenge owing to uneven physician distribution, notably affecting rural areas and community hospitals [1,2]. This maldistribution exacerbates community hospitals' challenges [3-5]. This issue is not confined to Japan; it impacts countries worldwide [6-11]. In 2019, the Ministry of Health, Labour and Welfare introduced the physician uneven distribution index as part of an intervention policy addressing prefectural geographical disparities in physician distribution [1,12-14]; it assesses the extent of physician maldistribution by evaluating prefectural medical supply and demand.

To combat the physician maldistribution, community hospital training has been integrated into second-year resident physicians' compulsory curriculum [15-17], highlighting the necessity of preparing future physicians with the competencies required to effectively meet rural communities' health care needs. Moreover, introducing community medicine principles early in medical education is an acknowledged need [18].

Japanese medical schools have begun to proactively adopt problem-based learning (PBL) as a foundational step before clinical rotations [19]. PBL emphasizes real-life medical scenarios, cultivating students' clinical reasoning and decision-making skills [20]. PBL prepares students for clinical rotations with an enriched understanding of community health care's challenges and prospects [21]. This approach bolsters medical students' clinical training and supports the alleviation of physician maldistribution by promoting community or rural medicine careers.

However, although PBL has been implemented in medical education settings, its integration with community-oriented medicine in addressing physician maldistribution remains underexplored. Furthermore, despite the global relevance of physician maldistribution, studies focusing on innovative educational interventions targeting this issue are limited [22-24].

We hypothesized that incorporating real-patient videos into community-focused PBL would significantly improve students' capacity to make well-informed career choices and identify with positive role models. This study addressed the aforementioned research gap by examining this approach's effectiveness within Japanese medical education.

Methods

Study Design

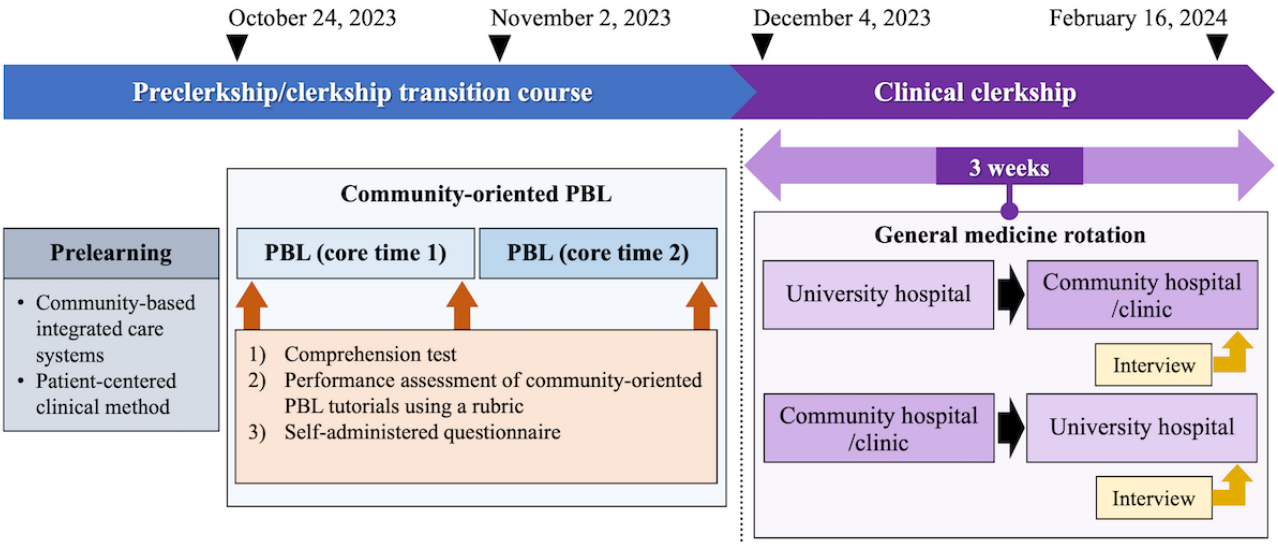
This study used an explanatory sequential mixed methods design following a pragmatic approach [25-27], capitalizing on quantitative and qualitative designs' strengths while minimizing their shortcomings. Furthermore, it allowed researchers to better understand experimental results while incorporating participants' perspectives. The National Institutes of Health advises a mixed methods approach "to improve the quality and scientific power of data" and to better address the complex issues facing health sciences today, including health professional education [28,29]. In the qualitative analysis, medical students' reflection papers were text-mined to analyze the word frequency in community-oriented PBL. Next, we conducted individual interviews with medical students during their clinical clerkship.

Participants and Trial Design

This study was conducted as part of the undergraduate medical curriculum at Chiba University, Japan. Community-oriented PBL was conducted from October 24 to November 2, 2023, as part of a preclerkship/clerkship transition course [30], a 5-week preparatory education period before clinical clerkship. Additionally, PBL is integrated into the Chiba University medical school's curriculum in years 1-4, and all students experience it. Participants were 113 fourth-year medical students who had attended lectures and received simulation training in basic and clinical medicine. To minimize potential biases and ensure an even distribution of student characteristics, participants were randomly divided into 16 groups of 7-8 each using the random number generation function in Microsoft Excel. The 16 groups received community-oriented PBL of a patient case using a community-based integrated care system with real-patient videos.

Quantitative data were gathered using the comprehension test, a tutor assessment with rubrics during the core time, and pre- and postintervention questionnaires to assess community health care perceptions (Figure 1). Additionally, a qualitative evaluation assessed community health care perceptions using a free-text reflection paper after PBL and follow-up interviews during clinical clerkships in community hospitals or clinics. The study used a mixed methods, sequential explanatory design to integrate the results [25,27,31].

Figure 1. Timeline from PBL in preclerkship/clerkship transition course to general medicine rotation in clinical clerkship. PBL: problem-based learning.



Community-Oriented PBL Educational Intervention

Real-patient videos were meticulously prepared to enhance the learning experience authenticity, simulating real-life scenarios in community health care settings (Multimedia Appendix 1). Prepared in collaboration with medical education experts, community health care professionals, and audiovisual production specialists, these videos aimed to accurately depict home health care characteristics, including medical interviews, physical examinations, and in-home patient interactions. Real patients participated in the production process under strict ethical guidelines to ensure authenticity and respect for patient privacy, emphasizing community health care's unique challenges and dynamics.

Each real-patient video was carefully scripted and filmed to represent common community health care situations, encompassing multiple scenes depicting different patient care stages and interactions, ranging from initial patient assessments in a community hospital setting to follow-up in-home visits. The video durations ranged from 3 to 5 minutes. Along with the videos, patient information sheets and tasks were presented to the students (Multimedia Appendix 2). One of the patient's primary conditions was underlying diabetes, and they presented with severe lower leg edema. The case involved transitioning from acute care to a chronic care hospital, followed by the introduction of home visit medical services. Patient consent was obtained for video use, and students were instructed to adhere to confidentiality guidelines.

During the community-oriented PBL sessions, students collectively viewed the real-patient videos in designated classrooms equipped with audiovisual facilities, allowing for simultaneous viewing on shared screens. Before watching the videos, students were divided into small groups and assigned a tutor to facilitate discussions and learning activities. The tutors observed whether the students could achieve the learning objectives and facilitated discussions. The tutors, randomly selected faculty members in medical education and community-oriented medical education, were given standardized instructions and materials before the sessions to ensure consistency and effectiveness [32].

Community-oriented PBL sessions were divided into 2 sessions per case, each lasting approximately 3 hours. In the first session (core time 1), students were presented with the patient's history and physical examination findings. In the second session (core time 2), the investigation findings and treatment plans were discussed. The same case scenario was used for all students.

Quantitative Measures

Comprehension Test

The comprehension test assessed the minimum essential knowledge required for problem solving in PBL, focusing on holistic medicine, patient-centered care, and the International Classification of Functioning, Disability, and Health (ICF). It comprised 5 multiple-choice questions (Q1-Q5, Multimedia Appendix 3). The test was administered as a pretest at the beginning of core time 1 and a posttest after core time 2, allowing for a comparison of test scores.

The comprehension test items were developed specifically for this study and have not been used in other contexts. They were based on the core learning objectives of the PBL sessions, which included understanding the structure of community health care systems, application of the ICF, and principles of patient-centered care. To ensure clarity and alignment with learning objectives, the questions underwent cognitive debriefing by faculty members in medical education. This process involved reviewing each question for relevance, accuracy, and comprehensibility, with feedback incorporated into the final version to enhance content validity.

Rubric-Based Performance Assessment

In addition to the comprehension test, each student's performance during the PBL sessions was assessed using a rubric. The rubric was determined based on previous studies after the authors discussed the validity of the criteria [33,34].

It evaluated 5 key dimensions: knowledge application, comprehensive care process, self-regulated learning, learning motivation, and communication skills with peers. Self-regulated learning refers to the ability of students to plan, monitor, and reflect on their learning process, fostering autonomy and adaptability in problem-solving contexts [35]. Each of the 5 dimensions was quantitatively evaluated on a 10-point scale (Multimedia Appendix 4). Performance assessments were conducted before and after the educational intervention to measure changes in the competencies.

Self-Administered Questionnaire

Students completed questionnaires before and after community-oriented PBL (Multimedia Appendix 5). They were assigned identification numbers to preserve their anonymity. Data were collected using a self-administered 5-point Likert-scale questionnaire ranging from 1 (strongly disagree) to 5 (strongly agree). The criteria were informed by previous studies and refined by the authors through discussions in focus groups [36,37]. After community-oriented PBL, 2 items ("I am interested in community health care" and "I can envision a community health care setting") were surveyed. The items assessed the students' interest in community health care and their ability to visualize a community health care setting.

Sample Size

This study also served as an educational program for fourth-year medical students in a basic clinical clerkship course. Altogether, 113 medical students from 12 groups were recruited. For quantitative data, the sample size required a 2-tailed *t* test of the difference between the pre- and post-PBL means, assuming a significance level of .05, a power of 0.8, and an effect size of 0.5. When the Mann-Whitney *U* test was conducted with those values, the required sample size was 54 in each group, totaling 108.

Data Analysis

All statistical analyses of quantitative data were conducted using SPSS Statistics for Microsoft Windows version 29.0 (IBM Corp), with a significance level under 5% for each analysis. The comprehension test results, including total scores and individual question responses (Q1-Q5), were analyzed using the Wilcoxon signed rank test for paired total scores. Additionally, the McNemar test was used to compare pre- and post-PBL correct response rates for individual questions. For rubric-based performance assessment, the Wilcoxon signed rank test was used to compare scores from core time 1 and core time 2 for total scores and individual rubric items. Effect sizes were calculated for all analyses: r values were derived from z scores for the Wilcoxon signed rank test, and the Cohen w value was calculated for the McNemar test.

Qualitative Measures

Follow-Up Interviews During Clinical Clerkships

Semistructured interviews (average duration: 20 minutes) with individual medical students were conducted by authors KS, KY, and NA. All sessions were recorded and transcribed verbatim, and interviews were conducted iteratively. An interview guide containing open-ended questions was constructed deductively based on the research question and thematic analysis findings. This guide was modified after the first 9 interviews to address emerging and previously unexplored themes in subsequent interviews. The interview participants received no gifts for participating.

Interview transcripts were analyzed using a template analysis approach [38,39]. An inductive code template was defined based on the research questions, thematic analysis findings, and interview guide. The initial template was developed through independent coding (performed by authors KS and IS) of the first 9 interviews. The template was further developed by coding the subsequent interviews. Regarding version 2 of the template, after coding 3 interviews, KS, NA, and IS agreed that the template adequately covered all texts. KS and IS individually coded the remaining transcripts using the template. At this stage, authors KY and SI discussed all further changes or additions to the template until they reached a consensus. After coding all 12 interviews, no additional changes were made to the templates. The final code template was further confirmed by analyzing the remaining 12 transcripts, which can be interpreted as a sign that code saturation was reached [40].

A qualitative evaluation was conducted to assess the acquisition of higher-order intellectual skills in which an interview was conducted after PBL and clinical clerkship. In the clinical clerkship interview, community-oriented PBL's effectiveness in improving clinical performance in home visit care was investigated. The interviewers (KS, KY, and NA) discussed the content and developed an interview guide. Students were asked the following open-ended question: "What is the effectiveness of community-oriented PBL for clinical clerkship?" The interviews were administered by 3 faculty members to 13 students from community-oriented PBL groups during their clinical clerkship. All target students had experienced home visits in their clinical clerkship 1-3 months after community-oriented PBL. The interviewers were trained facilitators from the faculty overseeing community-oriented PBL and conducted thematic analysis. Two researchers (KS and NA) independently read and coded the transcripts. Researcher triangulation was conducted in which the same 2 researchers conducted the analysis and consensus building.

KS and IS, who have extensive experience in qualitative research, defined and regularly discussed the themes and subthemes from the data to ensure the results' reliability. The cognitive process dimensions to which they corresponded were also evaluated.

Ethical Considerations

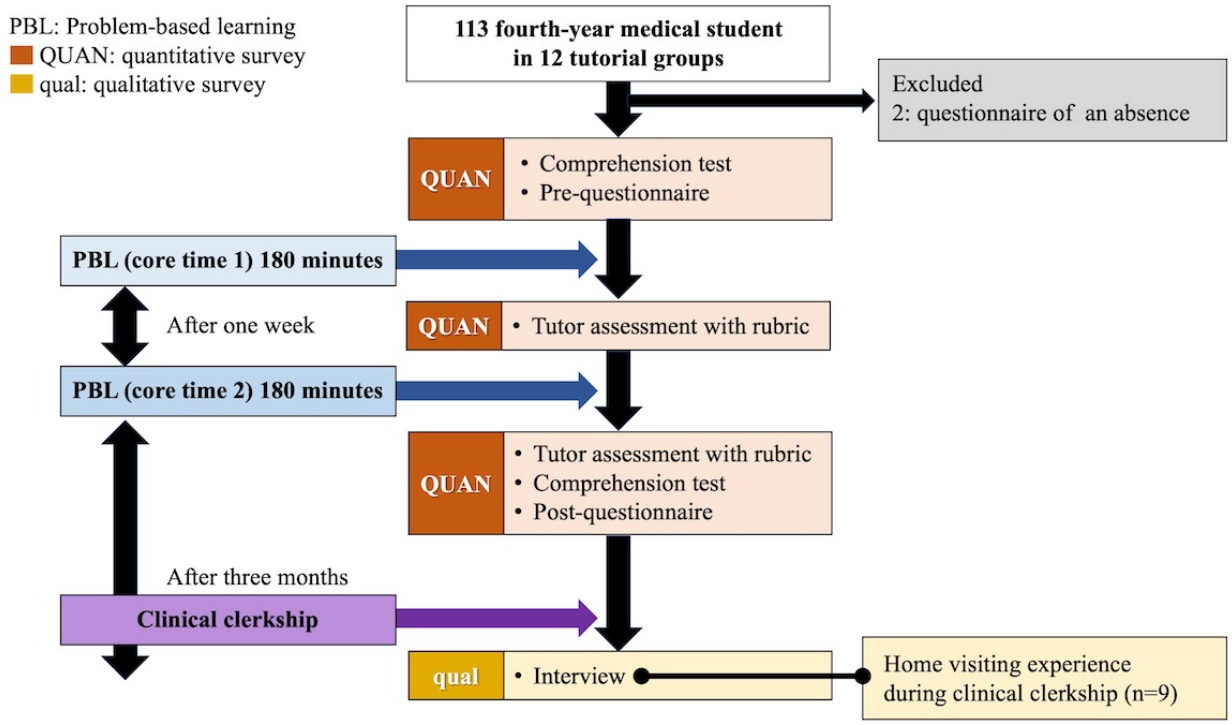
This study was approved by the Ethics Review Committee of the Graduate School of Medicine, Chiba University (approval number 3425). The procedures for obtaining informed consent were explained to the medical students, who were also informed that this study would not affect their grades. All data collected in this study were anonymized to ensure privacy and confidentiality. Participants did not receive any compensation for their participation in this study.

Results

Participant Characteristics

In total, 113 medical students participated in PBL. Of the 113 participants, 71 (62.8%) were male and 42 (37.2%) female. In addition, 2 (1.8%) students were excluded from the survey because they were absent for a core time session; 111 (98.2%) students participated in the quantitative evaluation. The study flowchart is shown in Figure 2.

Figure 2. Study flow diagram.



Quantitative Measures

Comprehension Test

The total comprehension test scores of the students significantly improved after the community-oriented PBL intervention (P<.001, effect size r=0.528). The total pretest scores had a median of 4.0 (IQR 2.5-5.0), whereas the posttest scores improved to a median of 5 (IQR 4-5). Table 1 provides the pre- and posttest analysis results of each question (Q1-Q5). All questions except Q4 showed significant improvements in the percentage of correct answers.

Table 1. Correct answer rates for individual questions in the comprehension test (N=113).

Question number	Pretest correct answers, n (%)	Posttest correct answers, n (%)	*P* value	Effect size (*r*)
1	104 (92.0)	112 (99.1)	.02	2.333
2	69 (61.1)	107 (94.7)	<.001	5.600
3	81 (71.7)	106 (93.8)	<.001	4.640
4	84 (74.3)	93 (82.3)	.19	2.180
5	77 (68.1)	99 (87.6)	<.001	4.310

Rubric-Based Performance Assessment of Community-Oriented PBL Tutorials

Table 2 presents the results of the rubric-based assessment of the students' performance in core time 1 and core time 2. The students' total scores significantly improved from core time 1 (median 38, IQR 33-43) to core time 2 (median 40, IQR 35-44), with P<.001 and an effect size (r) of 0.516. Regarding individual rubric items, significant improvements were observed in domains such as acquiring knowledge applicable to community health care, developing comprehensive care processes, and fostering self-directed learning. In contrast, no significant improvement was observed in the motivation to learn, whereas interpersonal skills showed a small but statistically significant enhancement.

Table 2. Performance assessment of community-oriented PBL[a] tutorials using a rubric.

Performance items	Core time 1, median (IQR)	Core time 2, median (IQR)	*P* value	Effect size (*r*)
Total score (0-50)	38 (33-43)	40 (35-44)	<.001	0.516
Domains				
1. Acquire knowledge that can be easily recalled and applied in community health care settings (0-10).	8 (7-9)	8 (8-8)	<.001	0.494
2. Develop an effective, comprehensive community care process (0-10).	8 (6-8)	8 (8-8)	<.001	0.532
3. Develop self-directed learning methods (0-10).	8 (7-9)	8 (8-8)	<.001	0.553
4. Motivate myself to learn (0-10).	8 (6-8)	8 (6-8)	.11	0.151
5. Acquire good interpersonal skills (0-10).	8 (7-9)	8 (7-9)	.03	0.201

[a]PBL: problem-based learning.

Self-Administered Questionnaire

Results indicated significant changes in students' perceptions of community health care after participating in community-oriented PBL. For the statement "I am interested in community health care," the preintervention median score was 3 (IQR 3-4), which increased to 4 (IQR 3-4) post intervention, with a statistical significance of $P<.001$ ($U=4446.5$). Similarly, the median score for the statement "I can envision a community health care setting" improved from 3 (IQR 3-4) preintervention to 4 (IQR 3-4) postintervention, also showing a significant difference, with $P<.001$ ($U=2589.5$).

Qualitative Measures

Follow-Up Interviews During Clinical Clerkships

We explored community-oriented PBL's impact on medical students' experiences during home visit consultations. In total, 12 (10.6%) medical students who had not experienced home visits before PBL consented to participate in the interview immediately after acquiring home visit experience during a community clinical clerkship. Through qualitative thematic analysis of the interviews, 7 main themes emerged: "building readiness for home care visit participation," "understanding and navigating the home care environment," "professional and personal growth," "interprofessional collaboration and team dynamics," "challenges and opportunities in home care," "community engagement and regional health care systems," and "ethical considerations and end-of-life care" (Table 3).

Table 3. Follow-up interviews and thematic analysis.

Theme	Subthemes
Building readiness for home care visit participation	• Broadened understanding of patient care beyond medical intervention • Enhanced preparedness for real-world clinical situations • Shift in perspective from theoretical knowledge to practical application
Understanding and navigating the home care environment	• Insights into the holistic approach required in home visits • Observations on the complexities of home care, including resource limitations and patient lifestyles • Recognition of the importance of patient and family communication
Professional and personal growth	• Aspiration to contribute meaningfully to patient care • Development of empathy and emotional intelligence • Recognition of the multifaceted role of health care providers in patient support
Interprofessional collaboration and team dynamics	• Importance of teamwork and a multidisciplinary approach in patient care • Learning from and contributing to the health care team • Navigating professional roles and patient relationships
Challenges and opportunities in home care	• Adapting PBL[a] knowledge to address specific patient needs • Confronting and managing unique patient care challenges • Opportunities for innovative care practices in constrained environments
Community engagement and regional health care system	• Enhancing community-oriented medical practice through targeted PBL • Gaining insights into community health care needs and resources • Understanding the impact of regional characteristics on health care delivery
Ethical considerations and end-of-life care	• Deepened understanding of end-of-life care preferences and practices • Navigating ethical dilemmas in patient care decisions • Valuing patient autonomy and quality of life in care planning

[a]PBL: problem-based learning.

Discussion

Principal Findings

This study demonstrated that integrating real-patient videos into community-oriented PBL improves medical students' knowledge, skills, and attitudes toward community health care. Comprehension test results showed significant improvements in students' understanding of core concepts, including community-based integrated care systems (Q1), the ICF framework (Q2 and Q3), and holistic, patient-centered care (Q5). These findings highlight students' enhanced theoretical knowledge essential for community health care practice.

Rubric-based performance assessments revealed notable improvements in 3 key domains:

- Knowledge application (item 1): Students showed improved abilities to recall and apply knowledge in community health care scenarios.
- Developing comprehensive care processes (item 2): Scores reflected stronger skills in designing patient-centered care plans tailored to community settings.
- Self-directed learning (item 3): Students demonstrated enhanced autonomy in planning, monitoring, and reflecting on their learning tasks.

Although interpersonal skills (item 5) improved slightly, no significant changes were observed in the motivation to learn (item 4), indicating areas for potential curriculum enhancement. Self-assessment questionnaires revealed increased interest in community health care and an improved ability to envision a community health care setting. These results suggest that the intervention can positively influence students' attitudes and readiness for community health care practice, potentially guiding their career interests toward rural areas.

Qualitative analysis of students' reflections underscored themes such as readiness for home care visits, professional and personal growth, and community engagement. Students reported a deeper understanding of the complexities of community health care, fostering empathy and patient-centered approaches essential for effective practice in underserved areas.

Implications of Findings

This study provided valuable insights into how community-oriented PBL, enhanced by real-patient videos, fosters medical students' ability to conceptualize their future professional roles. The qualitative data indicated that students develop a deeper understanding of the principles and complexities of rural care, including holistic approaches, patient-centered decision-making, and the importance of interprofessional collaboration. These findings suggest that the intervention successfully prepares students to engage with the challenges and rewards of rural and community-based practice.

Importantly, the qualitative analysis revealed that many students began to envision themselves as contributors to rural health care systems. Themes such as "professional and personal growth" and "community engagement" highlighted students' recognition of their potential roles in underserved areas. This shift was

supported by their increased interest in community health care, as measured by the "questionnaire for perceptions of community health care self-assessment." Postintervention, students reported greater interest and confidence in envisioning community health care settings.

However, although the data indicated a significant attitudinal shift, we lack sufficient evidence to confirm a direct impact on medical students' long-term career intentions to pursue rural care roles after graduation. Future studies should include longitudinal tracking to assess whether the observed changes in perceptions and interests translate into tangible career decisions. Additionally, research is needed to validate the predictive validity of the self-assessment questionnaire in forecasting students' career trajectories.

This intervention lays a strong foundation for addressing the global challenge of physician maldistribution by bridging theoretical knowledge with practical applications. However, further investigation is required to understand its long-term influence on medical workforce trends and rural health care outcomes.

Comparison With the Literature

Our findings will contribute to a growing body of evidence supporting the efficacy of PBL in medical education. Previous studies have demonstrated PBL's ability to enhance clinical reasoning and decision-making skills [19,21,41]. However, our research added a unique dimension by integrating real-patient videos, which provide authentic learning experiences and contextualize medical knowledge within the framework of community-oriented care. Similar studies have reported that experiential learning approaches, such as case-based learning with audiovisual materials, improve students' engagement and retention of knowledge [41].

The significant gains in self-directed learning observed in this study echo findings from Matsuyama et al [35], who emphasized the role of contextual attributes in promoting self-regulated learning. Additionally, the qualitative themes identified in our analysis, such as the importance of interprofessional collaboration and navigating the home care environment, align with the existing literature on the competencies required for effective community health care practice [42-48].

Limitations

Our study has some limitations. First, it was conducted as part of the curriculum of a single medical school, potentially limiting the findings' applicability to other institutions and geographical settings. Future research should involve multiple institutions to enhance the results' generalizability. Additionally, the demographic and cultural context of the participants may not fully represent broader populations, especially in countries with different health care challenges and educational frameworks. Second, the absence of a control group makes it difficult to attribute observed improvements solely to the intervention of incorporating real-patient videos into PBL. Our study evaluated a "bundle" of educational strategies, including real-patient videos, the PBL framework, faculty interventions, and testing conditions. We could not measure these components' differential effects or potential synergistic interactions. Although a randomized controlled trial (RCT) could offer stronger evidence, implementing RCTs in educational settings poses ethical and logistical challenges, such as withholding valuable learning resources from a control group or ensuring equivalent baseline characteristics. To address these limitations, we recommend that future research consider alternative designs to balance rigor and feasibility. Third, the scalability of real-patient videos poses a significant challenge. Producing high-quality real-patient videos requires substantial time, resources, and collaboration among medical educators, health care professionals, and audiovisual specialists. These demands may limit the feasibility of widespread adoption. We suggest collaborative efforts, such as interschool partnerships and the development of shared digital repositories, to distribute production costs and enhance scalability. Fourth, the mixed methods design relied on self-reported measures and reflections, which could introduce a response bias. Future studies could benefit from incorporating objective measures of clinical performance and patient care outcomes. Fifth, the effectiveness of clinical reasoning education via hybrid PBL may vary depending on instructors' teaching skills. Despite standardized training for tutors, differences in tutor effectiveness may have influenced the consistency of outcomes. We propose further methods to ensure uniformity in future implementations, such as advanced tutor workshops and peer evaluations. Sixth, because the study participants were fourth-year medical students at a single Japanese institution, the results may not be directly generalizable to other populations, such as residents or general physicians, or contexts outside Japan, underscoring the need for further validation. Seventh, the comprehension test used in this study, although developed specifically for the program's educational objectives, was not formally piloted with a separate cohort. Instead, the test underwent cognitive debriefing with faculty members to ensure clarity, relevance, and alignment with the intended learning objectives. Although this process enhanced content validity, the absence of a formal pilot test may limit the ability to fully validate the test's reliability and generalizability. Similarly, the 2 items in the self-assessment questionnaire—designed to evaluate medical students' interest in and ability to envision community health care—were developed based on previous studies and focus group discussions but have not been validated using established scales. This lack of formal validation for the comprehension test and self-assessment questionnaire limits the generalizability and robustness of the findings. Finally, although the qualitative data provided valuable insights into students' conceptualization of professional roles and their preparedness for community health care settings, the lack of longitudinal data limits our ability to assess the long-term impact of these interventions on career trajectories or practice in rural or underserved areas. Future research should include follow-up assessments to evaluate the sustained influence of such educational interventions on students' career decisions and professional development.

Conclusion

Integrating real-patient videos into a community-oriented PBL curriculum shows significant promise in fostering medical students' interest and competencies in community and rural medicine. Our study demonstrated improvements in knowledge

acquisition and application, as indicated by enhanced rubric and comprehension test scores. Moreover, qualitative analysis revealed PBL's effectiveness in developing essential skills and shaping medical students' perceptions toward community health care. Although these changes may not directly translate to career decisions, they represent an essential step toward fostering awareness of rural health care needs and aligning medical students' competencies with the demands of underserved areas. This approach highlights the potential of combining real-patient videos with PBL as an innovative educational strategy to address physician maldistribution and support rural health care systems.

Data Availability

The datasets generated and analyzed during the study are not publicly available, because they include participants' personal data, but they are available from the corresponding author upon reasonable request.

Authors' Contributions

KS, KY, and SI planned, designed, and conceived the study. KS drafted the manuscript. KS, KY, NA, and IS recruited participants. KS, KY, and NA conducted the interviews. KS and NA analyzed the initial coding. KS, KY, IS, and SI analyzed the final coding and interpreted the data. KS and KY performed statistical analyses. HK conceived the figures. KS, KY, NA, IS, TT, HT, LY, and SI participated as tutors. All authors have read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Community-oriented problem-based learning tutor guide.
[DOCX File , 23 KB - mededu_v11i1e68743_app1.docx]

Multimedia Appendix 2
Information and task sheet.
[DOCX File , 25 KB - mededu_v11i1e68743_app2.docx]

Multimedia Appendix 3
Comprehension test questions.
[DOCX File , 19 KB - mededu_v11i1e68743_app3.docx]

Multimedia Appendix 4
Rubric for community-oriented problem-based learning.
[DOCX File , 25 KB - mededu_v11i1e68743_app4.docx]

Multimedia Appendix 5
Questionnaires for perceptions of community health care self-assessment.
[DOCX File , 22 KB - mededu_v11i1e68743_app5.docx]

References

1. Takayama A, Poudyal H. Incorporating medical supply and demand into the index of physician maldistribution improves the sensitivity to healthcare outcomes. J Clin Med 2021 Dec 28;11(1):155 [FREE Full text] [doi: 10.3390/jcm11010155] [Medline: 35011896]
2. Hara K, Kunisawa S, Sasaki N, Imanaka Y. Examining changes in the equity of physician distribution in Japan: a specialty-specific longitudinal study. BMJ Open 2018 Jan 08;8(1):e018538 [FREE Full text] [doi: 10.1136/bmjopen-2017-018538] [Medline: 29317415]
3. Kobayashi Y, Takaki H. Geographic distribution of physicians in Japan. Lancet 1992 Dec 05;340(8832):1391-1393. [doi: 10.1016/0140-6736(92)92569-2] [Medline: 1360099]
4. Toyabe S. Trend in geographic distribution of physicians in Japan. Int J Equity Health 2009 Mar 03;8(1):5 [FREE Full text] [doi: 10.1186/1475-9276-8-5] [Medline: 19257879]
5. Tanihara S, Kobayashi Y, Une H, Kawachi I. Urbanization and physician maldistribution: a longitudinal study in Japan. BMC Health Serv Res 2011 Oct 08;11(1):260 [FREE Full text] [doi: 10.1186/1472-6963-11-260] [Medline: 21982582]
6. Wu J. Measuring inequalities in the demographical and geographical distribution of physicians in China: generalist versus specialist. Int J Health Plann Manage 2018 Oct 20;33(4):860-879. [doi: 10.1002/hpm.2539] [Medline: 29781216]

7. Karan A, Negandhi H, Nair R, Sharma A, Tiwari R, Zodpey S. Size, composition and distribution of human resource for health in India: new estimates using National Sample Survey and Registry data. BMJ Open 2019 May 27;9(4):e025979 [FREE Full text] [doi: 10.1136/bmjopen-2018-025979] [Medline: 31133622]
8. Morris S, Sutton M, Gravelle H. Inequity and inequality in the use of health care in England: an empirical investigation. Soc Sci Med 2005 Mar;60(6):1251-1266. [doi: 10.1016/j.socscimed.2004.07.016] [Medline: 15626522]
9. Guttmann A, Shipman SA, Lam K, Goodman DC, Stukel TA. Primary care physician supply and children's health care use, access, and outcomes: findings from Canada. Pediatrics 2010 Jun;125(6):1119-1126. [doi: 10.1542/peds.2009-2821] [Medline: 20498170]
10. Mick SS, Lee SD, Wodchis WP. Variations in geographical distribution of foreign and domestically trained physicians in the United States: 'safety nets' or 'surplus exacerbation'? Soc Sci Med 2000 Jan;50(2):185-202. [doi: 10.1016/s0277-9536(99)00183-5] [Medline: 10619689]
11. Winkelmann J, Muench U, Maier CB. Time trends in the regional distribution of physicians, nurses and midwives in Europe. BMC Health Serv Res 2020 Oct 12;20(1):937 [FREE Full text] [doi: 10.1186/s12913-020-05760-y] [Medline: 33046077]
12. Organization for Economic Co-operationDevelopment (OECD). Doctors. Paris: OECD Publishing; 2018.
13. Sato H. Demand, supply and shortages of physicians: a critical analysis on the current government?s method. J Health Welf Policy 2020;3:39-48.
14. Ministry of Health, Labour and Welfare Study Group on Supply and Demand of Medical Staff. Doctor Supply and Demand Subcommittee 4th interim report. Ministry of Health, Labour and Welfare. 2019. URL: https://www.mhlw.go.jp/content/12601000/000504403.pdf [accessed 2024-04-24]
15. Tago M, Shikino K, Hirata R, Watari T, Yamashita S, Tokushima Y, et al. General medicine departments of Japanese universities contribute to medical education in clinical settings: a descriptive questionnaire study. Int J Gen Med 2022;15:5785-5793 [FREE Full text] [doi: 10.2147/IJGM.S366411] [Medline: 35774114]
16. Teo A. The current state of medical education in Japan: a system under reform. Med Educ 2007 Mar;41(3):302-308. [doi: 10.1111/j.1365-2929.2007.02691.x] [Medline: 17316216]
17. Kozu T. Medical education in Japan. Acad Med 2006;81(12):1069-1075. [doi: 10.1097/01.acm.0000246682.45610.dd]
18. Ohde S, Deshpande GA, Takahashi O, Fukui T. Differences in residents' self-reported confidence and case experience between two post-graduate rotation curricula: results of a nationwide survey in Japan. BMC Med Educ 2014 Jul 12;14(1):141 [FREE Full text] [doi: 10.1186/1472-6920-14-141] [Medline: 25016304]
19. Davis MH. AMEE Medical Education Guide No. 15: problem-based learning: a practical guide. Med Teach 1999 Jul 03;21(2):130-140. [doi: 10.1080/01421599979743] [Medline: 21275726]
20. Haith-Cooper M. Problem-based learning within health professional education. What is the role of the lecturer? A review of the literature. Nurse Educ Today 2000 May;20(4):267-272. [doi: 10.1054/nedt.1999.0397] [Medline: 10827097]
21. Forbes H, Syed M, Flanagan O. The role of problem-based learning in preparing medical students to work as community service-oriented primary care physicians: a systematic literature review. Cureus 2023 Sep;15(9):e46074 [FREE Full text] [doi: 10.7759/cureus.46074] [Medline: 37900379]
22. Bhandary S. Problem-based learning curriculum and process assessment system for the undergraduate competency-based medical education: experiences from Nepal. Arch Med Health Sci 2021;9(2):331-336. [doi: 10.4103/amhs.amhs_282_21]
23. Kibret S, Teshome D, Fenta E, Hunie M, Taye MG, Fentie Y, et al. Medical and health science students' perception towards a problem-based learning method: a case of Debre Tabor University. AMEP 2021 Jul;12:781-786. [doi: 10.2147/amep.s316905]
24. Hou S. Integrating problem-based learning with community-engaged learning in teaching program development and implementation. Univ J Educ Res 2014 Jan;2(1):1-9. [doi: 10.13189/ujer.2014.020101]
25. Barbour RS. The case for combining qualitative and quantitative approaches in health services research. J Health Serv Res Policy 1999 Jan 01;4(1):39-43. [doi: 10.1177/135581969900400110]
26. Malterud K. The art and science of clinical knowledge: evidence beyond measures and numbers. Lancet 2001 Aug;358(9279):397-400. [doi: 10.1016/s0140-6736(01)05548-9]
27. Côté L, Turgeon J. Appraising qualitative research articles in medicine and medical education. Med Teach 2005 Jan 03;27(1):71-75. [doi: 10.1080/01421590400016308] [Medline: 16147774]
28. Dowding D. Review of the book best practices for mixed methods research in the health sciences. Qual Soc Work 2013 Jul 16;12(4):541-545. [doi: 10.1177/1473325013493540a]
29. Creswell JW, Plano CV. Designing and Conducting Mixed Method Research. 3rd Edition. Los Angeles: Sage Publications; 2017.
30. Ovitsh RK, Gupta S, Kusnoor A, Jackson JM, Roussel D, Mooney CJ, et al. Minding the gap: towards a shared clinical reasoning lexicon across the pre-clerkship/clerkship transition. Med Educ Online 2024 Feb 06;29(1):2307715. [doi: 10.1080/10872981.2024.2307715]
31. Malterud K. Qualitative research: standards, challenges, and guidelines. Lancet 2001 Aug;358(9280):483-488. [doi: 10.1016/s0140-6736(01)05627-6]
32. Savery JR. Overview of problem-based learning: definitions and distinctions. Interdiscip J Probl Based Learn 2006 May 22;1(1):3. [doi: 10.7771/1541-5015.1002]

33. Houlden RL, Collier CP, Frid PJ, John SL, Pross H. Problems identified by tutors in a hybrid problem-based learning curriculum. Acad Med 2001 Jan;76(1):81. [doi: 10.1097/00001888-200101000-00021] [Medline: 11154202]
34. Wang S, Yan D, Hu X, Liu J, Liu D, Wang J. Comparison of attitudes toward the medical student-led community health education service to support chronic disease self-management among students, faculty and patients. BMC Med Educ 2023 Jan 11;23(1):17 [FREE Full text] [doi: 10.1186/s12909-023-04008-7] [Medline: 36631772]
35. Matsuyama Y, Nakaya M, Okazaki H, Leppink J, van der Vleuten C. Contextual attributes promote or hinder self-regulated learning: a qualitative study contrasting rural physicians with undergraduate learners in Japan. Med Teach 2017 Nov 26;40(3):285-295. [doi: 10.1080/0142159x.2017.1406074]
36. Recker AJ, Sugimoto SF, Halvorson EE, Skelton JA. Knowledge and habits of exercise in medical students. Am J Lifestyle Med 2021 Oct 12;15(3):214-219 [FREE Full text] [doi: 10.1177/1559827620963884] [Medline: 34025308]
37. Trullàs JC, Blay C, Sarri E, Pujol R. Effectiveness of problem-based learning methodology in undergraduate medical education: a scoping review. BMC Med Educ 2022 Feb 17;22(1):104 [FREE Full text] [doi: 10.1186/s12909-022-03154-8] [Medline: 35177063]
38. Joseph N, Rai S, Madi D, Bhat K, Kotian S, Kantharaju S. Problem-based learning as an effective learning tool in community medicine: initiative in a private medical college of a developing country. Indian J Community Med 2016;41(2):133-140 [FREE Full text] [doi: 10.4103/0970-0218.177535] [Medline: 27051088]
39. King N. Using templates in the thematic analysis of text. In: Cassel C, Symon G, editors. Essential Guide to Qualitative Methods in Organizational Research. London, UK: Sage Publications; 2004:256-270.
40. Hennink MM, Kaiser BN, Marconi VC. Code saturation versus meaning saturation: how many interviews are enough? Qual Health Res 2017 Mar 26;27(4):591-608 [FREE Full text] [doi: 10.1177/1049732316665344] [Medline: 27670770]
41. Ikegami A, Ohira Y, Uehara T, Noda K, Suzuki S, Shikino K, et al. Problem-based learning using patient-simulated videos showing daily life for a comprehensive clinical approach. Int J Med Educ 2017 Feb 27;8:70-76 [FREE Full text] [doi: 10.5116/ijme.589f.6ef0] [Medline: 28245193]
42. Du X, Al Khabuli JOS, Ba Hattab RAS, Daud A, Philip NI, Anweigi L, et al. Development of professional identity among dental students - a qualitative study. J Dent Educ 2023 Jan 02;87(1):93-100. [doi: 10.1002/jdd.13092] [Medline: 36052467]
43. Orsmond P, McMillan H, Zvauya R. It's how we practice that matters: professional identity formation and legitimate peripheral participation in medical students: a qualitative study. BMC Med Educ 2022 Feb 09;22(1):91 [FREE Full text] [doi: 10.1186/s12909-022-03107-1] [Medline: 35139839]
44. Wang J, Wang B, Liu D, Zhou Y, Xing X, Wang X, et al. Video feedback combined with peer role-playing: a method to improve the teaching effect of medical undergraduates. BMC Med Educ 2024 Jan 19;24(1):73 [FREE Full text] [doi: 10.1186/s12909-024-05040-x] [Medline: 38243255]
45. Noverati N, R Naro G, J Fischer R, M Thompson B. Using video and virtual patients in problem-based learning: a scoping review. Med Sci Educ 2020 Dec 14;30(4):1685-1691 [FREE Full text] [doi: 10.1007/s40670-020-01108-7] [Medline: 34457832]
46. Berini CR, Bonilha HS, Simpson AN. Impact of community health workers on access to care for rural populations in the United States: a systematic review. J Community Health 2022 Jun 24;47(3):539-553. [doi: 10.1007/s10900-021-01052-6] [Medline: 34817755]
47. Xu Y, Koh XH, Chua YTS, Tan CGI, Aloweni FAB, Yap BEJ, et al. The impact of community nursing program on healthcare utilization: a program evaluation. Geriatr Nurs 2022 Jul;46:69-79. [doi: 10.1016/j.gerinurse.2022.04.024] [Medline: 35609434]
48. Jack HE, Arabadjis SD, Sun L, Sullivan EE, Phillips RS. Impact of community health workers on use of healthcare services in the United States: a systematic review. J Gen Intern Med 2017 Mar 5;32(3):325-344 [FREE Full text] [doi: 10.1007/s11606-016-3922-9] [Medline: 27921257]

Abbreviations

ICF: International Classification of Functioning, Disability, and Health
PBL: problem-based learning
RCT: randomized controlled trial

Edited by B Lesselroth; submitted 12.11.24; peer-reviewed by RA El Arab, M Agbede; comments to author 19.12.24; revised version received 08.01.25; accepted 14.01.25; published 31.01.25.

Please cite as:
Shikino K, Yamauchi K, Araki N, Shimizu I, Kasai H, Tsukamoto T, Tajima H, Li Y, Onodera M, Ito S
Understanding Community Health Care Through Problem-Based Learning With Real-Patient Videos: Single-Arm Pre-Post Mixed Methods Study
JMIR Med Educ 2025;11:e68743
URL: https://mededu.jmir.org/2025/1/e68743
doi:10.2196/68743
PMID:

Original Paper

Organizational Leaders' Views on Digital Health Competencies in Medical Education: Qualitative Semistructured Interview Study

Humairah Zainal[1*], PhD; Xin Xiao Hui[1*], MSocSci; Julian Thumboo[2,3*], MBBS, MMed; Fong Kok Yong[2,3*], MBBS, MMed

[1]Health Services Research Unit, Singapore General Hospital, Singapore, Singapore

[2]Department of Rheumatology and Immunology, Singapore General Hospital, Singapore, Singapore

[3]Duke-NUS Medical School, National University of Singapore, Singapore, Singapore

[*]all authors contributed equally

Corresponding Author:
Fong Kok Yong, MBBS, MMed
Department of Rheumatology and Immunology
Singapore General Hospital
10 Hospital Boulevard
Singapore, 168582
Singapore
Phone: 65 6908 8949
Email: fong.kok.yong@singhealth.com.sg

Abstract

Background: Digital technologies (DTs) have profoundly impacted health care delivery globally and are increasingly used in clinical practice. Despite this, there is a scarcity of guidelines for implementing training in digital health competencies (DHC) in medical schools, especially for clinical practice. A lack of sustained integration of DHC risks creating knowledge gaps due to a limited understanding of how DT should be used in health care. Furthermore, few studies have explored reasons for this lag, both within and beyond the medical school curriculum. Current frameworks to address these barriers are often specific to individual countries or schools and focus primarily on curriculum design and delivery. A comprehensive framework is therefore required to ensure consistent implementation of DHC across various contexts and times.

Objective: This study aims to use Singapore as a case study and examine the perspectives of doctors in organizational leadership positions to identify and analyze the barriers to DHC implementation in the undergraduate curriculum of Singapore's medical schools. It also seeks to apply the Normalization Process Theory (NPT) to address these barriers and bridge the gap between health care systems and digital health education (DHE) training.

Methods: Individual semistructured interviews were conducted with doctors in executive and organizational leadership roles. Participants were recruited through purposive sampling, and the data were interpreted using qualitative thematic analysis.

Results: A total of 33 doctors participated, 26 of whom are currently in organizational leadership roles and 7 of whom have previously held such positions. A total of 6 barriers were identified: bureaucratic inertia, lack of opportunities to pursue nontraditional career pathways, limited protective mechanisms for experiential learning and experimentation, lack of clear policy guidelines for clinical practice, insufficient integration between medical school education and clinical experience, and poor IT integration within the health care industry.

Conclusions: These barriers are also present in other high-income countries experiencing health care digitalization, highlighting the need for a theoretical framework that broadens the generalizability of existing recommendations. Applying the NPT underscores the importance of addressing these barriers to effectively integrate DHC into the curriculum. The active involvement of multiple stakeholders and the incorporation of continuous feedback mechanisms are essential. Our proposed framework provides concrete, evidence-based, and step-by-step recommendations for implementation practice, supporting the introduction of DHC in undergraduate medical education.

(JMIR Med Educ 2025;11:e64768) doi:10.2196/64768

KEYWORDS

technology; medical education; curriculum; clinical competence; digital competence; Singapore; digital health; qualitative study; medical school; risk; comprehensive framework; doctor; thematic analysis; information technology; evidence-based; undergraduate; healthcare systems; mobile phone

Introduction

Background

The integration of digital technologies (DTs) into clinical care is transforming health care worldwide [1], underscoring the need to prepare future health care professionals with digital health competencies (DHC) through digital health education (DHE). Despite widespread recognition of the importance of DHC, medical schools worldwide—including those in Singapore—have been lagging in their efforts to implement such training in a meaningful and systematic manner [2-8]. Countries like the United States, the United Kingdom, Canada, and Germany face similar challenges, such as fragmented integration efforts, limited faculty expertise, and curriculum overload, which hinder the consistent incorporation of DHE into undergraduate medical curricula [9-14]. Singapore, a high-income nation in Southeast Asia with advanced education systems and extensive digitalization, provides a compelling case study to explore these barriers. While Singapore's unique sociopolitical and cultural context informs this study, the challenges it faces mirror those encountered by other high-income nations, highlighting the broader international relevance of this research.

Existing efforts to integrate DHE in medical schools, especially for clinical practice, are often disconnected, lacking systematic frameworks and sustained engagement with key stakeholders [2-8]. For example, in Europe and the United States, DHE initiatives are often siloed, leading to significant variability in the quality and scope of training [2,4,6]. Similarly, Australia has faced barriers such as a lack of standardized frameworks for digital health training and challenges in aligning medical education with rapidly evolving health care technologies [3].

These gaps result in inconsistencies in training, with DHC frequently treated as elective content rather than as a core component of medical education. This study applies Normalization Process Theory (NPT)—a framework designed to examine how new practices become embedded within institutions—to provide insights into systematically normalizing DHE and ensuring its sustainable integration [15].

This study addresses the following research questions: (1) What are the institutional and structural barriers to integrating DHE into undergraduate medical curricula? (2) How can the medical school experience be aligned with technological advances? (3) How can NPT be systematically applied to facilitate the effective and sustainable incorporation of DHE? By focusing on Singapore, this study not only provides a deeper understanding of these challenges but also offers insights that can inform global efforts to strengthen DHE integration in medical education.

Exploring the Perspectives of Doctors in Organizational Leadership Roles

Despite their influence on governance and standards, the perspectives of doctors in organizational leadership roles are often overlooked. Existing research that evaluates the opinions of this group of stakeholders primarily addresses challenges in implementing DT in health care, characteristics of effective health systems, and key attributes for health care leaders [16-18].

This study recognizes that doctors in organizational leadership roles possess a strategic understanding of both clinical practice and medical education. Their insights are crucial as they can influence curriculum design, resource allocation, and policy formulation. By leveraging their dual perspectives, the study identifies unique barriers that may not be visible to frontline educators or students.

Furthermore, leaders in health care organizations have the authority to implement change and drive initiatives. Understanding their perspectives ensures that any proposed solutions are both feasible and likely to gain support at the highest levels of the institution. Their endorsement can facilitate smoother implementation and wider acceptance of DHE initiatives.

In addition, these leaders are often involved in broader system-wide decision-making. They are also likely familiar with challenges related to integrating new technologies and practices into established systems. Hence, their experience can provide valuable insights into overcoming any institutional inertia, aligning new initiatives with existing policies, and addressing systemic barriers.

Medical Education in Singapore

Medical education in Singapore is provided by 3 schools, which are Yong Loo Lin School of Medicine (YLL) at the National University of Singapore (NUS), Lee Kong Chian School of Medicine (LKCMedicine) at the Nanyang Technological University (NTU), and Duke-NUS Medical School (Duke-NUS). YLL, established in 1905, and LKCMedicine, founded in 2013, both provide a 5-year undergraduate program. In this program, students spend 2 years studying the basics of medical sciences before undergoing clinical clerkships from the third to the fifth year. YLL was formed to address the critical health care needs of the local population during the colonial period, whereas LKCMedicine was established to meet the increasing health care demands due to an aging population [19]. To boost Singapore's capabilities in translational medicine, Duke-NUS was founded in 2005 through a partnership between NUS and Duke University in the United States. Duke-NUS is a graduate medical school that offers a 4-year MD program, where the first year focuses on basic sciences and the second year on clinical postings. In their third year, students focus on developing research skills, and in their final year, they engage in clinical clerkships [20]. All medical schools in Singapore receive public

funding, and students' tuition fees are subsidized by the government.

Despite their cutting-edge facilities and innovation-driven educational technology, disparities persist between medical school training and clinical application [8,21]. Efforts to integrate DHE courses, such as virtual reality and point-of-care ultrasound, also vary in content and duration across these institutions [8]. Hence, standardizing the curriculum and ensuring consistent training across all medical schools is crucial to bridging the gap between theoretical knowledge and practical skills. This approach would enhance the overall competency of future health care professionals and improve the quality of patient care. In addition, ongoing assessment and adaptation of these programs to incorporate emerging technologies and methodologies are essential for keeping pace with the rapidly evolving medical field.

Methods

Data Collection

A qualitative study was conducted using individual semistructured interviews with doctors who are currently or have previously held organizational leadership positions. Participants were identified by our principal investigator (PI), FKY, based on their leadership roles within public health care organizations, ensuring they possessed the requisite knowledge and experience aligned with the research objectives. Selection criteria focused on senior leaders with expertise in research, clinical education, and development.

This cohort represented a niche group of chief physicians leading public tertiary hospitals that serve as teaching hospitals for undergraduate and postgraduate medical training. Specifically, participants included group chief executive officers from all 3 public health care clusters in Singapore, chairmen of medical boards at public health care institutions, senior administrators from the Ministry of Health, and Directors of Training and Education. To ensure a consistent depth of expertise, participants were required to have a minimum of 5 years of leadership experience within public health care organizations. Those with less than 5 years of such experience were excluded from the study.

Purposive sampling was used to ensure a diverse representation of organizational leaders based on factors such as organizational type, that is, public health care clusters and institutions and functional domains, ie, clinical services and administration. This approach enabled the collection of rich, varied perspectives, enhancing the study's credibility.

Data collection took place from January to April 2021. Participants were invited by the PI by email, which provided a detailed outline of the study's purpose, procedures, potential risks, and benefits. The email also included a consent statement for participants to review and acknowledge before proceeding. In reviewing issues of reflexivity, the threat of potential researcher biases due to the established professional relationship between the PI and the research participants was overcome by having the research fellow, who had no previous relationship with any of the participants, as the interviewer. During the Zoom (Zoom Video Communications) interview, participants were given the opportunity to ask questions, and their verbal consent was recorded at the beginning of each session to ensure informed and voluntary participation. They were reminded of their right to withdraw from the study at any point. It was clarified that data collected prior to withdrawal would still be retained and analyzed to enable a comprehensive evaluation of findings.

To protect participants' anonymity, we assigned code identifiers beginning with "OL" (organizational leader) to each of them. Any identifying information and audio recordings were stored separately from the main dataset in a secure, password-protected file, accessible only to authorized research team members. In reporting results, care was taken to remove or generalize any details that could potentially identify individuals. The data collected and analyzed was used exclusively to inform curriculum development, with no intention to disclose identifiable information.

The interview guide was developed based on the NPT constructs of coherence, cognitive participation, collective action, and reflexive monitoring (Textbox 1). We then adapted the interview guide iteratively to allow participants to share their views on matters that were not initially included in the guide. Generally, the questions sought participants' views on the clinical skills that are still relevant in the digital age (coherence), additional skills that medical students and doctors need for clinical practice amid increasing health care digitalization (coherence), and the clinical skills that are currently being covered in local medical schools (coherence). We also asked participants for their opinions on the clinical skills that should be emphasized more in the medical school curricula (coherence), the challenges of integrating DHE into the compulsory curricula, and suggestions for curriculum improvement to better prepare students for future clinical practice (cognitive participation). To explore how "Collective Action" could be operationalized, we asked how medical schools could improve their collaboration with other stakeholders, particularly professional bodies, health care institutions, and the health care system, to better prepare medical students for clinical practice in the digital era (collective action). We also raised the question of how participants would evaluate the impacts, benefits, and areas for improvement of DHE initiatives within the medical school curriculum (reflective monitoring).

Textbox 1. Interview questions.

- In general, what are the clinical skills that a medical doctor should have?
- Which of these skills are still relevant in the digital age?
- Are there any skills that have been replaced by digital technology, be it partially or completely?
- Against the backdrop of increasing digitalization of health care, what new skills, clinical or otherwise, should a doctor have in order to practice medicine?
- What clinical skills are currently being covered in the local medical schools?
- Which of these skills should be emphasized more in the medical school curriculum?
- In your opinion, how well do the current medical school curricula prepare medical students for the digital aspects of health care? What do you think are some of the challenges in implementing digital health education in medical schools?
- What other improvements can be made to our local medical school curriculum to better prepare the students for clinical practice in light of rapid advances in technology (for example, the advent of artificial intelligence, big data, imaging, smartphone applications, and digital equipment such as handheld ultrasound)?
- How can local medical schools improve their collaborations with professional bodies and health care institutions to prepare medical students for clinical practice in this era of new technology?
- What can the health care system do to support medical students and young doctors in this era of new technology?
- Do you have any other comments on the digital transformations of medicine or health care before we end this interview?

Challenges of contextual differences and stakeholder variation are crucial factors that need to be carefully considered when applying NPT in diverse settings. The study was conducted in Singapore, where the adoption of DT within health care settings has been gradual [7]. This presents challenges for students who may not have adequate exposure to digital systems during their clinical placements. In response to this, the application of NPT should be focused on building digital literacy and ensuring that any intervention is compatible with ongoing efforts to integrate digital solutions into clinical practice. Furthermore, there is also a limited innovation culture in Singapore's health care system [7]. To overcome this, interventions that adopt NPT should incorporate elements designed to stimulate collaboration and mentorship programs with industry professionals. This would help bridge the gap between academic training and the innovation needs of the health care sector.

The study included 33 participants, with the sample size determined based on theoretical and practical considerations. Data collection continued until saturation was reached, ensuring that no new themes or insights emerged from the interviews. This indicates that the sample size was sufficient to capture the relevant perspectives for the study. While practical constraints, such as time and resources, influenced the final number of participants, the primary focus was on ensuring data richness and diversity. This approach allowed for a comprehensive exploration of the research questions.

A total of 30 interviews were conducted and recorded over Zoom due to the physical restrictions brought about by the COVID-19 pandemic, while 3 in-person interviews were held with participants who were located in areas with fewer restrictions at the time or who specifically preferred in-person interaction. All in-person interviews were carried out in accordance with local health guidelines to ensure participant safety. Each interview lasted approximately 40 minutes and was audio-recorded. The transcriptions were derived from the audio recordings of the interviews, which were processed using Otter.ai software (Otter.ai, Inc) before being reviewed for accuracy by the PI and research fellow.

Data Analysis

Thematic analysis using Braun and Clarke's [22] 6-step framework was used to explore barriers that emerged from the data, while a deductive approach based on the constructs of NPT was used to map suggestions for curricula improvement to relevant NPT constructs. To overcome potential interpretive bias and selective perception, coding was conducted by 2 researchers independently. After the initial coding, discrepancies were discussed, and a consensus was reached to refine the codebook and ensure consistency in the application of codes. To enhance credibility and trustworthiness, data were triangulated by comparing the findings across participants from various public health care clusters to identify any consistencies and divergences in opinions. This helped to ensure that the themes captured diverse perspectives and were not unduly influenced by any single group.

In addition, we contextualized the findings by examining studies from other high-income countries undergoing similar digital transformations in health care. Furthermore, we analyzed recently published data reflecting the perspectives of other stakeholders in the health care industry, such as clinical educators and leaders of medical schools, regarding the digital competencies required for future clinical practice [8,21]. In the reporting of findings, we followed the Standards for Reporting Qualitative Research of O'Brien et al [23].

Ethical Considerations

This study was classified as a quality improvement (QI) project focusing on medical education curricula by the Research Integrity, Compliance, and Ethics (RICE) committee of SingHealth. In line with institutional guidelines, QI projects aimed at enhancing existing practices, processes, or programs, such as curriculum development in medical education, do not meet the criteria for human subjects research. As such, the study

was granted an ethical waiver by the SingHealth Centralized Institutional Review Board (2020/2880). This decision was based on the determination that the activities involved posed no more than minimal risk to participants. Despite this waiver, the research adhered strictly to the ethical principles outlined in the World Medical Association's Declaration of Helsinki and institutional guidelines.

Results

A total of 33 participants took part in the study. They included 19 chief medical officers from local public health care institutions, 3 chief executive officers from public health care clusters, 4 senior administrators, and 7 former organizational leaders. Each had at least 5 years of organizational leadership experience and represented various specialties (Table 1).

Participants shared that local medical schools have not yet revamped the curricula to incorporate relevant competencies for the digital age. They identified 6 reasons for the lag in DHC training, some of which extended beyond the medical schools. The analysis of codes, along with the generation of subthemes and themes, is summarized in Table 2. Illustrative quotes from the interviews are provided below.

Table 1. Demographics of participants (N=33).

Characteristics	Participants
Age (years)	
Mean	62
Median	60
Minimum age	44
Maximum age	82
Gender, n (%)	
Male	31 (94)
Female	2 (6)
Years in organizational leadership	
Mean	18.7
Median	18
Discipline, n (%)	
Gastroenterology and hepatology	5 (15.2)
Pediatrics (including pediatrics genetics, pediatric emergency medicine, and pediatric gastroenterology)	4 (12)
General surgery	3 (9.1)
Psychiatry	3 (9.1)
Renal medicine	3 (9.1)
Anesthesiology	2 (6.1)
Geriatric medicine	2 (6.1)
Respiratory medicine	2 (6.1)
Cardiology	2 (6.1)
Orthopedic surgery	2 (6.1)
General medicine	1 (3)
Medical oncology	1 (3)
Ophthalmology	1 (3)
Surgery and urology	1 (3)
Hand and reconstructive microsurgery	1 (3)

Table 2. Codes, subthemes, and themes identified from the coding process.

Codes	Subthemes	Themes
• Lack of time. • Hard to change. • Resistance. • Not open to new technologies. • Not willing to try new technologies. • Academics have to be open.	• Packed curriculum. • Preference for status quo. • Traditional mindset of senior clinicians and faculty.	• Bureaucratic inertia.
• Lack of alternative career pathways. • Lack of role models. • Mindset changes needed.	• Expectations for graduates to become doctors with patient-fronting roles.	• Limited opportunities to pursue traditional career pathways.
• Safe. • Safe sandbox. • Safety nets. • Patient safety. • Safe and creative space. • Nurture and protect. • Talk about the pitfalls and dangers of using technology.	• Lack of safety mechanisms to use DT[a] for educational purposes. • Limited opportunities to experiment with new technologies due to lack of creative space.	• Lack of protective mechanisms for experiential learning and experimentation.
• Clear guidelines. • Clear policies. • Clear intent. • Clear boundaries. • Help students navigate data, fake news, and misinformation. • Data abuse. • Medical ethics. • Respect privacy. • Ethical competency. • Schools presume these (ethical competencies) are common sense.	• Gaps in outlining guidelines and boundaries for technology use. • Gaps in teaching students the pitfalls of using technologies for clinical practice. • Gaps in equipping students with skills in handling data, medical information, and patients' privacy.	• Lack of clear policies and guidelines for clinical practice.
• Interface. • Incorporate teaching facilities within health care institutions. • Correlate. • String information.	• Limited integration of educational and research facilities for medical students within clinical settings. • Lack of feedback on students' performance outcomes. • Lack of compatible data encountered in medical school and residency.	• Lack of integration between medical school education and experience in the health care system.
• Gap between IT and health care. • Nonintegration. • Disorganized. • Slave to the system. • Need to redesign the system. • Put up robust systems. • Involve IT experts. • Facilitating platforms. • Support end users. • Internet separation.	• Health care industry should drive the IT industry.	• Lack of IT integration within the health care industry.

[a]DT: digital technology.

Bureaucratic Inertia

Participants suggested that bureaucratic inertia within both the health care system and medical schools contributed to sporadic and limited training in DT. They attributed this inertia to faculty members' lack of awareness regarding the evolution of clinical practice, their limited expertise in DT, and their resistance to incorporating new competencies, which would require sacrificing some traditional areas of expertise. As shared by OL8 and OL26:

There are senior clinicians who may not be so open to using DT. They are not willing to use different methodologies to solve the same problem. [OL8, Internal medicine, and Respiratory and Critical Care Medicine]

I tried to teach ultrasound in a medical school but with limited success… Unfortunately, it was met with great resistance from people who are traditional. [OL26, Cardiology]

Furthermore, participants perceived that policy makers and senior clinicians were hesitant to invest in DT due to concerns over higher health care costs, further hindering efforts to optimize DT in clinical settings. This perspective is illustrated by the following comment:

> *Some new technologies are almost invariably more expensive and will increase the cost of care.* [OL4, General Surgery]

The above excerpts highlighted systemic barriers to the integration of DT in clinical practice and medical education, emphasizing how institutional inertia and hesitation to invest in new technologies are contributing to the stagnation in clinical training and practice. The reluctance of policy makers and leaders to embrace change and allocate resources for DT exacerbates these challenges, ultimately hindering the evolution of medical education.

Lack of Opportunities to Pursue Nontraditional Career Pathways

Participants also identified limited opportunities to pursue alternative career paths and nonclinical roles, as well as the absence of role models in new technology fields, as significant barriers to implementation. As opined by OL26:

> *I've seen promising students and residents fall through the cracks and give up along the way because we don't have enough career pathways and role models for those in the medical innovation track.* [OL26, Cardiology]

In addition, OL8 highlighted the stigma within the medical community, where students who left medical school to explore nontraditional pathways were often perceived as failures.

> *We lack the definition of what kind of medical graduates we want to train. Other than basic clinical knowledge, I don't think we have defined anything further than that, like a clinician with knowledge of innovation. If a student decides to be an entrepreneur, for example, create a new start-up and drop out of medical school, we should still take that as a success and not a failure.* [OL8, Internal Medicine and Respiratory and Critical Care Medicine]

Without embracing alternative career paths and addressing the stigma associated with leaving traditional medical roles, the health care system risks alienating promising talent and limiting progress in medical innovation. Establishing clear pathways and celebrating diverse career outcomes is essential to cultivating a dynamic and adaptable health care profession.

Lack of Protective Mechanisms for Experiential Learning and Experimentation

In addition, participants noted limited protective mechanisms for experiential learning and experimentation in the health care system. The lack of a "safe and creative space" hindered trainees from engaging in innovative and secure experimentation with DT. Some participants proposed establishing sandboxes where trainees could test ideas with safeguards in place. This would enable them to contribute to clinical practice improvements while receiving proper guidance when mistakes occur. As articulated by OL8 and OL9:

> *The senior clinicians may not be so open to new things. As health care leaders ourselves, we need to embrace the idea of creating a safe sandbox where students [are] allowed to use their imagination to innovate, with all the safety nets in check for patient safety.* [OL8, Internal Medicine, and Respiratory and Critical Care Medicine]

> *What's lacking is a safe space for students and residents. A safe space is a space that offers professional, psychological, and personal safety for them. Measures need to be taken to train, nurture, and protect them rather than condemn them when they do something wrong. The health care system should give them that safe and creative space that ensures they are not bullied, harassed, and ridiculed.* [OL9, Anesthesiology]

Without the establishment of structured and supportive environments for experiential learning, the health care system risks stifling innovation and deterring the next generation of clinicians from engaging with DT. Proactively establishing protected and guided learning environments is essential for fostering a culture of experimentation and ensuring meaningful contributions to clinical advancements.

Lack of Clear Policies to Guide DT Integration in Clinical Practice

Another significant barrier articulated by participants was the lack of clear policies to guide the effective integration of DT in clinical practice. They emphasized the need for well-defined guidelines at both institutional and ministerial levels to support the ethical and professional use of DT. As noted by OL11:

> *The policies that govern digital technologies like telemedicine must be reasonable. Currently, the intent is unclear. At the institutional and ministerial level, there must be clear guidelines and policies that outline the learning and growth in the use of these technologies.* [OL11, Geriatric Medicine]

The lack of comprehensive policies limits awareness of the risks, pitfalls, and ethical considerations associated with DT, deterring its use, particularly among students. OL25 elaborated on the importance of training students in ethics and professionalism to prevent potential misuse of data.

> *In the world of AI and digital medicine, the role of ethics and professionalism are going to be even more important because it opens up easy channels to data abuse, and doctors will have so much data in their hands. So, you need to teach the students medical ethics and values related to patient information and treatment prescription. It's going to be so critical you need to enforce that.* [OL25, Medical Oncology]

The absence of clear, comprehensive policies to govern the use of DT in health care creates ethical and professional ambiguities, deterring adoption and proper training. Establishing well-defined guidelines is critical to mitigating risks, ensuring ethical use,

and preparing future clinicians to navigate the complexities of digital medicine responsibly.

Lack of Integration Between Medical School Education and Clinical Experience

Participants shared that the perceived lack of integration between medical school education and students' clinical experience in the health care system is another barrier to DHE. They attributed this gap to the lack of systems interoperability, which prevents students from accessing and using health care data used in clinical settings and receiving feedback from these systems. As one participant explained, a more integrated system would allow student performance data to correlate with hospital data, enabling continuous feedback and supporting learners' improvement:

> *The biggest gap is that we don't know how students are performing. The data that students are trained for should be similar to the place of practice. If the system is built such that medical school data correlates with say, hospital data, I can string all the information about your learning journey and see how that impacts your performance outcome. From that perspective, we can support the learners better because we give them an environment where they are constantly receiving feedback from the system and seeking new ways to improve themselves. I think that will probably be the most meaningful thing for our learners.* [OL15, Psychiatry]

More broadly, participants noted that the lack of integration between educational and research facilities within health care institutions limits students' clinical immersion. According to OL16, closer collaboration between medical schools and health care institutions is essential for strengthening this connection and enhancing experience, not just physically but through more active interaction between the institutions and health care professionals.

> *I think medical schools should be in the health care institutions. They should interface very closely. One way is to incorporate teaching and research facilities within health care institutions so that the immersion is useful. Currently, our medical schools are within the proximity of the hospital campus. It makes sense, but that's just the physical infrastructure. The people need to be interfaced quite a fair bit.* [OL16, Psychiatry]

The fragmented nature of medical education and clinical training suggests that a more integrated approach, leveraging data-driven feedback mechanisms and collaborative partnerships between academic and health care institutions, is necessary to foster a culture of continuous learning and improvement in health care.

Lack of IT Integration Within the Health Care Industry

Participants also suggested that an integrated IT infrastructure in health care institutions would increase DHE effectiveness and enhance clinical care. However, they highlighted the current lack of interoperability between systems, which hinders the optimization of technical needs. A recurring concern was the IT sector's lack of ability to understand and address the specific needs of health care, with participants noting disorganization and a disconnect between IT and health care practices. As one participant expressed:

> *The gap between the IT and health care industry has not been bridged yet. We have a lot of IT in the health care industry, but a lot of it is record-keeping. It does not integrate [and] information is coming from every direction that is totally disorganized. How, then, can we teach our medical students to be responsible for the patient as a whole? Somebody who has the ability to do IT programming has to follow the doctors on their rounds. I've ever asked my IT colleagues, "Look, is this an IT industry or a health care industry? When they said it's a health care industry, I said, okay, then you have to listen to me and make things work for me, not enslave me to your products."* [OL9, Anesthesiology]

Furthermore, participants emphasized the challenge of internet separation and the need for platforms that allow seamless cross-sharing of information, which they identified as crucial for effective learning environments. As shared by OL12:

> *One of the biggest challenges is Internet separation…The availability and cross-sharing of information are all important facilitating platforms that we have to provide for medical students.* [OL12, Pediatrics]

The lack of integrated IT infrastructure and disjointed systems within health care settings creates significant barriers to enhancing DHE and clinical care, often leaving medical practitioners frustrated with ineffective solutions. To bridge the gap between IT and health care, a more tailored approach is needed, where technological systems are designed to directly support clinical workflows, ensuring both efficiency and improved educational outcomes.

Discussion

Principal Findings

By interviewing doctors in organizational leadership, we gained insider perspectives on gaps in both the medical curricula and the health care system. A total of 6 barriers were identified: bureaucratic inertia, lack of opportunities to pursue nontraditional career paths, limited protective mechanisms for experiential learning, unclear policy guidelines, limited integration between education and clinical experience, and IT integration issues. The findings contributed to the existing literature by showing that DHE barriers were not limited to medical school curricula but involved broader systemic issues. Comprehensive strategies were needed to address these challenges.

By using qualitative interviews, our study uncovered nuances in leadership decision-making that are often missed in quantitative surveys, providing a richer understanding of the factors influencing leadership perspectives. While most studies suggest that organizational leaders prioritize efficiency and sustainability [16-18], our findings reveal that leaders in this

context place a higher emphasis on experimentation and innovation, a factor not traditionally associated with corporate leadership. Furthermore, our research also highlights the growing influence of digital transformation on leadership styles, an area that received limited attention in previous studies focused on traditional management structures. It underscores the importance of adaptive leadership in an era of constant change, suggesting the need for leadership training programs that focus on flexibility.

Many of the barriers identified in this study align with findings from other high-income countries. These include the lack of the necessary information and communication technology (ICT) skills and limited awareness of the potential benefits of DT among some clinicians. For example, in Germany, an empirical study by Ernstmann et al [24] revealed that some primary care doctors perceived eHealth cards as less useful due to their limited ICT expertise and lack of involvement in technological development. These eHealth cards, which store medical data, treatment plans, medications, and electronic patient files, rely on a telematics infrastructure for communication [24]. The study recognized that without robust IT support, comprehensive training for medical professionals, and a standardized national implementation procedure, the acceptance, adoption, and sustained use of eHealth technology by doctors are likely to be hindered [25].

In addition, other studies have shown an increasing proportion of medical school graduates pursuing careers outside full-time clinical practice in some countries [26]. However, findings from countries such as the United States and South Korea indicate that medical school curricula often fail to adequately address the need for programs providing information on nontraditional careers or nonclinical career pathways [27,28]. Despite expressed interest in these career options, medical students often lack awareness of available training opportunities. To attract students to such careers, early outreach programs, combined with appropriate indemnity and support for innovative projects, are essential. These initiatives could be implemented through elective classes, incentives from professional societies, or partnerships with experts [27].

Furthermore, research from countries such as Canada and Taiwan highlights how technological tools can be leveraged to foster experiential learning among medical students. At the University of Ottawa, social accountability experiential logs were developed for third-year medical students to address the social determinants of health, which are often overlooked in clinical learning objectives [29]. These logs guided students in reflecting on clinical encounters and targeting psychosocial skill development, improving clinical confidence, and demonstrating adaptability for other medical schools (Fung et al [29]). Similarly, a Taiwanese study by Liao et al [30] showcased how the mPath (KU Leuven) e-learning tool supported communication skills training by providing a flexible, technology-enhanced learning environment [30]. Features such as remote accessibility, session recordings, peer feedback mechanisms, and visualized analytical reports enabled learners to engage in self-reflection, adapt communication strategies, and enhance subverbal communication skills [30]. Together, these initiatives exemplify how experiential learning tools can address both biomedical and psychosocial challenges in medical education.

The lack of clear laws and policies to guide DT integration in clinical practice is also a barrier in other high-income countries. For instance, health care leaders in Sweden have acknowledged the need for updated policies [16]. They noted that existing laws and regulations have not kept pace with rapid technological advancements and the evolving organization of health care. These policies require revision to ensure clarity regarding liability and accountability, particularly in addressing how errors are managed when artificial intelligence (AI) systems play a role in clinical decision-making [16].

Furthermore, the limited integration between medical education and clinical experience has been highlighted in various studies and reviews. For instance, Pereira et al [31] describe the implementation of a single competency-based Epic onboarding process for medical students in certain US medical schools with rotations across multisystem training sites. This initiative has enabled learners to spend more time in clinical settings with optimized access to electronic health records (EHRs) [31]. While this approach reduces the training burden, curricula could be further enhanced by emphasizing the practical application of EHRs in clinical settings. This includes training students to maintain professionalism and establish rapport with patients while using EHR systems [31]. In addition, Chan and Zary [32] emphasize that providing immediate and formative feedback on students' performance can support the effective use of AI in medical education. However, delivering high-quality feedback in clinical contexts remains a challenge, as it depends on the underlying knowledge base and model of the AI system, which still requires refinement [32].

Previous systematic reviews have consistently identified infrastructure and technical barriers as the most frequently cited barriers to technology integration in health care [33]. These challenges include limitations in health care capacity for technology adoption, inadequate interconnectedness, insufficient network resources, and incompatibility with existing daily workflows [33]. Addressing these barriers requires the active involvement of health care professionals in the development and implementation of health technology tools, which can also enhance their capacity to effectively manage such applications. Furthermore, the reviews emphasize the critical importance of user engagement and collaboration with system developers throughout all phases of design, development, deployment, and continued use [33]. This collaborative approach ensures that the applications are fit for purpose, as they are designed to align with and address health care providers' needs and expectations.

Our findings highlighted structural and bureaucratic barriers beyond medical schools that hindered DHE implementation. Although they are common in high-income countries, no comprehensive framework has been proposed to address them to date. This study applies May and Finch's [15] NPT to suggest ways to bridge these gaps. A summary of how the 4 constructs of NPT can be applied to each of these barriers is found in Table 3.

Table 3. Addressing each identified barrier with the Normalization Process Theory (NPT).

Barriers	NPT contributions
Bureaucratic inertia	• Coherence: enhance understanding and sense-making among stakeholders about the importance and benefits of DHC[a]. This could be achieved by hiring prospective faculty with the skill sets that are relevant to the needs of up-and-coming developments in medicine. • Cognitive participation: engage key stakeholders to foster buy-in and commitment. For example, leaders of medical schools can engage individuals with influence to encourage the integration of digitalization in the core curriculum. These include engaging clinical educators, teachers, and innovators trained in DT[b] in knowledge exchange and talking with the faculty to facilitate the training of DHC and keep them abreast of the latest technological developments in clinical settings. • Collective action: develop strategies to streamline decision-making processes and reduce red tape. • Reflexive monitoring: continuously evaluate and adjust strategies to address bureaucratic resistance and demonstrate early successes to build momentum.
Lack of opportunities to pursue nontraditional career pathways	• Coherence: clarify the relevance of DHC to future career opportunities and the evolving landscape of health care. • Cognitive participation: involve influential faculty and practitioners, such as medical innovators, in promoting the value of alternative career pathways. Medical schools should also provide sufficient training and mentoring opportunities for students who wish to pursue alternative career pathways. • Sufficient resources should be invested in implementing a curriculum that provides students with opportunities to diversify their skill sets, such as skills in clinical informatics relevant to clinical practice. It should include collaborative mentorship where students can explore new fields in DT by forming partnerships with experts from both the clinical and nonclinical fields. • Collective action: integrate DHC into career development programs and highlight role models who have successfully incorporated digital skills. • Reflexive monitoring: gather feedback from students and professionals to continually refine the approach and address concerns about career impact. Relevant recognition should also be given to medical graduates who embark on alternative pathways to encourage the growth of the fields and normalize these pathways for them.
Lack of protective mechanisms for experiential learning and experimentation	• Coherence: emphasize the importance of experiential learning for mastering DHC. • Cognitive participation: foster a culture of experimentation and learning by involving faculty in the design and delivery of experiential learning opportunities. • Collective action: develop and implement policies and resources that support protected time and space for experiential learning and innovation. These include creating more sandboxes and expanding reasonable access to EHRs in clinical settings. • Reflexive monitoring: continuously assess and improve experiential learning programs based on feedback and outcomes.
Lack of clear policy guidelines for clinical practice	• Coherence: clearly articulate the need for and benefits of standardized DHC policies. • Cognitive participation: engage policy makers, clinical leaders, and educators in developing and endorsing clear guidelines. To ensure that the threat of litigation does not hinder technological adoption, professional bodies should establish clear policies that regulate the effective implementation of DT in clinical settings. A technology assessment committee could also be set up to develop guidelines that enable young trainees to use DT effectively and ethically, both for their safety as well as for their patients. • Collective action: implement training and support systems to ensure consistent application of policies across clinical settings. Professional bodies should also work with schools to equip students with knowledge of cybersecurity as well as the limitations and pitfalls of using DT in various circumstances. • Reflexive monitoring: regularly review and update policies based on clinical practice feedback and emerging best practices. Dedicating time to reflect on what can be improved along the way would be a crucial step for schools.
Insufficient integration between medical school education and clinical experience	• Coherence: highlight the importance of integrating DHC across the continuum of medical education. • Cognitive participation: involve both academic and clinical faculty in designing integrated curricula that seamlessly blend theory and practice. • Collective action: develop joint academic-clinical initiatives and placements that reinforce DHC training in real-world settings. • Reflexive monitoring: evaluate the effectiveness of integrated programs and make adjustments to enhance alignment between education and practice.
Limited IT integration within the health care industry	• Coherence: communicate the critical role of IT in supporting DHC and improving health care outcomes. • Cognitive participation: collaborate with IT professionals and health care administrators to prioritize IT integration. To ensure that digital health care technologies can be used safely and effectively by clinicians, new technology or equipment introduced for clinical practice needs to be installed by IT personnel with knowledge of the health care system and with input from health care professionals so that the latter's needs are met. • Collective action: advocate for investments in IT infrastructure and training to support DHC initiatives. At the national level, a move towards interoperability of systems that allow users to share data would also facilitate students' adaptation to new systems in different health care settings. • Reflexive monitoring: continuously assess the state of IT integration and address gaps through ongoing improvement efforts.

[a]DHC: digital health competencies.

[b]DT: digital technology.

Limitations of the Study

This study has several limitations that should be acknowledged. First, the focus on the perspectives of organizational leaders may not fully represent the experiences of frontline educators or students, limiting the generalizability of the findings. Furthermore, interviewing organizational leaders may introduce a bias toward presenting their organizations in a favorable light. They may be reluctant to express views that could be perceived as critical of their organizations. This concern may stem from the constraints they feel due to their roles or the public image of their organizations. As a result, their responses might reflect a more measured or politically cautious perspective. To address this, we incorporated triangulation by cross-referencing their responses with published articles on similar topics. This approach provided a more balanced perspective, though we recognize the inherent limitations in capturing the full organizational dynamics.

Second, the relatively small sample size, while sufficient to achieve thematic saturation, may constrain the breadth of insights. We also recognize that the unique sociopolitical, cultural, and economic context of Singapore may limit the generalizability of our findings to other settings. Singapore's centralized governance and relatively small population create conditions that may differ from other countries. Consequently, while the insights from our study provide valuable lessons, they should be interpreted with caution when applying them to contexts with different governance structures or cultural dynamics.

The third limitation was the gender imbalance among the organizational leaders interviewed, with 94% (31/33) male and only 6% (2/33) female participants. While this reflects the current leadership demographics within public health care institutions, the barriers and challenges identified in our research are rooted in institutional and structural factors rather than individual-level or gender-specific experiences. As such, we do not expect that the gender distribution significantly influenced the findings. However, future research could benefit from a more gender-diverse sample to explore whether different leadership perspectives might offer additional insights or nuances.

Another limitation of our study is the use of Zoom for interviews, which was necessitated by the COVID-19 pandemic and which might have influenced the depth and dynamics of the discussions compared to in-person interviews. In face-to-face settings, nonverbal cues such as body language, eye contact, and physical proximity play a significant role in building rapport and fostering a more comfortable environment for in-depth conversations. These subtle cues can often provide valuable insights into a participant's emotional state, engagement level, and willingness to share more personal or sensitive information. Nonetheless, the insights obtained through Zoom still offer valuable contributions to understanding the barriers to DHE integration.

In addition, we acknowledge that the NPT's focus on individual experiences may not fully capture the diversity of perspectives of multiple stakeholders. To address this, we triangulated our data by comparing the findings across participants from various public health care clusters to identify any consistencies and divergences in opinions. This helped to ensure that the normalization process was not unduly influenced by any single group. By addressing these challenges, we believe our study provides a more nuanced understanding of NPT, particularly in contexts where contextual variations and diverse stakeholder groups are at play. These adaptations strengthen the applicability of NPT and offer valuable insights for its broader use in similar settings.

While we have made considerable efforts to adapt NPT to our specific context, we recognize that there may still be limitations in generalizing our findings across very different settings. Thus, future research should explore how NPT applies in more varied environments with larger sample sizes to further validate our findings. Furthermore, given that normalization is a gradual process, further studies should also conduct longitudinal follow-up assessments to monitor changes over time.

Strengths of the Study

This qualitative study informs us about the institutional and structural barriers present in Singapore's medical school curricula. The diverse sample of this study, spanning various health care institutions and specialties, yielded rich data. Participants possessed extensive organizational leadership experience and were attuned to the needs of contemporary clinical practice. Unlike previous research focusing mainly on institutional inertia and pedagogical strategies [5,34-37], this study uncovered structural barriers as well.

While findings may seem limited to Singapore's context, applying relevant NPT constructs could render results applicable globally since many other high-income countries faced similar challenges in technological development and curriculum digitalization [3,12,38]. Furthermore, the identified barriers necessitated universal solutions extending beyond Singapore.

A potential line of future research would be to gather the views of medical innovators and entrepreneurs to explore other barriers to the effective adoption of DT in health care institutions. Another area would be to evaluate the ways in which DHC training among medical trainees and graduates influences the efficiency and cost-effectiveness of health care delivery. This research could provide valuable insights into how DHC in medical education affects not only the preparedness of new health care professionals but also the overall performance of health care organizations.

Conclusions

Focusing on the perspectives of doctors in organizational leadership roles provides a comprehensive understanding of the barriers to incorporating DHE into Singapore's medical curricula. Their strategic insight, policy influence, experience with system-wide challenges, understanding of the

education-practice gap, resource management capabilities, and expertise in innovation and change management are invaluable for developing practical, effective, and sustainable strategies to address these barriers.

Unlike previous studies focusing solely on gaps within schools, our findings underscored the importance of collaborations with professional bodies and health care institutions to overcome various barriers. By applying NPT, this study provides a structured approach to understanding and overcoming the barriers. It offers a roadmap for other countries facing similar challenges in DHE. However, NPT should be seen as adaptable, requiring regular reevaluation to accommodate dynamic changes in the field.

Acknowledgments

This study was funded by SingHealth Duke-NUS Medicine Academic Clinical Programme under Seah Cheng Siang Distinguished Professorship in Medicine. The authors would like to thank the participants of this study for their invaluable insights and the reviewers of this journal for their helpful comments. They are also grateful to Associate Professor Warren Fong Weng Sen for his helpful feedback on the draft of this manuscript.

Conflicts of Interest

None declared.

References

1. Gopal G, Suter-Crazzolara C, Toldo L, Eberhardt W. Digital transformation in healthcare - architectures of present and future information technologies. Clin Chem Lab Med 2019;57(3):328-335 [FREE Full text] [doi: 10.1515/cclm-2018-0658] [Medline: 30530878]
2. Aungst TD, Patel R. Integrating digital health into the curriculum-considerations on the current landscape and future developments. J Med Educ Curric Dev 2020;7:2382120519901275 [FREE Full text] [doi: 10.1177/2382120519901275] [Medline: 32010795]
3. Edirippulige S, Brooks P, Carati C, Wade V, Smith A, Wickramasinghe S, et al. It's important, but not important enough: eHealth as a curriculum priority in medical education in Australia. J Telemed Telecare 2018 Dec;24(10):697-702 [FREE Full text] [doi: 10.1177/1357633X18793282] [Medline: 30343657]
4. Khurana M. Keeping Pace: the need for digital health education in medical schools. Acad Med 2020;95(11):1629-1630 [FREE Full text] [doi: 10.1097/ACM.0000000000003672] [Medline: 33109967]
5. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019;5(2):e16048 [FREE Full text] [doi: 10.2196/16048] [Medline: 31793895]
6. Tudor Car L, Kyaw BM, Nannan Panday RS, van der Kleij R, Chavannes N, Majeed A, et al. Digital health training programs for medical students: scoping review. JMIR Med Educ 2021;7(3):e28275 [FREE Full text] [doi: 10.2196/28275] [Medline: 34287206]
7. Zainal H, Xiaohui X, Thumboo J, Yong FK. Exploring the views of Singapore junior doctors on medical curricula for the digital age: a case study. PLoS One 2023;18(3):e0281108 [FREE Full text] [doi: 10.1371/journal.pone.0281108] [Medline: 36862708]
8. Zainal H, Xiaohui X, Thumboo J, Kok Yong F. Digital competencies for Singapore's national medical school curriculum: a qualitative study. Med Educ Online 2023 Dec;28(1):2211820 [FREE Full text] [doi: 10.1080/10872981.2023.2211820] [Medline: 37186901]
9. Gillissen A, Kochanek T, Zupanic M, Ehlers J. Medical students' perceptions towards digitization and artificial intelligence: a mixed-methods study. Healthcare (Basel) 2022 Apr 13;10(4):A [FREE Full text] [doi: 10.3390/healthcare10040723] [Medline: 35455898]
10. Machleid F, Kaczmarczyk R, Johann D, Balčiūnas J, Atienza-Carbonell B, von Maltzahn F, et al. Perceptions of digital health education among European medical students: mixed methods survey. J Med Internet Res 2020;22(8):e19827 [FREE Full text] [doi: 10.2196/19827] [Medline: 32667899]
11. Sit C, Srinivasan R, Amlani A, Muthuswamy K, Azam A, Monzon L, et al. Attitudes and perceptions of UK medical students towards artificial intelligence and radiology: a multicentre survey. Insights Imaging 2020;11(1):14 [FREE Full text] [doi: 10.1186/s13244-019-0830-7] [Medline: 32025951]
12. Sorg H, Ehlers JP, Sorg CGG. Digitalization in medicine: are German medical students well prepared for the future? Int J Environ Res Public Health 2022;19(14):8308 [FREE Full text] [doi: 10.3390/ijerph19148308] [Medline: 35886156]
13. Chen M, Safdar N, Nagy P. Should medical schools incorporate formal training in informatics? J Digit Imaging 2011;24(1):1-5 [FREE Full text] [doi: 10.1007/s10278-009-9249-x] [Medline: 19908095]
14. Hurley KF, Taylor B, Postuma P, Grace P. What are Canadian medical students learning about health informatics. eJHI 2011;6 [FREE Full text]
15. May C, Finch T. Implementing, embedding, and integrating practices: an outline of normalization process theory. Sociology 2009 Jun 15;43(3):535-554 [FREE Full text] [doi: 10.1177/0038038509103208]

16. Petersson L, Larsson I, Nygren J, Nilsen P, Neher M, Reed JE, et al. Challenges to implementing artificial intelligence in healthcare: a qualitative interview study with healthcare leaders in Sweden. BMC Health Serv Res 2022 Jul 01;22(1):850 [FREE Full text] [doi: 10.1186/s12913-022-08215-8] [Medline: 35778736]
17. Enticott J, Braaf S, Johnson A, Jones A, Teede H. Leaders' perspectives on learning health systems: a qualitative study. BMC Health Serv Res 2020;20(1):1087 [FREE Full text] [doi: 10.1186/s12913-020-05924-w] [Medline: 33243214]
18. Alanazi A. Digital leadership: attributes of modern healthcare leaders. Cureus 2022;14(2):e21969 [FREE Full text] [doi: 10.7759/cureus.21969] [Medline: 35282530]
19. Kong Chiang L. FAQs. School of Medicine NTU. URL: https://www.ntu.edu.sg/medicine/about-us/faqs#Content_C020_Col00 [accessed 2022-12-22]
20. Samarasekera D, Ooi S, Yeo S, Hooi S. Medical education in Singapore. Med Teach 2015 Aug;37(8):707-713 [FREE Full text] [doi: 10.3109/0142159X.2015.1009026] [Medline: 25693792]
21. Zainal H, Xin X, Thumboo J, Fong KY. Medical school curriculum in the digital age: perspectives of clinical educators and teachers. BMC Med Educ 2022;22(1):428 [FREE Full text] [doi: 10.1186/s12909-022-03454-z] [Medline: 35659212]
22. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol Jan 2006;3(2):77-101 [FREE Full text] [doi: 10.1191/1478088706qp063oa]
23. O'Brien BC, Harris I, Beckman T, Reed D, Cook D. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014;89(9):1245-1251 [FREE Full text] [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
24. Ernstmann N, Ommen O, Neumann M, Hammer A, Voltz R, Pfaff H. Primary care physician's attitude towards the German e-health card project--determinants and implications. J Med Syst 2009;33(3):181-188 [FREE Full text] [doi: 10.1007/s10916-008-9178-0] [Medline: 19408451]
25. Safi S, Thiessen T, Schmailzl KJ. Acceptance and resistance of new digital technologies in medicine: qualitative study. JMIR Res Protoc 2018;7(12):e11072 [FREE Full text] [doi: 10.2196/11072] [Medline: 30514693]
26. Jeffe DB, Andriole DA, Hageman HL, Whelan AJ. The changing paradigm of contemporary U.S. allopathic medical school graduates' career paths: analysis of the 1997-2004 national AAMC graduation questionnaire database. Acad Med 2007;82(9):888-894 [FREE Full text] [doi: 10.1097/ACM.0b013e31812f797e] [Medline: 17726402]
27. Banerjee R, George P, Priebe C, Alper E. Medical student awareness of and interest in clinical informatics. JAMIA Internet 2015;22(1):e42-e47 [FREE Full text] [doi: 10.1093/jamia/ocu046] [Medline: 25726567]
28. Kim KJ, Park JH, Lee YH, Choi K. What is different about medical students interested in non-clinical careers? BMC Med Educ 2013;13:81 [FREE Full text] [doi: 10.1186/1472-6920-13-81] [Medline: 23731551]
29. Fung OW, Mulholland A, Bondy M, Driedger M, Kendall CE. Implementing experiential learning logs addressing social accountability into undergraduate medical clerkship education. Can Med Educ J 2023;14(2):146-149 [FREE Full text] [doi: 10.36834/cmej.73907] [Medline: 37304626]
30. Liao F, Murphy D, Wu JC, Chen CY, Chang CC, Tsai PF. How technology-enhanced experiential e-learning can facilitate the development of person-centred communication skills online for health-care students: a qualitative study. BMC Med Educ 2022;22(1):60 [FREE Full text] [doi: 10.1186/s12909-022-03127-x] [Medline: 35078482]
31. Pereira AG, Kim M, Seywerd M, Nesbitt B, Pitt MB, Minnesota Epic101 Collaborative. Collaborating for competency-a model for single electronic health record onboarding for medical students rotating among separate health systems. Appl Clin Inform 2018;9(1):199-204 [FREE Full text] [doi: 10.1055/s-0038-1635096] [Medline: 29564849]
32. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. JMIR Med Educ 2019;5(1):e13930 [FREE Full text] [doi: 10.2196/13930] [Medline: 31199295]
33. Borges do Nascimento IJ, Abdulazeem H, Vasanthan LT, Martinez EZ, Zucoloto ML, Østengaard L, et al. Barriers and facilitators to utilizing digital health technologies by healthcare professionals. NPJ Digit Med 2023;6(1):161 [FREE Full text] [doi: 10.1038/s41746-023-00899-4] [Medline: 37723240]
34. Foadi N, Koop C, Mikuteit M, Paulmann V, Steffens S, Behrends M. Defining learning outcomes as a prerequisite of implementing a longitudinal and transdisciplinary curriculum with regard to digital competencies at hannover medical school. J Med Educ Curric Dev 2021;8:23821205211028347 [FREE Full text] [doi: 10.1177/23821205211028347] [Medline: 34368455]
35. Behrends M, Steffens S, Marschollek M. The implementation of medical informatics in the national competence based catalogue of learning objectives for undergraduate medical education (NKLM). Stud Health Technol Inform 2017;243:18-22 [FREE Full text] [Medline: 28883161]
36. Hersh W, Biagioli F, Scholl G, Gold J, Mohan V, Kassakian S. From competencies to competence: model, approach, and lessons learned from implementing a clinical informatics curriculum for medical students. In: Health Professionals' Education in the Age of Clinical Information Systems, Mobile Computing and Social Networks. United States: Academic Press; 2017:269-287.
37. Haag M, Igel C, Fischer M, German Medical Education Society (GMA), Committee “Digitization – Technology-Assisted LearningTeaching”, Joint working group “Technology-enhanced TeachingLearning in Medicine (TeLL)” of the German Association for Medical Informatics, BiometryEpidemiology (gmds)the German Informatics Society (GI). Digital teaching

and digital medicine: a national initiative is needed. GMS J Med Educ 2018;35(3):Doc43 [FREE Full text] [doi: 10.3205/zma001189] [Medline: 30186953]
38. Cutrer WB, Spickard WA, Triola MM, Allen BL, Spell N, Herrine SK, et al. Exploiting the power of information in medical education. Med Teach 2021;43(sup2):S17-S24 [FREE Full text] [doi: 10.1080/0142159X.2021.1925234] [Medline: 34291714]

Abbreviations

AI: artificial intelligence
DHC: digital health competencies
DHE: digital health education
DT: digital technology
EHR: electronic health record
ICT: information and communication technology
LKCMedicine: Lee Kong Chian School of Medicine
NPT: Normalization Process Theory
NTU: Nanyang Technological University
NUS: National University of Singapore
OL: organizational leader
PI: principal investigator
QI: quality improvement
RICE: Research Integrity, Compliance, and Ethics
YLL: Yong Loo Lin School of Medicine

Edited by J Moen; submitted 25.07.24; peer-reviewed by J Walsh, SS Lee; comments to author 18.12.24; revised version received 29.12.24; accepted 20.01.25; published 07.03.25.

Please cite as:
Zainal H, Xiao Hui X, Thumboo J, Kok Yong F
Organizational Leaders' Views on Digital Health Competencies in Medical Education: Qualitative Semistructured Interview Study
JMIR Med Educ 2025;11:e64768
URL: https://mededu.jmir.org/2025/1/e64768
doi:10.2196/64768
PMID:40053774

Educational Effectiveness of a 5-Country Virtual Exchange Program for Internationalization in Occupational Therapy Education: Mixed Methods Study

Natsuka Suyama[1], MSci; Kaoru Inoue[1], PhD; Norikazu Kobayashi[1], PhD; Anuchart Kaunnil[2], PhD; Supatida Sorasak Siangchin[3], MSOT; Muhammad Hidayat Sahid[4], MEpid; Erayanti Saloko[5], PhD; Sk Moniruzzaman[6], MPH, MSc

[1]Department of Occupational Therapy, Graduate School of Human Health Sciences, Tokyo Metropolitan University, 7-2-10 Higashiogu, Arakawa-ku, Tokyo, Japan

[2]Department of Occupational Therapy, Faculty of Associated Medical Sciences, Chiang Mai University, Chiang Mai, Thailand

[3]Division of Occupational Therapy, Faculty of Physical Therapy, Mahidol University, Nakhon Pathom, Thailand

[4]Occupational Therapy Study Program, Department of Applied Health, Vocational Education Program, University of Indonesia, Depok, Indonesia

[5]Department of Occupational Therapy, The Health Polytechnic of Surakarta, Surakarta, Indonesia

[6]Department of Occupational Therapy, Bangladesh Health Professions Institute, Dhaka, Bangladesh

Corresponding Author:
Natsuka Suyama, MSci
Department of Occupational Therapy, Graduate School of Human Health Sciences, Tokyo Metropolitan University, 7-2-10 Higashiogu, Arakawa-ku, Tokyo, Japan

Abstract

Background: Global health care education that cultivates international orientation is important for providing medical care in consideration of diverse backgrounds and collaboration with foreign medical professionals. Virtual international exchange programs could be a new type of global education in the present postpandemic era.

Objective: This study aimed to examine the effectiveness of a virtual international exchange program in fostering quality academic and professional learning and international orientation from student perspectives across 5 countries. This research is expected to contribute to education for the development of global human resources in the health professions.

Methods: This quasi-experimental study used a before-and-after design using a convergent parallel mixed methods approach. In this study, a 5-day interactive virtual program was offered to occupational therapy students from Bangladesh, Indonesia, Japan, the Philippines, and Thailand. The students were asked about their expectations and international orientation before the program, and about their evaluation of the program and international orientation afterward. Numerical data from a questionnaire on program expectations and evaluations were analyzed using descriptive statistics. Data on international orientation were subjected to qualitative analysis using steps for coding and theorization.

Results: In total, 29 students participated in the program, out of which 12 students (response ratio 41.4%) answered the research questionnaires both before and after the program. Overall, the students' expectations of the program were met in terms of expertise, scientific learning skills, and group interactions. Comparing before and after the program, mean scores of how the program met expectations increased, and the mean scores after the program in all 12 items asking about program evaluation were from 3.8 (SD 1.19) to 4.9 (SD 0.67; range: score 1 [lowest]-5 [highest]). Even though their motivation for participating in the program was not specific before the program, after the program, they reported having a more concrete image and specific form of what they learned from an international perspective. The participants enjoyed communication with others from diverse backgrounds while recognizing the difficulty of understanding different values. They also expressed satisfaction with their understanding of occupational therapy professionals and diverse societies, including medical systems from other countries.

Conclusions: Even though the analyzed sample data were small, these findings suggest that the program in this study may provide the participants with valuable opportunities. The virtual exchange program could foster students to cultivate qualities such as problem-finding or problem-solving and having interactions with groups from diverse backgrounds.

Trial Registration: UMIN Clinical Trials UMIN000059629; https://center6.umin.ac.jp/cgi-open-bin/ctr/ctr_view.cgi?recptno=R000061896

(JMIR Med Educ 2025;11:e77564) doi:10.2196/77564

KEYWORDS

virtual exchange program; occupational therapy education; global education; international orientation; intercultural competence; health professions education

Introduction

Background

As globalization has expanded, education for global human resources in medical education is increasingly being promoted. The significance of international exchange activities in professional fields beyond cultural exchange is also expected. In addition, there is a growing need in the rehabilitation profession to provide medical care to people with diverse backgrounds and collaborate with foreign medical professionals for the further development of the professional field. However, due to travel restrictions, international programs faced serious challenges during the COVID-2019 pandemic, and virtual learning was pursued as an alternative approach [1-3]. The objectives of international exchange programs include not only exchanging academic ideas but also gaining experience with interactive communication in different cultural contexts. Even when restrictions are placed on traveling abroad, educational institutions are expected to manage international academic and cultural programs by synchronizing online programs. During the COVID-19 pandemic, programs were conducted in which students joined from 2 or more universities and participated in activities together for several weeks [4], as well as a successful interactive program to learn professional knowledge and cultivate an international perspective [5,6]. As global health requires significant societal and pedagogical transformations regardless of physical mobility, virtual collaborative international learning has the potential to transform students in the health professions into global human resources. This approach can offer feasible, meaningful, and cost-effective solutions to students in the health professions, thereby enriching cultural competence and global understanding of health through virtual knowledge exchange [6,7].

However, in the present postpandemic era, the restarting of in-person educational programs that involve traveling abroad has created a new situation for virtual international exchange programs. The rapid development of IT during the COVID-19 pandemic has promoted learning knowledge that supports international collaboration toward addressing increasingly complex societal issues, and as such, higher education needs to leverage virtual education while addressing issues such as access, equity, cost, and ecology. In addition, virtual international education promotes internationalization at home (IaH) and provides more opportunities for students on campus to experience internationalization and develop an international orientation. IaH through virtual environments may provide benefits such as fostering a collaborative and diverse online community as a source of social and professional support and networking [8]. In globalization, virtual education may be a part of the essential skills and experiences necessary for generations familiar with and required to acquire IT skills and literacy. Students can feel more open to people from diverse backgrounds and international careers while becoming more familiar with different online technologies [8]. In the health professional education area, pharmacy students have learned about similarities and differences in socioeconomic determinants of health as well as the structure, functioning, and financing of different health care systems. Technology enables more students in diverse geographic locations to be exposed to various perspectives and health care experiences [9]. Student experiences of diversifying and further integrating using virtual platforms can help promote adaptation to global society, develop novel skills and knowledge, and contribute to future development in the medical field.

Global Human Resource Development in Virtual Education

Effective virtual international education should not be considered a replacement for traditional exchange programs that include travel abroad, but rather, a different educational method for increasing international orientation and acquiring the global qualities needed for IaH. Beelen et al [10] describe IaH as the purposeful integration of international and intercultural aspects into students in their home countries. With improvements in IT, virtual collaborative international learning can be effectively managed for interactive communication. For instance, in 2006, the State University of New York developed a collaborative online international learning (COIL) program to engage in an international experience as a virtual type of exchange education based on a collaborative and social constructivist learning approach [11]. As interest in COIL increased, it was regarded as not only an alternative method, but also a method that considered the carbon footprint and environmental impact of physical mobility in regard to air travel [12]. The COIL program offers students an authentic learning experience at their home institution and helps them develop intercultural competencies while serving as a relatively environmentally friendly, sustainable method to internationalize the curriculum. COIL has several characteristics that facilitate learning effectiveness [12]: collaboration between two or more educators from different institutions of higher education in different countries, the co-design and cofacilitation of a joint collaborative course by educators, an emphasis on learning intercultural content and developing collaboration skills through collaborative learning assignments, and several synchronized online meeting times during the course. In the medical professional education field, some studies have been conducted through the COIL program [13,14]. For instance, through the COIL program, nursing students had meaningful, valued engagement with peers in another country to prepare students for diverse, multicultural work settings for their professional futures. The COIL program also helped faculty members conceptualize lessons that promoted intercultural respect and appreciation using online learning methods [13]. Even though English was a second language, the program allowed nursing students to increase their intercultural sensitivity, improve their English proficiency, and obtain more confidence in their interactions and communication with individuals with different cultural backgrounds [15]. Moreover, COIL experiences opened

their eyes to a new way of thinking about preconceived ideas developed without actual knowledge of the situations outside their country, indicating that these cultural encounters led to better cultural awareness, humility, knowledge, skills, and desires [16].

Within virtual education, effectiveness should be judged in terms of not only understanding diverse cultures and increasing cultural awareness, but also developing global human resources. The concept of global human resources includes not only language and communication skills, but also initiative, positivity, a spirit of challenge, cooperativeness, and a sense of responsibility, as well as cross-cultural understanding, a broad range of education, deep expertise, problem-finding and -solving skills, teamwork and leadership, and a sense of ethics. These are not merely limited to an understanding of other cultures and language skills, but also correspond to the professionalism and qualities required of medical professionals. Therefore, cultivating global qualities contributes to not only internationalization but also professionalism in the development of students and institutions. In the medical area [17], the acquisition of necessary skill sets to foster globally competent dental students was enhanced through international virtual team-working, problem-solving, and person-centered multidisciplinary care planning activities. An online program fostering these qualities was beneficial for students from a broader global perspective and demonstrated an appreciation of the importance of delivering culturally sensitive person-centered dental care [17]. Furthermore, by combining virtual exchange and clinical simulation-based experiences for nursing students, the program was associated with statistically significant gains in the cultural intelligence of nursing students to function effectively in situations where cultural diversity was present, which was considered to have a positive impact on their future competence as global health care workers [18]. Another study found that nursing students reported gaining a profound comprehension of, and broadened perspective on, global health and cultural awareness, thereby enhancing their cultural competence [19]. Furthermore, such investments in international virtual education programs have the substantial benefit of offering all students on campus the opportunity to acquire global qualities. As a matter of equity, resources devoted to virtual exchange programs pay substantial dividends for students in historically marginalized and underresourced groups that have been underrepresented in international curricular experiences [20]. Therefore, well-organized virtual programs can be effective for various students to cultivate professional qualities as demonstrated by improvements in global cultural intelligence and academic performance. However, to our knowledge, no studies on rehabilitation professionals, including those in occupational therapy (OT), have been reported. Furthermore, no studies have been conducted on virtual international exchange programs between multiple non–English-speaking countries.

Aim of This Study

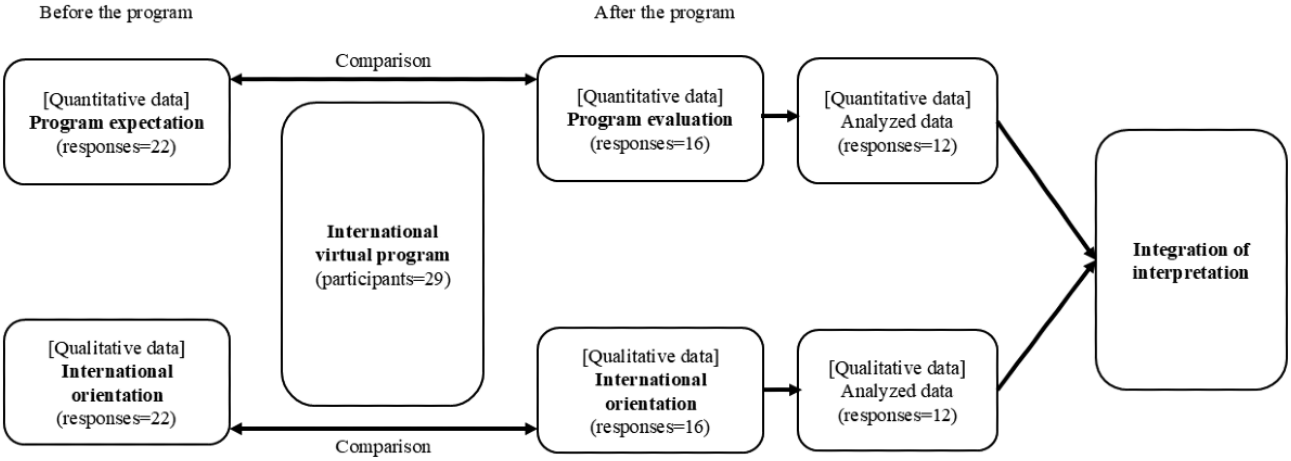
The development of virtual international exchange programs during the COVID-19 pandemic as a new method of fostering global human resources should have led to new adaptations, innovations, and equity considerations. Therefore, such programs need to be examined in terms of their feasibility and sustainability, not just as alternative options for in-person programs. Given this background, this study aimed to examine the effectiveness of virtual international exchange programs for improving academic learning skills, professional knowledge, and global communication abilities in group interactions from student perspectives and suggest meaningful ways to promote global education (global health knowledge and cultural intelligence). This report also aims to contribute to the educational effectiveness of virtual international programs in the rehabilitation professional education area in the postpandemic era, as well as to the development of global education, including OT education, via virtual learning among rehabilitation professionals. In addition, this research can be expected to contribute to the development of global human resources in various health professions through the exchange of information from multiple countries and various international backgrounds.

Methods

Study Design

This quasi-experimental study was conducted as intervention research using a before-and-after design and used a convergent parallel mixed methods approach (Figure 1) in which the researcher converges or merges quantitative and qualitative data to provide a comprehensive analysis [21]. In this study, both quantitative and qualitative data were collected before and after the intervention program and compared to evaluate students' international attitudes and perspectives. After the analysis, the quantitative and qualitative data were integrated into the interpretation of the overall results.

Figure 1. Study process.



Ethical Considerations

The questionnaire and methodology for this study were approved by the Research Ethics Committee of Arakawa Campus Tokyo Metropolitan University (ethics approval number 23060) and registered in the UMIN-Clinical Trials registry system (UMIN trial number UMIN000050884, Reception number R000056060). This study was performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki. Participation in this study was voluntary, and informed consent was obtained in writing from all participants before the study began. All participants' data were collected as anonymized data, and participants joined the research with no rewards.

Program Contents

The program was mainly organized by the host university (Tokyo Metropolitan University, Japan), with collaboration from 6 institutions from 4 countries (Angeles Foundation University, the Philippines; Bangladesh Health Profession Institute, Bangladesh; Chaing Mai University, Thailand; Mahidol University, Thailand; the Surakarta Ministry of Health Polytechnic, Indonesia; and the University of Indonesia, Indonesia). Tokyo Metropolitan University concluded a Memorandum of Understanding with all institutions to build relationships before the program was planned and primarily managed the program preparation through good collaboration. The goals of the program were to cultivate an international perspective, to understand OT in Asian countries through lectures and group work, and to learn basic research and presentation skills. During the development of the program, the Attention, Relevance, Confidence, and Satisfaction (ARCS) motivation model developed by Keller [22,23] was applied. The ARCS model proposes significant concepts regarding student attention in class, relevance to participants, student confidence in class, and student satisfaction. The ARCS model has been widely accepted and used to promote student motivation and active participation in class [24,25], and the online program was conducted effectively based on the ARCS model [5]. Hence, in the program, lecturers selected appropriate topics to achieve the program goals. In addition to the lecture contents, assignments and activities were organized to improve student understanding and maintain motivation in considering their academic progress and interests. All lecturers among the collaborative institutions involved in the management of the program agreed on the program contents.

As shown in Table 1, synchronized lectures from 4 countries were provided, and students were invited to join multimix group work on the last day. The group work topics included the role of OTs in the following areas: employment support in mental health, school, care for older adults, remote or isolated area care, assistive devices, stroke, addiction, and stigma. Students made 3 requests on the first day and were assigned to group work on the second. Lectures were provided by 4 countries, and students engaged in interactive group work within a small cultural program. A videoconferencing application (Zoom Communications, Inc) was used for the interactive live lectures, and a free cloud service (Google LLC) was used to share information and materials through an original website. The lecturers and students had access to the program contents and could provide or get handouts and recorded lecture materials for preparing and reviewing. During the program, students were encouraged to work together on their assignments and communicate with each other outside of the lectures through various tools and applications.

In all 7 institutions, the program participants were invited to participate in the virtual exchange program. In total, 30 - 35 students (n=5 from each institution) were targeted to join the program. The students were able to choose to participate in the program regardless of their involvement with the research data collection.

Table . Program contents from February 27 to March 7, 2024.

Session[a]	Date	Contents
1	February 27, 2024	Orientation: medical and welfare system in Japan, long-term care insurance (by Tokyo Metropolitan University lecturer)
2	February 28, 2024	Health and welfare system in Thailand (by Chiang Mai University lecturer)
3	February 29, 2024	Community care in Indonesia (by Surakarta Ministry of Health Polytechnic lecturer)
4	March 4, 2024	Community care in Indonesia (by Bangladesh Health Professions Institute lecturer)
5	March 7, 2024	Students' project presentations

[a]Each session was held at the following times: 10 AM–11:30 AM (Bangladesh), 11 AM-12:30 PM (Indonesia), 12 PM-1:30 PM (Philippines), 1 PM-2:30 PM (Japan).

Participants

The research participants were recruited from an exchange program. The students received no credits because the program was optional. Regarding recruitment, after obtaining ethical approval, Tokyo Metropolitan University lecturers requested research collaboration with joint researchers from each institution or the head of the department via a request letter. After obtaining approval from each institution, a coresearcher (KI) asked participating students by email to complete a questionnaire survey.

Data Collection

Questionnaires were prepared to examine the influence of virtual international learning and the students' international orientation. In addition to age, sex, and school year, before joining the program (within 5 d) and after joining the program (within 5 d), the research participants were asked to complete an online questionnaire survey that consisted of 2 sections: program expectations evaluations and international attitude. The online survey was conducted in accordance with the CHERRIES (Checklist for Reporting Results of Internet E-Surveys).

Program Expectations or Evaluations

The questionnaire asked about the students' expectations and evaluations based on the factors of academic experience, scientific learning skills, group interactions, learning value, program organization, assignments, and workload (refer to Multimedia Appendix 1). Responses were given on a 5-point Likert-type scale. These questionnaires on program expectations (before the program) and evaluations (after the program) were developed based on the student evaluation of educational quality questionnaire and other materials [26-28]. The student evaluation of educational quality questionnaire was originally developed by Marsh [26] as a student course evaluation in higher education and is now widely used around the world in multiple languages. The original questionnaire serves to evaluate the term course. In this study, we developed questionnaires that included items appropriate to the content of the program that could confirm the expected level of achievement. We selected questions that could compare the students' expectations and satisfaction before and after the program. In addition, some expressions were adjusted to fit the program. In total, 15 questions were used before the program to ask about program expectations, and 30 after the program to ask about program evaluations and satisfaction.

International Orientation

This questionnaire (free comments) asked about student motivation to join the international exchange program, the influence of the experience of joining the program, and international attitudes (refer to Multimedia Appendix 1). The questions on international attitudes were developed based on the revised version of IP [29,30], which was developed by Yashima [29,30], who devised an evaluation method for international orientation, from the work of Gardner [31] and Dörnyei [32] on the relationship between foreign language acquisition and international orientation. In this study, we used the IP factors that were appropriate for this study, including intercultural approach (avoidance) tendency, interest in international vocation, ethnocentrism (reaction to different customs or values or behaviors), interest in foreign affairs, and having things to communicate (willingness to communicate to the world).

Statistical Analysis

After the coresearcher (KI) removed all personally identifiable information, numerical data (questionnaire on program expectations or evaluations) were analyzed using descriptive statistics. Qualitative data (questionnaire on international orientation) were subjected to analysis using generative coding by means of steps for coding and theorization (SCAT), which is a qualitative analysis method developed based on the concept of grand theory [33,34]. This method is particularly effective for analyzing relatively small amounts of qualitative data, such as those from a single case or free comment sections of a questionnaire. SCAT consists of 4-step coding using a matrix accompanied by a procedure that describes the storyline and theory by spinning the constructs. The steps followed in SCAT are (1) a word or phrase in the data to focus on, (2) a phrase outside the data to paraphrase, (3) words and phrases to explain these, and (4) emerging themes and concepts. This 4-step coding process creates a storyline or theory that weaves together the themes or constitutive concepts in step 4. In particular, SCAT, as revised by Fukushi and Nago [34,35], is an effective method when there are many cases in which the linguistic data are very

short (eg, 1 or 2 lines), such as descriptions in the free comments column of the collected questionnaires. In the revised procedure, the intercepted data are grouped, paraphrased, and conceptualized. In this study, we used the revised version of SCAT for the data analysis. The first author (NS) analyzed the data, the validity of which was then confirmed by one of the coauthors (KI), who had experience with qualitative research and over 20 years of experience in verifying medical education data, followed by the other 3 coauthors (SSS, MHS, and SM). All data were analyzed and interpreted based on the participants' evaluations of the international program experience.

Results

Participant Characteristics

In total, 29 students (10 from Indonesia, 6 from the Philippines, 5 each from Japan and Bangladesh, and 3 from Thailand) joined the virtual exchange program from February 27 to March 7, 2024, among whom 22 completed a questionnaire before and 16 after the program. Data (from both before and after the program) were available from 12 participants (4 male and 8 female; mean 22.5, SD 2.6 years, age range 19‐28 years; response ratio: 41.4%) for analysis. Regarding the year of schooling, 3 participants were in their second year, 4 in their third, and 5 in their fourth (Figure 1)

Quantitative Data: Program Expectations or Evaluations

Table 2 shows the results of the questionnaires conducted before and after the program. The level of achievement before and after the program was compared for 15 items. The results indicated that the students were satisfied after the program. The students had high expectations before the program, and overall, these expectations were met in terms of academic experience, scientific learning skills, and group interactions. The students responded that they gained valuable knowledge and ideas from the program. The program contents were evaluated as being well prepared, and almost all the participants wanted to join a similar program if given the opportunity. However, some students expressed the need for more support from lecturers to engage in the program and the need to establish a good transmission of internet. In addition, the program schedule was a bit tight because regular classes were being managed at the same time. Overall, the program contents and workload seemed to be reasonable for the students. As a result, they gained new experience in deep expertise, scientific learning through problem-finding and problem-solving, and group interactions in an international context. Although significant differences were not found between before and after the program, the results indicated that the program met the students' high expectations.

Table . Results regarding student expectations and evaluations before and after the program (N=12).

Program expectation questionnaire (before program)	Student evaluation questionnaire (after program)	5-point Likert scale	Before program, mean (SD)	After program, mean (SD)
Academic experience		Strongly agree (5)-strongly disagree (1)		
I feel I will be able to increase my knowledge about OT[a] field.	I increased my knowledge about OT field.		4.5 (0.90)	4.5 (0.67)
I feel I will be more interested in academic learning.	I developed my interests in academic learning further.		4.1 (1.00)	4.3 (0.89)
I expect my knowledge of academia will change.	My knowledge of academia has changed.		4.4 (0.67)	4.0 (0.95)
Scientific learning skills		Strongly agree (5)-strongly disagree (1)		
I expect to learn about collecting and searching information skills.	I learned collecting and searching information skills.		4.8 (0.45)	4.9 (0.67)
I expect to learn about presentation and discussion skills.	I learned presentation and discussion skills.		4.8 (0.45)	4.6 (0.79)
I feel preparing for the presentation assignment will be good training for me.	Preparation for the presentation assignment was good training for me.		4.7 (0.49)	4.6 (0.79)
Group interaction		Strongly agree (5)-strongly disagree (1)		
I feel I will be able to understand my friends' thoughts in other countries.	I understood my friends' thoughts in other countries.		4.5 (0.52)	4.6 (0.67)
I feel I can communicate with OT students and OT lecturers from other countries.	I communicated with OT students and OT lecturers from other countries.		4.5 (0.52)	4.6 (0.67)
I feel I can learn from each other.	I learned from each other.		4.6 (0.51)	4.7 (0.65)
I expect to cultivate an international perspective through a student exchange program.	I cultivated an international perspective through a student exchange program.		4.5 (0.67)	4.6 (0.79)
—[b]	The atmosphere was good for sharing my ideas and thoughts in group discussion.		—	4.3 (0.78)
Learning value		Strongly agree (5)-strongly disagree (1)		
I expect to find the program intellectually challenging and stimulating.	I have found the program intellectually challenging and stimulating.		4.4 (0.67)	4.3 (0.75)
—	I have learned something that I consider valuable.		—	4.6 (0.67)
I feel my interest in the subject will increase as a consequence of this program.	My interest in the subject has increased as a consequence of this program.		4.5 (0.67)	4.2 (0.83)
—	I have learned and understood the subject materials of this program.		—	4.7 (0.65)

Program expectation questionnaire (before program)	Student evaluation questionnaire (after program)	5-point Likert scale	Before program, mean (SD)	After program, mean (SD)
Organization		Strongly agree (5)-strongly disagree (1)		
—	The lectures given by lecturers were very clear.		—	4.4 (0.79)
—	Program materials were well prepared and carefully explained.		—	4.5 (1.00)
I expect to deepen my knowledge about the areas of interest, and I will learn at my own pace through the program.	I learned interesting areas deeply at my own pace through the program.		4.4 (0.67)	4.2 (0.83)
—	The amount of support provided by lecturers was sufficient during the program.		—	3.9 (1.00)
—	Internet environment and technical support for the use of device and application were sufficient during the program.		—	3.9 (0.90)
Workload or difficulty		Too easy (5)-too difficult (1)		
—	How is the program difficulty?		—	3.3 (0.75)
—	How was the assignment workload?		—	3.1 (0.79)
—	How was the program pace?		—	2.6 (0.67)
I think the level of program content is suitable for me.	Overall, from Q23-25, the level of program content was suitable for me.	Strongly agree (5)-strongly disagree (1)	4.1 (0.67)	4.2 (0.72)
Assignment		Strongly agree (5)-strongly disagree (1)		
—	The required reading materials were valuable.		—	4.4 (1.00)
—	Required assignments contributed to appreciation and understanding of OT in the domestic and international fields.		—	4.6 (0.67)
Others		Strongly agree (5)-strongly disagree (1)		
—	Do you want to join a similar program if you have another opportunity?		—	4.7 (0.65)
—	The date and time schedule were reasonably good.		—	3.8 (1.19)
—	I actively participated.		—	4.3 (0.75)
How much do you expect out of this program overall?	The program met my expectation.	Very much (5)-not at all (1)	4.1 (0.90)	4.3 (0.75)

[a]OT: occupational therapy.

[b]Not applicable.

Qualitative Data: International Orientation

As a result of the SCAT coding of data before the program, 20 group concepts and 6 overall concepts emerged regarding international orientation (Table 3). In the text, [ ] indicate overall concepts and < > indicate group concepts.

Students decided to join the program because they were interested in [Learning new knowledge in the professional area] to <expect to know and learn about OT in other countries> and, they were interested in the [opportunity to have international experiences] and <seeking that opportunity>, but sometimes they had <no opportunity to exchange with people from other countries> or <they did not use that actively and usually>, even though they had the opportunity to communicate with people from other countries before joining the program. In addition, they were <interested in studying or working abroad>, but it might be <difficult for some of them due to financial and family concerns>. The virtual program was not concerned about these issues; therefore, it was easy to join to have international experiences. Furthermore, the students expected to [Have new international experiences with different cultures and values] from <making friends and having interactive communication with students in other countries> and <learning and experiencing something new>. In the international context, they <expected the program to lead to being more open-minded internationally>, as well as <learning about the cultures and values of other countries would make me change something> and being <interesting and fun to learn about diverse cultures and values>. On the other hand, some students were worried about <difficulties in understanding other cultures and values>. Similarly, in [international interactive communication], some students had <expectations of and interests in exchanging ideas and opinions>, <expectations of multicountry exchange experiences>, and <using English as a second language>; however, others experienced the <challenge of exchanging ideas and opinions>. Moreover, regarding [general global attitude sensitivity], depending on the individual, some <cared about international issues as usual interest>, but others did not or had <no opportunity to learn about that information>. They were interested in other values or study areas in other countries, but their general international awareness did not always explain their motivation to join the program.

Table . Results regarding concepts, group concepts, and participant descriptions of international orientation before the program (N=12).

Group concepts	Examples of participant descriptions
Concept: [Learning new knowledge in the professional area]	
Expectation to know and learn about OT[a] in other countries	• Good opportunity to know more about different dimensions of OT in different countries. • I want to know about the OT field in other countries.
Expectation to experience an international perspective in the medical health professional field	• From lectures given by other countries, I can compare their system with ours. • In the discussion session, my understanding about health care will be deepened from a student perspective. • I think this is a great opportunity to get a glimpse of the wider global professional perspective regarding OT.
Concept: [Having new international experiences with different cultures and values]	
Hope to make friends and have interactive communication with students in other countries	• I would like to participate in discussions with other students from other countries. • Participating in the virtual exchange program will improve my communication skills.
Expectation of learning and experiencing something new	• Sounds interesting, I will learn a lot from the program. • It can increase my vision about how I can become successful in the future.
Expectation of the program leading to a more open-minded international perspective	• It drives me to behave as politely as I can to accept diversity. • I will learn more about other perspectives, and it will also affect my international perspective.
Interesting and fun to learn about diverse cultures and values	• It is very fun to discover the diverse values and ideas of others. • I really enjoy and am open-minded about interacting with and discussing things with people who have different values and ideas. • I am very interested in learning about culture and other things from abroad.
Perceiving difficulties in understanding other cultures and values	• Sometimes it is hard to be on the other side of a position. • I feel like it takes effort to understand fully people from other cultures with different opinions.
Learning about other cultures and values changes me	• It encourages me to dive into other perspectives affected by my culture. • Exchange programs can open my mind and teach me lessons.
Concept: [Opportunity to have international experiences]	
Interested in or aspiration of studying or working abroad	• I have a big dream to study overseas. I am very motivated to participate in any cross-cultural opportunity. • I am really interested in studying abroad and later working abroad to increase my knowledge and experience. • I really want to continue my OT study abroad.
Difficulties of studying and working abroad	• I know that I will face many challenges abroad. • I have financial concerns.
Seeking more opportunities to communicate with people from other countries	• If I have an opportunity to contact people with other nationalities, I will be grateful. • I would like to have multicultural exchanges.
No previous exchange opportunities with people from other countries	• I am a little hesitant. • I have not had the opportunity to explore cultures in other countries.
Concept: [General global attitude sensitivity]	
Not interested in international issues or situations	• I rarely consume news from other countries. • I am not concerned about world affairs.

Group concepts	Examples of participant descriptions
Interested in international issues and situations	• We should know what is going on in other countries, this is our responsibility as human beings. • I always follow international news because I think everything that happens always affects everyone. • I am always curious about social conditions in other countries. • I can learn from updated information related to policy, health, and so on in other countries.
Not enough opportunities to know about international issues or situations	• I am not very informed about issues in other countries. • I don't have enough knowledge about international issues.
Concept: [International interactive communication]	
Expectations of and interest in exchanging ideas and opinions	• I would like to express my opinions on diverse positions. • I would like to share and discuss my perspectives with others. • I would like to exchange and receive opinions from different perspectives as it would help me to reflect and modify my knowledge or viewpoints.
Challenge of exchanging ideas and opinions	• I think it is challenging for me to exchange my ideas with others. • I feel like it will take effort to organize and tell people with other nationalities to understand ideas fully because of language barriers.
Expectation of a multicountry exchange experience	• I have the chance to work on projects and communicate with people from many countries. • We will share our perspectives in groups with students from different countries.
Using English as a second language	• It is a good opportunity to practice English skills. • The language barrier creates some difficulties while communicating with people from other countries.
Concept: [Motivation to engage in international experience through external encouragement]	
Recommendations from others	• Participating with some international OT students and teachers motivates me to learn more.

[a]OT: occupational therapy.

As a result of the SCAT coding of data after the program, 19 group concepts under 7 overall concepts regarding the participants' international orientation were extracted (Table 4). In the text, [ ] indicate overall concepts and <>indicate group concepts.

Through the program, students reported [knowing and learning new knowledge about OT and other topics in other countries in an international context]. They <learned about OT, including clinical and working situations in other countries> and <gained a wide knowledge of other countries>. With this [opportunity to have international experiences], they <enjoyed it by having international exchange experiences> and <multicultural exchange opportunities>. Some were <seeking the next opportunity for an international exchange experience> and were <interested in or aspiring to study or work abroad>, but they had <difficulties studying or working abroad> and did not have <much opportunity for international exchange after the program>. In the program, the students had the opportunity to [understand the cultures and values of others], which was good in terms of <learning from others and enjoying diverse values and ideas>, but it was <not easy to understand others> in some situations. The program involving multiple countries helped promote [mutual understanding in an international context] in terms of <understanding the viewpoints of others in an international context>, <knowing international perspectives>, and <gaining new ideas from others>. The students reported <enjoying sharing opinions, ideas, and interests with others> in [interactive communication with others], even though one student perceived a <language barrier>. After the program, some students reported the need <to know more about international issues and situations> and expressed <interest in and care about international issues and situations>, but others did <not care about international issues or situations> as part of [general global attitude sensitivity]. They had the experience of learning about the knowledge and values of other countries and enjoyed having interactive communication with people from other cultures and backgrounds.

The qualitative questionnaire was conducted from the perspective of intercultural approach tendencies; interest in international vocations; reactions to different customs, values, or behaviors; interest in foreign affairs; and having things to communicate; these corresponded with the group concepts. Regarding the comparison of data between before and after the program, before joining the program, participants were motivated by interests and curiosity about knowing about OT from overseas, not the concrete contents. However, after the

program, they had more concrete images about specific forms (eg, employment support and clinical skills) from what they learned from an international perspective. A lot of the participants sought opportunities for international exchange, but could not easily find them. Therefore, the chance to join the program was valuable for students, who reported that they wanted to participate the next time they had the opportunity. However, no participants indicated that they wanted to expand opportunities by themselves in the future. Some participants said that they would like to study or work abroad, but felt that it was difficult because of financial and family issues. As for attitudes toward different cultures and behaviors, the participants attempted to understand and enjoyed communicating with others from diverse backgrounds, even though they recognized that understanding different values was not always easy. On the other hand, interest in general international affairs was not necessarily a motivation for participation, and no changes were seen as a result of participating in the program.

Table . Results regarding concepts, group concepts, and participant descriptions of international orientation after the program (N=12).

Group concepts	Examples of participant descriptions
Concept: [Knowing and learning new knowledge about OT[a] and other topics in other countries in an international context]	
Learning about OT, including clinical and working situations in other countries	• We talked about the details of how OTs work and the OT process in every country. • I did not know anything about the system before. Now I know about OT situations in different countries. • I have learned so much about OT situations in other countries.
Gaining wide knowledge about other countries	• My knowledge horizons were widened with OT students from other countries. • I got to know things I did not know before.
Concept: [Understanding other cultures and values]	
Not easy to understand others	• I feel that it takes effort to understand and adapt to other people from different countries.
Learning from others and enjoying diverse values and ideas	• I really enjoyed hearing about and discussing differences in values and cultures. • I enjoy interacting with people who have different ideas. • I find it interesting that I can have discussions with students from other countries and see their perspectives.
Concept: [Mutual understanding in an international context]	
Understanding the viewpoints of others in an international context	• I gained a deep understanding about the different perspectives of other countries compared with us. • It allowed me to respect the other participants. • I know for a fact that no country has a perfect program, each has its own flaws and we can continue to grow by learning and appreciating them.
Knowing international perspectives	• It gave me a lot of new information about other countries so I can learn and know more about international situations. • I do not have that much knowledge about the OT field in other countries, so the program affected my international perspective. • My international perspective did not change.
Gaining new ideas from others	• I was able to improve my knowledge and see other perspectives. • I love to learn about how people think differently.
Concept: [Opportunity to have international experiences]	
Having and enjoying opportunities to have international exchange experiences	• I enjoy having the ability to communicate and participate in an international program. • This was the first time I had an experience like this.
Seeking new international exchange experience opportunities	• I would like to if there is another opportunity. • It was wonderful and I hope to participate in more programs like this if I get the chance in the future.
Multicultural exchange opportunities	• I would love to have the opportunity to communicate with students from other nationalities and cultures. • I really like multicultural exchanges. • I was usually a little hesitant to contact strangers, but now I am happy because I can contact students with other nationalities.
No opportunity for international exchange after the program	• I very rarely have this opportunity. • We didn't have any international program to facilitate us.
Interested in or aspiring to study or work abroad	• I always dream about working outside the country. • I had a dream to pursue higher education abroad. • My goal is to continue my studies abroad. I hope I have the chance someday.

Group concepts	Examples of participant descriptions
Difficulties studying or working abroad	• I have financial concerns. • I have some family concerns. • Working abroad is equally exciting and challenging.
Concept: [General global attitude sensitivity]	
Not interested in international issues or situations	• No time to see. • Normally I do not care about what is going on.
Need to know more about international issues and situations	• I think I need to have a look at situations happening around the world. • I am interested in other countries because I can compare and learn things from them. • I should be interested in it because it is called globalization.
Interested in and care about international issues and situations	• I am really interested as it gives me the opportunity to gain information about the world, which is very important in this era of digitalization. • Now, I am interested. • It is a very helpful example for us to develop.
Concept: [Interactive communication with others]	
Enjoying sharing opinions, ideas, and interests with others	• I really enjoy sharing my thoughts and interacting with people from other countries. • I really like to express my opinions on diverse positions. • I felt like they were close to me.
Language barriers	• It was a bit difficult to communicate with others while we were using our second language.
Concept: [Difficulty arranging the schedule of the international program as an optional program]	
Time management	• I did not have time to participate in our class.

[a]OT: occupational therapy.

Discussion

Principal Findings

This study examined the educational effectiveness of a virtual international exchange program focusing on fostering global qualities involving academic and professional learning and international orientation from students' perspectives. The research participants expressed satisfaction with the program in regard to cultivating an international perspective, understanding OT professionals from other countries, learning academic skills through international communication, and learning about diverse cultures and societies, including medical systems. Even though the results of the quantitative survey did not indicate significant differences, the program participants had high expectations and indicated that they were satisfied with the contents. In addition, the program gave participants a valuable opportunity to cultivate global qualities such as problem-finding and problem-solving, and to have group interactions with people from diverse backgrounds.

The results could indicate that the virtual program was capable of meeting student expectations for international experience through the learning of professional knowledge and communication skills, in line with previous studies [6,7]. This program consisted of lectures, understanding medical situations in other countries through OT professional knowledge, and interactive group work followed by presentations. This program was developed to maintain student motivation and cultivate global qualities based on the ARCS model. The participating students expected to gain new knowledge and ideas about professional areas in international contexts and communicate with others with diverse opinions and backgrounds; from this perspective, their expectations were met. Even though the results of the quantitative survey did not show significant differences between before and after the program because of the participants' high expectations (ceiling effect), the results indicated a mutual understanding of diverse opinions and a recognition that the participants' differences and similarities were valuable. In addition, the vague image of international exchange before the program became more concrete and practical after the program. Thus, the participants had the opportunity to foster their international orientation. On the other hand, the program period may have affected workload or difficulty; for example, the program pace was a bit fast, so the student might not have had time to absorb and review knowledge or brainstorm with international friends in their groups before presenting an assignment. However, the schedule for the long-term program is difficult to structure because of the busy curriculum, including practical placement in health education in multiple institutions. Therefore, even in the short term, the program should make more improvements from the students' perspective.

Implications of the Findings

The virtual program in this study provided students with meaningful experiences, in terms of cultural awareness, competence, and intelligence, for growing as global human resources, in line with previous studies [13,16,19]. Participants showed interest in international communication with professional knowledge both before and after the program, including having friends with different values and learning new things about other countries, even though some of them felt difficulties in communicating with those with different backgrounds. After the program, students reported having more concrete knowledge about OT in other countries and communicating with people with various values through professional knowledge. This may indicate the need to cultivate the ability to build a global community with a common understanding of global health and raise awareness of safety and cultural competency for people from diverse backgrounds in health care. Global institutional collaboration fosters international collaboration between not only students but also educators and researchers [19]. In addition, the virtual interactive program could enhance the development of student competency by contextualizing knowledge, fostering collaboration and innovation among universities, creating an international professional network for students and instructors, and promoting professional skills [36].

Therefore, it is essential to manage communication in virtual programs and foster solid relationships among educational institutions. Student experiences could lead to a global campus environment, which includes faculty and administrators. The establishment of virtual programs relies on existing relationships, clear communication, and a commitment to collaborate [37]. In addition, communication within each institution and with participating students is key to a successful program. Some participants reported that they might have needed more support from faculty members. Interuniversity communication and collaboration among administrators, faculty, prospective students, and partner universities are also important [36,37]. Furthermore, a previous study demonstrated that in practice, these strategies need significant modifications, at least in part, to suit local contexts. Regarding logistical support to achieve effective internationalization “at home,” the present program, which included participants from 5 different countries, should have given more consideration to mutual understanding [38]. Other concerns in regard to the management of virtual programs include the online environment and program schedules [1,5]. In this study, the satisfaction score was not bad, but the satisfaction level was lower than the other items. Therefore, communication among institutions to manage and arrange schedules is important, and a solid, reliable relationship is essential to improve virtual educational programs.

Due to its short duration and contents, the program did not lead to significant changes in global attitudes; however, participants seeking an opportunity were able to learn and think about different values and conditions in other countries. As some students expressed concerns about financial and family issues in terms of studying or working abroad, virtual education was an easy option that allowed them to cultivate global qualities at home [8,20]. Furthermore, IaH programs based on virtual exchange and simulation have been shown to improve general student self-efficacy in the short term [39]. Virtual programs such as the one described in this study could therefore help foster an awareness of international orientation in global human resources. In OT education, OT knowledge and skills are mainly required; however, the experience of learning about and understanding different cultures and values is very helpful to support the lives of others and contribute to development in this field through collaborative communication with other countries [40]. Regarding virtual environments, even though students have adequate communication tools, interactive communication can be difficult, and thus, facilitating interactive activities such as group work is needed [6]. In this study, the participants expressed satisfaction with interactive communication, even in English as a second language. Therefore, the COIL program helps students cultivate meaningful interactive communication skills as a global quality to develop and coordinate team management skills among individuals with various values.

In this postpandemic era, some students may regard virtual education as an alternative option to in-person learning by traveling abroad. However, educators need to demonstrate the value and benefits of virtual education as IaH and clarify how virtual learning abroad programs should be promoted to students [41]. It was clear from the literature that international virtual education can not only maintain and create sustainable ties with international partners, which adds depth and richness, but also provides opportunities to create meaningful, lasting collaborative spaces for the ongoing expansion of global activities [42]. In addition, this type of program can be a positive influence as it gives all students on campus the opportunity to participate in global human resource development, even though the result of this study might indicate only the possibility due to the small sample size.

Limitations and Future Research

This study has some limitations. First, the sample size was small, and the program was only conducted over a short period (1 wk). Therefore, the results may not fully reflect the students’ perspectives. On the other hand, conducting a virtual program with a much larger number of participants would make its quality difficult to maintain. Furthermore, at the moment, international programs are not mandatory in medical education in general, so it remains difficult to recruit many participants for practical reasons. Second, there is no comparison group in this study; therefore, the degree to which this program influenced student perspectives remains unclear. In the future, the effectiveness of the virtual program in each country should be assessed, and the program structure and preparation should be tailored to each country’s situation. This should lead to richer research data.

Conclusions

As globalization has been increasing, it is important to provide medical care to people with diverse backgrounds and collaborate with foreign medical professionals for the development of the professional field. Effective virtual international education should not replace traditional exchange programs, but rather, offer a different educational method to increase international orientation and acquire global qualities as IaH. Within virtual

education, effectiveness should be judged based not only on understanding diverse cultures and increasing cultural awareness, but also on developing global human resources. This study aimed to develop a meaningful virtual international exchange program to promote global education and examine academic learning skills, professional knowledge, and global communication ability from student perspectives. The results would be expected to contribute to education for the development of global human resources in the health professions through the exchange of information by people from various countries and international backgrounds. The present virtual international program was conducted using a quasi-experimental before-and-after design that used a convergent parallel mixed methods approach among 5 countries (Bangladesh, Indonesia, Japan, the Philippines, and Thailand). The program could provide students with meaningful experiences as global human resources in terms of cultural awareness, competence, and intelligence. In this postpandemic era, some students may regard virtual education as an alternative to in-person learning combined with traveling abroad. However, educators need to show the benefits and value of virtual education as IaH and clarify how these benefits should be promoted to students. Assigning projects that necessitate teamwork across different countries can compel students to engage more deeply with their peers, thereby fostering stronger communication skills and a better understanding of diverse perspectives. These collaborative tasks can simulate real-world scenarios where multidisciplinary and multicultural teams work together to solve complex problems. This approach can not only enhance learning experiences but also prepare students for professional environments where international collaboration is often essential. By requiring students to navigate language barriers, cultural differences, and varied working styles, such tasks can significantly enhance their global competence and teamwork abilities. These deeper interactions could bridge the gap between mere exposure to international elements and the development of a truly global perspective.

Acknowledgments

We wish to thank all the faculty members at Tokyo Metropolitan University, Angeles Foundation University, Bangladesh Health Profession Institute, Chiang Mai University, Thailand; Mahidol University, Thailand; The Surakarta Ministry of Health Polytechnic, Indonesia; and the University of Indonesia, as well as all participating students. In particular, Ms. Mohuya Akter and Ms. Cahya Ramadani Renhoran were very supportive of the progress of this program. The authors also thank FORTE Science Communications [43] for English language editing. This work was partially supported by the General Research Fund of Tokyo Metropolitan University for manuscript writing and publication.

Data Availability

The data used to support the findings of this study are included within the article. However, for more information, the datasets generated and analyzed in this study are available from the corresponding author upon reasonable request.

Authors' Contributions

All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by NS, KI, SSS, MHS, and SM. The first draft of the manuscript was written by NS, and all authors reviewed and commented on the previous version of the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questionnaires for before and after joining the program.
[DOCX File, 21 KB - mededu_v11i1e77564_app1.docx]

References

1. Adedoyin OB, Soykan E. Covid-19 pandemic and online learning: the challenges and opportunities. Interact Learn Environ 2023 Feb 17;31(2):863-875. [doi: 10.1080/10494820.2020.1813180]
2. Rojek NW, Madigan LM, Seminario-Vidal L, et al. A virtual faculty exchange program enhances dermatology resident education in the COVID-19 era: a survey study. Dermatol Online J 2021 Mar 15;27(3):13030/qt1zt0q59g. [Medline: 33865275]
3. Martinez PGV, Pineda RC, Sy MP, Vera CKM, Galang M, Cayanan KKS, et al. Experiences and reflections of clinical supervisors on online occupational therapy internship during the COVID-19 pandemic. Phil J Health Res Dev 2021;25(Supplement 1):S86-S95 [FREE Full text]
4. Kim RE, Morningstar-Kywi N, Romero RM, et al. An online international pharmacy summer course during the COVID-19 pandemic. Pharm Educ 2021;20(2):136-144. [doi: 10.46542/pe.2020.202.136144]

5. Suyama N, Inoue K, Hidayat MS, Sasaki C, Shioji R. Effectiveness and feasibility of virtual international exchange program for occupational therapy students to develop the international perspective and professional skills -Mix-method study. JPKI 2024;13(1):53. [doi: 10.22146/jpki.89952]
6. Suyama N, Inoue K, Sorasak S, Thawisuk C, Watanabe M. Reflection on feasibility and usability of interactive online international exchange program for occupational therapy students. Discov Educ 2023;2(1):7. [doi: 10.1007/s44217-023-00031-4] [Medline: 36741295]
7. Jung D, De Gagne JC, Choi E, Lee K. An online international collaborative learning program during the COVID-19 pandemic for nursing students: mixed methods study. JMIR Med Educ 2022 Jan 24;8(1):e34171. [doi: 10.2196/34171] [Medline: 34982035]
8. Keshishi N, Seal A, Jicha K, Gaustad Shantz B, Slovic AD. Attempts to replicate the skills, attributes and capabilities associated with international mobility in an online world: a case study. JUTLP 2023;20(4). [doi: 10.53761/1.20.4.11]
9. Perumal-Pillay VA, Bangalee V, Oosthuizen F, Andonie G, Rotundo H. Shared learning experiences: pilot study of an online exchange project between pharmacy students in South Africa and the United States. Curr Pharm Teach Learn 2023 Oct;15(10):896-902. [doi: 10.1016/j.cptl.2023.06.021] [Medline: 37507312]
10. Beelen J, Jones E. In: Curaj A, Matei L, Pricopie R, Salmi J, Scott P, editors. Redefining Internationalization at Home: Springer Cham; 2015:59-72.
11. Rubin J. Embedding Collaborative Online International Learning (COIL) at Higher Education Institutions: Internationalisation of Higher Education; 2017, Vol. 2:27-44 URL: https://www.handbook-internationalisation.com/en/handbuch/gliederung/#/Beitragsdetailansicht/905/1192/Embedding-Collaborative-Online-International-Learning-(COIL)-at-Higher-Education-Institutions [accessed 2025-10-30]
12. Hackett S, Dawson M, Janssen J, van Tartwijk J. Defining collaborative online international learning (COIL) and distinguishing it from virtual exchange. TechTrends 2024 Nov;68(6):1078-1094. [doi: 10.1007/s11528-024-01000-w]
13. de Castro AB, Dyba N, Cortez ED, Pe Benito GG. Collaborative online international learning to prepare students for multicultural work environments. Nurse Educ 2019;44(4):E1-E5. [doi: 10.1097/NNE.0000000000000609] [Medline: 30339556]
14. Saftner MA, Ayebare E. Using collaborative online international learning to support global midwifery education. J Perinat Neonatal Nurs 2023;37(2):116-122. [doi: 10.1097/JPN.0000000000000722] [Medline: 37102558]
15. Hua J, Kondo A, Moross J. Enhancing intercultural sensitivity in Japanese nursing students through international online nursing courses: a quasi-experimental study. Nurse Educ Today 2023 Sep;128(105870):105870. [doi: 10.1016/j.nedt.2023.105870] [Medline: 37385149]
16. Jenssen U, Bochenek JM, King TS, Steindal SA, Hestvold IV, Morrison-Beedy D. Impact of COIL: learning from student nurses in Norway who collaborated with U.S. students. J Transcult Nurs 2024 Jan;35(1):74-82. [doi: 10.1177/10436596231209043] [Medline: 37933746]
17. Kanamori Y, Seki N, Foxton R, et al. Fostering globally competent dental students through virtual team-working, problem-solving and person-centred multi-disciplinary care planning. J Dent Sci 2023 Jan;18(1):95-104. [doi: 10.1016/j.jds.2022.07.004] [Medline: 36643270]
18. Galan-Lominchar M, Roque IMS, Cazallas CDC, Mcalpin R, Fernández-Ayuso D, Ribeiro AS. Nursing students' internationalization: virtual exchange and clinical simulation impact cultural intelligence. Nurs Outlook 2024;72(2):102137. [doi: 10.1016/j.outlook.2024.102137] [Medline: 38340388]
19. Örtlund OM, Andersson I, Osman F. Promoting global health knowledge and cultural competence of Swedish and Somali nursing students through collaborative virtual seminars: a qualitative descriptive study. J Transcult Nurs 2024 Nov;35(6):491-500. [doi: 10.1177/10436596241271088] [Medline: 39148417]
20. Lee J, Leibowitz J, Rezek J, Millea M, Millea M, Saffo G. Impact of international virtual exchange on student success. J Int Stud 2022;12(S3):77-95. [doi: 10.32674/jis.v12iS3.4593]
21. Creswell JW. Research Design Qualitative, Quantitative, and Mixed Methods Approaches: SAGE Publications, Inc; 2014.
22. Keller JM. Development and use of the ARCS model of instructional design. J Instr Dev 1987 Sep;10(3):2-10. [doi: 10.1007/BF02905780]
23. Keller JM. The ARCS Model of Motivational Design In: Motivational Design for Learning and Performance: Springer; 2010:43-74.
24. Daugherty KK. ARCS motivation model application in a pharmacy elective. Curr Pharm Teach Learn 2019 Dec;11(12):1274-1280. [doi: 10.1016/j.cptl.2019.09.009] [Medline: 31836153]
25. Luo X, Liu L, Li J. The effects of ARCS motivational instruction in physical education on learning cognition and the health-related physical fitness of students. Front Psychol 2022;13(53):786178. [doi: 10.3389/fpsyg.2022.786178] [Medline: 35734461]
26. Marsh HW. SEEQ: a reliable, valid, and useful instrument for collecting students' evaluations of university teaching. Brit J Edu Psychol 1982 Feb;52(1):77-95. [doi: 10.1111/j.2044-8279.1982.tb02505.x]
27. Coffey M, Gibbs G. The evaluation of the student evaluation of educational quality questionnaire (SEEQ) in UK higher education. Assess Eval Higher Education 2001 Jan;26(1):89-93. [doi: 10.1080/02602930020022318]

28. Tsubakimoto Y. Students' consciousness before the start of first year classes -a comparison of online and face-to-face. Presented at: 10th Meeting on Japanese Institutional Research; Nov 12-14, 2021. [doi: 10.50956/mjir.10.0_90_18]
29. Yashima T. Willingness to communicate in a second language: the Japanese EFL context. Modern Lang J 2002 Jan;86(1):54-66. [doi: 10.1111/1540-4781.00136]
30. Yashima T. International posture and the ideal L2 self in the Japanese EFL context. In: Dörnyei Z, Ushioda E, editors. Motivation, Language Identity and the L2 Self, Bristol, Blue Ridge Summit: Multilingual Matters 2009:144-163. [doi: 10.2307/jj.30945943.10]
31. Gardner RC. In: Arnold E, editor. Social Psychology and Second Language Learning: The Role of Attitudes and Motivation 1985.
32. Dörnyei Z. Conceptualizing motivation in foreign‐language learning. Lang Learn 1990 Mar;40(1):45-78. [doi: 10.1111/j.1467-1770.1990.tb00954.x]
33. Otani T. SCAT a qualitative data analysis method by four-step coding: easy startable and small scale data-applicable process of theorization. Bull Grad Sch Educ Hum Dev 2008;54(2):27-44. [doi: 10.18999/nueduca.54.2.27]
34. SCAT: steps for coding and theorization. J Japan Soc Kansei Eng 2011;10(3):155-160. [doi: 10.5057/kansei.10.3_155]
35. Fukushi M, Nago N. Clinical educators unable to accept the clinical medical training system and residents with no sense of belonging: results of needs assessment involving clinical educators in faculty development workshops. Med Educ 2011;42(2):65-73. [doi: 10.11307/mededjapan.42.65]
36. Vázquez-Villegas P, Gómez-Guerrero D, Mejía-Manzano LA, Morales-Veloquio G, Montaño-Salinas LP, Membrillo-Hernández J. Evaluation of good practices and opportunity areas of a collaborative online international learning (COIL) program: global shared learning classroom. Educ Inf Technol 2024 Nov;29(16):22247-22286. [doi: 10.1007/s10639-024-12739-3]
37. Enkhtur A, Zhang (张希西) X, Li (李明) M, Chen (陈丽兰) L. Exploring an effective international higher education partnership model through virtual student mobility programs: a case study. ECNU Rev Education 2024 Dec;7(4):971-990. [doi: 10.1177/20965311241232691]
38. Ambrose M, Murray L, Handoyo NE, Tunggal D, Cooling N. Learning global health: a pilot study of an online collaborative intercultural peer group activity involving medical students in Australia and Indonesia. BMC Med Educ 2017 Jan 13;17(1):10. [doi: 10.1186/s12909-016-0851-6] [Medline: 28086875]
39. Galan-Lominchar M, Roque IMS, Del Campo Cazallas C, Mcalpin R, Fernández-Ayuso D, Zerolo BE. Internationalization at home program significantly increases the self-efficacy of nursing students: a pre-post study. Nurse Educ Today 2024 Dec;143:106361. [doi: 10.1016/j.nedt.2024.106361] [Medline: 39190959]
40. Online overseas fieldwork initiatives and student evaluation. Yokohama J Nurs 2023;16(1):14-21. [doi: 10.15015/00002449]
41. Kosman BA, Castro de Jong D, Knight-Agarwal CR, Chipchase LS, Etxebarria N. The benefits of virtual learning abroad programs for higher education students: a phenomenological research study. Nurse Educ Today 2024 May;136(106133):106133. [doi: 10.1016/j.nedt.2024.106133] [Medline: 38387211]
42. Weaver GC, McDonald PL, Louie GS, Woodman TC. Future potentials for international virtual exchange in higher education post COVID-19: a scoping review. Education Sci 2024;14(3):232. [doi: 10.3390/educsci14030232]
43. FORTE Science Communications. URL: https://www.forte-science.co.jp/ [accessed 2025-10-29]

Abbreviations

ARCS: attention, relevance, confidence, and satisfaction
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
COIL: collaborative online international learning
IaH: internationalization at home
OT: occupational therapy
SCAT: steps for coding and theorization

Edited by J Gentges; submitted 15.05.25; peer-reviewed by J Moreno-Chaparro, S Kolla; revised version received 13.09.25; accepted 17.09.25; published 06.11.25.

Please cite as:
Suyama N, Inoue K, Kobayashi N, Kaunnil A, Siangchin SS, Sahid MH, Saloko E, Moniruzzaman S
Educational Effectiveness of a 5-Country Virtual Exchange Program for Internationalization in Occupational Therapy Education: Mixed Methods Study
JMIR Med Educ 2025;11:e77564
URL: https://mededu.jmir.org/2025/1/e77564
doi:10.2196/77564

Exploration and Practice of the First Clinical Medical Postdoctoral Program in China: Retrospective, Nonrandomized, Controlled Study

Lingda Zhang, MPH; Lianghong Sun, MEd; Honglei Li, MD, PhD

[1]Department of Education, The Second Affiliated Hospital, Zhejiang University School of Medicine, 88 Jiefang Rd, Hangzhou, China

Corresponding Author:
Honglei Li, MD, PhD
Department of Education, The Second Affiliated Hospital, Zhejiang University School of Medicine, 88 Jiefang Rd, Hangzhou, China

Abstract

Background: To further optimize the clinical and scientific training of high-level doctoral graduates, the Office of the National Postdoctoral Administration launched a clinical postdoctoral program in 2015. This program provides postdoctoral clinical medicine trainees with 3 years of individualized, intensive training through a full mentorship system, interdisciplinary collaboration, and a multiteam teaching platform.

Objective: This study aimed to compare the effectiveness of this novel clinical postdoctoral training program against conventional doctoral training, using specific, quantifiable metrics. Our primary research questions were as follows: (1) Does the program lead to superior clinical performance, as measured by theoretical examination scores and Case Mix Index (CMI)? (2) Does it enhance scientific research productivity, measured by publication output and success rates in procuring provincial and national NSFC (National Natural Science Foundation of China) funds? (3) Are there differences in teaching capacity and overall career advancement?

Methods: This was a retrospective, nonrandomized controlled study. Doctoral graduates who entered the hospital for standardized residency training between 2015 and 2019 were enrolled and divided into a postdoctoral training group (n=23) and a doctoral training group (n=106).

Results: The postdoctoral group demonstrated significantly higher clinical performance, as indicated by higher theoretical examination scores (445.70, SD 14.67 vs 435.12, SD 15.29; *P*=.003) and a higher median CMI (1.14 vs 0.92, *P*=.03), reflecting greater ability to manage complex clinical cases. In terms of research productivity, the postdoctoral group outperformed the doctoral group in the number of published papers (2.35, SD 2.39 vs 1.11, SD 1.47; *P*=.002) and the proportion of approved provincial-level NSFC projects (47.83% vs 17.00%, *P*=.001). However, no significant differences were observed in the acquisition of national-level NSFC funding (60.87% vs 44.34%, *P*=.15), teaching capacity (0.22, SD 0.518 vs 0.10, SD 0.306; *P*=.16), or overall competency indicators such as the rate of professional title promotion (100% vs 94.34%, *P*=.53) and the attainment of a master's-degree supervisor qualification (13.04% vs 5.67%, *P*=.42).

Conclusions: The clinical postdoctoral training program demonstrates promising effectiveness in enhancing both clinical performance and scientific innovation among medical trainees. These findings support the value of integrating clinical practice, research, and mentorship in advanced postgraduate medical education and suggest that this model is worth promoting in more medical institutions.

(JMIR Med Educ 2025;11:e65622) doi:10.2196/65622

KEYWORDS

clinical medicine postdoctoral program; standardized residents training; clinical competence; scientific research ability; teaching skills; comprehensive ability

Introduction

Background

China’s medical education system has undergone significant reform in recent years, evolving into a 3-stage model: medical school education, standardized residency training (SRT), and continuing professional development. SRT serves as the cornerstone of clinical training, aiming to produce physicians who are competent in managing common diseases with strong ethical standards and clinical skills [1-3]. However, while SRT focuses on standardized clinical practice, it often falls short in cultivating research literacy, academic leadership, and

innovation capabilities—key competencies required for high-level clinical scholars.

To address this gap, the Office of the National Postdoctoral Management Committee launched a clinical postdoctoral training program in 2015 [4]. As the first pilot institution, Zhejiang University introduced a novel training model that builds upon SRT but goes beyond it by integrating advanced clinical training with rigorous scientific research and teaching development [5].

This program is specifically designed for highly qualified medical doctors who have completed their doctoral studies and seek to become leaders in both clinical practice and biomedical research. Unlike traditional SRT, which emphasizes competency in routine diagnosis and treatment, the clinical postdoctoral program enhances trainees' abilities in clinical reasoning and management of complex cases, original research and evidence-based medicine, medical education and mentorship, and innovation and translational application of clinical findings [6].

The curriculum combines structured clinical rotations with independent research projects under the supervision of multidisciplinary mentors. It also includes formal teaching experiences, international exchanges, and interdisciplinary collaboration, all aimed at fostering a new generation of clinicians who can lead in academia, contribute to global health, and drive clinical innovation.

Ultimately, the program seeks to cultivate clinician scientists and clinician educators who possess not only excellent clinical expertise but also the capacity to conduct high-impact research, mentor future generations of physicians, and translate scientific discoveries into clinical practice. This represents a significant departure from conventional postgraduate medical training models seen worldwide and aligns more closely with integrated clinical-academic pathways found in some elite programs in North America and Europe [7-9].

This study focuses on the pioneering program at Zhejiang University to evaluate its effectiveness in bridging the gap between clinical training and academic development, thereby providing evidence for its potential role in advancing China's medical education system.

Objectives

While Zhejiang University School of Medicine and its Second Affiliated Hospital have a long-standing tradition of excellence in medical education, traditional postresidency training often fails to provide integrated development in advanced clinical practice, research, and teaching. To address this, we launched the Clinical Medical Postdoctoral Training Program in 2015. This study aims to evaluate the effectiveness of this program's first 8 years of implementation (2015 - 2023) through a retrospective, nonrandomized, controlled design.

Specifically, we sought to compare the postdoctoral trainees with conventionally trained counterparts using a set of predefined, quantifiable metrics as proxies for key competencies: (1) clinical performance measured based on standardized theoretical examination scores and the Case Mix Index (CMI) of treated patients, (2) research productivity measured by publication output and success rates in obtaining provincial and national-level NSFC (National Natural Science Foundation of China) funds, (3) teaching engagement measured by the number of formal teaching sessions conducted, and (4) career advancement measured by the rate of professional title promotion and qualification as a master's-degree supervisor.

We hypothesized that this program would significantly improve participants' clinical reasoning, research productivity, and teaching capabilities compared to conventional postresidency pathways. The findings are expected to be of particular interest to medical educators, hospital administrators, and policymakers involved in designing advanced clinical training programs, especially in countries seeking to strengthen their clinical academic workforce.

Methods

Ethical Considerations

This study was approved by the Ethics Committee of The Second Affiliated Hospital, Zhejiang University School of Medicine (approval number 2023-0974). Informed consent was waived by our Institutional Review Board because of the retrospective nature of our study.

Study Design and Participants

We conducted a retrospective, nonrandomized, controlled study. The study population consisted of medical doctoral graduates who entered the hospital between 2015 and 2019 and were required to undergo SRT. Individuals without a numeric score on the SRT completion test were excluded to ensure the quantitative evaluation of training outcomes.

Participants were divided into 2 groups based on their training pathway:

- Doctoral group (control): graduates who completed only SRT.
- Postdoctoral group (intervention): graduates who successfully applied for and completed the additional 3-year Clinical Medicine Postdoctoral Training Program.

Admission to the postdoctoral program was competitive, based on academic merit, clinical performance, and personal motivation. The program provided enhanced, structured training in clinical practice, scientific research, and teaching under a dedicated mentorship team [10,11].

Outcome Measures

To evaluate the program's effectiveness, we selected a comprehensive set of quantitative metrics designed to proxy competencies in clinical practice, scientific research, teaching, and career advancement (Table 1). Data for all metrics were collected over the 3-year training period. All metrics were analyzed with stratification by age, gender, clinical discipline, and prior training duration.

Table . Primary outcome measures and their operational definitions.

Metric category	Operational definition and data source	Construct measured
Clinical performance		
Theoretical examination score	Final examination score; source: Education Department	Mastery of clinical knowledge
SRT[a] completion pass rate	Binary outcome (pass/fail) for the standardized national residency completion examination; source: Education Department	Attainment of the minimum competency standard required for independent clinical practice
Annual assessment score	Composite annual review score; source: Education Department	Overall clinical performance and professionalism
Case Mix Index	The average diagnosis-related group (DRG) weight of patients treated; source: Medical Record Room	Complexity of clinical experience
Scientific research ability		
Number of papers published	Total count of peer-reviewed research articles published; source: Scientific Research Department	Scientific output and research ability
The number of achievement awards won as a participant	The number of scientific research achievement awards (at the provincial/ministerial level or above) won by the trainee as a contributor; source: Scientific Research Department	Peer recognition and impact of research output
National/provincial NSFC[b] projects	Total number of grants obtained as a principal investigator; source: Scientific Research Department	Competitiveness in securing research funding
Teaching ability		
Teaching engagement	Total number of teaching articles/teaching reform projects/teaching competitions; source: Education Department	Contribution to medical education
Comprehensive ability		
Professional title promotion	Binary outcome (yes/no) for promotion from a junior to an intermediate title; source: Human Resources	Institutional recognition of competence
Master's degree supervisor qualification	Binary outcome (yes/no) for being approved as a supervisor; source: Education Department	Recognition of academic standing

[a]SRT: standardized residency training.

[b]NSFC: National Natural Science Foundation of China.

Data Collection and Processing

Data were collected from the hospital's departmental databases (Education Department, Human Resources, Scientific Research Department, and Medical Record Room). The collection process involved extracting structured data from electronic source files, including examination records, personnel files, grant management systems, and publication databases. To ensure data quality, we handled missing data through listwise deletion and checking for implausible values or outliers. This process was conducted by 2 independent researchers to minimize error.

Statistical Analysis

SPSS (version 27.0; IBM Corp) was used for statistical evaluation of the data. Numerical/count data were expressed as the n (%) values, and the *t* test or chi-square test was used for comparison between groups. Normally distributed data are expressed as mean (SD) values, and the *t* test was used for comparison between groups. Nonnormally distributed data are expressed as median (IQR) values, and between-group comparisons were made using the Mann-Whitney *U* test. *P* values of <.05 were considered statistically significant.

Results

Study Participants

A total of 251 trainees (admitted from 2015 - 2019) met the inclusion criteria of this study. After excluding 17 trainees with less than 1 year of training, 6 trainees with more than 3 years of training, and 99 trainees without a specific score on the SRT completion test, 129 trainees were finally enrolled, including 23 postdoctoral trainees and 106 control trainees (Figure 1). Table 2 shows that there was no statistically significant difference between the two groups of trainees in terms of gender, age, distribution of disciplines, and other aspects. The difference in training years was statistically significant (*P*<.001).

Figure 1. Flowchart for the study. A total of 251 trainees (admitted from 2015 to 2019) met the inclusion criteria of this study. Exclusion criteria: 1 year of training, more than 3 years of training, and not having a specific score on the SRT completion test. In total, 129 trainees were finally enrolled, including 23 postdoctoral trainees and 106 control trainees. SRT: standardized residency training.

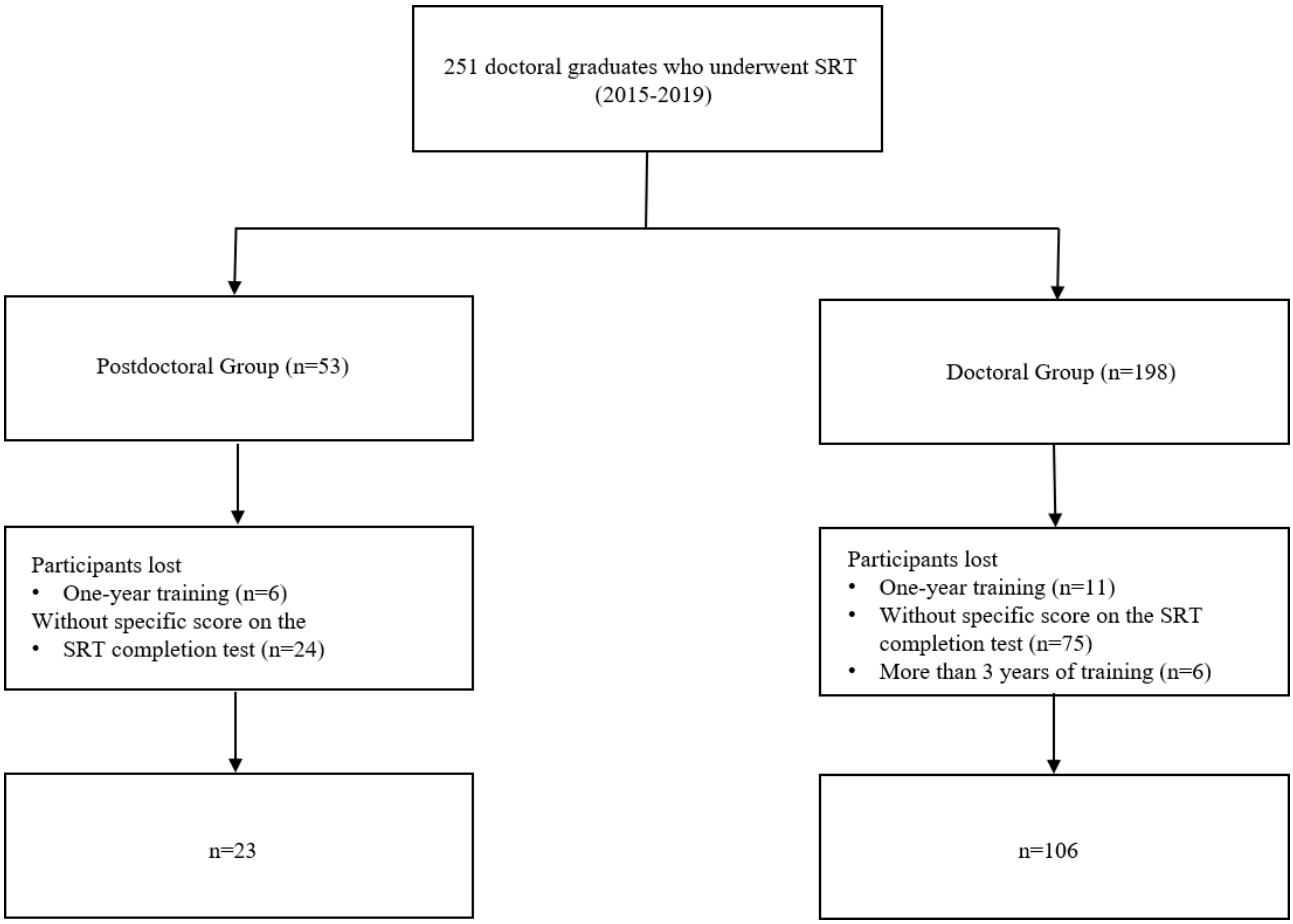


Table . General information of all the trainees enrolled in the study (N=129).

Parameter	Doctoral (n=106)	Postdoctoral (n=23)	Statistic (*df*)	*P* value
Gender, n (%)			0.706 (1)[a]	.40
Male	59 (55.66)	15 (65.22)		
Female	47 (44.34)	8 (34.78)		
Age (years), mean (SD)	28.05±1.73	27.85±1.91	0.485 (127)[b]	.63
Discipline, n (%)			1.346 (2)[a]	.51
Internal medicine	31 (29.25)	4 (17.39)		
Surgery	36 (33.96)	9 (39.13)		
Others	39 (36.79)	10 (43.48)		
Training years, n (%)			18.612 (1)[a]	<.001
2 years	16 (15.09)	13 (56.52)		
3 years	90 (84.91)	10 (43.48)		

[a]Chi-square values.

[b]*t* values.

Clinical Performance

The postdoctoral group demonstrated significantly higher clinical performance, as measured based on the standardized theoretical examination scores (445.70, SD 14.67 vs 435.12, SD 15.29; *P*=.003). The CMI was also significantly higher in the postdoctoral group (1.14 vs 0.92, *P*=.03), indicating exposure to and management of more complex clinical cases. However, analysis revealed no statistically significant differences between the 2 groups regarding the three key metrics: graduation

examination pass rate, annual theoretical assessment scores, and annual skill assessment scores (as shown in Table 3).

Table . Comparison between the doctoral and postdoctoral groups.

Parameter	Doctoral (n=106)	Postdoctoral (n=23)	Statistic (*df*)	*P* value
Clinical performance				
Theoretical scores of SRT[a] completion test, mean (SD)	435.12 (15.29)	445.70 (14.76)	–3.024 (127)[b]	.003
Pass rate of graduation examination, n (%)	99 (93.40)	20 (86.96)	0.380 (1)[c]	.54
Annual assessment of theoretical scores, mean (SD)	81.84 (10.44)	80.70 (12.67)	0.455 (127)[b]	.65
Annual assessment of skill scores, mean (SD)	86.24 (4.81)	87.85 (4.72)	–1.455 (127)[b]	.15
CMI[d], median (IQR)	0.92 (0.39)	1.14 (0.46)	–2.130[e]	.03
Scientific research ability				
Number of papers published, mean (SD)	1.11 (1.47)	2.35 (2.39)	–3.225 (127)[b]	.002
Number of achievement awards won as a participant, mean (SD)	0.06 (0.303)	0.09 (0.288)	–0.439 (127)[b]	.66
Proportion of national NSFC[f] approvals, n (%)	47 (44.34)	14 (60.87)	2.072 (1)[c]	.15
Number of provincial NSFC applications approved, n (%)	18 (17.00)	11 (47.83)	10.318 (1)[c]	.001
Teaching ability				
Teaching articles/teaching reform projects/teaching competitions, mean (SD)	0.10 (0.306)	0.22 (0.518)	–1.402 (127)[b]	.16
Comprehensive ability, n (%)				
Promotion of professional title	100 (94.34)	23 (100)	0.387 (1)[c]	.53
Master's degree supervisor	6 (5.67)	3 (13.04)	0.654 (1)[c]	.42

[a]SRT: standardized residency training.

[b]*t* value.

[c]Chi-square value.

[d]CMI: Case Mix Index.

[e]Mann-Whitney *U* test value; df value not applicable.

[f]NSFC: National Natural Science Foundation of China.

Scientific Research Ability

According to the statistics provided by the Scientific Research Department of the hospital, the number of published papers per capita was significantly higher in the postdoctoral group than in the doctoral group (2.35, SD 2.39 vs 1.11, SD 1.47; P=.002). In the doctoral group, 18 of 106 (17.00%) participants received a provincial NSFC fund, while in the postdoctoral group, 11 of 23 (47.83%) participants received a provincial NSFC fund, and the difference was significant between the 2 groups (*P*=.001). In terms of the number of provincial NSFC funds and the number of published papers per capita, the postdoctoral group had significantly better metrics than the doctoral group. As for NSFC applications, 14 of 23 (60.87%) postdoctoral participants and 47 of 106 (44.34%) doctoral trainees received approval. Although there was no significant difference between the 2 groups, the postdoctoral group may have been granted more funding if the sample size was large enough.

Teaching Skills

In this study, the trainees' teaching ability was described by the number of approved teaching improvement projects, number of published teaching papers, and whether they had participated in teaching competitions and awards. These indicators allow for the assessment of the trainees' participation and creativity in clinical teaching. As most trainees did not fully participate in clinical teaching and few papers were published or approved, there was no significant difference between the 2 groups.

Comprehensive Ability

In this study, there was no significant difference between the 2 groups on comparing whether or not an intermediate title was obtained within 3 years. On the other hand, trainees are eligible to become graduate supervisors, which requires a strong scientific research background. Applicants must meet the following criteria: they must chair at least 1 ongoing national or provincial research project, publish academic papers in high-level journals as first author or corresponding author, or receive a research award at the provincial or ministerial level or higher. A certified supervisor's evaluation is additionally necessary. Due to the very limited number of supervisors authorized to supervise graduate trainees, there was no significant difference between the postdoctoral and doctoral groups.

Discussion

Principal Findings

In this study, we compared the competencies of medical graduates who completed SRT (the doctoral group) with those who undertook the advanced clinical postdoctoral program (the postdoctoral group). The principal findings reveal distinct advantages for the postdoctoral group in key areas. Specifically, these trainees demonstrated a superior capacity to manage complex clinical cases, as evidenced by a significantly higher median CMI (1.14 vs 0.92, *P*=.03). Their research productivity was also markedly greater, reflected in a higher number of published papers (2.35, SD 2.39 vs 1.11, SD 1.47; *P*=.002) and a significantly greater success rate in securing provincial-level NSFC-funded projects (47.83% vs 17.00%, *P*=.001). Furthermore, the postdoctoral group achieved significantly higher scores on standardized theoretical examinations over the final 3 years of the study period (445.70, SD 14.67 vs 435.12, SD 15.29; *P*=.003), indicating a deeper mastery of medical knowledge. Collectively, these quantitative findings support the hypothesis that the integrated, postdoctoral training model more effectively enhances clinical readiness, scientific innovation, and theoretical understanding among medical graduates.

Implications of Findings

The findings from this study have significant implications for advancing medical education and workforce development. The superior clinical performance of the postdoctoral group, as quantified by higher theoretical examination scores and a greater ability to manage complex cases (reflected in the significantly higher CMI), suggests that such integrated training can produce clinicians with enhanced diagnostic acumen and decision-making confidence, which may directly translate to improved patient care and clinical outcomes. Furthermore, the program's effectiveness in cultivating physician scientists is clearly demonstrated by the marked increase in research output and success in securing provincial-level grants, bridging a critical gap between clinical practice and scientific innovation. While teaching competence showed limited quantitative difference, the structured mentorship and teaching experiences are posited to foster essential communication skills and educational confidence, building a foundation for future academic leaders. Collectively, these results provide a strong evidence-based rationale for other institutions and policymakers to adopt similar integrated clinical academic training models, underscoring the value of a unified framework that concurrently develops clinical, research, and teaching competencies to prepare physicians for multifaceted roles in modern health care.

Comparison to the Literature

Our findings align with those of previous studies highlighting the importance of structured postdoctoral training in fostering clinical and academic excellence [6,10-12]. For instance, integrated clinical academic training models in Western countries—such as clinical fellowships and combined MD/PhD programs—have been shown to enhance research output and leadership roles among physicians [13,14].

However, unlike traditional postresidency research fellowships, the clinical postdoctoral program described here places equal emphasis on clinical service, academic inquiry, and teaching, reflecting a unique model tailored to China's evolving health care and educational systems. Our results contribute to the growing body of evidence supporting innovative postdoctoral medical training models that extend beyond conventional residency programs.

Strengths and Limitations

This study represents one of the first efforts to evaluate the effectiveness of a clinical postdoctoral training program in mainland China, offering a comprehensive assessment of trainees' clinical, research, and teaching competencies. A key strength lies in the use of objective performance indicators, such as the CMI and research output metrics, which provide quantifiable evidence of clinical complexity management and academic productivity. Additionally, the inclusion of multiple competency domains allows for a more holistic understanding of the program's impact on professional development. However, several limitations should be acknowledged. As a single-center study with a relatively small sample size (23 postdoctoral participants), the findings may lack generalizability and statistical power, increasing the risk of type II error. It should be noted that this analysis only included participants who completed the entire program and possessed complete assessment data. Exclusions from the final analysis were primarily for 2 reasons: voluntary withdrawal due to personal circumstances and the absence of a quantitative theoretical score in the completion examination, which precluded a fair comparative assessment. These exclusions may introduce elements of attrition and selection bias. The 3-year observation period may also be insufficient to capture the long-term developmental trajectory of medical professionals. Furthermore, selection bias cannot be ruled out, as participation in the postdoctoral program was self-selected, potentially reflecting preexisting differences in motivation or academic background. Some outcome measures, particularly those related to teaching and comprehensive abilities, relied on qualitative or semiquantitative assessments, limiting their objectivity and reproducibility. To address these limitations, future studies should adopt multicenter designs with larger cohorts and extended follow-up periods to better assess the longitudinal impact of the program. In addition, more standardized and

validated tools should be developed to objectively measure teaching and comprehensive competencies. Comparative analyses with international postgraduate training models could further inform best practices and opportunities for cross-learning in clinician scientist development.

Conclusion

As a new medical education mode in China, the clinical medical postdoctoral training program enhances the trainees' clinical performance and scientific research innovation ability, which is helpful to cultivate the top innovative talents in clinical medicine. This training mode is effective and can be promoted in various medical teaching institutions as an important supplement to the current medical degree system in China.

Acknowledgments

We appreciate the hospital's Human Resources, Scientific Research Department, and Medical Record Room for providing data for this study.

Funding

This study was supported in part by a grant from Ministry of Education industry-university-research cooperation project (220604942134732) and a grant from the Teaching Reform Project of the School of Medicine, Zhejiang University (jgyb2025025).

Data Availability

All data generated or analyzed during this study are included in this published article. The raw data that support the findings of this study are available on request from the corresponding author HL upon reasonable request.

Authors' Contributions

Conceptualization: HL
Data curation: LZ and LS
Formal analysis: HL and LZ
Investigation: LZ and LS
Writing – original draft: LZ
Writing – review & editing: HL, LZ, and LS

Conflicts of Interest

None declared.

References

1. National Health Commission of the People's Republic of China. NHCotPsRo: China officially launched the construction of standardized training system for resident doctors. URL: https://www.nhc.gov.cn/qjjys/c100015/201402/b55fda83401e4f6ebbc942d77e9e2a58.shtml [accessed 2014-02-13]
2. Zhu J, Li W, Chen L. Doctors in China: improving quality through modernisation of residency education. Lancet 2016 Oct 15;388(10054):1922-1929. [doi: 10.1016/S0140-6736(16)00582-1] [Medline: 27339756]
3. National Health Commission of the People's Republic of China. NHCotPsRo: Notice of the General Office of the National Health and Family Planning Commission on the Issuance of the Criteria for the Recognition of Residency Standardization Training Bases (for Trial Implementation) and the Contents and Standards of Residency Standardization Training (for Trial Implementation). URL: https://www.nhc.gov.cn/qjjys/c100016/201408/3c0b15d947aa412dabbc1275d22c8052.shtml [accessed 2014-08-26]
4. Foundation SCfOSaEotMoHRaSS-CPS: Notice of the Office of the National Postdoctoral Administration Committee on the Training of Postdoctoral Fellows in Clinical Medicine. Ministry of Human Resources and Social Security. URL: https://www.chinapostdoctor.org.cn/article?mid=eb19886fa6c44c5b25c6a0d75e237c78&catname=%E9%80%9A%E7%9F%A5%E5%85%AC%E5%91%8A&catid=8c892b1c-4ade-4a5f-9a87-5e736cb5e9f9 [accessed 2020-09-12]
5. Medicine ZUSo: The Postdoctoral Expert Committee of the School of Medicine, Zhejiang University was established, and its first plenary meeting was successfully held. School of Medicine, Zhejiang University. URL: http://www.cmm.zju.edu.cn/2015/0906/c38670a1629252/page.htm [accessed 2015-09-06]
6. Lei C, Shaohua C, Xiangming F, Zhi C, Jianhong L, Denian B. Clinical Postdoctoral Training Program: exploration and practice of Zhejiang University. China Higher Medical Education 2019 Apr 15(4):17-18. [doi: 10.3969/j.issn.1002-1701.2019.04.009]
7. Eshel N, Chivukula RR. Rethinking the physician-scientist pathway. Acad Med 2022 Sep 1;97(9):1277-1280. [doi: 10.1097/ACM.0000000000004788] [Medline: 35731582]

8. Harding CV, Akabas MH, Andersen OS. History and outcomes of 50 years of physician-scientist training in Medical Scientist Training Programs. Acad Med 2017 Oct;92(10):1390-1398. [doi: 10.1097/ACM.0000000000001779] [Medline: 28658019]
9. Buckley S, Smith M, Patel J, Gay S, Davison I. Enhanced model for leadership development for trainees and early career health professionals: insights from a national survey of UK clinical scientists. BMJ Lead 2022 Sep;6(3):212-218. [doi: 10.1136/leader-2021-000465] [Medline: 36170475]
10. Lei C, Hengchao R, Lingxiao X, Xiangming F, Denian B. Multidisciplinary online case discussion: the exploration of online medical education in clinical postdoctoral training program. China Higher Medical Education 2020 Apr 15(4):1-2. [doi: 10.3969/j.issn.1002-1701.2020.04.001]
11. Xiangming F, Wei H, Shiqi X, et al. Iterative innovation empowers the training of high-level interdisciplinary medical talents in the new era: experience from School of Medicine, Zhejiang University. Medical Journal of Peking Union Medical College Hospital 2022 Jan 27;13(1):9-12. [doi: 10.12290/xhyxzz.2021 0718]
12. Shuyang Z. Exploration and practice of the training system of comprehensive medical talents at Peking Union Medical College Hospital. Medical Journal of Peking Union Medical College Hospital 2021 Aug 4;13(1):5-8. [doi: 10.12290/xhyxzz.2021 0535]
13. Williams CS, Rathmell WK, Carethers JM, et al. A global view of the aspiring physician-scientist. Elife 2022 Sep 13;11:e79738. [doi: 10.7554/eLife.79738] [Medline: 36098684]
14. Steer CJ, Jackson PR, Hornbeak H, McKay CK, Sriramarao P, Murtaugh MP. Team science and the physician-scientist in the age of grand health challenges. Ann N Y Acad Sci 2017 Sep;1404(1):3-16. [doi: 10.1111/nyas.13498] [Medline: 28981971]

Abbreviations

CMI: Case Mix Index
NSFC: National Natural Science Foundation of China
SRT: Standardized Residents Training

Edited by B Lesselroth; submitted 20.08.24; peer-reviewed by S Mohanadas, X Qi; accepted 30.10.25; published 24.11.25.

Please cite as:
Zhang L, Sun L, Li H
Exploration and Practice of the First Clinical Medical Postdoctoral Program in China: Retrospective, Nonrandomized, Controlled Study
JMIR Med Educ 2025;11:e65622
URL: https://mededu.jmir.org/2025/1/e65622
doi:10.2196/65622

Comparison of Physician Assistant and Medical Students' Clinical Reasoning Processes Using an Online Patient Simulation Tool to Support Clinical Reasoning (eCREST): Mixed Methods Study

Alistair Thorpe[1,2], PhD; Angelos P Kassianos[3], PhD; Ruth Plackett[1,4], PhD; Vinodh Krishnamurthy[5], PhD; Maria A Kambouri[6], PhD; Jessica Sheringham[1], FFPH, PhD

[1]Institute of Epidemiology and Health Care, University College London, 1-19 Torrington Place, London, United Kingdom

[2]Department of Population Health Sciences, Spencer Fox Eccles School of Medicine, University of Utah, Salt Lake City, UT, United States

[3]Cyprus University of Technology, Limassol, Cyprus

[4]Department of AI in Preventative Medicine School of Life Course & Population Sciences, King's College London, London, United Kingdom

[5]Royal Free Hospital, Royal Free London NHS Foundation Trust, London, UK

[6]Institute of Education, University College London, London, United Kingdom

Corresponding Author:
Jessica Sheringham, FFPH, PhD
Institute of Epidemiology and Health Care, University College London, 1-19 Torrington Place, London, United Kingdom

Abstract

Background: Clinical reasoning is increasingly recognized as an important skill in the diagnosis of common and serious conditions. eCREST (electronic Clinical Reasoning Educational Simulation Tool), a clinical reasoning learning resource, was developed to support medical students to learn clinical reasoning. However, primary care teams now encompass a wider range of professional groups, such as physician assistants (PAs), who also need to develop clinical reasoning during their training. Understanding PAs' clinical reasoning processes is key to judging the transferability of learning resources initially targeted to medical students.

Objective: This exploratory study aimed to measure the processes of clinical reasoning undertaken on eCREST by PA students and compare PAs' reasoning processes with previous data collected on medical students.

Methods: Between 2017 and 2021, PA students and medical students used eCREST to learn clinical reasoning skills in an experimental or learning context. Students undertook 2 simulated cases of patients presenting with lung symptoms. They could ask questions, order bedside tests, and select physical exams during the case to help them form, reflect on, and reconsider diagnostic ideas and management strategies while completing a case. Exploratory analysis was undertaken by comparing students' data gathering, flexibility in diagnosis, and diagnostic ideas between medical and PA students.

Results: In total, 159 medical students and 54 PA students completed the cases. PAs were older (mean 27, SD 7 y vs mean 24, SD 4 y; $P<.001$) and more likely to be female (43/54, 80% vs 84/159, 53%; $P<.001$). Medical and PA students were similar in the proportion of essential questions asked (Case 1: mean 70.1 vs mean 73.2; $P=.33$; Case 2: mean 74.6 vs mean 70.9; $P=.27$), physical examinations requested (Case 1: mean 54.7 vs mean 54.0; $P=.59$; Case 2: mean 69.3 vs mean 67.5; $P=.59$), bedside tests selected (Case 1: mean 74.4 vs mean 83.3; $P=.05$; Case 2: mean 47.9 vs mean 50.0; $P=.69$), and number of times they changed their diagnoses (Case 1: mean 2.8 vs mean 2.8; $P=.99$; Case 2: mean 2.8 vs mean 2.5; $P=.81$). Both student groups improved in their diagnostic accuracy during the cases.

Conclusions: These results provide suggestive evidence that medical and PA students had similar clinical reasoning styles when using an online training tool to support their diagnostic decision-making.

(JMIR Med Educ 2025;11:e68981) doi:10.2196/68981

KEYWORDS

clinical reasoning; medical education; primary care; educational technology; patient simulation tool; physician assistant; medical student

Introduction

Diagnostic error has been identified as the most common cause of avoidable harm in primary care [1]. Efforts to improve the clinical reasoning processes of health care staff who investigate and refer patients are critical for limiting the burden of disease by reducing diagnostic errors [2]. Clinical reasoning broadly encompasses the thought processes and strategies that underlie

clinical judgments. Clinical reasoning is a core skill with relevance to many facets of clinical practice [3] and with a particular influence on clinicians' judgments when attending to patients with symptoms that might be serious disease [4-6]. As a result, a need for formal education around clinical reasoning at all stages of the journey—from student to advanced professional—has been articulated [7,8].

eCREST (electronic Clinical Reasoning Educational Simulation Tool), an online patient simulation training tool, was developed for medical students to address this need [9]. It was tested with final-year medical students in UK medical schools, demonstrating good efficacy and receiving positive feedback from student testers about its value as an educational resource in improving clinical reasoning skills [10]. Interviews with medical student learners using eCREST, combined with analysis of their actions whilst using eCREST, demonstrated they displayed a range of data gathering strategies, with those more experienced on eCREST displaying more thorough data gathering strategies and identifying more essential diagnostic information [11].

Since eCREST was first developed, the landscape for primary care provision changed. In England, for example, primary care now encompasses a wider range of professionals beyond physicians and nurses to include pharmacists and social prescribers [12]. In addition, many primary care organizations now use physician assistants (PAs), with numbers working in primary care set to expand further [13]. PAs are a relatively new role in the United Kingdom, comparable to that of the PA in the United States, which have existed for over 50 years [14]. PA roles typically include taking patients' medical histories, conducting physical exams, formulating differential diagnoses, and proposing management plans, including potential referrals [15]. Qualified PAs are master's-level graduates who have completed a 2-year postgraduate course following a medical curriculum.

The expanded primary care team and the expansion of clinical roles have resulted in a need for clinical reasoning training suitable for a range of future clinical professionals, not just medical students. In a recent review of clinical reasoning across health professions, no resources were identified targeted toward PAs, as students or qualified professionals [7]. We adapted eCREST for PA students, in consultation with PA faculty, which involved ensuring the case description referred to PAs not just general practitioners (GPs). Adaptations were minimal but included changing the introduction at the start of cases to frame the cases for PAs and selecting from existing eCREST cases those that best fit the PA curriculum. The usefulness of eCREST as a meaningful and useful tool for PAs rests on the assumption that PAs approach clinical reasoning in a similar way to medical students. However, there is very limited understanding of how PA students in the United Kingdom acquire clinical reasoning skills, and whether they would exhibit similar reasoning processes to medical students. Insight into their reasoning processes while learning could be useful in designing multiprofessional education, for both current and future clinical professionals.

This exploratory study aimed to measure the processes of clinical reasoning undertaken on eCREST by PA students and compare PAs' reasoning processes with previous data collected on medical students. It tests the a priori hypothesis that PA learners exhibit similar reasoning processes to those of medical students when using eCREST.

Methods

Theoretical Framework

In common with Young et al [16], we recognize that clinical reasoning is not a homogenous construct, nor is it understood in the same way by different individuals or professional groups. How clinical reasoning is defined has implications both for the focus of learning resources and how the effectiveness of such resources is evaluated.

Young et al [17] have mapped how different reasoning constructs and theories influence approaches to teaching and assessment, distinguishing between theories focused on acquisition of knowledge, knowledge organization, cognitive processes, or meta-cognitive processes. Informed by Plackett et al's [18] review of the effectiveness of online simulated patient tools, eCREST was designed to focus on dimensions of clinical reasoning amenable to improvement and that are associated with common diagnosis-related biases in reasoning [9]. In this paper, therefore, we focus on clinical reasoning theories related to cognitive and meta-cognitive processes and display how these relate to eCREST in Table 1. Aligned with Young et al [16,17], we consider the process of clinical reasoning focusing on cognitive skills, with the outcome being reduction in cognitive biases supporting a goal of improving diagnostic reasoning.

Table . Application of clinical reasoning theory taken from Young et al [17] to the design and content of eCREST[a].

Theoretical focus: good clinical reasoning requires...	Approaches to support students to develop clinical reasoning	Application to design of eCREST to develop reasoning skills	Reasoning measures derived from students' completion of eCREST
...an ability to recognize relevant features in a clinical presentation and test hypotheses	Case-based learning using real or simulated cases	eCREST comprises a series of simulated patients through which students work.	• Data gathering • Diagnostic accuracy
...an ability to monitor one's own processes for possible errors or biases and reflect on one's own reasoning	Prompted reflection, clinical justification (eg, in patient notes)	eCREST seeks to influence learning through prompting reflection.	• Diagnostic flexibility • Reflections after the case
...sufficient knowledge base	Lectures and readings	eCREST was not designed to impart knowledge, but it starts with a knowledge quiz as a self-diagnostic, so students can revise key content before the case starts if needed.	Multiple-choice questions

[a]eCREST: electronic Clinical Reasoning Educational Simulation Tool.

Design

We conducted a retrospective comparison of clinical reasoning processes and outcomes between medical and PA students using the eCREST online educational platform to teach clinical reasoning.

Platform

eCREST is an online patient simulation training tool that has been tested with final-year medical students in medical schools in the United Kingdom [10,11]. eCREST [19] is proprietary software but the platform is available for use free to educational settings that are willing to contribute anonymized data and provide feedback to support its evaluation.

A workflow through eCREST is shown in Figure 1, with illustrative screengrabs from eCREST's website during a simulated case. Briefly, students log on to the eCREST website and enter a virtual "waiting room." They select a "patient" by clicking on a still of a patient actor and name, which opens a video of a patient actor who gives a brief account of the problem from the patient's perspective for which they are seeking medical advice (Figure 1, panel 1). The student is then prompted to provide their ideas about likely diagnosis, through selecting 5 differentials from a long list of potential diagnoses; indicating on a scale of 1 - 5 how worried they are about the patient, their ideas about likely diagnosis and typing as free text their reasoning behind their choices (Figure 1, panel 2). The student can then click on options to obtain further information—as videos or in text form—about the patient's disclosed symptom, the presence or absence of other symptoms, and wider health (Figure 1, panel 3). At regular points during the "consultation," the student is prompted to revisit their initial differential diagnosis choices. Finally, once students decide they have sought all the information they need, they are required to finalize their differential diagnosis and propose a management plan by selecting from further investigation and referral options and entering their rationale as free text (Figure 1, panel 4). Students can then access feedback, in the form of videos of GPs discussing why they prioritized certain differentials over other options. Once the case is completed, they are asked for their reflections and have the option to download a summary of their reasoning processes and GP feedback for each case (Figure 1, panel 5).

Figure 1. A student's typical workflow through eCREST. Author JS is pictured in this figure. eCREST: electronic Clinical Reasoning Educational Simulation Tool; GP: general practitioner.

The content for the simulated cases included in this study was developed with clinicians, researchers, and patients informed by literature on diagnosis of common symptoms that could indicate lung cancer. A panel of GPs with a range of experience was convened to discuss and agree on the form of feedback. For each case, a GP in training developed the initial iteration of the case and proposed which information (ie, which questions, tests, and examinations) would be "essential" and what

diagnoses would be most clinically important to consider. Each case was then reviewed by experts in primary care, diagnostics, and the disease systems as relevant to the case, to agree on the case content, including the options for differentials and the relevance of information sought.

Recruitment

We worked with higher educational institutions (HEIs) in the United Kingdom to recruit participants. HEIs were sought with a diversity of teaching approaches and identified through personal contacts of the research team. Between 2017 and 2021, 4 universities providing PA programs and 3 medical schools used eCREST with their students. In the United Kingdom, undergraduate medicine courses typically have a duration of at least 5 years. In some cases, courses last 6 years, where students take an intercalated BSc between years 2 and 3. In the first years of training, students are typically mainly classroom-based. In later years of training, students have longer and more concentrated clinical placements, which include general practice, where they are required to be present in a general practice environment, primarily for educational purposes. During clinical placements in undergraduate medicine, students are mainly observing clinical interactions with opportunities for case-based discussion before and afterwards, though they may have some small direct care tasks under supervision, which increase in their volume and complexity, to prepare undergraduates for the foundation stage in their medical career, in clinical environments [20]. PAs typically undertake a 2-year master's program, which covers similar content to medicine (eg, anatomy, physiology, and pharmacology) and provides exposure to clinical environments, including primary care, through clinical placements. Master's graduates who pass the national Faculty of PA exams are then permitted to work as PAs throughout the United Kingdom's National Health Service.

Medical students in this study in all 3 HEIs were in their final year of undergraduate training, PAs were in their final year of their master's program. Students at each HEI were mainly recruited through advertisement in faculty newsletters, lecture "shout-outs," social media, and tutor promotion between 2017 and 2018 and subsequently through the medical school promotion of eCREST via the teaching tutors. For medical students, participation was part of a feasibility trial. It was outside of timetabled lectures and entirely voluntary. However, for PA students, eCREST was introduced in relevant modules (eg, respiratory) by the study team in collaboration with teaching faculty during a scheduled session with the option for discussion of cases during later teaching sessions. Educators in each setting were consulted as to the best way to introduce eCREST to maximize possible educational value to students and educators and minimize burden to both.

As eCREST is an educational resource implemented within educational settings, the sample size was determined by uptake among students from the schools that had agreed to use eCREST. We did not perform a priori sample size calculations as part of this exploratory project. The content on eCREST was presented in English and accessed online by students.

Procedure

On the advice of educators, UK medical students had access to 4 cases on eCREST, while PA students only had access to 2 cases. Analysis was undertaken on the 2 simulated cases that were common to all students. The cohorts completing each case were slightly different, but the results from both cases did not differ substantively, so we present data for the first case in the main text and the second case in Multimedia Appendix 1.

Simulated Case

Students were presented with a 58-year-old female of Asian ethnicity presenting with chest pain (Case 1) and a 91-year-old male of White British ethnicity presenting with a cough (Case 2). Other information on the cases, including other symptoms experienced by the patient but not volunteered, was available to the student through selecting questions to ask the "patient" (to which they received a video reply) and by clicking on medical history or examination results. A full description of eCREST is available in Kassianos et al [9].

Measures

When registering for eCREST, all students self-reported their program [medical student or physician assistant student], gender identity [male or female], and age [in y].

We selected quantitative outcome measures to align with three domains of clinical reasoning identified in the literature: data gathering, flexibility in thinking about diagnoses, and diagnostic accuracy [18]. We also included measures coded from student reflections obtained from free text data. These dimensions are:

1. Data gathering: students' ability to elicit the necessary information to make an informed judgment.
2. Flexibility in thinking about diagnoses: students' capacity to remain open-minded and to consider alternative views.
3. Diagnostic accuracy: students' ability to identify the most relevant potential diagnoses for the patient.

Our previous work identified a gap in validated measures of clinical reasoning processes that are sensitive to change from short-term courses or training [10,11]. We have therefore developed our own measures of each of these dimensions, specific to eCREST.

In addition, we compare students' free-text reflections about their own performance and transferable learning from the case to clinical settings.

Data Gathering

Three measures were selected to reflect students' case-specific clinical reasoning processes:

1. Essential questions asked: an expert panel of GPs determined that 20 of the 32 available questions were essential for coming to an informed diagnosis. For data analyses, this measure was calculated as (number of essential questions asked/20) × 100. For case 2, the expert panel identified 14 essential questions.
2. Essential bedside tests requested: an expert panel of GPs determined that for case 1, three of the 9 available questions were essential for coming to an informed diagnosis and for data analyses it was calculated as (number of essential

bedside tests requested/3) × 100. For case 2, the expert panel identified 2 essential bedside tests.

3. Essential physical exams selected: an expert panel of GPs determined that for both cases, 6 of the 11 available physical exams were essential for coming to an informed diagnosis. For data analyses, this measure was calculated as (number of essential physical exams selected/6) × 100.

Flexibility in Thinking About Diagnoses

Students were asked to select 6 specific diagnoses at the very beginning of the cases. They could then review their diagnoses at any point (with the option to change any of the diagnoses or leave it the same). They were also prompted to review their list of diagnoses every time they asked 6 questions, after they had asked 8 tests or exams, and then after they had asked for 2 more types of information (eg, question answer or a test or exam). We measured flexibility in thinking about diagnoses in 2 ways as the average number of times students changed diagnosis and the proportion of students who changed diagnoses at least once. We defined changing diagnosis as any one of the following: (1) introducing a new diagnosis they had not previously considered, (2) removing a diagnosis that they had previously considered, and (3) revising the order in which they prioritized the diagnoses they were considering.

Relevant Diagnostic Accuracy

This was measured as the proportion of relevant diagnoses considered. An expert panel of GPs determined that, for both cases, 6 of the 13 available diagnoses were relevant to the patient case. This measure was calculated as (number of relevant diagnoses selected/6) × 100.

Student Reflections

At the end of the case, students responded to four reflective questions on the case they had just completed, comprising: (1) What do you think you did well? (2) What do you think you need to improve? (3) What changed during the case? (4) What will you take forward from doing these simulated cases to your clinical work?

Data Analysis

All analyses were conducted in R Studio (version 1.4.1106; Posit PBC) [21]. We compared whether the proportion of essential questions asked, bedside tests requested, physical exams selected, number of times changed diagnoses, and relevant diagnoses considered differed according to student type (medical student vs PA students). Qualitative responses to the four reflection questions were reviewed and coded to indicate whether the student's reflection related to any of 6 themes, selected in part deductively based on eCREST's goals of challenging cognitive biases, and in part inductively informed by ideas expressed in many of the student reflections: Knowledge, open-mindedness, awareness of cognitive biases, confidence or uncertainty, content of the case, no meaningful response.

These are retrospective analyses from a larger study of the eCREST tool. They were not preregistered, and a sample size calculation was not conducted. Significance was set at 2-sided α=.05 and we did not perform any adjustment for multiple comparisons in this exploratory study and instead report unadjusted effect sizes, 95% CIs, and raw *P* values [22].

Ethical Considerations

This study was approved by University College London and the participating medical schools (15453/002). Invited student participants consented to take part and for their data to be analyzed as part of research on eCREST.

The undergraduate medical students in the United Kingdom were final-year students in a 6-year course, whereas the PAs were final-year students on a 2-year postgraduate training course, with an undergraduate degree in biomedical sciences. Information sheets and consent forms were shared through each student's course lead and were also available for download on eCREST. Medical students participated as part of a trial and were offered compensation for completing cases (a maximum of £30 [US $40.54 at the time of the study; US $39.47 in 2025] in vouchers). PAs participated as part of their course. The information sheets confirmed that data would be stored in secure servers at University College London and used for research purposes and confirmed that their performance would not be shared with their course organizers.

Results

Participant Characteristics

A total of 654 students accessed and registered on eCREST during the study period. Of these, 213 students completed the patient case and were included in the analyses. Most were medical students (159/213, 75%), evenly distributed between 3 medical schools (School A: 51/159, 32.1%; School B: 60/159, 37.7%; School C: 48/159, 30.2%). Of the 54 physician assistant students, the majority came from one medical school (School D: 40/54, 74.1%; School E: 7/54, 13%; School F: 4/54, 7.4%; School G: 3/54, 5.6%; Table 2).

Table . Student institutions according to student type.

		Student participants	Completion[b]	
	Recruitment years[a]	Values, n (%)	Rate	Percent
Medical students				
School A	2017 and 2020	51 (32.1)	51/162	31.5
School B	2017 - 18 and 2020 - 21	60 (37.7)	60/168	35.7
School C	2017 - 18 and 2020 - 21	48 (30.2)	48/159	30.2
Physician assistant students				
School D	2021	40 (74.1)	40/126	31.7
School E	2020 - 21	7 (13)	7/15	46.7
School F	2020	4 (7.4)	4/18	22.2
School G	2020	3 (5.6)	3/5	60

[a]All 2017 cohorts were part of a larger feasibility trial evaluation. Any other cohorts were recruited openly.

[b]Completion was calculated with all students who registered for eCREST from each school as the denominator.

The mean age of students in our sample was 25 (SD 5; range 19-48) years and most were female (n=127, 60%). The PA student group was older on average than the medical student group (*P*=.001) and contained a greater proportion of females (*P*=.001; Table 3).

Table . Student demographics overall and according to student type.

	Overall (n=213)	Medical students (n=159)	Physician assistant students (n=54)	Medical students versus physician assistant students	
				Values	*P* value
Age in years					
Mean (SD)	25 (5)	24 (4)	27 (7)	−3.48 (−5.53 to −1.42)[a]	<.001
Median (range)	24 (19-48)	23 (19-42)	25 (19-48)		
Age (years), n (%)					
19-24	142 (67)	117 (74)	25 (46)	—[b]	
25-34	56 (26)	35 (22)	21 (39)	—	
35-44	7 (3)	4 (3)	3 (6)	—	
45 and older	4 (2)	—	4 (7)	—	
Did not respond	4 (2)	3 (2)	1 (2)	—	
Gender, n (%)					
Female	127 (60)	84 (53)	43 (80)	10.9 (0.24)[c]	<.001
Male	86 (40)	75 (47)	11 (20)	—	

[a]Difference (95% CI).

[b]Not applicable.

[c]X^2 value (ϕ). Degrees of freedom=1.

Clinical Reasoning Measures

Full results are displayed in Table 4 and summarized below by outcome measure.

Table . Clinical reasoning outcome measures overall and according to student type.

	Overall (n=209)	Medical students (n=159)	Physician assistant students (n=54)	Medical students versus physician assistant students	
				Difference (95% CIs)	*P* value
Data gathering					
Essential questions				−3.1 (−9.4 to 3.2)	.33
Mean (SD)	70.9 (22.2)	70.1 (23.2)	73.2 (18.9)		
Median (range)	75 (0 to 100)	75 (0 to 100)	73 (10 to 100)		
Physical exams				−1.8 (−8.8 to 4.7)	.59
Mean (SD)	55.2 (19.2)	54.7 (18.6)	54.0 (21.1)		
Median (range)	50 (0 to 100)	50 (0 to 100)	50 (0 to 100)		
Bedside tests				−8.9 (−17.9 to 0.1)	.53
Mean (SD)	76.7 (29)	74.4 (29.1)	83.3 (28.8)		
Median (range)	100 (0 to 100)	67 (0 to 100)	100 (0 to 100)		
Flexibility in thinking about diagnoses					
Times changed diagnoses				0.0 (−0.5 to 0.5)	.99
Mean (SD)	2.8 (1.4)	2.8 (1.4)	2.8 (1.5)		
Median (range)	3 (0 to 7)	3 (0 to 7)	3 (0 to 6)		
0 times, n (%)	10 (4.7)	8 (5)	2 (3.7)	0.0 (0.03)[a]	.98
1 times, n (%)	33 (15.5)	23 (14.5)	10 (18.5)	—	—
2 times, n (%)	45 (21.1)	33 (20.8)	12 (22.2)	—	—
3 times, n (%)	60 (28.2)	49 (30.8)	11 (20.4)	—	—
4 times, n (%)	41 (19.2)	27 (17)	14 (25.9)	—	—
5 times, n (%)	20 (9.4)	17 (10.7)	3 (5.6)	—	—
6 times, n (%)	3 (1.4)	1 (0.6)	2 (3.7)	—	—
7 times, n (%)	1 (0.5)	1 (0.6)	2 (3.7)	—	—
Relevant diagnoses					
Initial				−8.2 (−12.6 to −3.7)	<.001
Mean (SD)	45.5 (14.8)	43.4 (14.4)	51.5 (14.2)		
Median (range)	50 (17 to 83)	50 (16.7 to 66.7)	50 (16.7 to 83.3)		
Change				5.3 (0.1 to 10.5)	.046
Mean (SD)	18.2 (16.7)	19.5 (16.6)	14.2 (16.64)		
Median (range)	17 (−17 to 67)	17 (−16.7 to 66.7)	17 (−16.7 to 50.0)		
Final				−2.85 (−7.5 to 1.8)	.23
Mean (SD)	63.6 (15.3)	62.9 (15.5)	65.7 (14.6)		
Median (range)	67 (17 to 83)	67 (16.7 to 83.3)	67 (33.3 to 83.3)		

[a]Represents X^2 value with (ϕ) comparing students who changed diagnosis 0 times with those who changed at least 1 time by student type. Degrees of freedom=7.

Data Gathering

Essential Questions

Overall, students asked around 70% (14/20) of essential questions on average. We found no evidence of differences between PA students (mean 73.2, SD 18.9) and medical students (mean 70.1, SD 22.2) in the proportion of essential questions asked by each group (*P*=.33). Across all 20 essential questions, we found no evidence of differences between PA students and medical students in the proportion who asked each question (see Multimedia Appendix 1). For case 2, it was PAs (mean 74.6, SD 22.0) and medical students (mean 70.9, SD 22.9; *P*=.27; Multimedia Appendix 1).

Physical Exams

Overall, both medical and PA students selected around 55% (3.3/6) of essential physical exams on average. We found no evidence of differences between PA students (mean 54.0, SD 21.1) and medical students (mean 54.7, SD 18.6) in the proportion of essential physical exams selected by each group (*P*=.59). For case 2, it was PAs (mean 67.5, SD 17.9) and medical students (mean 69.3, SD 17.9; *P*=.59; Multimedia Appendix 1).

Bedside Tests

Overall, students selected around 77% (2.3/3) of essential bedside tests on average. On average, PA students (mean 83.3, SD 28.8) appeared to have selected a greater proportion of essential bedside tests than medical students (mean 74.4, SD 29.1), though the *P* value (*P*=.053) suggests there is only weak evidence for this difference. For case 2, it was PAs (mean 50, SD 50.0) and medical students (mean 47.9, SD 34.6; *P*=.69; Multimedia Appendix 1).

Flexibility in Thinking About Diagnoses

Number of Times Students Changed Diagnoses

Overall, we found students changed diagnoses 2.8 times on average. We found no evidence of differences between physician assistant students and medical students in the number of times they changed diagnoses (*P*=.99). For case 2, the average number of times the diagnosis was changed was 2.5, and there was no difference between PAs (mean 2.8, SD 1.1) and medical students (mean 2.5, SD 1.1; *P*=.08).

Changed Diagnosis at Least Once

Overall, 203 of 213 (95%) students changed their diagnosis at least once. We found no evidence of differences in the proportion of PA students (151/159, 96%) and medical students (52/54, 95%) who changed their diagnosis at least once (*P*=.98). For case 2, all students changed diagnosis at least once.

Diagnostic Accuracy

Overall, in their initial diagnosis, students included around 45% (2.7/6) of the relevant diagnoses on average. Compared with medical students (mean 43.4, SD 14.4), PA students (mean 51.5, SD 14.2) included a greater proportion of relevant diagnoses in their initial diagnosis (*P*<.001). For case 2, it was 43% (2.6/6) overall and did not differ between student groups (PAs: mean 41.5, SD 14.1 vs medical students: mean 44.4, SD 15.0; *P*=.19).

Both PA students and medical students had more relevant diagnoses in their final diagnosis than they did initially in case 1, with the overall proportion of relevant diagnoses included rising to 64% (3.8/6) on average overall. Notably, the average improvement was steeper among medical students (mean 19.5, SD 16.6) than PA students (mean 14.2, SD 16.7; *P*=.05). In their final diagnoses, both PA students (mean 65.7, SD 14.6) and medical students (mean 62.9, SD 15.5) were similar in the proportion of relevant diagnoses included (*P*=.23). For case 2 overall, it rose to 48% (SD 14.2; 2.9/6). Both PAs (mean 47.7, SD 13.9, +6.1 from initial) and medical students (mean 48.1, SD 14.3, +3.73 from initial; *P*=.84) were similar.

Student Reflections

The reflections of both PA students and medical students were similar overall, with the following themes emerging from the data:

1. Knowledge-based: Reflections that focus on acquiring, applying, or deepening clinical or biomedical knowledge relevant to the case content.
2. Open-mindedness: Comments indicating a willingness to consider alternative diagnoses, questions, tests, exams, or approaches, and to revise initial assumptions.
3. Awareness of cognitive biases: Reflections that explicitly mention or imply recognition of cognitive biases (eg, anchoring, confirmation bias) and their potential impact on clinical reasoning.
4. Confidence: Statements related to the student's self-assessed confidence in their clinical decision-making, diagnostic reasoning, or use of the eCREST tool.
5. Content: Reflections that comment on the structure, usability, or educational value of the eCREST tool itself, rather than the clinical case or reasoning process.
6. No meaningful response: Responses that lacked substantive reflection or were too brief or vague to be categorized under other themes.

Student responses are shown in Table 5.

Table . Reflection measures overall and according to student type.

	Overall (n=236), n (%)	Medical students (n=282), n (%)	Physician assistant students (n=46), n (%)	Medical students versus physician assistant students	
				Chi-square (*df*)	*P* value
What do you think you did well?					
Knowledge-based	8 (3)	8 (3)	0 (0)	0.61 (5)	.44
Open-mindedness	79 (28)	67 (28)	12 (26)	0.02 (5)	.89
Awareness of cognitive biases	2 (1)	2 (1)	0 (0)	0.00 (5)	.99
Confidence	7 (2)	7 (3)	0 (0)	0.44 (5)	.51
Case content	253 (90)	212 (90)	41 (89)	0.00 (5)	.99
No meaningful response	5 (2)	3 (1)	2 (4)	0.70 (5)	.40
What do you think you need to improve?					
Knowledge-based	49 (17)	46 (19)	3 (7)	3.65 (5)	.06
Open-mindedness	75 (27)	68 (29)	7 (15)	2.98 (5)	.08
Awareness of cognitive biases	21 (7)	21 (9)	0 (0)	3.23 (5)	.07
Confidence	13 (5)	12 (5)	1 (2)	0.23 (5)	.63
Case content	203 (72)	162 (69)	41 (89)	7.03 (5)	.008
No meaningful response	10 (4)	8 (3)	2 (4)	0.00 (5)	.99
What changed during the case?					
Knowledge-based	2 (1)	1 (0)	1 (2)	0.11 (5)	.74
Open-mindedness	91 (32)	73 (31)	18 (39)	0.84 (5)	.36
Awareness of cognitive biases	8 (3)	7 (3)	1 (2)	0.00 (5)	.99
Confidence	5 (2)	5 (2)	0 (0)	0.15 (5)	.70
Case content	154 (55)	132 (56)	22 (48)	0.72 (5)	.40
No meaningful response	55 (20)	44 (19)	11 (24)	0.39 (5)	.53
What will you take to your clinical work?					
Knowledge-based	17 (6)	14 (6)	3 (7)	0.00 (5)	.99
Open-mindedness	125 (44)	113 (48)	12 (26)	6.55 (5)	.01
Awareness of cognitive biases	17 (6)	17 (7)	0 (0)	2.37 (5)	.12
Confidence	12 (4)	9 (4)	3 (7)	0.19 (5)	.66
Case content	205 (73)	165 (70)	40 (87)	4.81 (5)	.03
No meaningful response	14 (5)	12 (5)	2 (4)	0.00 (5)	.99

Both groups of students focused most of their reflections on the content of the case (ie, describing their strategy or perceptions of how "accurate" they were) when asked what they did well, could improve, and would take into their clinical work. Recognizing the importance of being open-minded and flexible was also a common reflection by both groups of students.

However, there was moderate evidence of a difference in the learning students reported they would apply to their clinical work, in two respects. A greater proportion of medical student reflections related to open-mindedness and flexibility (48% vs 26%; *P*=.01). For example, one medical student reflected that, from now on, they would "always keep an open mind, not to be fixed on one diagnosis. Ask questions to rule out all possible differential diagnosis."

A greater proportion of PA student reflections related to improving knowledge related to the clinical content of the case

(87% vs 70%; *P*=.03). For instance, one PA student noted that they would take into their clinical work that they "have learnt which factors to consider" for the patient case, while another mentioned that "I will need to develop my knowledge of gastrointestinal differentials that cause respiratory-related symptoms."

Discussion

Principal Results

This study yielded three key findings. First, we found no differences in data gathering through eCREST between medical students and PA students. Overall, both groups of students performed well on the proportion of essential questions asked (14/20, 70.9%), with scope for improvement on the essential physical exams selected (3.3/6, 55%). Second, we also found no differences between medical and PA students on our measures reflecting their flexibility in thinking about diagnoses. Third, we found that over the course of the cases, both PAs and medical students improved in the number of relevant diagnoses that they considered for the patients. These findings provided suggestive evidence that PAs and medical students used eCREST in similar ways to apply and develop diagnostic reasoning skills.

Limitations

There are several methodological limitations to consider when interpreting the present findings. The use of just 2 simulated cases where differential diagnoses largely concerned respiratory and cardiovascular causes significantly limits the extent to which it is possible to generalize from these findings to cases seeking to develop diagnostic reasoning relevant to other bodily systems. The low sample size limited our ability to make strong statistical inferences or to adjust for possible confounders. Additionally, the medical students in our study self-selected to participate, and completion rates among those who registered were generally low. As we did not collect overall class cohort numbers, we were also unable to calculate response rates for all eligible students. Thus, there is a clear need for additional studies that can achieve greater recruitment and retention of student participants, particularly for the PA students, to yield larger and more representative samples. Notably, there was disproportionate representation of participants from School D, which may have introduced institutional bias and limited the generalizability of the findings across other PA programs.

The education context in which our two student groups also differed. As the medical students in our study had access to more cases (n=4) than their physician assistant counterparts (n=2), it is possible that they may have benefitted from being more familiar with the eCREST platform or from having practice with other cases. The context differed also in that medical students were completing cases before the COVID-19 lockdowns where online teaching was relatively unusual. In contrast, PAs were completing the cases in 2021, when online educational delivery was much more normalized. The difference in context, however, does not appear to have affected students' responses to eCREST.

Comparison With Prior Work

The findings of our study are broadly aligned with a review of performance of qualified PAs compared with physicians, which found performance on specific diagnostic tasks related to cancer was similar across both professional groups [23]. This review, however, is limited in its applicability to this study for two reasons. First, it was restricted to practicing clinicians, rather than those in training. Second, most of the literature included came from outside of the United Kingdom, in countries where primary care systems are not the same as the United Kingdom. As Barnhill et al found with respect to PAs in the United States, the role of PAs in primary care is different from the role of PAs in other health care settings [24].

We reflect also on the qualitative reflections of medical students and PAs in this study. We note this was a small exploratory study, so we are cautious about drawing firm conclusions regarding the underlying reasons for these differences but instead have reflected on key reflections from the groups. It is perhaps not surprising that a high proportion of PAs recognized in themselves a knowledge deficit (87% of PAs vs 70% of medical students; *P*=.03). At the point at which they commenced eCREST, they had experienced approximately 1.5 years of clinical training. In contrast, medical students would have had approximately 5 years of exposure to medical curricula. However, we interpret it as encouraging that a high proportion of both groups expressed a motivation to address knowledge gaps given the consistent evidence that knowledge deficits are a cause of diagnostic error [25].

It was interesting that nearly half of the medical students reported a need to develop open-mindedness and flexibility versus a quarter of PA students (48% vs 26%; *P*=.01). In light of evidence suggesting that confidence is a poor predictor of diagnostic accuracy [26], and that high levels of confidence lead to less information gathering, fewer changes in diagnosis, and biased evaluations [27], these reflections suggesting lower confidence and an increased awareness of the need to develop their reasoning could lead them to be better diagnosticians. The students' reflections accord with the educational experiences of tutors using eCREST with a range of student groups who have observed that eCREST exposed students to clinical uncertainty and welcomed the stimulus provided by eCREST to discuss diagnostic complexity and uncertainty (unpublished data). We therefore consider it as encouraging for students' future diagnostic reasoning that they identified the need for more open-mindedness as one of their key reflections.

Conclusions

Developing students' clinical reasoning skills is critical for improving patient care and reducing diagnostic error. These exploratory results provide suggestive and positive evidence that medical and PA students had similar clinical reasoning styles when using an online patient simulation training tool to support their diagnostic decision-making, and both groups reported some changes in their reasoning styles through using eCREST. As primary care teams widen to include a range of clinical professionals, further evidence is now needed to understand and compare how different clinical groups at all

stages of training develop and apply essential clinical reasoning skills.

Acknowledgments

This research was part of the programme of the National Institute for Health and Care Research [NIHR] Policy Research Unit in Cancer Awareness, Screening and Early Diagnosis, 106/0001. The Policy Research Unit in Cancer Awareness, Screening, and Early Diagnosis receives funding for a research programme from the Department of Health Policy Research Programme. This report is independent research supported by the NIHR ARC North Thames. JS is funded by an NIHR Population Health Career Scientist Fellowship (303616). The views expressed in this publication are those of the author(s) and not necessarily those of the National Institute for Health and Care Research or the Department of Health and Social Care.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data for the second case.
[DOCX File, 25 KB - mededu_v11i1e68981_app1.docx]

References

1. Avery AJ, Sheehan C, Bell B, et al. Incidence, nature and causes of avoidable significant harm in primary care in England: retrospective case note review. BMJ Qual Saf 2021 Dec;30(12):961-976. [doi: 10.1136/bmjqs-2020-011405]
2. Singh H, Schiff GD, Graber ML, Onakpoya I, Thompson MJ. The global burden of diagnostic errors in primary care. BMJ Qual Saf 2017 Jun;26(6):484-494. [doi: 10.1136/bmjqs-2016-005401]
3. Norman G. Research in clinical reasoning: past history and current trends. Med Educ 2005 Apr;39(4):418-427. [doi: 10.1111/j.1365-2929.2005.02127.x] [Medline: 15813765]
4. Kostopoulou O, Delaney BC, Munro CW. Diagnostic difficulty and error in primary care--a systematic review. Fam Pract 2008 Dec 1;25(6):400-413. [doi: 10.1093/fampra/cmn071]
5. Kostopoulou O, Sirota M, Round T, Samaranayaka S, Delaney BC. The role of physicians' first impressions in the diagnosis of possible cancers without alarm symptoms. Med Decis Making 2017 Jan;37(1):9-16. [doi: 10.1177/0272989X16644563] [Medline: 27112933]
6. Sheringham J, Sequeira R, Myles J, et al. Variations in GPs' decisions to investigate suspected lung cancer: a factorial experiment using multimedia vignettes. BMJ Qual Saf 2017 Jun;26(6):449-459. [doi: 10.1136/bmjqs-2016-005679] [Medline: 27651515]
7. Young ME, Thomas A, Lubarsky S, et al. Mapping clinical reasoning literature across the health professions: a scoping review. BMC Med Educ 2020 Apr 7;20(1):107. [doi: 10.1186/s12909-020-02012-9] [Medline: 32264895]
8. Cooper N, Bartlett M, Gay S, et al. Consensus statement on the content of clinical reasoning curricula in undergraduate medical education. Med Teach 2021 Feb;43(2):152-159. [doi: 10.1080/0142159X.2020.1842343] [Medline: 33205693]
9. Kassianos A, Plackett R, Schartau P, et al. eCREST: a novel online patient simulation resource to aid better diagnosis through developing clinical reasoning. BMJ Simul Technol Enhanc Learn 2020 Jul;6(4):241-242. [doi: 10.1136/bmjstel-2019-000478] [Medline: 32832102]
10. Plackett R, Kassianos AP, Kambouri M, et al. Online patient simulation training to improve clinical reasoning: a feasibility randomised controlled trial. BMC Med Educ 2020 Jul 31;20(1):245. [doi: 10.1186/s12909-020-02168-4] [Medline: 32736583]
11. Plackett R, Kassianos AP, Timmis J, Sheringham J, Schartau P, Kambouri M. Using virtual patients to explore the clinical reasoning skills of medical students: mixed methods study. J Med Internet Res 2021 Jun 4;23(6):e24723. [doi: 10.2196/24723] [Medline: 34085940]
12. NHS England. We are the NHS: people plan for 2020/21 – action for us all. 2020 Jul 30. URL: https://www.england.nhs.uk/publication/we-are-the-nhs-people-plan-for-2020-21-action-for-us-all/ [accessed 2024-11-10]
13. NHS long term workforce plan. NHS England. URL: https://www.england.nhs.uk/long-read/nhs-long-term-workforce-plan-2/ [accessed 2024-11-10]
14. What is a PA? AAPA. URL: https://www.aapa.org/about/what-is-a-pa/ [accessed 2023-09-14]
15. Physician assistants. NHS Employers. URL: https://www.nhsemployers.org/articles/physician-assistants [accessed 2025-09-18]
16. Young M, Thomas A, Gordon D, et al. The terminology of clinical reasoning in health professions education: implications and considerations. Med Teach 2019 Nov;41(11):1277-1284. [doi: 10.1080/0142159X.2019.1635686] [Medline: 31314612]
17. Young ME, Dory V, Lubarsky S, Thomas A. How different theories of clinical reasoning influence teaching and assessment. Acad Med 2018 Sep;93(9):1415. [doi: 10.1097/ACM.0000000000002303] [Medline: 29847325]

18. Plackett R, Kassianos AP, Mylan S, Kambouri M, Raine R, Sheringham J. The effectiveness of using virtual patient educational tools to improve medical students' clinical reasoning skills: a systematic review. BMC Med Educ 2022 May 13;22(1):365. [doi: 10.1186/s12909-022-03410-x] [Medline: 35550085]
19. eCREST - electronic Clinical Reasoning Educational Simulation Tool. 2025. URL: https://ecrest.uk/ [accessed 2025-09-18]
20. Introduction to our guidance on undergraduate clinical placements. General Medical Council. URL: https://www.gmc-uk.org/education/standards-guidance-and-curricula/guidance/undergraduate-clinical-placements/guidance-on-undergraduate-clinical-placements/introduction [accessed 2025-09-18]
21. RStudio: integrated development environment for R. 2021. URL: http://www.rstudio.com [accessed 2025-11-06]
22. Althouse AD. Adjust for multiple comparisons? it's not that simple. Ann Thorac Surg 2016 May;101(5):1644-1645. [doi: 10.1016/j.athoracsur.2015.11.024] [Medline: 27106412]
23. Sheringham J, King A, Plackett R, Khan A, Cornes M, Kassianos AP. Physician associate/assistant contributions to cancer diagnosis in primary care: a rapid systematic review. BMC Health Serv Res 2021 Jul 3;21(1):644. [doi: 10.1186/s12913-021-06667-y] [Medline: 34217265]
24. Barnhill GC, Dallas AD, Mauldin SG, Hooker RS. PA practice analysis: multidisciplinary tasks, knowledge, and skills. JAAPA 2018 Dec;31(12):34-40. [doi: 10.1097/01.JAA.0000547750.31052.5a] [Medline: 30399009]
25. Norman GR, Monteiro SD, Sherbino J, Ilgen JS, Schmidt HG, Mamede S. The causes of errors in clinical reasoning: cognitive biases, knowledge deficits, and dual process thinking. Acad Med 2017 Jan;92(1):23-30. [doi: 10.1097/ACM.0000000000001421] [Medline: 27782919]
26. Meyer AND, Payne VL, Meeks DW, Rao R, Singh H. Physicians' diagnostic accuracy, confidence, and resource requests: a vignette study. JAMA Intern Med 2013 Nov 25;173(21):1952-1958. [doi: 10.1001/jamainternmed.2013.10081] [Medline: 23979070]
27. Kourtidis P, Nurek M, Delaney B, Kostopoulou O. Influences of early diagnostic suggestions on clinical reasoning. Cogn Research 2022;7(1):103. [doi: 10.1186/s41235-022-00453-y]

Abbreviations

eCREST: electronic Clinical Reasoning Educational Simulation Tool
HEI: higher educational institution
PA: physician assistant

Edited by B Lesselroth; submitted 19.11.24; peer-reviewed by S Ito, S Kotwal; revised version received 18.09.25; accepted 07.10.25; published 01.12.25.

Please cite as:
Thorpe A, Kassianos AP, Plackett R, Krishnamurthy V, Kambouri MA, Sheringham J
Comparison of Physician Assistant and Medical Students' Clinical Reasoning Processes Using an Online Patient Simulation Tool to Support Clinical Reasoning (eCREST): Mixed Methods Study
JMIR Med Educ 2025;11:e68981
URL: https://mededu.jmir.org/2025/1/e68981
doi:10.2196/68981

A Brief Web-Based Person-Centered Care Group Training Program for the Management of Generalized Anxiety Disorder: Feasibility Randomized Controlled Trial in Spain

Vanesa Ramos-García[1,2,3,4], BSc; Amado Rivero-Santana[1,3,4], PhD; Wenceslao Peñate-Castro[2], PhD; Yolanda Álvarez-Pérez[1,3,4], PhD; Andrea Duarte-Díaz[1,3,4], BSc; Alezandra Torres-Castaño[1,3,4], PhD; María del Mar Trujillo-Martín[1,3,4], PhD; Ana Isabel González-González[5,6], PhD; Pedro Serrano-Aguilar[3,4,7], PhD; Lilisbeth Perestelo-Pérez[3,4,7], PhD

[1]Canary Islands Health Research Institute Foundation, Santa Cruz de Tenerife, Spain

[2]Department of Clinical Psychology, Psychobiology and Methodology, University of La Laguna (ULL), Santa Cruz de Tenerife, Spain

[3]Network for Research on Chronicity, Primary Care, and Health Promotion (RICAPPS), Tenerife, Spain

[4]The Spanish Network of Agencies for Health Technology Assessment and Services of the National Health System (RedETS), Tenerife, Spain

[5]Network for Research on Chronicity, Primary Care, and Health Promotion (RICAPPS), Madrid, Spain

[6]Área de Fomento de la Innovación e Internacionalización de la Investigación Sanitaria, Subdirección General de Investigación Sanitaria y Documentación, Dirección General Investigación y Docencia, Consejería de Sanidad, Madrid, Spain

[7]Evaluation Unit (SESCS), Canary Islands Health Service (SCS), Santa Cruz de Tenerife, Spain

Corresponding Author:
Lilisbeth Perestelo-Pérez, PhD
Network for Research on Chronicity, Primary Care, and Health Promotion (RICAPPS), Tenerife, Spain

Abstract

Background: Shared decision-making (SDM) is a crucial aspect of patient-centered care. While several SDM training programs for health care professionals have been developed, evaluation of their effectiveness is scarce, especially in mental health disorders such as generalized anxiety disorder.

Objective: This study aims to assess the feasibility and impact of a brief training program on the attitudes toward SDM among primary care professionals who attend to patients with generalized anxiety disorder.

Methods: A feasibility randomized controlled trial was conducted. Health care professionals recruited in primary care centers were randomized to an intervention group (training program) or a control group (waiting list). The intervention consisted of 2 web-based sessions applied by 2 psychologists (VR and YA), based on the integrated elements of the patient-centered care model and including group dynamics and video viewing. The outcome variable was the Leeds Attitudes Towards Concordance scale, second version (LATCon II), assessed at baseline and after the second session (3 months). After the randomized controlled trial phase, the control group also received the intervention and was assessed again.

Results: Among 28 randomized participants, 5 withdrew before the baseline assessment. The intervention significantly increased their scores compared with the control group in the total scale (b=0.57; P=.018) and 2 subscales: communication or empathy (b=0.74; P=.036) and shared control (ie, patient participation in decisions: b=0.68; P=.040). The control group also showed significant pre-post changes after receiving the intervention.

Conclusions: For a future effectiveness trial, it is necessary to improve the recruitment and retention strategies. The program produced a significant improvement in participants' attitude toward the SDM model, but due to this study's limitations, mainly the small sample size, more research is warranted.

(JMIR Med Educ 2025;11:e50060) doi:10.2196/50060

KEYWORDS

person-centered care; primary care; shared decision-making; anxiety disorder; training program; SDM

Introduction

About 264 million people in the world are affected by anxiety disorders, according to the latest estimates of the World Health Organization [1]. In Spain, around 2 million people (4.1% of the population) suffer from anxiety disorders [1]. In primary care (PC) settings, the generalized anxiety disorder (GAD) is one of the most prevalent anxiety disorders [2]. GAD is characterized by a continuous state of worry and alertness most

of the time [3] and sometimes, its high comorbidity with other psychiatric and somatic disorders makes diagnosis difficult [4]. GAD has a tendency to chronicity, due to its specific characteristics, leading to the person being worried and alert most of the time [3]. Information on the causes of the disorder and the available treatments is an unmet need in this population, given that some patients with GAD are willing to have an active or collaborative role in their health care [5].

Person-centered care (PCC) is considered the gold standard for medical care in health care settings because it humanizes the person and places him or her at the center of clinical decision-making [6]. The PCC model consists of several components, one of which is shared decision-making (SDM), whose goal is to create a collaborative dialogue between patients and health care professionals, in which patients' values, preferences, and concerns about the different available treatment options are taken into account and incorporated into the decision-making process [7-9].

Patient decision aids are tools designed to facilitate SDM. Its use can help patients participate in the clinical decisions, improving the decision-making process and promoting informed decisions that are concordant with patients' values and preferences [10]. On the part of professionals, it is important to develop communication skills and empathy to help patients participate in the decisions [11-13]. Research has shown that interventions and training programs aimed to promote the PCC model may improve professionals' knowledge and the ability to communicate with patients [12,14] as well as patients' satisfaction [15]. However, there are some barriers to apply the PCC model related to time constraints, clinical uncertainty, poor expectations, patients' characteristics (eg, age, comorbidity, and attitude), lack of continuity of care, or knowledge about SDM [16-19]. Despite some SDM training programs have been developed for health care professionals, very few of them have been evaluated [20-22]. Therefore, despite the growing acceptance of interventions to implement SDM in health care settings, several gaps remain in the demand, perception, and clinical application of the PCC model [23,24]. In mental health care, and specifically in GAD, interventions to promote the SDM process are still very limited [25,26]. A recent qualitative study with patients with GAD concluded that there is scarce orientation to elicit patients' preferences and values throughout the process of care [27], emphasizing the need of interventional studies aimed at promoting SDM in the clinical encounter.

The aim of this study is to evaluate the feasibility and effect of a brief training program on the attitudes toward SDM for professionals in PC who attend patients with GAD.

Methods

Design

A feasibility randomized controlled trial (RCT) was conducted, in which participants were allocated to a PCC training program or a control group (waiting list). It was carried out in 13 PC centers in Tenerife (Canary Islands, Spain), from January 2021 to February 2022.

Ethical Considerations

The study was approved by the ethics committee of the Hospital Universitario Nuestra Señora de La Candelaria (reference: CHUNSC_2019_58). The study was not registered because participants were health professionals and not patients, the intervention was educational, and the only outcome measured was attitudinal. Participants who agreed to participate signed a web-based informed consent form.

Participants

Participants were health care professionals working in PC centers (ie, physicians and nurses) or community mental health units (ie, psychiatrists, psychologists, and nurses) for at least 1 year before the start of the study, who attend patients with GAD in the Canary Islands, Spain. There were no exclusion criteria.

Procedure, Randomization, and Allocation Concealment

The directors of the health centers were contacted and informed about the study. They were asked to invite the professionals from their centers to participate. The invitation included an infographic, graphically describing the study and a link to a web platform, where health professionals could register their willingness to participate and contact information. Then, they were contacted by telephone to provide a full explanation of the study. Those who agreed to participate signed a web-based informed consent form (reference: CHUNSC_2019_58). Participants were randomly assigned to either the intervention or control group (waiting list), using a computer-generated random number table. The randomization process was conducted by an independent researcher who was not involved in the recruitment or assignment of participants. In addition, the researcher who recruited the professionals was blinded to the group assignments in order to maintain allocation concealment. Due to the nature of the intervention, the study participants could not be blinded.

Intervention

Intervention group participants received 2 training sessions via Zoom (version 5.15.7. [21404]) based on the integrated elements of the PCC model [28]. The training was originally intended to be applied in person, in a group format, but this was not possible due to the COVID-19 pandemic, so it was finally applied on the web. Sessions were conducted by 2 researchers (VR and YA [psychologists]). The first session lasted approximately 2 hours and was focused on presenting the principal elements of intervention: (1) introduction, which included a description of common clinical relationship models (first 20 minutes); (2) basic characteristics of the basic PCC model, through group dynamics and video viewing of a role-play in the clinical practice with a patient with GAD; this included a description of the Feelings, Ideas, Function, and Expectations model [29] (60 minutes), which was developed at the University of Western Ontario and explores the patient's emotions, his or her ideas on what caused the problem, the effects of the illness on his or her functioning and relationships, and his or her expectations for the future and from medical care [29,30]; and (3) presentation of the Three-Talk Model for SDM, a multistage consultation process developed by Elwyn et al [31] (30 minutes). The

Three-Talk Model for SMD is a theoretical approach that describes collaborative deliberation. It outlines 3 broad steps that form the core elements of SDM [31]. The last 10 minutes of the session were aimed at the resolution of doubts. The detailed contents of this first SDM training session are shown in Table 1.

Table . Content of first shared decision-making (SDM) training session.

Module and content	Form of communication	Learning objectives
Introduction		
Clinical relationship models	• Lecture • Video examples • Interactive live • Feedback with group dynamic	• Be able to know the characteristics of the paternalistic, informative or contractual, interpretive or personalized, and deliberative or friendly models
Characteristics of a basic PCC[a] model		
Explore the disease	• Lecture	• Acquire skills in active listening and directed anamnesis in the use of SDM
Know the patient's perspective (beliefs, fears, expectations, repercussions, etc)	• Lecture • Video examples • Interactive live • Feedback with group dynamic	• Acquire skills in how to prepare the ground and how to explore the personal experience of the disease in terms of SDM • Be able to use the FIFE[b] model to improve the quality of communication in terms of SDM
Know the person ("moving from patient to person")	• Lecture • Video examples	• Acquire skill about how exploring the personal and social context of the disease in terms of SDM
Involve the patients in their disease	• Lecture • Video examples • Interactive live • Feedback with group dynamic	• Acquire information skills to reach agreements on problem solving, to seek shared solutions, and to involve the patient in the use of SDM
Three-Talk Model for SDM		
Team dialogue	• Lecture	• Acquire skills to establish a team dialogue based on the needs for change on beliefs and preferences
Dialogue on options	• Lecture	• Acquire skills to discuss the treatment options that exist for the disease
Dialogue on the decision	• Lecture	• Acquire skills to help the patient decide on which option to choose

[a]PCC: person-centered care.

[b]FIFE: Feelings, Ideas, Function and Expectations.

The second session was carried out 3 months later (review session), with an approximate duration of 1 hour. The structure of the session included (1) the review of the main contents of the first training module, together with comments on participants' potential and sharing their experiences applying the SDM model since then (30 minutes), and (2) the discussion on the main barriers and facilitators for patients and professionals in applying the SDM process in the clinical practice (30 minutes). Detailed content of this session is present in Table 2.

Control group participants did not receive any intervention. They were informed that they could access the training program after the feasibility RCT was completed. Participants completed the baseline and 3-month (postintervention) assessments. Subsequently, participants in the control group received the intervention and were reevaluated 3 months later (second postintervention measure).

Table . Content of second shared decision-making (SDM) training session.

Unit and content	Form of communication	Learning objectives
(1) Introduction and (2) characteristics of a basic PCC[a] model		• Review the characteristics of the paternalistic model,: informative or contractual, interpretive or personalized, and the deliberative or friendly models • Review tasks in active listening and directed anamnesis in the use of SDM: how to prepare the ground and how to explore the personal experience of the disease; how to explore the personal and social context of the disease; and how to reach agreements on problem solving, to seek shared solutions, to involve the patient-shared solutions, and to involve the patient in the use of SDM
Clinical relationship models Explore the disease; know the patient's perspective (beliefs, fears, expectations, repercussions, etc); know the person ("moving from patient to person"); and involve the patients in their disease	Lecture	
(3) Characteristics of the Three-Talk Model for SDM		• Be able to apply the principal components of the Three-Talk Model for SDM • Have knowledge about how to apply this model in clinical practice to support SDM
Fifteen characteristics total of a Three-Talk Model for SDM are described: *First step*: Take a step back, present the possibility of choice, justify the choice, personalizing preference, uncertainty, check the reaction, and postpone closure *Second step*: Check knowledge, list of options, provide decision support to the patient, and summaries *Third step*: Focus on preferences, elicit a preference, lead toward a decision, and offer review	Lecture	
(4) Barriers and enablers to apply Three-Talk Model for SDM		• Invite to participate by presenting the experience from a professional point of view in clinical practice • Openly share and discuss observations of the professional communication • Offer, explicitly and without judging, feedback on implementation
Identification of barriers from a professional point of view that can condition the application of the 3-step model for SDM	Identification of professionals' own barriers to communication with their patients	
Identification of barriers from a patient's point of view that can condition the application of the 3-step model for SDM	Identification of patients' own barriers to communication with their care team	
Identification of facilitators who may exist to carry out the 3-step model for SDM	Identify the individual facilitators in communication to implement a SDM model	

[a]PCC: person-centered care.

Measures

The outcome measure was the professionals' attitude toward PCC. It was assessed with the Leeds Attitudes Towards Concordance scale, second version (LATCon II) [32]. This self-report instrument includes 20 items with a 4-point Likert format from strongly disagree (0) to strongly agree (3). Although the original instrument includes 5 subscales, we used the 3 components identified by means of principal component analysis in the Spanish validation [33], carried out with psychiatrists and psychiatry residents. These subscales were labeled "communication/empathy" (CE, 12 items about the importance of a good communication and the consideration of patient's feelings and beliefs), "shared control" (SC, 4 items reflecting a positive attitude toward equality and SDM), and "eventual paternalistic style" (EPS, 4 items stating that sometimes a paternalistic style is necessary; these items are reverse-coded, and therefore higher scores indicate lower agreement with EPS) [33]. Scores on the total scale and the subscales are divided by the corresponding number of items, thus ranging 0 - 3. The LATCon II has shown good internal consistency in previous studies [33-35].

The following sociodemographic and professional variables were measured at baseline: age, gender, specialty (medicine or nursing), years of professional experience and work in the health care center, level of perceived workload (low, medium, and high), and previous training on PCC or SDM.

Statistical Analysis

We calculated that a mixed model with 2 repeated measures per participant (cluster) requires 38 subjects (19 in each group) in order to detect a significant moderate-to-strong between-group effect (standardized mean difference of 0.80), assuming type I and II errors of 0.05 and 0.20, respectively, and an intraclass correlation of 0.50 [35].

Descriptive statistics were calculated for continuous and categorical variables (means, SDs, and percentages). Cronbach α was calculated for the LATConII scale and its 3 subscales, as well as the correlations between the subscales (Spearman ρ). The effect of the intervention was analyzed with mixed lineal models, including fixed effects for time (pre, post), group (intervention, control) and its interaction, and the participant as a random effect (assuming an unstructured covariance matrix). Successive models were carried out adjusting for 1 covariate at a time (ie, sociodemographic and professional variables). Unstandardized β values and effect sizes (Hedges *g*) are reported.

Changes from baseline to postintervention were evaluated analyzing the effect of time in a mixed model separately for each group. The same test was used to analyze the change in the control group after receiving the intervention upon completion of the RCT. Analyses were performed with SPSS (version 25; IBM Corp) and STATA (version 17; StataCorp LLC).

Results

Thirty-four health professionals were interested in participating and were contacted by phone. After being informed in detail, 6 declined participation and 28 accepted, signing informed consent and being randomly allocated to the intervention or control group (14 each). However, 5 of them (4 in the intervention group) withdrew from the study before completing the baseline assessment (Figure 1). Table 3 shows the characteristics of the 23 participants. There were 18 women (18/23, 78.3%) and the mean age was 48.3 (range: 26 - 64) years. They had an average of 22.3 years of professional experience, and 52% (12/23) considered having a high caseload. Only 5 (21.7%) had had previous training in PCC.

Figure 1. CONSORT flow diagram. *The analysis includes 3 control participants lost at postintervention.

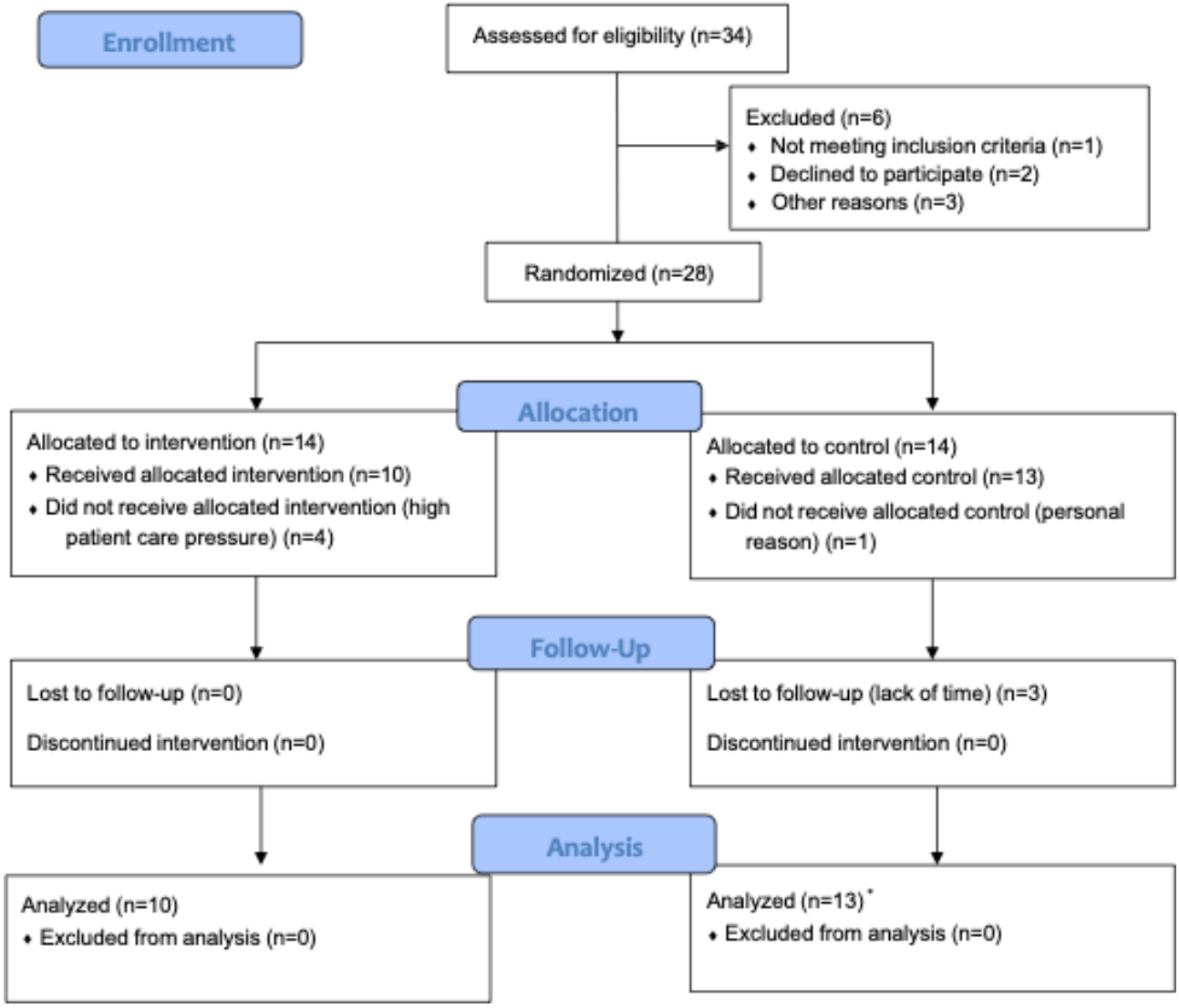


Table . Characteristics of participants.

	Intervention (n=10)	Control (n=13)	Total (N=23)
Female, n (%)	10 (100)	8 (61.54)	18 (78.26)
Age, mean (SD)	45.60 (10.82)	50.38 (9.77)	48.30 (10.24)
Specialty, n (%)			
Nursing	3 (30)	2 (15.38)	5 (21.74)
Medicine	7 (70)	11 (84.62)	18 (78.26)
Years of professional experience, mean (SD)	19.70 (9.44)	24.33 (9.42)	22.32 (9.51)
Years working in the center, mean (SD)	7.05 (7.15)	6.40 (7.02)	6.68 (6.92)
Previous training in PCC[a], n (%)	1 (10)	4 (30.77)	5 (21.74)
Self-perceived care load, n (%)			
Low-medium	6 (60)	5 (38.46)	11 (47.83)
High	4 (40)	8 (61.54)	12 (52.17)

[a]PCC: person-centered care.

At baseline, internal consistency (Cronbach α) was 0.94 for the total LATConII scale, and 0.97 (CE), 0.88 (SC), and 0.25 (EPS) for the subscales. The total mean score was 2.08 (SD 0.60), and the mean scores were 2.29 (SD 0.78), 1.77 (SD 0.85), and 1.78 (SD 0.41) for the subscales CE, SC, and EPS, respectively (Table 3). CE and SC were significantly correlated (ρ=0.49; *P*=.01), whereas EPS was not significantly associated with CE (ρ=0.11; *P*=.62) or SC (ρ=0.29; *P*=.180) (Table 4).

Table . Effect of the intervention.

Time[a]	Intervention (n=10), mean (SD)	Control (n=10), mean (SD)	Time × group interaction, *b* (*P*)[b]	Between-group effect size, Hedges *g* (95% CI)
LATCon II[c] total (range: 0 - 3)			0.57 (.018)	0.92 (0.13 to 1.71)
Pre	1.87 (0.76)	2.25 (0.40)[d]		
Post	2.27 (0.51)[e]	2.08 (0.61)		
Post2	—[f]	2.60 (0.24)[g]		
Communication/empathy (range: 0 - 3)			0.74 (.036)	0.86 (0.06 to 1.65)
Pre	1.98 (1.02)	2.52 (0.46)[d]		
Post	2.57 (0.70)[e]	2.34 (0.86)		
Post2	—	2.84 (0.20)[e]		
Shared control (range: 0 - 3)			0.68 (.040)	0.76 (0.01 to 1.52)
Pre	1.55 (1.06)	1.94 (0.63)[d]		
Post	1.80 (0.44)	1.52 (0.70)[e]		
Post2	—	2.28 (0.43)[h]		
Eventual paternalistic style (range: 0 - 3)			−0.04 (.856)	0.08 (−0.93 to 0.93)
Pre	1.83 (0.44)	1.75 (0.41)[d]		
Post	1.83 (0.57)	1.83 (0.54)		
Post2	—	2.18 (0.64)[e]		

[a]Pre-post: randomized controlled trial (intervention vs waiting list); post2: intervention period for the control group, after the randomized controlled trial.

[b]Unstandardized β coefficients (*P* value) from mixed lineal models analyzing the randomized controlled trial (pre-post), including the participant as a random effect (the analysis includes 3 control participants lost at postintervention).

[c]LATCon II: Leeds Attitudes Towards Concordance scale, second version.

[d]n=13.

[e]*P*<.05.

[f]Not applicable.

[g]*P*<.001, compared with the previous assessment (effect of time in mixed models separately by group).

[h]*P*<.01.

Three control participants were lost at postintervention (3 months), but their baseline scores were included in the mixed models on an intention-to-treat basis (postintervention scores were not imputed). The time × group interaction was statistically significant for the total scale, showing a differential increment in scores favoring the intervention (*b*=0.57; *P*=.01) (Table 4). The same occurred with the subscales CE (*b*=0.74; *P*=.036) and SC (*b*=0.68; *P*=.04). The inclusion of potential confounders in the model did not change the results (see Table S1 in Multimedia Appendix 1 for the total scale). The intervention group significantly increased their scores compared with baseline in the total scale (*b*=0.4; *P*=.033) and CE (*b*=0.58; *P*=.030), whereas the control group significantly decreased in SC (*b*=−0.43; *P*=.037) (Table 4).

After the trial was completed, the control group received the intervention and showed significant increments in the total score (*b*=0.52; *P*<.001) and the 3 subscales: CE (*b*=0.50; *P*=.020), SC (*b*=0.75; *P*=.002), and EPS (*b*=0.35; *P*=.02) (Table 4).

Discussion

Principal Findings

This study aimed to evaluate the feasibility and effect of a brief web-based training program on the attitudes toward SDM and PCC of PC professionals who treat patients with GAD. The program was initially intended to be conducted in person at the professionals' centers, but due to the pandemic context, it was shifted to a web-based format. The 2 sessions went smoothly and the professionals actively participated, asking questions

and describing their experiences related to SDM. Previous studies evaluating learning programs for health professionals or university students have not shown relevant differences between web-based and in-person formats [36-39], although in some cases better results have been observed with the face-to-face intervention [40]. Given the brevity of our program, we do not expect that there will be relevant differences between both formats.

The recruitment and retention rate were low, only 33 eligible professionals showed interest in the study (2.5 per center) during the 5-month recruitment period, and 5 declined participation when they were fully informed about the study. It is possible that direct contact with professionals, instead of the general call that was made through center directors, would have improved the recruitment rate to some extent. Among the 28 randomized participants, 5 more did not start the trial and 3 did not complete the study. The high workload, a common situation in the Spanish public health system even in a nonpandemic context, was the main reported cause of these withdrawals. On the other hand, the group format enriches the training process by enabling the interaction of professionals, but it also represents a difficulty when coordinating their schedules and availability. In summary, the participation and retention rates were not satisfactory, and for future trials it is necessary to develop more structured and intensive strategies. Theoretical frameworks as proposed by Solberg [41] that identified 7 factors that influence the recruitment of health care professionals (ie, relationships, reputation, requirements, rewards, reciprocity, resolution, and respect) could help to this aim.

Regarding effectiveness, the results showed significant moderate-to-strong effects (although with very wide confidence intervals) on the total scale and the CE and SC subscales. The pre-post change in the intervention group was greater on the former, and the similar between-group effect size was due in part to a significant decrease in SC in the control group. The EPS dimension was not affected by the intervention, but this result is unclear given the low internal consistency of this subscale (future studies should confirm the factorial structure of the instrument). After the RCT was completed, the control group received the intervention and showed significant before-after improvements of similar magnitude in the 3 dimensions. Due to the wide confidence intervals, the results should be interpreted with caution and verified in studies with greater statistical power.

Baseline scores indicated a positive attitude (values above the midpoint of the scale) for the total scale and the 3 subscales, although scores on CE and SC suggest that, comparatively, participants seemed more favorable to empathetically communicate with their patients than sharing decisions with them. This result has also been observed in several studies that applied the Patient - Practitioner Orientation Scale [42], the most frequently used instrument to assess health professionals' attitudes toward PCC, showing higher scores on the *caring* subscale of the questionnaire (ie, empathy, warmth, and treating patients as whole persons) than on the *sharing* one (ie, sharing information, decisions, and power) [43-47].

Other studies also have shown significant benefits of different training programs on professionals' and medical students' attitudes toward SDM and PCC and their intention to apply it in the future, showing high levels of satisfaction with the program [48-52]. A positive attitude toward the PCC model is an obvious requisite for the professionals' learning and demonstration of behaviors aimed at promoting SDM in consultation. Validation studies with the Patient - Practitioner Orientation Scale showed that more favorable attitudes were significantly associated with more patient - centered behaviors in consultations [53], and that concordance of patients and physicians' attitudes was associated with greater patient's satisfaction [53-55], trust, and endorsement of physicians [53], as well as fewer referrals to specialized care [56]. Nonetheless, for the implementation of SDM it is necessary to have not only a positive attitude toward PCC but also the appropriate knowledge and communication skills required by this model, for which training programs have been developed. However, the effect of interventions targeting health professionals on the actual promotion of SDM in consultation remains uncertain. The last update of a Cochrane systematic review reported a significant effect of these interventions (eg, educational meetings and materials, outreach visits, and reminders), compared with usual care when SDM in consultation was assessed by external observers, but not by patients, even when the intervention is directed to both patients and professionals [11]. Observational studies have also shown a lack of association between patients' and external observers' perception of SDM [57-59], but the causes of this discrepancy have not been investigated. Furthermore, the evidence about the effects of SDM interventions targeting health professionals on patients' cognitive, affective, behavioral, and health outcomes is also scarce [10].

Although the PCC and SDM models are a paradigm to be applied to every patient regardless of his or health problems, patients with GAD could present specific psychological characteristics that might affect the decision-making process. In experimental settings involving stimulus reinforcement, these patients have shown greater intolerance to uncertainty and impaired decision-making [55,57-59]. Nonetheless, this does not translate into a preference for a passive role in decision-making, since a recent study showed that more than 80% research participants desired to play an active or collaborative role when making decisions about treatment, although one-third of them perceived more involvement than they preferred [60]. Therefore, professionals should adapt the SDM process to the patients' preference for involvement and manage the unavoidable uncertainty about the potential adverse effects of treatment and the likelihood and intensity of symptoms' improvement.

Limitations

The study has important limitations. First, feasibility of in-person group sessions could not be evaluated due to the emergence of the COVID-19 pandemic, but that allowed us to check the web-based application of the program, which was delivered without problems. However, the recruitment and retention rates were low. The recruited sample was small and there were some relevant differences in baseline variables,

including the scores on the LATCon, and therefore a high risk of selection bias is present. The intervention group was 5 years younger and less experienced, included more nurses and less participants with prior experience on SDM training, and showed a less favorable attitude toward SDM. These characteristics suggest a greater margin for potential benefit in this group. Although the inclusion of these covariates in the model did not change the results, this analysis is strongly underpowered. Nonetheless, given the strong effects sizes obtained and the similar ones showed by the control group after receiving the training, it is reasonable to think that the intervention could produce a real improvement in attitudes, although effects sizes are probably inflated due to the mentioned confounders. The small sample size and the fact that participants were voluntary also challenges the external validity of the results, since it is probable that they were more motivated or favorable to the SDM model.

On the other side, this was a pilot study and we did not assess other professionals' outcomes (eg, knowledge of SDM, satisfaction with the program, and intention to apply SDM in the future), whether the observed effect is maintained over time or its influence on professionals' behavior in consultation as well as on patients' outcomes, which is the ultimate aim of these interventions. An RCT with an adequate sample size is warranted to confirm the results on professionals' attitude and to investigate the mentioned issues.

Acknowledgments

The authors would like to thank all those who contributed to the realization of this study: health managers, health care professionals, and all the researchers who made this study possible.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Results of the models including covariates on the LATCon total score.
[PDF File, 62 KB - mededu_v11i1e50060_app1.pdf]

Checklist 1.
CONSORT-EHEALTH checklist (V 1.6.1).
[PDF File, 961 KB - mededu_v11i1e50060_app2.pdf]

References

1. Depression and Other Common Mental Disorders: Global Health Estimates: World Health Organization; 2017.
2. Bados López A. Trastorno de ansiedad generalizada: naturaleza, evaluación y tratamiento. Generalized anxiety disorder: nature, assessment and treatment [Article in Spanish]. : University of Barcelona; 2017 URL: https://diposit.ub.edu/dspace/handle/2445/115724 [accessed 2024-12-12]
3. Diagnostic and Statistical Manual of Mental Disorders, 5th edition: American Psychiatric Association; 2013.
4. Parmentier H, García-Campayo J, Prieto R. Comprehensive review of generalized anxiety disorder in primary care in Europe. Curr Med Res Opin 2013 Apr;29(4):355-367. [doi: 10.1185/03007995.2013.770731] [Medline: 23356728]
5. Ramos-García V, Rivero-Santana A, Duarte-Díaz A, et al. Shared decision-making and information needs among people with generalized anxiety disorder. Eur J Investig Health Psychol Educ 2021 May 21;11(2):423-435. [doi: 10.3390/ejihpe11020031] [Medline: 34708821]
6. Lidsaar Powell RC, Bu SMK. Paternering with and involving patients. In: Leslie R, Martin M, DiMatteo R, editors. The Oxford Handbook of Health Communication: Oxford University Press; 2013:84-108.
7. Stiggelbout AM, Van der Weijden T, De Wit MPT, et al. Shared decision making: really putting patients at the centre of healthcare. BMJ 2012 Jan 27;344:e256. [doi: 10.1136/bmj.e256] [Medline: 22286508]
8. Makoul G, Clayman ML. An integrative model of shared decision making in medical encounters. Patient Educ Couns 2006 Mar;60(3):301-312. [doi: 10.1016/j.pec.2005.06.010] [Medline: 16051459]
9. Tilburt JC, Wynia MK, Montori VM, et al. Shared decision-making as a cost-containment strategy: US physician reactions from a cross-sectional survey. BMJ Open 2014 Jan 14;4(1):e004027. [doi: 10.1136/bmjopen-2013-004027] [Medline: 24430879]
10. Goldwag J, Marsicovetere P, Scalia P, et al. The impact of decision aids in patients with colorectal cancer: a systematic review. BMJ Open 2019 Sep 12;9(9):e028379. [doi: 10.1136/bmjopen-2018-028379] [Medline: 31515416]
11. Légaré F, Adekpedjou R, Stacey D, et al. Interventions for increasing the use of shared decision making by healthcare professionals. Cochrane Database Syst Rev 2018 Jul 19;7(7):CD006732. [doi: 10.1002/14651858.CD006732.pub4] [Medline: 30025154]

12. Geiger F, Hacke C, Potthoff J, et al. The effect of a scalable online training module for shared decision making based on flawed video examples—a randomized controlled trial. Patient Educ Couns 2021 Jul;104(7):1568-1574. [doi: 10.1016/j.pec.2020.11.033] [Medline: 33334633]
13. Tan ASL, Mazor KM, McDonald D, et al. Designing shared decision-making interventions for dissemination and sustainment: can implementation science help translate shared decision making into routine practice? MDM Policy Practice 2018 Jul;3(2):2381468318808503. [doi: 10.1177/2381468318808503]
14. Mahmood S, Hazes JMW, Veldt P, et al. The development and evaluation of personalized training in shared decision-making skills for rheumatologists. J Rheumatol 2020 Feb;47(2):290-297. [doi: 10.3899/jrheum.180780] [Medline: 30936289]
15. Ritter S, Stirnemann J, Breckwoldt J, et al. Shared decision-making training in internal medicine: a multisite intervention study. J Grad Med Educ 2019 Aug;11(4 Suppl):146-151. [doi: 10.4300/JGME-D-18-00849] [Medline: 31428272]
16. Légaré F, Witteman HO. Shared decision making: examining key elements and barriers to adoption into routine clinical practice. Health Aff (Millwood) 2013 Feb;32(2):276-284. [doi: 10.1377/hlthaff.2012.1078]
17. Hernández-Leal MJ, Pérez-Lacasta MJ, Feijoo-Cid M, Ramos-García V, Carles-Lavila M. Healthcare professionals' behaviour regarding the implementation of shared decision-making in screening programmes: a systematic review. Patient Educ Couns 2021 Aug;104(8):1933-1944. [doi: 10.1016/j.pec.2021.01.032] [Medline: 33581968]
18. Frerichs W, Hahlweg P, Müller E, Adis C, Scholl I. Shared decision-making in oncology—a qualitative analysis of healthcare providers' views on current practice. PLoS One 2016;11(3):e0149789. [doi: 10.1371/journal.pone.0149789] [Medline: 26967325]
19. Coates D, Clerke T. Training interventions to equip health care professionals with shared decision-making skills: a systematic scoping review. J Contin Educ Health Prof 2020;40(2):100-119. [doi: 10.1097/CEH.0000000000000289] [Medline: 32433322]
20. Diouf NT, Menear M, Robitaille H, Painchaud Guérard G, Légaré F. Training health professionals in shared decision making: update of an international environmental scan. Patient Educ Couns 2016 Nov;99(11):1753-1758. [doi: 10.1016/j.pec.2016.06.008] [Medline: 27353259]
21. Légaré F, Politi MC, Drolet R, et al. Training health professionals in shared decision-making: an international environmental scan. Patient Educ Couns 2012 Aug;88(2):159-169. [doi: 10.1016/j.pec.2012.01.002] [Medline: 22305195]
22. Kienlin S, Nytrøen K, Stacey D, Kasper J. Ready for shared decision making: pretesting a training module for health professionals on sharing decisions with their patients. J Eval Clin Pract 2020 Apr;26(2):610-621. [doi: 10.1111/jep.13380] [Medline: 32114700]
23. Pollard S, Bansback N, Bryan S. Physician attitudes toward shared decision making: a systematic review. Patient Educ Couns 2015 Sep;98(9):1046-1057. [doi: 10.1016/j.pec.2015.05.004] [Medline: 26138158]
24. Légaré F, Stacey D, Brière N, et al. Healthcare providers' intentions to engage in an interprofessional approach to shared decision-making in home care programs: a mixed methods study. J Interprof Care 2013 May;27(3):214-222. [doi: 10.3109/13561820.2013.763777] [Medline: 23394265]
25. Fisher A, Sharpe L, Anderson J, Manicavasagar V, Juraskova I. Development and pilot of a decision-aid for patients with bipolar II disorder and their families making decisions about treatment options to prevent relapse. PLoS One 2018;13(7):e0200490. [doi: 10.1371/journal.pone.0200490] [Medline: 29990368]
26. Tlach L, Wüsten C, Daubmann A, Liebherz S, Härter M, Dirmaier J. Information and decision-making needs among people with mental disorders: a systematic review of the literature. Health Expect 2015 Dec;18(6):1856-1872. [doi: 10.1111/hex.12251] [Medline: 25145796]
27. Hurtado MM, Villena A, Vega A, Amor G, Gómez C, Morales-Asencio JM. "I have anxiety, but I have values and preferences" experiences of users with generalized anxiety disorder: a qualitative study. Int J Ment Health Nurs 2020 Jun;29(3):521-530. [doi: 10.1111/inm.12690] [Medline: 31908140]
28. Scholl I, Zill JM, Härter M, Dirmaier J. An integrative model of patient-centeredness—a systematic review and concept analysis. PLoS One 2014;9(9):e107828. [doi: 10.1371/journal.pone.0107828] [Medline: 25229640]
29. Stewart M, Brown JB, Weston WW, McWhinney IR, McWilliam CL, Freeman TR. Patient-Centered Medicine: Transforming the Clinical Method: Sage Publisher; 1995.
30. Weston WW, Brown JB, Stewart MA. Patient-centred interviewing part I: understanding patients' experiences. Can Fam Physician 1989 Jan;35:147-151. [Medline: 21253278]
31. Elwyn G, Durand MA, Song J, et al. A three-talk model for shared decision making: multistage consultation process. BMJ 2017 Nov 6;359:j4891. [doi: 10.1136/bmj.j4891] [Medline: 29109079]
32. Knapp P, Raynor DK, Thistlethwaite JE, Jones MB. A questionnaire to measure health practitioners' attitudes to partnership in medicine taking: LATCon II. Health Expect 2009 Jun;12(2):175-186. [doi: 10.1111/j.1369-7625.2009.00545.x] [Medline: 19538648]
33. de Las Cuevas C, Rivero-Santana A, Perestelo-Perez L, et al. Mental health professionals' attitudes to partnership in medicine taking: a validation study of the Leeds Attitude to Concordance Scale II. Pharmacoepidemiol Drug Saf 2012 Feb;21(2):123-129. [doi: 10.1002/pds.2240] [Medline: 21956875]

34. He W, Bonner A, Anderson D. Translation and psychometric properties of the Chinese version of the Leeds Attitudes to Concordance II scale. BMC Med Inform Decis Mak 2015 Aug 1;15:60. [doi: 10.1186/s12911-015-0184-0] [Medline: 26232245]
35. Hsieh FY, Lavori PW, Cohen HJ, Feussner JR. An overview of variance inflation factors for sample-size calculation. Eval Health Prof 2003 Sep;26(3):239-257. [doi: 10.1177/0163278703255230]
36. Rowe L. Comparing learning outcomes and student and instructor perceptions of a simultaneous online versus in-person biochemistry laboratory course. J Chem Educ 2024 Mar 12;101(3):882-891. [doi: 10.1021/acs.jchemed.3c00571] [Medline: 38495613]
37. Holmes G, Clacy A, Hamilton A, Kõlves K. Online versus in-person gatekeeper suicide prevention training: comparison in a community sample. J Ment Health 2024 Oct;33(5):605-612. [doi: 10.1080/09638237.2024.2332811] [Medline: 38602188]
38. Iyer P, Mok V, Sehmbi AS, et al. Online versus in-person surgical near-peer teaching in undergraduate medical education during the COVID-19 pandemic: a mixed-methods study. Health Sci Rep 2024 Feb;7(2):e1889. [doi: 10.1002/hsr2.1889] [Medline: 38357488]
39. Lelutiu-Weinberger C, Clark KA, Pachankis JE. Mental health provider training to improve LGBTQ competence and reduce implicit and explicit bias: a randomized controlled trial of online and in-person delivery. Psychol Sex Orientat Gend Divers 2023 Dec;10(4):589-599. [doi: 10.1037/sgd0000560] [Medline: 38239562]
40. Bos-van den Hoek DW, van Laarhoven HWM, Ali R, et al. Blended online learning for oncologists to improve skills in shared decision making about palliative chemotherapy: a pre-posttest evaluation. Support Care Cancer 2023 Mar;31(3):184. [doi: 10.1007/s00520-023-07625-6]
41. Solberg LI. Recruiting medical groups for research: relationships, reputation, requirements, rewards, reciprocity, resolution, and respect. Implement Sci 2006 Oct 26;1:25. [doi: 10.1186/1748-5908-1-25] [Medline: 17067379]
42. Krupat E, Hiam CM, Fleming MZ, Freeman P. Patient-centeredness and its correlates among first year medical students. Int J Psychiatry Med 1999;29(3):347-356. [doi: 10.2190/DVCQ-4LC8-NT7H-KE0L] [Medline: 10642908]
43. Ishikawa H, Son D, Eto M, Kitamura K, Kiuchi T. Changes in patient-centered attitude and confidence in communicating with patients: a longitudinal study of resident physicians. BMC Med Educ 2018 Jan 25;18(1):20. [doi: 10.1186/s12909-018-1129-y] [Medline: 29370796]
44. Mudiyanse RM, Pallegama RW, Jayalath T, Dharmaratne S, Krupat E. Translation and validation of patient-practitioner orientation scale in Sri Lanka. Educ Health (Abingdon) 2015;28(1):35-40. [doi: 10.4103/1357-6283.161847] [Medline: 26261112]
45. Wang J, Zou R, Fu H, Qian H, Yan Y, Wang F. Measuring the preference towards patient-centred communication with the Chinese-revised Patient-Practitioner Orientation Scale: a cross-sectional study among physicians and patients in clinical settings in Shanghai, China. BMJ Open 2017 Sep 18;7(9):e016902. [doi: 10.1136/bmjopen-2017-016902] [Medline: 28928188]
46. Ahmad W, Ashraf H, Talat A, et al. Association of burnout with doctor-patient relationship and common stressors among postgraduate trainees and house officers in Lahore-a cross-sectional study. PeerJ 2018;6:e5519. [doi: 10.7717/peerj.5519] [Medline: 30221087]
47. Akkafi M, Sajadi HS, Sajadi ZS, Krupat E. Attitudes toward patient-centered care in the mental care services in Isfahan, Iran. Community Ment Health J 2019 Apr;55(3):548-552. [doi: 10.1007/s10597-018-0357-2] [Medline: 30535891]
48. Körner M, Ehrhardt H, Steger AK, Bengel J. Interprofessional SDM train-the-trainer program “Fit for SDM”: provider satisfaction and impact on participation. Patient Educ Couns 2012 Oct;89(1):122-128. [doi: 10.1016/j.pec.2012.04.008] [Medline: 22647558]
49. Leblanc A, Légaré F, Labrecque M, et al. Feasibility of a randomised trial of a continuing medical education program in shared DECISION-making on the use of antibiotics for acute respiratory infections in primary care: the DECISION+ pilot trial. Implement Sci 2011 Jan 18;6:5. [doi: 10.1186/1748-5908-6-5] [Medline: 21241514]
50. Volk RJ, Shokar NK, Leal VB, et al. Development and pilot testing of an online case-based approach to shared decision making skills training for clinicians. BMC Med Inform Decis Mak 2014 Nov 1;14:95. [doi: 10.1186/1472-6947-14-95] [Medline: 25361614]
51. Rusiecki J, Schell J, Rothenberger S, Merriam S, McNeil M, Spagnoletti C. An innovative shared decision-making curriculum for internal medicine residents. Acad Med 2018;93:937-942. [doi: 10.1097/ACM.0000000000001967]
52. Hoffmann TC, Bennett S, Tomsett C, Del Mar C. Brief training of student clinicians in shared decision making: a single-blind randomized controlled trial. J Gen Intern Med 2014 Jun;29(6):844-849. [doi: 10.1007/s11606-014-2765-5] [Medline: 24481686]
53. Shaw WS, Woiszwillo MJ, Krupat E. Further validation of the Patient-Practitioner Orientation Scale (PPOS) from recorded visits for back pain. Patient Educ Couns 2012 Nov;89(2):288-291. [doi: 10.1016/j.pec.2012.07.017] [Medline: 22954491]
54. Krupat E, Bell RA, Kravitz RL, Thom D, Azari R. When physicians and patients think alike: patient-centered beliefs and their impact on satisfaction and trust. J Fam Pract 2001 Dec;50(12):1057-1062. [Medline: 11742607]
55. Krupat E, Yeager CM, Putnam S. Patient role orientations, doctor-patient fit, and visit satisfaction. Psychol Health 2000 Sep;15(5):707-719. [doi: 10.1080/08870440008405481]

56. Carlsen B, Aakvik A, Norheim OF. Variation in practice: a questionnaire survey of how congruence in attitudes between doctors and patients influences referral decisions. Med Decis Making 2008;28(2):262-268. [doi: 10.1177/0272989X07311751] [Medline: 18349435]
57. Diendéré G, Farhat I, Witteman H, Ndjaboue R. Observer ratings of shared decision making do not match patient reports: an observational study in 5 family medicine practices. Med Decis Making 2021 Jan;41(1):51-59. [doi: 10.1177/0272989X20977885] [Medline: 33371802]
58. Evong Y, Chorney J, Ungar G, Hong P. Perceptions and observations of shared decision making during pediatric otolaryngology surgical consultations. J Otolaryngol Head Neck Surg 2019 Jun 17;48(1):28. [doi: 10.1186/s40463-019-0351-x] [Medline: 31208462]
59. Williams D, Edwards A, Wood F, et al. Ability of observer and self-report measures to capture shared decision-making in clinical practice in the UK: a mixed-methods study. BMJ Open 2019 Aug 18;9(8):e029485. [doi: 10.1136/bmjopen-2019-029485] [Medline: 31427333]
60. Meterko M, Wright S, Lin H, Lowy E, Cleary PD. Mortality among patients with acute myocardial infarction: the influences of patient-centered care and evidence-based medicine. Health Serv Res 2010 Oct;45(5 Pt 1):1188-1204. [doi: 10.1111/j.1475-6773.2010.01138.x] [Medline: 20662947]

Abbreviations

CE: communication/empathy
EPS: eventual paternalistic style
GAD: generalized anxiety disorder
LATConII: Leeds Attitudes Towards Concordance scale, second version
PC: primary care
PCC: person-centered care
RCT: randomized controlled trial
SC: shared control
SDM: shared decision-making

Edited by B Lesselroth; submitted 18.06.23; peer-reviewed by C Timm, J Rubel; revised version received 27.07.24; accepted 19.08.24; published 16.01.25.

Please cite as:
Ramos-García V, Rivero-Santana A, Peñate-Castro W, Álvarez-Pérez Y, Duarte-Díaz A, Torres-Castaño A, Trujillo-Martín MDM, González-González AI, Serrano-Aguilar P, Perestelo-Pérez L
A Brief Web-Based Person-Centered Care Group Training Program for the Management of Generalized Anxiety Disorder: Feasibility Randomized Controlled Trial in Spain
JMIR Med Educ 2025;11:e50060
URL: https://mededu.jmir.org/2025/1/e50060
doi:10.2196/50060

Faculty Perceptions on the Roles of Mentoring, Advising, and Coaching in an Anesthesiology Residency Program: Mixed Methods Study

Sydney Nykiel-Bailey[1*], DO; Kathryn Burrows[2*], PhD; Bianca E Szafarowicz[1*], DO; Rachel Moquin[1*], EdD, MA

[1]Department of Anesthesiology, Washington University School of Medicine, 660 S Euclid Avenue, Saint Louis, MO, United States

[2]National Coalition of Independent Scholars, Independent Scholar 432 Division, Oregan City, OR, United States

[*]all authors contributed equally

Corresponding Author:
Sydney Nykiel-Bailey, DO
Department of Anesthesiology, Washington University School of Medicine, 660 S Euclid Avenue, Saint Louis, MO, United States

Abstract

Background: Mentoring, advising, and coaching are essential components of resident education and professional development. Despite their importance, there is limited literature exploring how anesthesiology faculty perceive these practices and their role in supporting residents.

Objective: This study aims to investigate anesthesiology faculty perspectives on the significance, implantation strategies, and challenges associated with mentorship, advising, and coaching in resident education.

Methods: A comprehensive survey was administrated to 93 anesthesiology faculty members at Washington University School of Medicine. The survey incorporated quantitative Likert-scale questions and qualitative short-answer responses to assess faculty perceptions of the value, preferred formats, essential skills, and capacity for fulfilling multiple roles in these support practices. Additional areas of focus included the impact of staffing shortages, training requirements, and the potential of these practices to enhance faculty recruitment and retention.

Results: The response rate was 44% (n=41). Mentoring was identified as the most important aspect, with 88% (n=36) of faculty respondents indicating its significance, followed by coaching, which was highlighted by 78% (n=32) of respondents. The majority felt 1 faculty member can effectively hold multiple roles for a given trainee. The respondents desired additional training for roles and found roles to be rewarding. All roles were seen as facilitating recruitment and retention. Barriers included faculty burnout; confusion between roles; time constraints; and desire for specialized training, especially in coaching skills.

Conclusions: Implementing structured mentoring, advising, and coaching can profoundly impact resident education but requires role clarity, protected time, culture change, leadership buy-in, and faculty development. Targeted training and operational investments could enable programs to actualize immense benefits from high-quality resident support modalities. Respondents emphasized that resident needs evolve over time, necessitating flexibility in appropriate faculty guidance. While coaching demands unique skills, advising hinges on expertise and mentoring depends on relationship-building. Systematic frameworks of coaching, mentoring, and advising programs could unlock immense potential. However, realizing this vision demands surmounting barriers such as burnout, productivity pressures, confusion about logistics, and culture change. Ultimately, prioritizing resident support through high-quality personalized guidance can recenter graduate medical education.

(JMIR Med Educ 2025;11:e60255) doi:10.2196/60255

KEYWORDS

coaching; faculty perceptions; mentoring; perception; medical education; anesthesia; modality; support; Washington University; university; coaching skills; training; culture change; culture; flexibility; systematic framework

Introduction

The current landscape of medical education is influenced by both medical culture and shifting demographics among learners. Factors such as medical provider burnout [1], a nationwide shortage of medical staff [2], and the evolving characteristics of different generations of learners are reshaping medical education [3]. It is imperative that the well-being and guidance of learners, both personally and professionally, are recentralized as the core of medical education. Emphasizing principles such as advising, mentoring, and coaching is crucial to support learners in their journey toward academic and personal fulfillment. These principles should be thoroughly examined

and reevaluated to empower learners to pursue paths of academic and personal success, foster self-assessment, ensure a nurturing learning environment, and encourage a commitment to lifelong learning [1,2]. The objective of this paper is to examine the attitudes and experiences of clinical-academic anesthesiology faculty with respect to their understanding and practice of mentoring, advising, and coaching. Our aim is to identify key themes that more clearly define these roles within medical education, as well as to elucidate potential barriers to their implementation and sustainability. Furthermore, we seek to understand faculty perspectives on the need for formalized educational support in these areas. We anticipate that the insights gained from this study could be broadly applicable across the graduate medical education spectrum, particularly as the focus in education increasingly shifts toward professionalism and well-being.

The education and welfare of medical residents hinge upon a multifaceted network of connections. Residents at different stages of their training will necessitate varying forms of engagement: mentoring, advising, or coaching. While these 3 avenues are distinct, they all share the common aim of nurturing education, wellness, and career progression [2,3]. Each approach serves its unique purpose and uses diverse methodologies [2]. Identifying the most suitable modality for the learner is paramount. Facilitators must adeptly involve themselves and customize sessions to ensure that expectations and objectives resonate with the learner [2].

Traditionally, mentoring has been the primary means of providing guidance [4]. It entails a sustained personal relationship between mentor and mentee, with the learner's overarching aspirations guiding the interaction. Conversations, career mapping, and counsel are derived from the mentor's experiences and expertise [2,3]. Typically, mentors possess knowledge in the pertinent field and share their insights with the learner. The mentor guides sessions, posing direct questions with long-term goals as the focal point. In residency education, mentoring often follows a structured format, though informal mentorships may naturally evolve. Institutions may request mentors to provide feedback or document these sessions for accreditation purposes [2,3].

Advising typically comprises a single, informal session focused on a specific issue or inquiry. The advisor leads the session and provides solutions or strategies based on their own experiences. The learner has the autonomy to decide whether to heed the advice. Unlike mentoring, a sustained relationship is not necessarily a prerequisite for advising, and subsequent follow-up is usually with independence and self-driven by the needs of the advisee [5]. Advisors may possess limited insight into the learner's personal or academic strengths and weaknesses, resulting in advice limited to specific scenarios [6].

Academic coaching differs from advising and mentoring in that it prioritizes the learner's agency. Coaches refrain from offering advice or engaging in decision-making. Instead, their role is to facilitate self-discovery and create a supportive atmosphere for self-assessment and future planning [2]. Coaches assist learners in identifying actions that may lead to success or failure. Unlike mentors and advisors, coaches may not necessarily possess expertise in the medical field. Coach engagement is supported by actively listening to the learner and offering questions to encourage self-awareness. Coaching fosters a consistent, enduring relationship characterized by an educational partnership between coach and learner [2].

No single form of guidance is adequate to meet the needs of today's students, and students' needs evolve as they move through residency [7]. Faculty must be facile in their ability to intuit what type of guidance is appropriate for a specific student or situation, and be able to provide that guidance or refer the student to someone who can [8]. For this reason, faculty development programs play a crucial role in supporting faculty as they rise to meet the challenges of guiding trainees, and faculty training in these support modalities may be lacking [9]. Training educators on how to target student needs by using the most effective guidance strategy will help decrease role confusion [8]. Training and developing faculty in advising, mentoring, and coaching help cultivate an ongoing culture of scholarship [10] and can help faculty navigate the competing challenges of their clinical and nonclinical roles [11]. Faculty report that lack of support from leadership and lack of proper training are barriers to their role as advisors, coaches, and mentors [11], and training and assessment tools for faculty members are crucial [7,9].

Methods

Study Design

A survey (Multimedia Appendix 1) was sent to 93 Washington University School of Medicine Anesthesiology clinical educator faculties. This target population was used as a convenience sample, representing a cohesive cohort with consistent interactions with trainees. This survey was developed based on core competencies and conceptual differentiations outlined for the roles of advisors, coaches, and mentors in medical education [5,6,8,9]. Drawing from Wolff et al [9], support modality definitions and key characteristics were designed to reflect critical distinctions regarding focus, relationship context, longevity, skill sets, and objective alignment [9]. Survey questions were formulated to assess physician perspectives across these theoretical domains for each resident support role.

A group of coaching experts within the Department of Anesthesiology was selected to create a novel survey tool. To facilitate the design and construction of the survey instrument, the research team used a modified Delphi technique, a widely recognized method for achieving consensus among experts. A subset of academic faculty was invited to participate in a pilot study aimed at testing multiple dimensions of the survey's implementation. This pilot study served several purposes: (1) to ensure the clarity and comprehensibility of the survey questions, (2) to evaluate the technical functionality of the survey platform, and (3) to assess the feasibility of applying inductive thematic analysis to the pilot data. Through iterative revisions and rounds of expert feedback, the survey underwent several modifications to enhance both face validity and content validity. The final version of the survey, which reflects the culmination of this rigorous development process, is presented in Multimedia Appendix 1. The survey is composed of 2 Likert

5-point scale quantitative items and 11 qualitative open-ended questions.

Quantitative items examined perceptions of importance and optimal configurations applying the principles of Wolff et al [9] regarding situational demands and need for role clarity [9]. Quantitative data were collected using the REDCap (Research Electronic Data Capture) Consortium platform (Vanderbilt University), a secure web-based application designed to support data capture for research studies. Faculty received the voluntary survey through department email, no incentives offered, and faculty log-in prevented duplicate entries. Data were analyzed using descriptive statistics. Qualitative questions elicited feedback on specialized skills, training interests, and implementation barriers grounded in advising, coaching, and mentoring competency frameworks [5-9]. The sequence of survey topics reflects established theory comparing and contrasting these support avenues [6-8]. An inductive qualitative analysis was conducted, using the Braun and Clarke [12] 6-phase approach to thematic analysis. This methodological framework, widely used in qualitative research, ensures both the flexibility and rigor required for the interpretative analysis of complex datasets. The 6 stages—familiarization with data, generating initial codes, searching for themes, reviewing themes, defining and naming themes, and writing up—provide a structured yet adaptable framework for data interpretation [12]. Qualitative data were collected through open-response questions included in the REDCap survey. The text from these open-response questions was analyzed using the Dedoose coding themes platform.

The process begins with open coding to identify initial patterns within the data. Codes were then further examined to uncover relationships, allowing for the grouping of related codes into broader thematic categories. Subsequently, these groups were analyzed to identify overarching themes that reflect deeper insights into the data. This iterative process was designed to ensure a comprehensive exploration of the qualitative data and enhance the interpretive depth of the analysis.

To ensure the reliability of the findings, at least two independent researchers reviewed and coded all data. The initial coding and preliminary analysis of the qualitative data was conducted by 1 author (SN-B), using Dedoose—a cloud-based software platform designed to facilitate mixed methods analysis. After the initial coding phase, 2 members of the research team engaged in collaborative discussions to reconcile coding discrepancies and synthesize their interpretations. This process of researcher triangulation not only strengthens the credibility of the findings but also helps to ensure that the emerging themes are robust and reflect the nuances present in the data [13].

The survey contained a brief textual description of the difference between the roles of mentor, advisor, and coach, respondents. Respondents were asked how important they thought each role is in graduate medical training, whether 1 individual can fulfill all 3 roles, what kind of training is needed for faculty to perform these roles, and whether resident needs for different forms of faculty relationship change over time. In addition, faculty were asked if they had ever performed any of the 3 roles. Questions were both quantitative (responses on a 5-point Likert scale) and qualitative (open-ended short responses). A total of 41 surveys were completed (44% response rate).

Ethical Considerations

This study used both quantitative and qualitative data collection and analysis. This study was approved by the Institutional Review Board at Washington University (202310164). Faculty were informed about this study via an initial email announcement, followed by 2 reminder emails. Informed consent was obtained with faculty selecting “accept” on the survey; the ability to opt out was provided. Electronic data were password protected, encrypted, and transmitted using recognized security for electronic submission. No compensation was provided.

Results

Roles

Respondents had varying opinions about the importance of mentoring, advising, and coaching in graduate medical education. Mentoring was seen as most important, with 88% of respondents indicating that they agreed or strongly agreed that it was important, and coaching was seen as less important, with only 78% of respondents indicating agreement or strong agreement that it was important (Table 1).

Table . Importance of mentoring, advising, and coaching.

	Agree or strongly agree, n (%)
Coaching	32 (78)
Advising	34 (83)
Mentoring	36 (88)

In total, 90% of respondents agreed that 1 faculty member could fulfill two or more roles for a single resident. For example, respondent 2 explained, “Faculty can possess more than one skill set and/or the relationship between a faculty and resident may benefit from a multi-faceted focus once trust has been developed.” However, others noted that there may be conflicts between roles and that the unique skills required for each role are not always possessed by the same person. Respondent 3 noted, “This works sometimes, I think, but can’t dependably work all the time. Some faculty are better at one role or another. Obviously, some coaching and advising can only be done by faculty with certain skills or areas of expertise.”

Additionally, respondents noted the role faculty mentoring and coaching play in recruitment and retention efforts for faculty and trainees. For example, respondent 3 noted, “if it were made clear that we offered thoughtful assignment of each of these roles, with examples for coaching and advising, I think that would likely be seen as a significant benefit.” Others agreed

that providing these roles to residents in a systematic way would be beneficial for recruitment, but noted barriers to implementation, as respondent 33 explained, "I think that these three roles are important to recruit residents for fellowships and faculty. Fostering a supportive environment through these roles is very important for recruitment; however, other factors such as the job market and hours worked often overshadow these aspects in recruiting."

Training

Most respondents agreed that specialized training in all 3 roles was important, especially for coaching, which was seen as requiring a unique skill set. Formal training for all 3 roles was endorsed, especially for coaching. Respondents noted that the skills required for the roles came naturally to some faculty. For the advising role, having career experience and expertise in the graduate education process was seen as especially useful. For example, respondent 10 noted, "Knowing the residency experience well and knowing what challenges residents face. Additionally, it's important to know career options after." Mentoring was regarded as being based on relationship building and interpersonal skills, as well as necessitating emotional intelligence. Respondents reflected that mentorship involves skill sets not necessarily embedded in clinical training. Respondent 21 explained, "Teach the teacher/instructor courses are helpful. Being a good clinician and/or researcher do not provide us the skills of being a good teacher. A bit of more understanding, empathy, and psychological support are necessary for knowing ourselves better and using these abilities for others. Patience, more listening, time, sharing experiences, sometimes coming up with a challenging scenario to discuss, widen the horizon, show other possibilities never thought of before as options."

Respondents indicated that they would be interested in targeted training. Coaching (63%) was the highest, however, respondents were less interested in specialized training in advising and mentoring skills (Table 2).

Table . Interest in specialized training in coaching, advising, and mentoring.

	Interest in specialized training, n (%)
Coaching	29 (63)
Advising	17 (41)
Mentoring	16 (39)

Experience

Nearly 88% of respondents had fulfilled one or more of these roles in their career, and they noted that holding all 3 roles was personally and professionally rewarding. Of the 36 faculty members who reported fulfilling these roles in the past, 15 (42%) mentioned the satisfaction of watching students progress through their training and career. Coaching was noted as being the most challenging, but also the most rewarding. For example, respondent 22 said, "Honestly, I think that serving in this role for strong residents is one of the most rewarding parts of my job. I love to see people be successful in their careers."

Barriers

Respondents identified barriers to faculty engaging in quality mentoring, coaching, and advising, which included faculty burnout, time limitations, and confusion about roles, responsibilities, and expectations. Respondent 10 said, "The residents have so many rotations. It's rare to have consistent clinic time to coach and mentor/advise. Coaching off hours is very time consuming." Lack of time was mentioned by 68% of respondents, for example as respondent 29 explained "I was a terrible mentor. Never could find time to meet with my mentee."

Respondents had mixed responses about whether the national anesthesia provider shortage had impacted their engagement with or performance of any of these roles. Respondents noted lack of time in general, and lack of protected time more specifically, as factors influencing their ability to engage in these roles, and some attributed the challenge with time to provider shortage. For example, respondent 17 said, "The shortage has decreased faculty time to provide these aspects, may be important for departments to assign a subgroup of faculty to serve these roles so time is protected."

Table 3 presents the results of the thematic analysis, offering a detailed synthesis of the emergent themes and subthemes derived from the qualitative data. The richness of respondent narratives facilitated a comprehensive exploration, allowing for nuanced insights into the key thematic categories. These findings provide a robust framework for understanding the underlying patterns and relationships within the data, supporting the depth and validity of the analysis.

Table . Main themes and representative quotes.

Theme and subtheme		Representative quotes
Roles		
	Faculty can perform multiple roles	• "Faculty can possess more than one skill set and/or the relationship between a faculty and resident may benefit from a multi-faceted focus once trust has been developed." [Respondent 2] • "Different skill sets are needed and faculty may possess one or many of the skill sets needed." [Respondent 9] • "I believe the necessary skills can be learned and employed by a single person. It also depends upon the mentee's needs and the qualities of their relationship with the mentor/advisor/coach." [Respondent 5] • "A faculty member can take different roles throughout the 4 years that a trainee is counseled. I find that interns need mentoring and advising, as the resident progresses coaching and mentoring is important." [Respondent 16]
	Faculty cannot perform multiple roles	• "This works sometimes, I think, but can't dependably work all the time. Some faculty are better at one role or another. Obviously, some coaching and advising can only be done by faculty with certain skills or areas of expertise." [Respondent 3] • "Sometimes the line between just providing feedback for a specific case as an advisor can be hard if you are also a mentor to that person." [Respondent 37] • "Different goals and different time frames over which those goals are realized. The trainee asking advising may be frustrated by a "mentoring" approach. Some great career mentors may not have the specific subspecialty background for focused advising." [Respondent 7]
Training		

Theme and subtheme		Representative quotes
	Request for formal education or faculty development	• "I think at least some sort of education on how to be an advisor would be helpful." [Respondent 1] • "Teach the teacher/instructor courses are helpful. Being a good clinician and/or researcher do not provide us the skills of being a good teacher. A bit of more understanding, empathy, and psychological support are necessary for knowing ourselves better and using these abilities for others. Patience, more listening, time, sharing experiences, sometimes coming up with a challenging scenario to • discuss, widen the horizon, show other possibilities never thought of before as options." [Respondent 21] • "Coaching should require some training/knowledge of professional coaching, which is more structured that mentorship or career advising which can be more informal." [Respondent 4] • "Didactics/workshops/peer mentoring needed." [Respondent 31] • "Training focused to the knowledge and skillset as well as teaching techniques and current best practices." [Respondent 2] • "Structured professional coaching training." [Respondent 6]
Experiences		

Theme and subtheme		Representative quotes
	Mentor role	• "I was a terrible mentor. Never could find time to meet with my mentee." [Respondent 29] • "Mentoring has been the most rewarding, coaching second. Advising feels limited and one-directional." [Respondent 5]
	Coach role	• "All three - coaching seems to be the most challenging." [Respondent 7] • "I have played all 3 roles during my time as an educator. The coaching roles are always the most rewarding. The ability to guide residents through self-discovery is extremely rewarding. I find that coaching residents later in their training prepares them for being faculty and having a successful trajectory." [Respondent 17] • "I have been a coach and an advisor. Coaching is extremely rewarding." [Respondent 39] • "Primarily coaching, which I found rewarding when a trainee felt our interaction was beneficial through skill-based or confidence improvements." [Respondent 41]
	Advisor role	• "Advising in clear goal-directed tasks, such as a conference, abstract, paper." [Respondent 8] • "I have served as a mentor and advisor, both of which were very rewarding. I felt that it made it easier to discuss topics at work that we may otherwise would not have brought up. I also felt satisfaction getting to know the trainees better and become more a part of their lives." [Respondent 27] • "Clinical teaching while supervising trainees fulfills the "advisor" role. I was also a designated faculty mentor for a clinical fellow." [Respondent 34] • "Clinical teaching while supervising trainees fulfills the advisor role." [Respondent 33]
	Combination of roles	• "Have provided all three of these roles in different capacities. I enjoy fostering learning with the goal of being the attending I wish I had as a trainee." [Respondent 33] • "Yes, I feel that I serve as an advisor to residents and mentor to fellows." [Respondent 20] • "I would say informally on day-to-day basis interactions with residents and fellows, yes for all 3. Advisor more than mentor more than coach. It is rewarding when it seems welcomed and appreciated by the residents and fellows and I can see them grow and improve. It is frustrating when I am putting in the effort/trying to do these things and the trainees are not receptive, not appreciative, or feel as though I am being too particular or micromanaging." [Respondent 35] • "I have provided all 3. The coaching roles are always the most rewarding. The ability to guide residents through discovery is extremely rewarding." [Respondent 20]
	Recruitment role	

Theme and subtheme	Representative quotes
	• "The biggest drivers right now for recruitment are time and money. The biggest long-term satisfaction will come from deeper meaning. Using the relationships in these roles may help highlight some of these deeper meanings and may help recruit fellows and faculty if they have the sense that this is best for themselves and their families. At the same time, there has to be felt and sustained room for the individual to act on these deeper meaningful insights. Solving for individual growth requires commitment from the system as well as the individual." [Respondent 9] • "Yes. When residents can see faculty care about their education and also enjoy working here it's easier to recruit. " [Respondent 20] • "Mentorship and coaching require a relationship, that may be beneficial for recruitment." [Respondent 17] • "A structured mentor/coaching program would be very appealing to most applicants." [Respondent 31]
Barriers	

Theme and subtheme		Representative quotes
	Discrete roles	• "If role/project is not clearly defined, could cause some confusion. Time." [Respondent 1] • "Mentorship is often a friendly and personal relationship, which could make it harder to, for example, challenge the mentee in a coaching scenario. Very specific example - perhaps a mentee would feel uncomfortable doing mock oral boards with their mentor, if they're relatively advanced in training, but early in the oral boards prep process." [Respondent 3] • "Different goals and different time frames over which those goals are realized. The trainee asking for advising may be frustrated by a "mentoring" approach. Some great career mentors may not have the specific subspecialty background for focused advising." [Respondent 7]
	Time	• "Time and lack in continuous interactions with the resident." • [Respondent 18] • "Time to meet with the trainee and to establish a relationship." • [Respondent 14] • "Time and managing the balance btw one's professional responsibilities and taking on additional responsibilities that the above would entail." [Respondent 6] • "The residents have so many rotations. It's rare to have consistent clinic time to coach and mentor/advise. Coaching off hours is very time-consuming." [Respondent 10]
	Burnout	• "It would be a good recruitment tool but difficult to deliver in near future with current staffing shortages and burn-out among faculty members. In practice, it would require significant training, time, and effort to optimize and ensure an equal experience among trainees. Remuneration could increase participation but doesn't get around the issue of lack of time." [Respondent 12] • "Yes, particularly for faculty. Relatively little resources currently to develop faculty. More investment needed to reduce the chance of burnout/disengagement/attrition to other practices." [Respondent 31] • "We are all strapped for time and burnt out." [Respondent 40]
	Anesthesia shortage	• "These 3 are probably even more important for our trainees and may be beneficial to expand these past trainees and onto faculty as well. The shortage has decreased faculty time to provide these aspects, may be important for departments to assign a subgroup of faculty to serve these roles so time is protected." [Respondent 17] • "I think that with the shortages, faculty have taken on more solo assignments and have overall less contact with the residents and don't get to know them as well." [Respondent 36]

Discussion

Overview

This study explored perceptions of anesthesia faculty regarding the roles of mentoring, advising, and coaching in graduate medical education. The results highlight the perceived benefits of these practices as well as barriers to implementation. Anesthesia residency is unique in its internship, and a vast majority of education and interactions with faculty occurs at bedside in the operating room. Medical training and trainee progression differ across disciplines. This study focuses specifically on anesthesia faculty and a single institution, which overall limits generalizability.

Principal Findings

The survey results indicate that faculty view mentoring, advising, and coaching as important for resident education and development. These practices have been shown to improve resident well-being, promote career planning, facilitate reflection and self-assessment, and identify knowledge gaps [5,6]. Furthermore, implementing structured programs in these areas can aid recruitment and retention of both residents and faculty.

Of the 3 roles faculty partake in, there is a consensus on the importance of mentoring throughout training and prioritizing this role over advising and coaching. However, the data suggests a significant interest in specialized training for coaching versus roles in advising and mentoring. Investigating the differences in practice versus desire, recurrent themes of time and experience were identified. Although the roles as a mentor, advisor, and coach can overlap, a majority of the cohort indicated they prioritize mentoring given the noted constraints of time and experience.

Implications of Findings

To actualize these practices, each department must clearly define the roles of mentor, advisor, and coach. Expectations, training requirements, and time commitments should be delineated. Assignments of roles can be made between faculty and residents based on alignment of career goals, personalities, and logistics. Protected nonclinical time should be designated for these meetings separate from clinical work. Success stories and positive impacts on residents should be tracked and celebrated.

Comparison to the Literature

Recurrent themes were identified when comparing to other literature, such as the establishment of a clear definition and terms of each role. This would help faculty facilitate their approach to the learner needs [5,8]. Additional repeated themes of the overlap in roles, limitations in time, and experiences were highlighted in other studies in reference to mentorship, advising, and coaching [5,10]. Anesthesiology training presents challenges specific to the discipline, which can be generalized to medical training programs at other institutions. There has been an increased productivity within the academic institutions leading to less bedside education opportunities and difficulty establishing dedicated time for routine meetings with trainees.

Limitations

This study has several limitations. First, this study was based on a single survey with a 44% response rate, which may limit the generalizability of the findings. Nonresponders may have had different perspectives on the importance and implementation of mentoring, advising, and coaching. Second, this study was conducted at a single academic medical center, so the results may not be representative of other institutions. Additionally, this study was solely conducted with anesthesia faculty. Other specialties may not portray the same obstacles and constraints in fulfilling the roles of mentorship, advising, and coaching. The learning environment and progression through training also differ between anesthesiology and other specialties, which limits the generalizability across disciplines. The limited time and consistency with faculty may lead to less specific demands from trainees and unfulfillment from educators. Third, the survey relied on self-reported perceptions and experiences, which are subject to recall bias and social desirability bias. Fourth, this study did not explore the perspectives of residents themselves on these support modalities. Future research should examine resident experiences with and preferences for mentoring, advising, and coaching. Finally, while this study identified perceived barriers to implementing these practices, it did not evaluate specific strategies for overcoming these obstacles. Further work is needed to develop and test interventions to enhance faculty engagement in resident support roles.

Conclusion

Addressing barriers such as faculty burnout, role ambiguity, time constraints, and the need for specialized training is critical for the success of mentoring, advising, and coaching initiatives. Implementing comprehensive faculty development programs aimed at enhancing skills in these domains is essential, particularly for coaching, which requires distinct pedagogical approaches. The recruitment and retention of faculty, as well as their career longevity, may be positively influenced by the intrinsically rewarding nature of relationships with trainees.

To facilitate meaningful faculty engagement, institutional leadership must ensure protected time for participation in these activities without detriment to clinical productivity. Moreover, a cultural shift may be necessary in programs that place disproportionate emphasis on service obligations, potentially at the expense of educational and developmental support for residents. Prioritizing resident education and well-being can contribute to improved morale and overall program satisfaction.

By investing in faculty development, enhancing institutional infrastructure, and fostering a culture that values educational alliance, graduate medical education programs can realize significant benefits from high-quality mentoring, advising, and coaching relationships. Such investments are pivotal for advancing the professional development of both faculty and trainees, ultimately enhancing the overall quality of medical education.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Coaching, mentoring, and advising survey.
[DOCX File, 17 KB - mededu_v11i1e60255_app1.docx]

References

1. ăranu SM, tefăniu R, Rotaru T, et al. Factors associated with burnout in medical staff: a look back at the role of the COVID-19 pandemic. Healthcare (Basel) 2023 Sep 13;11(18):2533. [doi: 10.3390/healthcare11182533] [Medline: 37761730]
2. GlobalData Plc. The complexities of physician supply and demand: projections from 2019 to 2034. : AAMC; 2024.
3. Plochocki JH. Several ways generation Z may shape the medical school landscape. J Med Educ Curric Dev 2019 Oct 31;6:2382120519884325. [doi: 10.1177/2382120519884325] [Medline: 31701014]
4. Farkas AH, Allenbaugh J, Bonifacino E, Turner R, Corbelli JA. Mentorship of US medical students: a systematic review. J Gen Intern Med 2019 Nov;34(11):2602-2609. [doi: 10.1007/s11606-019-05256-4] [Medline: 31485967]
5. Marcdante K, Simpson D. Choosing when to advise, coach, or mentor. J Grad Med Educ 2018 Apr;10(2):227-228. [doi: 10.4300/JGME-D-18-00111.1] [Medline: 29686766]
6. Deiorio NM, Hammoud MM. Coaching in medical education: accelerating change in medical education consortium. : AMA; 2024.
7. Alisic S, Boet S, Sutherland S, Bould MD. A qualitative study exploring mentorship in anesthesiology: perspectives from both sides of the relationship. Can J Anesth/J Can Anesth 2016 Jul;63(7):851-861. [doi: 10.1007/s12630-016-0649-3]
8. Santiesteban L, Young E, Tiarks GC, et al. Defining advising, coaching, and mentoring for student development in medical education. Cureus 2022 Jul;14(7):e27356. [doi: 10.7759/cureus.27356] [Medline: 36043012]
9. Wolff M, Deiorio NM, Miller Juve A, et al. Beyond advising and mentoring: competencies for coaching in medical education. Med Teach 2021 Oct;43(10):1210-1213. [doi: 10.1080/0142159X.2021.1947479] [Medline: 34314291]
10. Reid MB, Misky GJ, Harrison RA, Sharpe B, Auerbach A, Glasheen JJ. Mentorship, productivity, and promotion among academic hospitalists. J Gen Intern Med 2012 Jan;27(1):23-27. [doi: 10.1007/s11606-011-1892-5] [Medline: 21953327]
11. Jha P, Quinn B, Durbin S, Bhandari S. Perceptions of junior faculty in general internal medicine regarding mentoring medical students and residents in scholarly projects. J Gen Intern Med 2019 Jul;34(7):1098-1099. [doi: 10.1007/s11606-019-04937-4] [Medline: 30887433]
12. Peel KL. A beginner's guide to applied educational research using thematic analysis. Pract Assess Res Eval 2020 Jan;25(1):2. [doi: 10.7275/ryr5-k983]
13. Patton M. Qualitative Research and Evaluation Methods: Sage; 2015.

Abbreviations

REDCap: Research Electronic Data Capture

Edited by B Lesselroth; submitted 06.05.24; peer-reviewed by K Kagkelaris, M Ilaghi; revised version received 23.09.24; accepted 03.12.24; published 21.01.25.

Please cite as:
Nykiel-Bailey S, Burrows K, Szafarowicz BE, Moquin R
Faculty Perceptions on the Roles of Mentoring, Advising, and Coaching in an Anesthesiology Residency Program: Mixed Methods Study
JMIR Med Educ 2025;11:e60255
URL: https://mededu.jmir.org/2025/1/e60255
doi:10.2196/60255

Performance of Plug-In Augmented ChatGPT and Its Ability to Quantify Uncertainty: Simulation Study on the German Medical Board Examination

Julian Madrid[1], MD; Philipp Diehl[1], MD, PhD; Mischa Selig[2,3], PhD; Bernd Rolauffs[2,3], MD, PhD; Felix Patricius Hans[2,4], MD, MSc; Hans-Jörg Busch[2,4], MD, PhD; Tobias Scheef[2,5*], MD; Leo Benning[2,4*], MPH, MD

[1]Department of Cardiology, Pneumology, Angiology, Acute Geriatrics and Intensive Care, Ortenau Klinikum, Klosterstrasse 18, Lahr, Germany

[2]Faculty of Medicine, University of Freiburg, Freiburg, Germany

[3]G.E.R.N. Research Center for Tissue Replacement, Regeneration and Neogenesis, Department of Orthopedics and Trauma Surgery, University of Freiburg, Freiburg, Germany

[4]University Emergency Center, Medical Center, University of Freiburg, Freiburg, Germany

[5]Department of Diagnostic and Interventional Radiology, Medical Center, University of Freiburg, Freiburg, Germany

[*]these authors contributed equally

Corresponding Author:
Julian Madrid, MD
Department of Cardiology, Pneumology, Angiology, Acute Geriatrics and Intensive Care, Ortenau Klinikum, Klosterstrasse 18, Lahr, Germany

Abstract

Background: The GPT-4 is a large language model (LLM) trained and fine-tuned on an extensive dataset. After the public release of its predecessor in November 2022, the use of LLMs has seen a significant spike in interest, and a multitude of potential use cases have been proposed. In parallel, however, important limitations have been outlined. Particularly, current LLMs encounter limitations, especially in symbolic representation and accessing contemporary data. The recent version of GPT-4, alongside newly released plugin features, has been introduced to mitigate some of these limitations.

Objective: Before this background, this work aims to investigate the performance of GPT-3.5, GPT-4, GPT-4 with plugins, and GPT-4 with plugins using pretranslated English text on the German medical board examination. Recognizing the critical importance of quantifying uncertainty for LLM applications in medicine, we furthermore assess this ability and develop a new metric termed “confidence accuracy” to evaluate it.

Methods: We used GPT-3.5, GPT-4, GPT-4 with plugins, and GPT-4 with plugins and translation to answer questions from the German medical board examination. Additionally, we conducted an analysis to assess how the models justify their answers, the accuracy of their responses, and the error structure of their answers. Bootstrapping and CIs were used to evaluate the statistical significance of our findings.

Results: This study demonstrated that available GPT models, as LLM examples, exceeded the minimum competency threshold established by the German medical board for medical students to obtain board certification to practice medicine. Moreover, the models could assess the uncertainty in their responses, albeit exhibiting overconfidence. Additionally, this work unraveled certain justification and reasoning structures that emerge when GPT generates answers.

Conclusions: The high performance of GPTs in answering medical questions positions it well for applications in academia and, potentially, clinical practice. Its capability to quantify uncertainty in answers suggests it could be a valuable artificial intelligence agent within the clinical decision-making loop. Nevertheless, significant challenges must be addressed before artificial intelligence agents can be robustly and safely implemented in the medical domain.

(JMIR Med Educ 2025;11:e58375) doi:10.2196/58375

KEYWORDS

medical education; artificial intelligence; generative AI; large language model; LLM; ChatGPT; GPT-4; board licensing examination; professional education; examination; student; experimental; bootstrapping; confidence interval

Introduction

The GPT—recently updated to its fourth iteration (GPT-4)—is a generative and autoregressive large language model (LLM). It is pretrained on a vast corpus of internet text and fine-tuned on a labeled dataset using a transformer architecture [1-3]. GPT generates coherent and contextually appropriate text. It likely discovered a semantic grammar of language (ie, semantic regularities), enabling it to construct semantically and syntactically correct sentences [4,5]. However, GPT does not perform meaningful computations on symbolic representations [4-8]. The Wolfram language, a Turing-complete computational language, in contrast, allows such symbolic representation. GPT and the Wolfram language combined hence cover 2 different aspects of human cognition [4,9,10]. Combining these features, particularly when computation and symbolic representations are needed, represents a significant step toward general artificial intelligence (AI). This combination has already been successfully used to examine contradictions in Einstein Special Theory of Relativity equations [11].

In the light of these technological advances, LLMs show increasing promise in supporting medical training and practice. However, the models must acquire an in-depth and accurate representation of medical knowledge to be used in these sensitive domains. A medical board examination exemplifies these domains well, as it determines the qualification of medical students to obtain their license to practice medicine.

Our primary outcome is the model's ability to achieve the minimum required score for passing the 2 written parts of the German medical licensing examination. This task poses a different challenge to an LLM than medical board examinations in the English language [12,13], as the performance of such models in other languages and in combination with more recent GPT versions and available plugins has not been explored. In the medical field, where mistakes can have severe consequences, assessing the amount of uncertainty is of paramount importance [14]. It is therefore crucial to gain insights into the depth and structure the LLMs have of the medical knowledge representation and where its limitations lie [15]. Hence, our secondary outcomes were the total correct answer rates, the presence of logical justification of the answer, the presence of information internal to the question, the presence of information external to the question, the confidence GPT displays in its answers, the difficulty of the question, information errors, logical errors, reasoning errors, and the correctness of a second try answer when the first answer was wrong. Insights into these 2 dimensions of outcomes can contribute to facilitating a meaningful use of novel LLM technologies in the medical domain.

Methods

Medical Board Examination Dataset

The German medical board examination consists of 3 steps. The first board examination, taken after 2 years of study, primarily covers basic natural sciences. It comprises 320 questions, which students answer over 2 consecutive days. The second board examination takes place after 6 years of study. It likewise consists of 320 medical questions, which students answer over 3 consecutive days. The third board examination, also after 6 years of study, is an oral examination and was, hence, excluded from this study. The German medical board examination takes place biannually, once in spring and once in fall. As a representative sample, we used the medical board examination from spring 2023. We excluded questions the medical board examination committee deemed inconsistent with the medical literature in the regular post examination review of the content. Additionally, we did not include questions displaying images, as GPT models could not analyze them at the time of our analysis. Furthermore, LLMs are not able to analyze images, GPT4vision which became broadly accessible in the second half of 2023 combines computer vision algorithms—which generate a text description of images—and LLMs to analyze this text. All questions and answers were exported from AMBOSS SE, a German medical education content creator and service provider.

GPT Models and Prompt Engineering

We evaluated several GPT models with varying characteristics using OpenAI's web interface. The models tested included GPT-3.5, GPT-4, GPT-4 integrated with the Wolfram, ScholarAI, and Web Request (WeGPT.ai) plugins, and GPT-4 integrated with the Wolfram, ScholarAI, Web Request plugins, and an additional feature for translating German inputs into English. We did not investigate earlier versions of GPT as they demonstrated lower performance in a similar study on the American medical board examination [12].

Creating a precise and adequate context is crucial for generating expected results [16,17]. Thus, we aimed to be as specific as possible, simulating the context of a medical student taking the medical board examination. The prompts hence included the request to answer each respective question with 5 possible answers, where only 1 answer was correct. We asked the models to justify their choices based on the provided patient case information, and to estimate their confidence in the answer's accuracy as a percentage of maximal confidence (ie, 100%). If the selected answer was incorrect, the GPT models were asked to explain their mistake in a second attempt. For the GPT-4 model with plugin integration, we asked the model to use the available plugins (Wolfram, ScholarAI, and Web Request). For the GPT-4 model with plugin integration and English translation, we first asked the model to translate the input into the English language, and then to use the translated text to perform the abovementioned tasks. All used prompts are available in Multimedia Appendix 1.

Model Testing and Outcome Parameters

For each GPT model, we used the appropriate prompt followed by the question and the possible answers. The investigators then analyzed the GPT's answer to assess the defined primary and secondary outcomes, which were either binary or in proportions. In cases of uncertainty, the investigators (JM, TS, and LB) convened to resolve the issue.

First, the correctness of the answer was recorded (binary variable), followed by the presence of logical justification, the presence of information internal to the question, and the

presence of information external to the question (binary variables).

Next, we recorded the model's confidence in its answer (proportion), and the difficulty of the question, derived from the number of students who answered correctly on the AMBOSS platform (proportion).

To enhance our understanding of where GPT models falter, we sought to classify potential errors. As literature on error types is limited, we conducted a formal analysis to determine distinctive error types and established a formal definition. We propose a classification into 3 categories: information error, logical error, and reasoning error.

The GPT response can be formalized as "answer A" is given "link" because of "information B." There are only three possibilities for errors: (1) "answer A" is incorrect because "information B" is incorrect—termed an information error; (2) "answer A" is incorrect while "information B" is correct, but the link between them is incorrect—termed a logical error; (3) "answer A" is incorrect, "information B" is correct, and the link between "answer A" and "information B" is correct—termed a reasoning error (Figure 1). If the answer provided was incorrect, the investigator informed the GPT of its faulty answer, recorded whether it understood its mistake, and provided the correct answer in a second attempt. In the models with integrated plugin use, the active use of plugins was documented for Wolfram, ScholarAI, and Web Requests (binary variables).

Figure 1. Formal definition of error types; we propose a classification into 3 categories: information error, logical error, and reasoning error. The GPT response can be formalized as “answer A” is given “link” because of “information B.” There are only three possibilities for errors: (1) “answer A” is incorrect because “information B” is incorrect—termed an information error; (2) “answer A” is incorrect while “information B” is correct, but the link between them is incorrect—termed a logical error; and (3) “answer A” is incorrect, “information B” is correct, and the link between “answer A” and “information B” is correct—termed a reasoning error.

Data Analysis

Summary statistics were calculated for the outcome variables (Table 1 and Multimedia Appendices 2 and 3). Dichotomous variables were represented by frequency and proportions with 95% CIs, while continuous variables were expressed as mean values with 95% CIs. Uncertainty calculations displayed as 95% CIs were computed via bootstrapping [18].

Table . Characteristics of GPT model answers (N=541).

	GPT-3.5	GPT-4	GPT-4 + plugin	GPT-4 + plugin + translation
Correct answer (proportion±95% CI)	373 (0.69±0.65 to 0.73)	493 (0.91±0.89 to 0.93)	493 (0.91±0.89 to 0.94)	486 (0.9±0.87 to 0.92)
Logical justification (proportion±95% CI)	479 (0.89±0.86 to 0.91)	526 (0.97±0.96 to 0.98)	529 (0.98±0.96 to 0.99)	527 (0.97±0.96 to 0.99)
Question's difficulty mean (±95% CI)	0.288 (0.272 to 0.303)	0.288 (0.272 to 0.303)	0.288 (0.272 to 0.303)	0.288 (0.272 to 0.303)
Error overall (proportion±95% CI)	168 (0.31±0.27 to 0.35)	48 (0.09±0.07 to 0.11)	48 (0.09±0.06 to 0.11)	55 (0.1±0.08 to 0.13)
Presence of internal information (proportion±95% CI)	521 (0.96±0.95 to 0.98)	537 (0.99±0.98 to 1)	537 (0.99±0.98 to 1)	537 (0.99±0.98 to 1)
Presence of external information (proportion±95% CI)	538 (0.99±0.99 to 1)	540 (1±0.99 to 1)	541 (1±1 to 1)	541 (1±1 to 1)
Information error (proportion±95% CI)	37 (0.22±0.16 to 0.29)	5 (0.1±0.02 to 0.19)	5 (0.1±0.02 to 0.2)	7 (0.13±0.05 to 0.22)
Logical error (proportion±95% CI)	61 (0.36±0.29 to 0.43)	18 (0.38±0.25 to 0.52)	12 (0.25±0.125 to 0.375)	19 (0.35±0.22 to 0.47)
Confidence mean (±95% CI)	0.912 (0.904 to 0.918)	0.938 (0.934 to 0.942)	0.919 (0.915 to 0.924)	0.919 (0.915 to 0.923)
Use of plugin Wolfram (proportion±95% CI)	N/A[a]	N/A	50 (0.09±0.07 to 0.12)	47 (0.09±0.06 to 0.11)
Reasoning error (proportion±95% CI)	72 (0.42±0.36 to 0.51)	26 (0.54±0.4 to 0.69)	30 (0.63±0.48 to 0.75)	29 (0.53±0.4 to 0.65)
Correct answer in second try (proportion±95% CI)	90 (0.54±0.46 to 0.61)	32 (0.67±0.52 to 0.79)	36 (0.75±0.63 to 0.88)	33 (0.6±0.47 to 0.73)
Use of plugin ScholarAI (proportion±95% CI)	N/A	N/A	107 (0.2±0.16 to 0.23)	47 (0.09±0.06 to 0.11)
Use of plugin web requests (proportion±95% CI)	N/A	N/A	2 (0.003±0 to 0.01)	25 (0.05±0.03 to 0.06)

[a]N/A: not applicable.

The primary outcome was determined by comparing the performance of the GPT-4 model, integrated with the plugins and the English translation, to the required passing score for the medical board examination, which is 60%. The difference of proportions was calculated with 95% CI using bootstrapping (Multimedia Appendix 4).

Subsequently, secondary outcomes were calculated: the final examination rate for each GPT model was compared to both chance and the required passing score for the medical board examination. The difference of proportions was calculated with 95% CI using bootstrapping (Multimedia Appendix 4).

The proportions of logical justification within the answer, information internal to the answer, and information external to the answer were compared between correct and incorrect responses. The difference of proportions was calculated with 95% CI using bootstrapping (Table 2 and Multimedia Appendix 5).

Table . Analysis of plugin-integrated GPT-4 model answers.

		All correct answers (n=493)	All incorrect answers (n=48)	Difference in proportions or Cohen *d* or Pearson *r* (±95% CI)	Confidence accuracy (±95% CI)
Comparison of GPT models justifications between correct and incorrect answers					
GPT-4 + plugin (N=541)					
	Logical justification (proportion ±95% CI)	493 (1±1 to 1)	36 (0.75±0.63 to 0.88)	0.25 (0.13 to 0.38)[a]	—[b]
	Internal information (proportion ±95% CI)	489 (0.99±0.983 to 998)	48 (1±1 to 1)	0 (-0.01 to 0)[a]	—
Comparison of GPT models justifications between correct and incorrect answers					
	External information (proportion ±95% CI)	493 (1±1 to 1)	48 (1±1 to 1)	0 (0 to 0)[a]	—
Confidence of GPT models compared between correct and incorrect answers					
GPT-4 + plugin (N=541)					
	Confidence mean (±95% CI)	0.923 (0.918 to 0.928)	0.886 (0.87 to 0.901)	-0.69 (-0.99 to -0.39)[c]	0.037 (0.021 to 0.053)
Comparison of question's difficulty of GPT models between correct and incorrect answers					
GPT-4 + plugin (N=541)					
	Question's difficulty mean (±95% CI)	0.279 (0.263 to 0.295)	0.379 (0.327 to 0.438)	0.57 (0.27 to 0.86)[c]	—
Correlation of confidence and question's difficulty for all answers					
GPT-4 + plugin (N=541)				-0.0874 (-0.176 to 0.004)[d]	
	Confidence mean (±95% CI)	0.920 (0.916 to 0.924)	—		—
	Question's difficulty mean (±95% CI)	—	0.288 (0.273 to 0.304)		—
Comparison of correct answers between GPT models (N=541)					
GPT-4 + plugin vs GPT-3.5					
	Correct answer rate (proportion ±95% CI)	373 (0.69±0.65 to 0.73)	493 (0.91±0.89 to 0.94)	0.22 (0.18 to 0.27)[a]	—
GPT-4 + plugin vs GPT-4					
	Correct answer rate (proportion ±95% CI)	493 (0.91±0.89 to 0.94)	493 (0.91±0.89 to 0.94)	0 (-0.03 to 0.03)[a]	—
GPT-4 + plugin vs GPT-4 + plugin + translation					
	Correct answer rate (proportion ±95% CI)	493 (0.91±0.89 to 0.94)	486 (0.9±0.87 to 0.92)	-0.01 (-0.05 to 0.02)[a]	—

[a]Difference in proportions.

[b]Not available.

[c]Cohen *d*.

[d]Pearson *r*.

The model's confidence in its answers was compared between correct and incorrect responses. Additionally, the relationship between the model's confidence in its answers and the difficulty of the question was assessed. Cohen *d* values and 95% CI were computed using a linear regression model and bootstrapping (Table 2 and Multimedia Appendices 6 and 7).

To evaluate the accuracy of the model's confidence in its answers, we developed a parameter termed confidence accuracy (CA). It is defined as follows:

CA = (confidence of correct answers in percentage – confidence of incorrect answers in percentage)/100

Consequently, this parameter can take values from −1 to 1, where 1 accurately reflects the model's uncertainty, 0 indicates no ability to quantify uncertainty, and −1 suggests incorrect quantification.

The difficulty of the question was assessed using real correct response proportions from students available on the AMBOSS platform. The difficulty was assessed as follows:

Difficulty=1 – correct answer proportion

Then, the difficulty of the question was compared between correct and incorrect answers, with Cohen *d* calculated using a linear regression model (Table 2 and Multimedia Appendix 7).

Furthermore, we compared the proportion of correct answers between models (Table 2 and Multimedia Appendix 8).

We compared the proportion of correct answers in the GPT-4 models with the proportion of correct answers in the answers where a plugin has been used. We compared the proportion of plugin usage in GPT models with German and English input. We compared the confidence of the model when using plugins to the confidence of the model overall. We compared the proportion of correct answers when averaging the 4 different models to each model in particular (Multimedia Appendix 9).

In instances where questions were accompanied by images, GPT models sometimes responded by describing the image, although the models could not access the respective images. This phenomenon is known as a type of hallucination [19]. Therefore, we compared the proportion of hallucinations present in each model when answering questions, including image questions. We calculated the proportion of correct answers for each model when keeping the questions with pictures (Multimedia Appendix 9).

We compared the different error proportions between different models. We compared the proportion of logical errors when using the Wolfram plugin to the proportion of errors when using the entire model. We compared correct second-try answers between different models (Multimedia Appendix 9).

The 95% CIs were calculated using bootstrapping. Where necessary, parametric assumptions were tested using quantile-quantile plots for normality and Levene tests for the homogeneity of variances. The independence of question answers was assumed. All statistical analyses were performed in RStudio (version 2023.06.0+421). The significance level for all tests was set a priori at 95% CI.

Results

All tests were performed on the 541 questions of the German medical board examination from spring 2023. Sub analyses were performed on other subgroups, the respective sample sizes are indicated in the appropriate tables. All results for GPT-3.5, GPT-4, GPT-4 + plugin (GPT4P), and GPT-4 + plugin + translation (GPT4PT) are listed in full detail in the tables and the supplementary materials. To ensure legibility, only relevant results are addressed in the results section.

Descriptive statistics with CIs for the first board examination, second board examination, and the overall examination are displayed in Table 1 and Multimedia Appendices 2 and 3.

All models performed significantly better than chance. Furthermore, all GPT models were significantly better than the required proportion to pass the final medical board examination.

All GPT models had a significantly higher proportion of providing a logical justification for correct answers compared to incorrect answers (Table 2 and Multimedia Appendix 5). Yet, there was no statistical significance for the proportion of used internal information for correct and incorrect answers (Table 2 and Multimedia Appendix 5). Similarly, there was no statistical significance for the proportion of used external information for correct and incorrect answers (Table 2 and Multimedia Appendix 5).

Although generally high for both incorrect and correct answers, models had a confidence mean which was significantly higher for correct answers than incorrect answers (Table 2 and Multimedia Appendix 6). This is reflected in CA values significantly different from zero: GPT-3.5 (0.028, 95% CI 0.011 to 0.048), GPT-4 (0.041, 95% CI 0.023 to 0.062), GPT4P (0.037, 95% CI 0.021 to 0.053), and GPT4PT (0.043, 95% CI 0.028 to 0.059).

From all models, only GPT4P made significantly more reasoning errors than logical errors (0.37, 95% CI 0.125 to 0.60). All models made significantly more reasoning errors than information errors: GPT-3.5 (0.21, 95% CI 0.11 to 0.30), GPT-4 (0.44, 95% CI 0.27 to 0.60), GPT4P (0.52, 95% CI 0.31 to 0.71), and GPT4PT (0.40, 95% CI 0.20 to 0.58). All models but GPT4P made significantly more logical errors than information errors: GPT-3.5 (0.14, 95% CI 0.029 to 0.26), GPT-4 (0.27, 95% CI 0.10 to 0.44), and GPT4PT (0.22, 95% CI 0.05 to 0.38). GPT-4 (0.12, 95% CI 0.05 to 0.22) and GPT4P (0.12, 95% CI 0.02 to 0.22) made significantly less information errors than GPT3.5.

The GPT4-based models all performed better than the GPT 3.5 model in providing correct answers as reflected in the difference of correct answer proportions (Table 2 and Multimedia Appendix 8). However, no GPT4-based model was better than another GPT4-based model, as reflected in the difference of correct answer proportions (Table 2 and Multimedia Appendix 8).

Discussion

Primary Outcome

All GPT models assessed performed above the minimum required score of 60%. The GPT-4 models performed particularly well, outperforming most students in the given examinations. Specifically, for the first board examination, all GPT-4 models performed better than 98.6% of students. For

the second board examination, they surpassed 95.8% of students, as detailed in the records of the examining body [20].

In general, there was a significant gap between GPT-3.5 and the GPT-4 models. The more recent models, with substantially more parameters and the capacity to remember longer prompts, appear to increase the accuracy of responses. However, we observed no additional benefit when GPT-4 models were paired with plugins.

The use of plugins did not yield a higher proportion of correct answers than the standard model. It is possible that GPT-4 already achieves a very high rate of accuracy, resulting in a ceiling effect. Hence, the addition of plugins may not offer a significant advantage for the questions prompted.

During our study, we noted that the Wolfram plugin was frequently used for more complex calculations. Yet, in the context of clinically applicable questions, complex mathematical procedures are typically not required and the use of symbolic language is usually not required. Thus, using the Wolfram Alpha plugin is likely more beneficial for questions that involve extensive computations or advanced mathematical problems requiring symbolic representations. The ScholarAI plugin was activated for complex informational queries, but the resulting papers were not consistently useful. Surprisingly, the Internet Access plugin (WeGPT.ai) was the least used. This may be because answering medical questions typically demands expert-level knowledge, and general internet searches do not provide sufficiently specific information. Moreover, since the model has been trained on a vast amount of internet data, it likely already encompasses the knowledge available from the world wide web within its parameters.

We speculated that posing questions in German might hinder the model's access to the broader body of knowledge available in English. However, this was not the case; the GPT model equipped with translation capabilities did not outperform the GPT-4 models without translation features. The GPT model likely abstracts high-level concepts and is not impeded by the language of the queries. This aligns with the LLMs' transformer architecture, which accesses higher-level concepts prior to translating text into another language [21].

Interestingly, the GPT-4 model with translation invoked plugins less frequently than the model without translation. We hypothesize that plugin calls occur at a lower level in the neural network, making them less necessary in English due to the larger available language corpus. In German, the model might need to delve deeper into the latent representation of concepts not tied to a specific language. However, this remains speculative and warrants further research.

Secondary Outcomes

While all models provided a very high proportion of logical justification for correct answers, it was significantly less extensive for incorrect answers. However, upon further analysis, we did not detect a significant difference in the proportion of internal information from the question in the answer or in the use of external information not contained in the question between correct and incorrect answers. One study already assessed the presence of logical justification in answers to United States Medical Licensing Examination questions, where all answers exhibited logical justification regardless of their accuracy [12]. Hence, this metric could not be used as a discriminator for correctness.

We were unable to demonstrate a significant correlation between the model's confidence in an answer and the difficulty level of the question for humans. This suggests that the model's interpretation of question difficulty differs from that of humans. However, as with humans, the model showed improved performance on easier questions compared to more challenging ones. Thus, it appears that the representation of question difficulty is distinct between LLMs and humans.

Conceptual Implications

Use for Medical Education

This performance suggests that LLMs such as GPT could assume a greater role in medical education, as their integration could significantly change the conventional approach to medical education, which has traditionally emphasized the acquisition and maintenance of medical knowledge. The emergence of AI agents with superior information retention abilities, however, prompts a reevaluation of our educational focus. In this light, teaching methodologies could shift toward navigating and structuring available information with respective AI agents. The approach could hence shift from retaining information to learning how to efficiently access information and deeply understand these systems, along with their benefits and drawbacks.

Use in Clinical Practice

The utility of LLMs is not limited to educational settings but also extends to clinical practice. Although LLMs may not be as effective in highly specialized tasks where dedicated machine learning algorithms excel—for instance, XGBoost in identifying pulmonary embolisms [22-24] — LLMs are highly proficient in text processing and information integration from diverse algorithms [25]. This positions them as intelligent medical assistants, capable of transforming complex data into narratives that are comprehensible in a human context. Currently, clinicians have a limited understanding of AI agents and their functions. Clinicians must, therefore, gain a thorough understanding of how various AI agents function, including their strengths and weaknesses.

With insufficient knowledge on the principles of LLM-based assistants, clinicians are at risk of blindly following such assistant's guidance without fully understanding its operations [26,27]. Due to the inherent complexity of LLMs, which often function as a black box, we can only partially monitor their operations at varying levels of complexity and behavior [26]. Given the marginal uncertainty intrinsic to such complex models, the AI agent should not supplant clinicians in decision-making, but rather provide additional informed perspectives.

To serve as a useful assistant, however, the assessment of uncertainty for any output provided by such is crucial. The key attribute enabling this evaluation is the ability to quantify uncertainty, a trait humans are presumed to possess [14]. For

LLM-based assistants to provide a comparable estimate, a standardized measure is needed to gauge the confidence in an AI agent's output. For binary outcomes such as healthy or diseased, metrics such as specificity, sensitivity, and area under the curve are effective. For more complex queries with multiple potential answers—as managed by LLMs—traditional measures such as sensitivity and specificity are inadequate. We therefore developed a new metric called "confidence accuracy" (CA) which correlates the confidence assigned to an answer with its empirical accuracy. This allows for the quantification of uncertainty, crucial for clinical decision-making. Although our work showed that all GPT models have the ability to quantify uncertainty, the expression in percentage does not seem to reflect the confidence for any specific decision (ie, the models were overall largely overconfident). Although statistically different from zero, CA values were consistently close to zero. New LLM methodologies aim to enhance this by incorporating uncertainty estimation [28]. Future AI agents should be fine-tuned using the CA metric in order to improve uncertainty quantification, a critical objective for implementing AI as a supportive tool for physicians in clinical environments.

Identified Errors

We observed that GPT models commit different types of errors, particularly reasoning errors. Reasoning errors typically occur in situations where multiple options are correct, but one is more critical than the other. GPT models over proportionately make reasoning errors likely because this skill is acquired through human experience and is challenging to learn from text-based web sources. The second most common error type in GPT models was logical errors. Since LLMs use a statistical approach to reconstruct human-written text, we anticipated difficulties with logic and mathematics, which require formal symbolic representation [4-8]. We hypothesized that the Wolfram plugin, using the Wolfram language, would mitigate these challenges. Yet, using the Wolfram plugin did not reduce the number of logical errors. Finally, fewer information errors were observed compared to other error types across all GPT models. This likely reflects the strength of these LLMs, which have assimilated a vast corpus of knowledge. In addition to the 3 error types derived from the informational and logical structure of GPT's answers, there are 2 sources of bias that arise prior to answer generation. First, due to the stochastic nature of token generation, there is likely a stochastic bias inherent in all GPT responses. Second, due to in-context generation conditioned by the prompting strategy, a systematic bias probably occurs as well. We attempted to mitigate the stochastic bias by averaging the results from all models and selecting the most common outcome. However, the performance of such averaged models did not surpass that of the GPT-4 models.

To assess whether the GPT models could recognize and correct their own mistakes, we prompted them to attempt another answer after providing incorrect responses. In most instances, the model would acknowledge the mistake and provide the correct answer along with a new explanation. This phenomenon could likely be attributed to the differing mechanics of forward and backward reasoning in LLMs. With forward reasoning, the LLM calculates the probability of the next token without a specific reasoning goal [29]. In contrast, backward reasoning enables the LLM to better contextualize the information. It is crucial to note, however, that we did not request the model to immediately reassess the answer; instead, we informed it of the answer's incorrectness before asking for a reevaluation [29]. Future studies could further investigate the model's ability to self-correct without prior notification of its errors.

In instances where questions were accompanied by images (ie, the model did not have access to the images), GPT models, particularly GPT-3.5, often responded by describing the image that the model had not actually seen. This unexpected information error, known as a hallucination [19], persisted in the GPT-4 models, albeit at a significantly reduced frequency compared to GPT-3.5. Nevertheless, the propensity for overconfidence in entirely fabricated information remains a challenge for the latest generation of LLMs and is a phenomenon not fully understood [30].

Limitations

Technological Limitations of LLMs

Although the results were impressive with GPT outperforming most students in the German medical board examination, it is crucial to remember that these models still possess significant limitations. At the time of our data collection, GPT-4 was incapable of interpreting medical images, such as chest x-rays or histological samples. This is a considerable drawback, given that medical information is inherently multimodal, and the ability to integrate multimodal data will be essential for the adoption of such models in academic and clinical settings. It is anticipated that future GPT iterations and other LLMs will be fully multimodal, which necessitates additional research to evaluate their performance across a more diverse array of questions.

A second concern relates to the stochastic nature of token generation, meaning that answers may vary slightly when questions are posed multiple times [31].

A third concern pertains to the prompt sensitivity of LLMs. This trait can be advantageous as it allows the incorporation of context into the generation of meaningful output and may contribute to the models' Bayesian characteristics [32]. However, prompt sensitivity also increases the risk of systematic errors with repetitive use of the same prompt. Prompt engineering is a discipline that emerged in trying to minimize systematic errors [33,34].

Within the extensive volume of data available online, there are significant risks of bias. Given that LLMs are trained on vast datasets, there is an inherent risk of adopting biases from the underlying data structures. However, fine-tuning through supervised learning on labeled data can help mitigate these risks [35,36].

Limitations of the Use of LLMs in a Medical Context

Despite the seemingly immediate promise of using LLMs in both educational and clinical contexts, the current ethical and regulatory environment needs to be considered to advance the use of these novel technologies safely.

As the representation of medical information of an LLM must not be confused with medical knowledge from a medical professional, it remains crucial to enable students and medical professionals alike to identify LLM-generated outputs as such in order to interpret them very carefully. Different to, for example, a senior medical colleague providing guidance for a clinical decision, an LLM-generated output is neither based on clinical knowledge, nor experience. The risk of such confusion has been described as anthropomorphic projection and efforts for advancing these novel technologies in the medical field need to simultaneously foster the awareness of such phenomena. This differentiation resonates with the provisions of the European Union (EU) on a risk-based assessment approach [37] and, more recently, with the Bletchley Declaration [38]. The latter emphasizes the risks at the "frontier" of AI, at which we operate with the presented project.

While the concerns discussed in the context of medical education—and, more widely, training—are mainly within the realm of AI ethics, more specific limitations apply to the clinical use of these technologies. At the time of our analysis, no commercially available LLM in the EU—including the GPT versions assessed in this work—have an assigned intended medical use, a basic regulatory prerequisite for their use in a clinical context. Without such intended medical use, the Medical Device Regulation (MDR), the regulatory framework for medical devices in the EU, is not applicable. Hence, such a device would not be a medical device in the regulatory sense and could, therefore, not be used in a clinical context without irresponsible safety and liability risks. While it is not the user (eg, researchers or clinicians), but the manufacturer (eg, OpenAI for the ChatGPT models) who assigns an intended medical use—which itself comes with further regulatory requirements—the clinical use of the currently available and mostly all-purpose LLMs remains challenging.

Yet, even developing an LLM with an intended medical use and fulfilling all adjacent regulatory requirements would—as of now—not necessarily resolve the challenge centering around the clinical use of such program, as a key requisite for software as a medical device outlined in the MDR ("devices that incorporate electronic programmable systems, including software, or software that are devices in themselves, shall be designed to ensure repeatability, reliability and performance in line with their intended use." MDR Annex I, Rule 17.1 [39]) is currently considered to be violated, although this question remains subject to debate.

However, the rapid development of technological advances and the concurrent establishment of respective regulations should not be perceived as a "race to get to grips with AI" [40], but should be viewed as a co-evolution to eventually yield the best population-wide benefit from these technological advances. In this light, the emphasis of a "pro-innovation and proportionate governance," as proposed in the Bletchley Declaration, is equally as crucial as the implementation of regulatory frameworks.

Limitations of This Study

Our study has several limitations. We used a specific German medical board examination as a sample to represent the general distribution of medical questions. While it is acknowledged that questions evolve over time and may introduce bias, the objective of the medical board examination is to maintain a consistent level of difficulty, reflecting the minimum required knowledge to attain board approval for medical practice. The distribution of student grades has remained relatively stable over time, leading us to believe that this potential bias is minimal. In the model with translation, we used GPT to translate the questions before applying them to the model. Although we did not observe any, it is possible that translation errors occurred, potentially acting as a confounder in this study. In the context of the medical board examination, multiple-choice questions are posed to elicit clear answers that can be quantitatively assessed. By contrast, in a clinical setting, questions tend to be open-ended, which introduces a different dynamic. Nevertheless, we asked the model to justify its answers to glean insight into its computational process, thus rendering the questions more comparable to open-ended inquiries.

Conclusion

The performance of GPT models in the German medical board examination have surpassed both the passing threshold and the performance of most students. While GPT appears to possess a latent representation of uncertainty, it currently exhibits a significant degree of overconfidence. The introduced metric of CA could facilitate the appropriate measurement and fine-tuning of models to improve this aspect. However, there are numerous limitations that clinicians should be aware of. Challenges such as hallucinations, the stochastic nature of token generation, and prompt sensitivity are highlighted, indicating areas for further research and development. Further, we see the remaining open questions regarding the ethical and regulatory use of LLMs in the educational and clinical context, which need to be addressed on a policy level.

Authors' Contributions

JM participated in the conceptualization, data acquisition, data curation, formal analysis, investigation, methodology, project administration, software, validation, visualization, writing of the original draft, and review and editing of the writing, and should be considered the first author. PD, MS, BR, FPH, and HJB participated in the methodology and review editing and should be considered as second authors. LB and TS participated in the conceptualization, data acquisition, formal analysis, investigation, methodology, validation, review and editing of the writing, and should be considered last authors. Correspondence should be addressed to JM and LB.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Prompting strategies for different GPT models.
[DOCX File, 14 KB - mededu_v11i1e58375_app1.docx]

Multimedia Appendix 2
Question's difficulty and error structure of GPT model answers.
[XLSX File, 10 KB - mededu_v11i1e58375_app2.xlsx]

Multimedia Appendix 3
Correct answers of GPT models compared with required and random scores.
[XLSX File, 9 KB - mededu_v11i1e58375_app3.xlsx]

Multimedia Appendix 4
Comparison of correct answers between GPT models.
[XLSX File, 9 KB - mededu_v11i1e58375_app4.xlsx]

Multimedia Appendix 5
Supplementary analysis of GPT models answers (statistically significant results are highlighted in blue and statistically nonsignificant results are highlighted in brown).
[XLSX File, 12 KB - mededu_v11i1e58375_app5.xlsx]

Multimedia Appendix 6
Confidence of GPT models compared between correct and incorrect answers.
[XLSX File, 9 KB - mededu_v11i1e58375_app6.xlsx]

Multimedia Appendix 7
Relationship between question's difficulty, performance, and confidence of GPT model answers.
[XLSX File, 10 KB - mededu_v11i1e58375_app7.xlsx]

Multimedia Appendix 8
Comparison of GPT models justifications between correct and incorrect answers.
[XLSX File, 10 KB - mededu_v11i1e58375_app8.xlsx]

Multimedia Appendix 9
Performance, information content, confidence, and plugin usage of GPT model answers.
[XLSX File, 10 KB - mededu_v11i1e58375_app9.xlsx]

References

1. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. arXiv. Preprint posted online on Jun 12, 2017. [doi: 10.48550/arXiv.1706.03762]
2. Korngiebel DM, Mooney SD. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. NPJ Digit Med 2021 Jun 3;4(1):93. [doi: 10.1038/s41746-021-00464-x] [Medline: 34083689]
3. Bubeck S, et al. Sparks of artificial general intelligence: early experiments with GPT-4. arXiv. Preprint posted online on Mar 22, 2023. [doi: 10.48550/arXiv.2303.12712]
4. Wolfram S. What Is ChatGPT Doing...and Why Does It Work?: Wolfram Media, Inc; 2023.
5. Traylor A, Feiman R, Pavlick E. AND does not mean OR: using formal languages to study language models' representations. Presented at: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2; Aug 1-6, 2021. [doi: 10.18653/v1/2021.acl-short.21]
6. Misra K, Rayz J, Ettinger A. COMPS: conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models. Presented at: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics; May 2-6, 2023; Dubrovnik, Croatia URL: https://aclanthology.org/2023.eacl-main [accessed 2025-02-28] [doi: 10.18653/v1/2023.eacl-main.213]

7. Kim N, Linzen T. COGS: a compositional generalization challenge based on semantic interpretation. Presented at: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); Nov 16-20, 2020 URL: https://www.aclweb.org/anthology/2020.emnlp-main [accessed 2025-02-28] [doi: 10.18653/v1/2020.emnlp-main.731]
8. Ettinger A. What BERT is not: lessons from a new suite of psycholinguistic diagnostics for language models. Trans Assoc Comput Linguist 2020 Dec;8:34-48. [doi: 10.1162/tacl_a_00298]
9. Goertzel B. Generative AI vs AGI: the cognitive strengths and weaknesses of modern llms. arXiv. Preprint posted online on Sep 19, 2023. [doi: 10.48550/arXiv.2309.10371]
10. Vzorin G, Bukinich AM, Sedykh A, Vetrova I, Sergienko E. Emotional intelligence of GPT-4 large language model. PsyArXiv Preprints. Preprint posted online on Oct 20, 2023. [doi: 10.31234/osf.io/b6vys]
11. Bryant S. Assessing GPT-4's role as a co-collaborator in scientific research: a case study analyzing Einstein's special theory of relativity. Research Square. Preprint posted online on Apr 12, 2023. [doi: 10.21203/rs.3.rs-2808494/v1]
12. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? the implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
13. Nori H, King N, McKinney SM. Capabilities of GPT-4 on medical challenge problems. arXiv. Preprint posted online on Mar 20, 2023. [doi: 10.48550/arXiv.2303.13375]
14. Cosmides L, Tooby J. Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. Cognition 1996 Jan;58(1):1-73. [doi: 10.1016/0010-0277(95)00664-8]
15. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023 Mar 30;388(13):1233-1239. [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
16. Wang J, et al. Prompt engineering for healthcare: methodologies and applications. arXiv. Preprint posted online on Apr 28, 2023. [doi: 10.48550/arXiv.2304.14670]
17. Liu J, Shen D, Zhang Y, Dolan B, Carin L, Chen W. What makes good in-context examples for GPT-3? Presented at: Proceedings of Deep Learning Inside Out (DeeLIO 2022); May 27, 2022; Dublin, Ireland URL: https://aclanthology.org/2022.deelio-1 [accessed 2025-02-28] [doi: 10.18653/v1/2022.deelio-1.10]
18. Haukoos JS, Lewis RJ. Advanced statistics: bootstrapping confidence intervals for statistics with "difficult" distributions. Acad Emerg Med 2005 Apr;12(4):360-365. [doi: 10.1197/j.aem.2004.11.018] [Medline: 15805329]
19. Wang J, et al. Evaluation and analysis of hallucination in large vision-language models. arXiv. Preprint posted online on Aug 29, 2023. [doi: 10.48550/ARXIV.2308.15126]
20. Archiv medizin. IMPP. URL: https://www.impp.de/pruefungen/medizin/archiv-medizin.html [accessed 2025-02-28]
21. Belinkov Y, Glass J. Analysis methods in neural language processing: a survey. Trans Assoc Comput Linguist 2019 Apr 1;7:49-72. [doi: 10.1162/tacl_a_00254]
22. Ryan L, Maharjan J, Mataraso S, et al. Predicting pulmonary embolism among hospitalized patients with machine learning algorithms. Pulm Circ 2022 Jan;12(1):e12013. [doi: 10.1002/pul2.12013] [Medline: 35506114]
23. Dua R, Wallace GR, Chotso T, Raj VFD. Classifying pulmonary embolism cases in chest CT scans using VGG16 and xgboost. In: Lecture Notes on Data Engineering and Communications Technologies: Springer; 2023, Vol. 131:273-292. [doi: 10.1007/978-981-19-1844-5_22]
24. Ding R, Ding Y, Zheng D, et al. Machine learning-based screening of risk factors and prediction of deep vein thrombosis and pulmonary embolism after hip arthroplasty. Clin Appl Thromb Hemost 2023;29:10760296231186145. [doi: 10.1177/10760296231186145] [Medline: 37394825]
25. Wu Q, et al. AutoGen: enabling next-gen LLM applications via multi-agent conversation. arXiv. Preprint posted online on Aug 16, 2023. [doi: 10.48550/arXiv.2308.08155]
26. Verdicchio M, Perin A. When doctors and AI interact: on human responsibility for artificial risks. Philos Technol 2022;35(1):11. [doi: 10.1007/s13347-022-00506-6] [Medline: 35223383]
27. Xu J. Overtrust of robots in high-risk scenarios. Presented at: AIES '18; Feb 2-3, 2018; New Orleans, LA, United States URL: https://dl.acm.org/doi/proceedings/10.1145/3278721 [accessed 2025-02-28] [doi: 10.1145/3278721.3278786]
28. Sankararaman KA, Wang S, Fang H. BayesFormer: transformer with uncertainty estimation. arXiv. Preprint posted online on Jun 2, 2022. [doi: 10.48550/arXiv.2206.00826]
29. Jiang W, Shi H, Yu L. Forward-backward reasoning in large language models for mathematical verification. Presented at: Findings of the Association for Computational Linguistics ACL 2024; Aug 11-16, 2024; Bangkok, Thailand URL: https://aclanthology.org/2024.findings-acl [accessed 2025-02-28] [doi: 10.18653/v1/2024.findings-acl.397]
30. Yao JY, Ning KP, Liu ZH, Ning MN, Yuan L. LLM lies: hallucinations are not bugs, but features as adversarial examples. arXiv. Preprint posted online on Oct 2, 2023. [doi: 10.48550/arXiv.2310.01469]
31. Bender EM, Gebru T, Mcmillan-Major A, Shmitchell S, Shmitchell SG. On the dangers of stochastic parrots: can language models be too big? Presented at: FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency; Mar 3-10, 2021. [doi: 10.1145/3442188.3445922]
32. Xie SM, Raghunathan A, Liang P, Ma T. An explanation of in-context learning as implicit bayesian inference. arXiv. Preprint posted online on Nov 3, 2021. [doi: 10.48550/arXiv.2111.02080]

33. Zhao TZ, Wallace E, Feng S, Klein D, Singh S. Calibrate before use: improving few-shot performance of language models. arXiv. Preprint posted online on Jul 1, 2021. [doi: 10.48550/arXiv.2102.09690]
34. Zheng C, Zhou H, Meng F, Zhou J, Huang M. Large language models are not robust multiple choice selectors. arXiv. Preprint posted online on Sep 7, 2023. [doi: 10.48550/ARXIV.2309.03882]
35. Jin X, Barbieri F, Kennedy B, Mostafazadeh Davani A, Neves L, Ren X. On transferability of bias mitigation effects in language model fine-tuning. Presented at: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics; Jun 6-11, 2021. [doi: 10.18653/v1/2021.naacl-main.296]
36. Chu T, Song Z, Yang C. Fine-tune language models to approximate unbiased in-context learning. arXiv. Preprint posted online on Oct 5, 2023. [doi: 10.48550/arXiv.2310.03331]
37. Regulation of the european parliament. European Commission. URL: https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF [accessed 2025-01-15]
38. The Bletchley Declaration. GOV.UK. URL: https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023 [accessed 2025-02-28]
39. ANNEX I medical device regulation. Medical Device Regulation. URL: https://www.medical-device-regulation.eu/2019/07/23/annex-i-general-safety-and-performance-requirements/ [accessed 2025-02-28]
40. Regulating the machine. POLITICO. URL: https://www.politico.eu/article/regulate-europe-race-artificial-intelligence-ai-drugs-medicines/ [accessed 2025-02-28]

Abbreviations

AI: artificial intelligence
CA: confidence accuracy
EU: European Union
GPT4P: GPT-4 + plugin
GPT4PT: GPT-4 + plugin + translation
LLM: large language model
MDR: Medical Device Regulation

Edited by B Lesselroth; submitted 14.03.24; peer-reviewed by G Gill, T Nakao; revised version received 29.07.24; accepted 23.11.24; published 21.03.25.

Please cite as:
Madrid J, Diehl P, Selig M, Rolauffs B, Hans FP, Busch HJ, Scheef T, Benning L
Performance of Plug-In Augmented ChatGPT and Its Ability to Quantify Uncertainty: Simulation Study on the German Medical Board Examination
JMIR Med Educ 2025;11:e58375
URL: https://mededu.jmir.org/2025/1/e58375
doi:10.2196/58375

Visual Learning in Electrocardiography Training for Medical Residents: Comparative Intervention Study

Heng-You Sung[1], MD; Feng-Ching Liao[1], MD; Shu-I Lin[1], MD; Han-En Cheng[2], PharmD; Chun-Wei Lee[1], PhD

[1]Cardiovascular Division, Department of Internal Medicine, Mackay Memorial Hospital, No. 92, Sec. 2, Zhongshan N. Rd., Zhongshan Dist., Taipei, Taiwan

[2]School of Pharmacy, National Yang Ming Chiao Tung University, Taipei, Taiwan

Corresponding Author:
Chun-Wei Lee, PhD
Cardiovascular Division, Department of Internal Medicine, Mackay Memorial Hospital, No. 92, Sec. 2, Zhongshan N. Rd., Zhongshan Dist., Taipei, Taiwan

Abstract

Background: Although electrocardiogram (ECG) interpretation training begins early in medical school, achieving accuracy in interpretation of 12-lead ECG remains a persistent challenge. We conducted a pilot educational program to compare the effectiveness of a conventional didactic lecture, self-drawing, and self-drawing following a flipped classroom (SDFC) approach.

Objectives:: This study aimed to evaluate the effectiveness of three instructional strategies—traditional didactic lecture, self-drawing, and SDFC approach—in improving ECG interpretation skills among first-year postgraduate (PGY-I) medical residents.

Methods: This study was conducted among postgraduate-year PGY-I residents at MacKay Memorial Hospital over 3 years. The study enrolled 76 PGY-I residents, who were randomized into three groups: conventional control (group 1), self-drawing (group 2), and SDFC (group 3). All participants were provided with the same learning material and didactic lectures. Knowledge evaluation was performed using pre- and posttests, which were administered using questionnaires.

Results: The groups involving self-drawing, both combined with and without a flipped classroom approach, demonstrated better performance on the written summative examination. These findings highlight the benefits of self-drawing in integrating theoretical knowledge with practical approaches to ECG interpretation.

Conclusion: Our study demonstrated promising effects of self-drawing on the recognition of ECG patterns, which could address the inadequacies of traditional classroom teaching. It can be incorporated into routine teaching after validation in a larger cohort.

(JMIR Med Educ 2025;11:e73328) doi:10.2196/73328

KEYWORDS

flipped classroom; electrocardiogram learning; lecture; postgraduate education; junior residents

Introduction

Electrocardiogram (ECG) interpretation remains one of the most important diagnostic tools in health care for screening, early diagnosis, and treatment of cardiovascular diseases such as arrythmias and acute coronary syndrome [1]. ECG interpretation is a cognitive skill that requires considerable time and effort to master [2].

Inaccurate ECG interpretation can lead to missed critical diagnoses, resulting to failure in providing life-saving treatments (eg, complete atrioventricular block, ventricular arrhythmias, and acute coronary syndrome) or unnecessary medical interventions [3,4]. Bogun et al [3] reported that misdiagnosing atrial fibrillation in 4% of patients led to inappropriate treatments such as the unwarranted use of anticoagulant or antiarrhythmic therapies.

Teaching ECG interpretation poses significant challenges due to its complexity, often leading to reluctance among students while engaging with the subject. Although a variety of pedagogical materials (eg, textbooks, research articles, quizzes, videos) are available, and teaching methods vary across medical schools and countries (eg, self-directed learning, workshop-based training, lecture-based instruction), the most effective approach remains unclear [4-7]. It is important to note that self-directed learning refers to student-led engagement with content, while self-drawing is a structured activity aimed at enhancing visual-spatial learning through active reconstruction of ECG patterns. Flipped classroom, another active learning strategy, involves learners reviewing instructional content (eg, videos or readings) before class, allowing in-class time for discussion, problem-solving, and interactive application.

Research indicates that self-directed learning correlates with lower interpretation competence, whereas summative assessment

is associated with improved interpretation competence compared to formative assessment [4]. Studies have suggested that self-drawing enhances memory retention, particularly by reinforcing the characteristic waveform patterns of specific cardiac conditions [5]. Additionally, flipped classroom models have been reported to increase student engagement and facilitate active learning in ECG education [6-8].

This study aimed to evaluate the effectiveness of different ECG teaching strategies, including lecture-based learning (LBL) alone, LBL combined with self-drawing, and LBL combined with self-drawing and a flipped classroom model. By assessing the impact of these approaches, this study aimed to identify effective pedagogical strategies for improving ECG interpretation skills among medical trainees.

Methods

Study Design and Participants

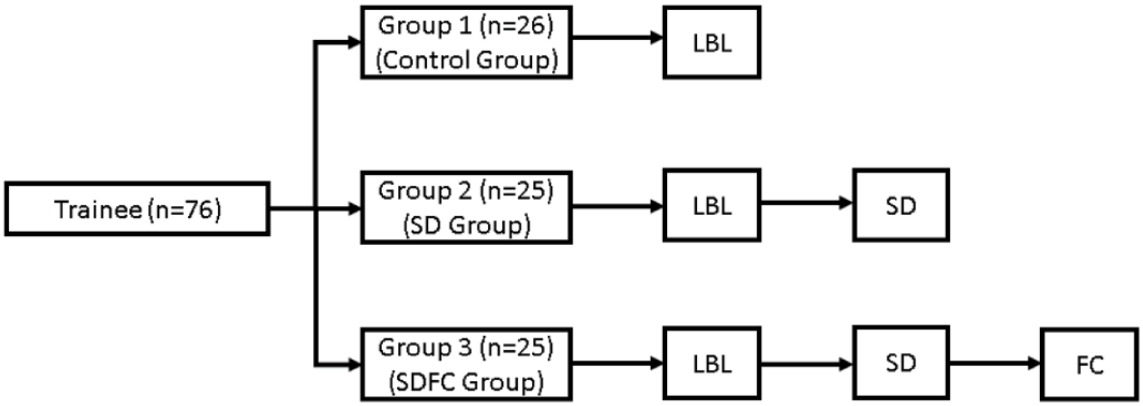
This retrospective study was conducted between September 2020 and June 2023 and included three consecutive cohorts of first-year postgraduate residents at MacKay Memorial Hospital. Participants were divided into three instructional groups based on their training year, with each group drawn from a distinct cohort:

- Group 1 (LBL only) included residents from the 2020 - 2021 academic year (September 2020 to June 2021)
- Group 2 (lecture + self-drawing) from the 2021 - 2022 academic year (July 2021 to June 2022), and
- Group 3 (lecture + self-drawing + flipped classroom) from the 2022 - 2023 academic year (July 2022 to June 2023)

Participants were categorized into three groups: group 1 (control group) underwent conventional LBL; group 2 (self- drawing; SD group) received the same LBL but incorporated self-drawing exercises; and group 3 (self- drawing after flipped classroom; SDFC group) engaged in LBL and the same self-drawing exercises as group 2, followed by a flipped classroom approach (Figure 1).

All participants completed a pretest at the beginning of their cardiovascular ward rotation and a posttest on the final day. The interval between the pre- and posttest was approximately 2 weeks, corresponding to the duration of each participant's cardiovascular ward rotation.

Figure 1. Schematic demonstration of the process of teaching activities. Group 1: Control group: traditional lecture-based learning (LBL) method; Group 2: traditional LBL and self-drawing (SD) method; Group 3: traditional LBL+SD+flipped classroom (SDFC) method.



Conventional LBL Approach

The control group (ie, group 1) was provided preclass learning materials 3 weeks before the teaching session. These materials included textbook chapters, supplementary electrocardiogram resources, online materials, and ECG exercises. During the teaching session, instruction was delivered in a traditional LBL format, consisting of a 120-minute lecture followed by a 15-minute question-and-answer session. After the session, students were given access to the lecture slides for self-directed learning.

Self-Drawing Method

In group 2, students engaged in self-drawing exercises to enhance cognitive processing of ECG patterns. The instructor first illustrated a characteristic ECG pattern for a specific disease, highlighting key features. After the explanation, students were instructed to replicate the ECG from memory without assistance after a 10-minute interval. Upon completion, students' drawings were compared against the standard reference ECG, and differences were discussed to reinforce understanding.

The ECG drawings were required to include the following key elements: (1) rate and rhythm of the ECG; (2) identification of P waves, QRS complexes, T waves, PR interval, QT interval, and QRS duration; (3) characteristic ECG findings for the specific disease being studied.

In group 3, the self-drawing component was introduced following the flipped classroom session.

Flipped Classroom Method

The flipped classroom approach in group 3 consisted of two phases:

- Preclass preparation: (1) Instructor preparation: Teachers underwent microteaching training and developed microvideo lessons, each featuring a real clinical case with guided learning instructions; (2) Student preparation: Three weeks before the teaching session (ie, immediately after the pretest), students were informed of the course structure and FC requirements. They were provided with microlessons and assigned reading materials, including textbook sections, lecture slides, and supplementary ECG resources. One day before their cardiovascular ward rotation, students submitted questions, which were compiled and reviewed by instructors.
- Classroom phase: Each 70-minute session was divided into three stages: (1) Stage 1 (20 minutes): Instructors addressed students' presubmitted questions, clarifying challenging concepts; (2) Stage 2 (20 minutes): An interactive discussion between students and instructors facilitated deeper engagement with ECG interpretation; (3) Stage 3 (30 minutes): A concise lecture was delivered emphasizing on the core and complex concepts.

ECG Competency Assessment

To evaluate students' comprehension and ability to apply acquired knowledge, all three groups completed the same standardized examination. To ensure consistency in difficulty levels, the questions in the pre- and posttests were reviewed and validated by two independent instructors.

The ECG cases used in the pre- and posttests featured the same diagnoses but were derived from different patients, ensuring similar ECG morphology while varying the sequence of questions. Each test consisted of 10 ECG interpretation questions, with each correct response awarded 1 point, yielding a maximum possible score of 10 points. The time limit for the test was 25 minutes.

Statistical Analysis

Baseline characteristics of the three groups were analyzed using appropriate statistical tests. Continuous variables such as age and pretest scores, were expressed as mean (SD) and compared using one-way ANOVA. Gender distribution was presented as a proportion and analyzed using the χ^2 test.

Ethical Considerations

This study was reviewed and approved by the institutional review board of MacKay Memorial Hospital, Taiwan (IRB number: 24MMHIS128e) in accordance with the ethical standards of the Declaration of Helsinki. We prioritized the rights and welfare of our participants. For secondary analyses using existing data, the original consent or institutional review board approval encompasses secondary analysis without the need for additional consent. To ensure participant privacy, all data collected were either anonymized or deidentified. If data could not be fully anonymized or deidentified, we implemented robust protective measures, including data encryption, restricted access, and data use agreements, to safeguard participant information and maintain confidentiality. Participants in this research study are resident physicians, therefore, they did not receive monetary compensation for their involvement. Instead, we provided nonmonetary incentives such as access to professional development resources, educational materials, or opportunities for mentorship. This approach ensured transparency and fairness in the compensation process while acknowledging the valuable time and contributions of the participants.

Within-Group Comparisons

To evaluate whether there was a significant improvement in scores from pretest to posttest within each group, paired, two-tailed *t* tests were conducted. The Shapiro-Wilk test was first applied to assess the normality of the data distribution. Since all groups met the normality assumption ($P>.05$), a parametric paired *t* test was deemed appropriate. The test was conducted separately for each group, with a significance level set at $\alpha=.05$.

Between-Group Comparisons

To determine whether the posttest scores differed significantly among the three groups, a nonparametric approach was chosen due to violations of normality assumptions. Specifically, Mann-Whitney *U* tests were performed for pairwise comparisons of posttest scores between groups. Effect sizes for each between-group comparison were calculated using the formula, $r=z/\sqrt{N}$, where z is the standardized test statistic from the Mann-Whitney *U* test and N is the total number of observations across the two groups. The resulting values were interpreted according to Cohen criteria (small: $r=0.1$; medium: $r=0.3$; large: $r=0.5$). To account for multiple comparisons across the three groups, a Bonferroni correction was applied, setting the significance threshold at $P<.0167$ (.05/3). The following comparisons were analyzed.

Pairwise comparisons were performed with one-tailed Mann-Whitney *U* tests**,** as the hypothesis was directional, aiming to determine whether higher posttest scores were observed in the latter group of each comparison.

While within-group differences (pretest vs posttest) were normally distributed and permitted the use of parametric paired *t* tests, the distribution of posttest scores across the three independent groups violated normality, warranting a nonparametric approach (Mann-Whitney *U* test) for between-group comparisons.

Visualization and Interpretation

Results were visualized using bar plots, where the mean values for pre- and posttest were displayed for each group. Pretest values were represented in blue, while posttest values were depicted in red. Statistical significance was indicated directly on the plots, with *P* values reported as follows: $P<.05$ was considered statistically significant; conversely, $P\geq.05$ was reported as nonsignificant (n.s.), with exact *P* values provided where relevant.

The visualization approach aimed to facilitate the interpretation of both within-group improvements and between-group differences, ensuring clarity in the presentation of statistical findings.

All statistical analyses were performed using SPSS (version 25.0; IBM Corp). Continuous variables such as age and pretest scores were presented as mean (SD) and analyzed using one-way ANOVA. Categorical variables such as gender distribution were expressed as percentages and analyzed using the χ^2 test. The dataset consisted of three independent groups, each undergoing pre- and posttest assessments. Descriptive statistics including mean and SD were calculated for each group to summarize the central tendency and variability of the scores.

Results

Baseline Participants

A total of 76 trainees were enrolled in this study, with 26 students in group 1 (control group), 25 trainees in group 2 (SD group), and 25 students in group 3 (SDFC group). There were no significant differences in age, gender, or baseline ECG test scores among the three groups prior to training (Table 1). Specifically, the findings showed that there were no significant differences in age (F=0.134, P=.88) or pretest scores (F=0.024, P=.98) among the three groups, indicating a similar baseline distribution. The gender distribution (% of men) was also not significantly different among the groups (χ^2=0.126, P=.94). These findings suggest that all three groups were well-balanced in terms of demographic characteristics before the intervention. (Table 1)

Table . Baseline characteristics of enrolled trainees.

Characteristics	Group 1 (n=26)	Group 2 (n=25)	Group 3 (n=25)	F test (df)/χ^2 (df)	P value
Age (years), mean (SD)	26.5 (1.1)	26.4 (0.6)	26.4 (0.6)	0.134 (2,73)[a]	.88
Pretest score, mean (SD)	2.96 (1.31)	2.92 (1.44)	2.88 (1.20)	0.024 (2)[b]	.98
Gender (men), n (%)	80.8	84.0	84.0	0.126 (2)[b]	.94

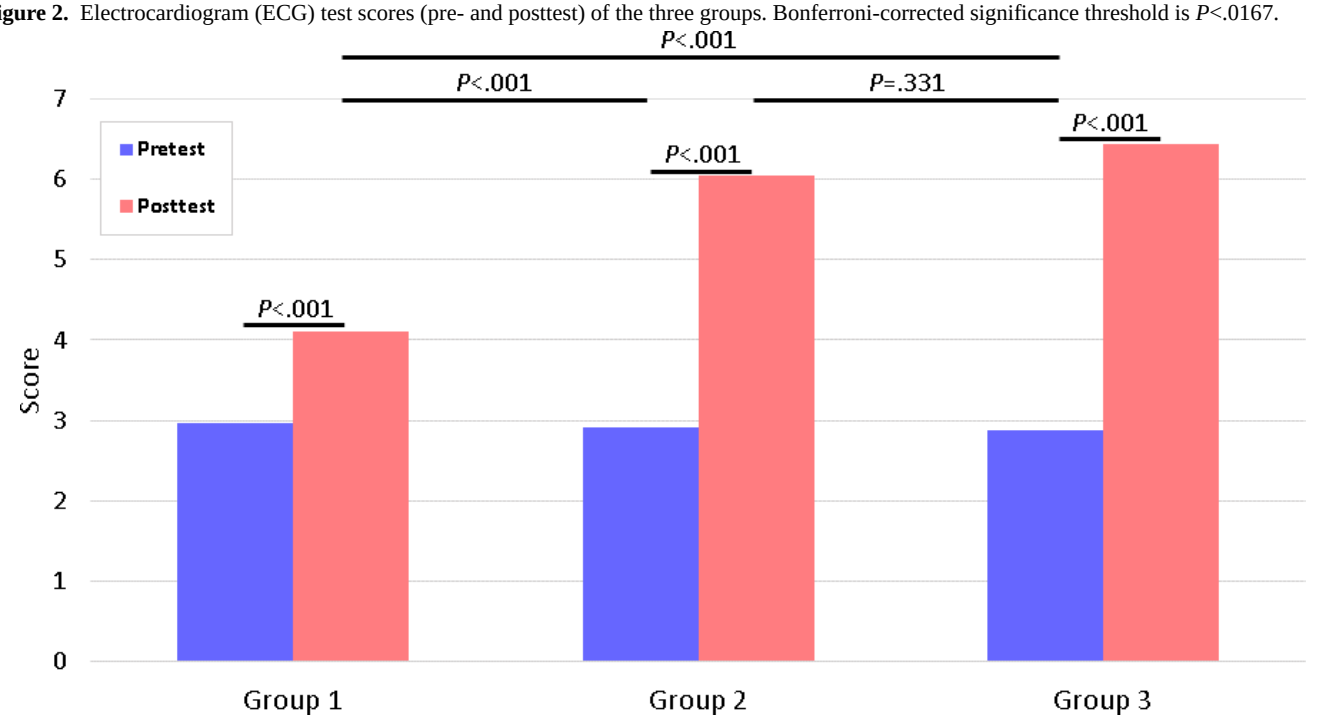
[a]ANOVA.

[b]Chi-square (χ^2) test.

Within-Group Comparisons (Pretest vs Posttest)

The findings demonstrated a statistically significant improvement in posttest scores across all three groups. Group 1 exhibited a significant increase from pretest to posttest. Group 2 also showed a significant improvement, with posttest scores being higher than pretest scores. Group 3 followed a similar trend, with a significant increase from pretest to posttest. These findings suggest that all three groups benefited from the training, with group 1 exhibiting the most substantial improvement (Figure 2).

Figure 2. Electrocardiogram (ECG) test scores (pre- and posttest) of the three groups. Bonferroni-corrected significance threshold is P<.0167.



Between-Group Comparisons (Posttest Scores)

Group 2 outperformed group 1 significantly (U=120.0, z=−3.86, P<.001, r=0.541), which remained statistically significant after applying Bonferroni correction (adjusted α=.0167). Group 3 also significantly outperformed group 1 (U=75.5, z=−4.70, P<.001 , r=0.658), meeting the corrected significance level. However, the difference between group 2 and group 3 was not statistically significant (U=263.0, z=−0.96, P=.33 , r=0.136), even before correction. These results indicate that while group 2 demonstrated a statistically significant advantage over group 1, the difference between group 2 and group 3 was not significant. (Figure 2).

Discussion

Principal Findings

The results of this study demonstrate that all three groups exhibited significant improvements in their ECG interpretation skills following training, suggesting that the instructional methods employed were effective in enhancing students' understanding and application of ECG principles. Among these methods, self-drawing emerged as a particularly impactful learning strategy, as evidenced by the superior posttest performance of students in group 2 (SD group) and group 3 (SDFC group) compared to group 1 (control group). These statistically significant differences were further substantiated by effect size metrics, which demonstrated medium-to-large magnitude improvements in ECG interpretation performance in the intervention groups compared to the control. Statistical differences between groups remained robust after applying Bonferroni correction for multiple comparisons (adjusted α=.0167), confirming that the observed advantages of the self-drawing and flipped classroom interventions were not attributable to chance or multiple testing artifacts.

A key reason why self-drawing may have contributed to improved learning outcomes is its ability to engage sensorimotor processes, spatial reasoning, and active recall. Blended learning yields significantly better ECG competence and confidence among medical students compared to conventional teaching. A stepwise approach to ECG analysis combined with deliberate practice and feedback may serve as an effective complement to lectures for electrocardiography education [9]. Unlike passive learning methods such as reading textbooks, watching videos, or participating in discussions, self-drawing requires students to actively reconstruct ECG waveforms, compelling them to process the anatomical and electrophysiological principles in a highly interactive manner. Prior research suggests that hand-drawn illustrations enhance memory encoding and retrieval by reinforcing visual-spatial associations [10,11]. This effect is particularly relevant for ECG interpretation, where recognizing waveform morphology and associating it with clinical conditions is crucial. The act of manually reproducing ECG patterns may thus create stronger mental representations, leading to enhanced retention and diagnostic accuracy.

Interestingly, the findings indicate that adding a flipped classroom component to self-drawing (group 3) did not lead to statistically significant gains over self-drawing alone (group 2; P=.18). While flipped learning has been widely recognized for enhancing engagement and promoting deeper conceptual understanding, its benefits appear to be diminished when self-drawing is already incorporated as a core learning strategy [12-17]. A previous study revealed that students using a flipped classroom format outperformed their classmates in the ability to interpret ECGs; however, this advantage was not evident when compared to lecture-delivered content. The students were able to prioritize their time when making decisions about attendance, based on teaching modality [18]. One possible explanation is that self-drawing is already an inherently active learning process, requiring students to deconstruct, analyze, and synthesize ECG patterns in a way that other instructional formats—such as passive content review or peer discussion—may not fully replicate. A previous prospective randomized trial indicated summative assessments significantly affect midterm retention of ECG interpretation skills. More intensive teaching showed no advantage over self-directed learning in retention test performance, and the substantial performance decline over eight weeks occurred independently of overall performance levels. These findings have implications for the design of ECG teaching and assessment intervention [19]. As a result, while flipped classroom elements may provide additional instructional support, their relative impact on learning may be reduced when students have already engaged in self-directed drawing exercises.

Moreover, the self-drawing process itself may offer an engaging and enjoyable learning experience, further reinforcing knowledge retention. The act of physically sketching ECG waveforms may create a sense of personal involvement that fosters intrinsic motivation, a factor that has been shown to positively influence long-term knowledge retention [5]. Student feedback on the teaching model indicated that they found self-directed activities beneficial, reinforcing the notion that active participation enhances learning engagement.

Despite these promising findings, certain challenges and limitations should be acknowledged. First, while self-drawing appears to be a highly effective tool for ECG interpretation training, its applicability to other domains of medical education remains to be explored. Second, although efforts were made to standardize the pre- and posttest difficulty levels, individual variations in artistic ability or drawing confidence could have influenced the results. Future studies should investigate whether guidance in sketching techniques or structured drawing templates may further optimize learning outcomes. Additionally, although self-drawing with flipped classroom instruction did not show a statistically significant advantage over self-drawing alone, further research is needed to explore whether longer-term retention benefits emerge when flipped learning is combined with self-drawing over extended periods.

Limitations

Several limitations should be acknowledged in this study. First, the improvement in ECG interpretation skills was assessed by comparing pre- and posttest scores using the same exam questions, albeit with a six-week interval between assessments. It is possible that some participants may have memorized the exam questions, potentially influencing the evaluation of their true ECG interpretation ability. Second, the posttest was

conducted immediately after the teaching program, which may not fully reflect long-term retention and competency in ECG interpretation. Future studies should incorporate delayed posttests to assess whether the observed improvements persist over time.

Third, although self-drawing was found to be an effective tool for ECG training, its applicability to other areas of medical education remains to be explored. Although efforts were made to standardize learning quality by using a checklist for discussions following ECG drawing, the drawing process itself is inherently subjective. Variations in artistic ability or drawing confidence among participants may have influenced the subsequent discussions and, consequently, the learning outcomes. Future research could examine whether structured guidance in sketching techniques or the use of standardized drawing templates might further optimize learning effectiveness.

Last, although self-drawing with flipped classroom did not demonstrate a statistically significant advantage over self-drawing alone, further research is required to explore whether longer-term retention benefits emerge when flipped learning is combined with self-drawing over an extended period. Additionally, investigating the interactive dynamics between different active learning modalities may provide insights into how to better integrate self-drawing within broader medical education frameworks.

Conclusion and Future Directions

This study highlights the effectiveness of self-drawing as a learning tool in ECG education, particularly in enhancing students' ability to recognize and interpret waveform characteristics. The findings suggest that self-drawing offers cognitive and experiential advantages that may not be easily replicated through passive learning modalities such as textbooks, videos, or discussions. While the flipped classroom remains a valuable pedagogical approach, its benefits appear to be moderated when self-drawing is already implemented as a central learning strategy. Given these insights, future research should explore the role of self-drawing across different medical education contexts and investigate how best to integrate flipped learning techniques to maximize knowledge retention and engagement.

Conflicts of Interest

None declared.

References

1. Auer R, Bauer DC, Marques-Vidal P, et al. Association of major and minor ECG abnormalities with coronary heart disease events. JAMA 2012 Apr 11;307(14):1497-1505. [doi: 10.1001/jama.2012.434] [Medline: 22496264]
2. Kashou A, May A, DeSimone C, Noseworthy P. The essential skill of ECG interpretation: How do we define and improve competency? Postgrad Med J 2020 Mar;96(1133):125-127. [doi: 10.1136/postgradmedj-2019-137191] [Medline: 31874907]
3. Bogun F, Anh D, Kalahasty G, et al. Misdiagnosis of atrial fibrillation and its clinical consequences. Am J Med 2004 Nov 1;117(9):636-642. [doi: 10.1016/j.amjmed.2004.06.024] [Medline: 15501200]
4. Fent G, Gosai J, Purva M. Teaching the interpretation of electrocardiograms: which method is best? J Electrocardiol 2015;48(2):190-193. [doi: 10.1016/j.jelectrocard.2014.12.014] [Medline: 25573481]
5. Deci EL, Ryan RM. The "What" and "Why" of goal pursuits: human needs and the self-determination of behavior. Psychol Inq 2000 Oct;11(4):227-268. [doi: 10.1207/S15327965PLI1104_01]
6. Sierra-Fernández CR, Alejandra HD, Trevethan-Cravioto SA, Azar-Manzur FJ, Mauricio LM, Garnica-Geronimo LR. Flipped learning as an educational model in a cardiology residency program. BMC Med Educ 2023 Jul 17;23(1):510. [doi: 10.1186/s12909-023-04439-2] [Medline: 37460995]
7. Allenbaugh J, Spagnoletti C, Berlacher K. Effects of a flipped classroom curriculum on inpatient cardiology resident education. J Grad Med Educ 2019 Apr;11(2):196-201. [doi: 10.4300/JGME-D-18-00543.1] [Medline: 31024653]
8. Chen HH, Pan HH, Hsu YD, et al. Flipped classroom as an effective model to teach electrocardiography to undergraduate medical students. J Med Educ 2016;20(4):276-284. [doi: 10.6145/jme201628]
9. Viljoen CA, Millar RS, Manning K, Burch VC. Effectiveness of blended learning versus lectures alone on ECG analysis and interpretation by medical students. BMC Med Educ 2020 Dec 3;20(1):488. [doi: 10.1186/s12909-020-02403-y] [Medline: 33272253]
10. Van Meter P, Garner J. The promise and practice of learner-generated drawing: literature review and synthesis. Educ Psychol Rev 2005 Dec;17(4):285-325. [doi: 10.1007/s10648-005-8136-3]
11. Wyss C, Rosenberger K, Bührer W. Student teachers' and teacher educators' professional vision: findings from an eye tracking study. Educ Psychol Rev 2021 Mar;33(1):91-107. [doi: 10.1007/s10648-020-09535-z]
12. Kim MK, Kim SM, Khera O, Getman J. The experience of three flipped classrooms in an urban university: an exploration of design principles. Internet High Educ 2014 Jul;22:37-50. [doi: 10.1016/j.iheduc.2014.04.003]
13. Sosa Díaz MJ, Narciso D. The impact of the flipped classroom in higher education: a case study. Aloma 2019;37(2):15-23. [doi: 10.51698/aloma.2019.37.2.15-23]
14. Phillips J, Wiesbauer F. The flipped classroom in medical education: a new standard in teaching. Trends Anaesth Crit Care 2022 Feb;42:4-8. [doi: 10.1016/j.tacc.2022.01.001] [Medline: 38620968]

15. Chen F, Lui AM, Martinelli SM. A systematic review of the effectiveness of flipped classrooms in medical education. Med Educ 2017 Jun;51(6):585-597. [doi: 10.1111/medu.13272] [Medline: 28488303]
16. Hew KF, Lo CK. Flipped classroom improves student learning in health professions education: a meta-analysis. BMC Med Educ 2018 Mar 15;18(1):38. [doi: 10.1186/s12909-018-1144-z] [Medline: 29544495]
17. Mishall PL, Meguid EMA, Elkhider IA, Khalil MK. The application of flipped classroom strategies in medical education: a review and recommendations. Med Sci Educ 2025 Feb;35(1):531-540. [doi: 10.1007/s40670-024-02166-x] [Medline: 40144088]
18. Henry M, Clayton S. Attendance improves student electrocardiography interpretation skills using the flipped classroom format. Med Sci Educ 2023 Feb;33(1):39-47. [doi: 10.1007/s40670-022-01689-5] [Medline: 37008425]
19. Raupach T, Harendza S, Anders S, Schuelper N, Brown J. How can we improve teaching of ECG interpretation skills? Findings from a prospective randomised trial. J Electrocardiol 2016;49(1):7-12. [doi: 10.1016/j.jelectrocard.2015.10.004] [Medline: 26615874]

Abbreviations

ECG: electrocardiogram
LBL: lecture-based learning
PGY-I: first-year postgraduate
SDFC: self-drawing following a flipped classroom

Edited by J Gentges; submitted 02.03.25; peer-reviewed by C Shah, C Lo, MC Tsai, XY Zhang; revised version received 16.04.25; accepted 30.04.25; published 13.06.25.

Please cite as:
Sung HY, Liao FC, Lin SI, Cheng HE, Lee CW
Visual Learning in Electrocardiography Training for Medical Residents: Comparative Intervention Study
JMIR Med Educ 2025;11:e73328
URL: https://mededu.jmir.org/2025/1/e73328
doi:10.2196/73328

e-Learning in Phoniatrics and Speech-Language Pathology: Exploratory Analysis of Free Access Tools in Augmentative and Alternative Communication

Jessica Büchs, MA; Christiane Neuschaefer-Rube, MD

Clinic for Phoniatrics, Pedaudiology & Communication Disorders, University Hospital and Medical Faculty, RWTH Aachen University, Pauwelsstraße 30, Aachen, Germany

Corresponding Author:
Jessica Büchs, MA
Clinic for Phoniatrics, Pedaudiology & Communication Disorders, University Hospital and Medical Faculty, RWTH Aachen University, Pauwelsstraße 30, Aachen, Germany

Abstract

Background: Augmentative and alternative communication (AAC) is a therapeutic approach and modality of expression for patients with limited or no expressive language. Speech-language pathologists and phoniatricians need to be competent in AAC to treat patients with complex communication needs. For knowledge acquisition and enhancement in AAC, a significant number of e-learning tools are available. To improve e-learning in AAC, it is essential to understand the attributes of these tools, such as formats, content areas, learning styles, or learning goals. However, these structures have yet to be investigated.

Objective: With this study, we aimed to (1) explore free access AAC e-learning tools that are appropriate for students and professionals of phoniatrics and speech-language pathology; (2) gain insight into formats, content areas, learning styles, and learning goals; and (3) investigate structural differences within and between basic and advanced learner level.

Methods: In 2023, we conducted a systematic web-based search with defined search terms in PubMed, peDOCS, Google Scholar, Google, the Apple App Store, and the Google Play Store in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines and piloting a protocol for data abstraction and validation. Inclusion criteria were free access, a mandatory minimum AAC content, and the use of the English or the German language. Social networks, video-sharing platforms, blogs, and forums were excluded. We analyzed formats (websites, online courses, apps, and podcasts), content areas (types of AAC, diagnostics, therapy, and other content areas), learning styles (visual, auditory, and audio-visual), and learning goals (receptive and performative) within and between basic and advanced level tools.

Results: We identified 131 tools, of which 57 (43.5%) were basic level and 74 (56.5%) were advanced level. Of these 131 tools, 105 (80.2%) were websites, 21 (16%) were online courses, 3 (2.3%) were apps and 2 (1.5%) were podcasts. Only 12 out of 74 (16.2%) tools for advanced learners offered performative tasks. For basic learners no such tasks could be identified. For learning style, all basic tools and most of the advanced level tools were “visual (text)” (57/57, 100% basic vs 66/74, 89.2% advanced). In terms of content, advanced level tools pertained more often to “diagnostics” (28/57, 49.1% basic vs 65/74, 87.8% advanced) and “therapy” (17/57, 29.8% basic vs 64/74, 86.5% advanced). Advanced level courses were more likely online courses (2/57, 3.5% basic vs 19/74, 25.7% advanced) and more often showed audio-visual learning styles compared with basic level tools (5/57, 8.8% basic vs 27/74, 36.5% advanced).

Conclusions: Our study showed that free-access AAC tools for phoniatrics and speech-language pathology varied in formats, content areas, learning styles, and learning goals. Furthermore, we found differences within and between learner levels. Thus, we established a basis for future research in e-learning in AAC.

(JMIR Med Educ 2025;11:e63392) doi:10.2196/63392

KEYWORDS

E-learning; digital learning; augmentative and alternative communication; speech-language pathology; phoniatrics; communication disorders; complex communication needs; communication aid

Introduction

Background

Augmentative and alternative communication (AAC) describes ways to support or replace spoken words for people who are unable to speak or communicate effectively using natural speech. Types of AAC include facial expressions, gestures, signs, cards with symbols, letterboards, or the use of electronic communication aids such as voice output devices [1-3]. Some published works concur that sign languages belong to AAC

[4,5] while others state that they are not considered AAC [6]. The types of AAC are categorized into "unaided" and "aided," depending on whether the patient uses solely their body to communicate or a communication aid [7]. Therefore, patients with limited expressive language may use various types of AAC to communicate.

A significant number of patients may benefit from AAC. Numerous medical conditions are known to be the cause of severe speech and language impairments that require AAC. These medical conditions are genetic disorders (eg, Down syndrome [8,9], Rett syndrome [10], and Angelman syndrome [11]), neurological impairments (eg, cerebral palsy [7,12-14], aphasia [7,15,16], dysarthria [7], apraxia of speech [7], Parkinson's disease [14]), motor neuron diseases (eg, amyotrophic lateral sclerosis and progressive muscular atrophy) [14,17,18], intellectual and developmental disabilities [19], severe hearing loss or deafness [20,21], different states of consciousness [22-24], postsurgical states affecting speech (eg, laryngectomy and tracheostomy) [7,25], and other medical conditions such as dementia [14], multiple sclerosis [14], autism spectrum disorder [7,14,26,27], visual impairment [28-30], locked-in-syndrome [31], and treatment in intensive care [32,33]. Creer et al [14] estimate that approximately 0.5% of the population of the United Kingdom are potential AAC users. Thus, AAC is a common therapeutic approach for various kinds of patients.

Patients with special communication needs, who may benefit from AAC, seek treatment in hospitals and established practices for phoniatrics and speech-language therapy [4,34]. Zinkevich et al [35] demonstrated the importance of the implementation of AAC in medical service delivery. Accordingly, phoniatricians and speech-language pathologists need to be competent in AAC. In preparation of this study, we formulated 3 main competences that phoniatricians and speech-language therapists may need in order to provide service to potential AAC users. First, phoniatricians and speech-language therapists should be able to identify patients that benefit from AAC. Second, they should be able to differentiate the types of AAC. Third, they should be able to decide which patients may benefit from what types of AAC. Thus, clinical professionals in phoniatrics and speech-language pathology (SLP) require specific knowledge in AAC to treat patients who may benefit from AAC.

What are sources of knowledge when it comes to AAC? Medical students are not obligatorily taught AAC at university. In Germany, even university students of SLP may have limited knowledge of AAC. In view of this fact, students as well as clinical professionals may search the web for information on AAC. Phoniatricians and speech-language pathologists use e-learning tools to acquire and enhance their expertise in their respective field [36,37], among other learning methods. Consequently, AAC e-learning tools are relevant in phoniatrics and SLP.

e-Learning tools can be described by their structures including formats, content areas, presentation modes, sensory modes, learning goals, target groups, and other describing structures [38]. In terms of formats, a tool can be a website, an online course, an app, a podcast, or another digital format. The learning styles of e-learning tools can be visual, auditory, or audio-visual [39,40]. Furthermore, e-learning tools have either a receptive or a performative learning goal [40]. In addition, e-learning tools may vary in content. Consequently, there are various possibilities to classify e-learning tools. This study focusses on the following e-learning structures: formats, content areas, learning styles, and learning goals.

Why is it essential to gain insight into the nature of AAC e-learning tools? Certain attributes such as "online course," "audio-visual," or "performative" characterize e-learning tools. This characterization may attract a specific group of learners (eg, learners who prefer a visual learning style may use websites with diagrams). A quantitative analysis could identify predominant or lacking structures. A "baseline" or status quo of the e-learning structures of AAC tools would be a starting point for further investigations, understanding and improving e-learning in AAC in the long term. However, the nature of AAC e-learning tools has yet to be investigated.

Goals of This Study

With this study, we aim to (1) explore free access AAC tools that are appropriate for e-learning in phoniatrics and SLP, (2) gain insight into the e-learning features of these tools (formats, content areas, learning styles, and learning goals), and (3) investigate structural differences within and between basic and advanced level tools. Furthermore, our goal is to establish a basis for future research in which we plan to test and evaluate a newly developed AAC e-learning tool for students of medicine and SLP.

Previous Work and Contribution

The study of Lin and Neuschaefer-Rube [38] in 2021 was about the onset of e-learning studies that discussed the improvement of e-learning in SLP, phoniatrics, and otolaryngology. They investigated the e-learning structures of tools in SLP, phoniatrics, and otolaryngology. Differences within and between academic-level learners and clinical-professional learners were found in terms of formats, content areas, and learning goals. Thus, their study presented an initial overview of existing e-learning tools in the interdisciplinary field of SLP and phoniatrics.

Our study contributes to the improvement of e-learning in AAC as being one of the many fields of interest in phoniatrics and SLP. By systematically searching the web for AAC e-learning tools, we gain an understanding of the overall quantity of AAC tools and their availability. An analysis of the e-learning tools in AAC provides further insight into the formats, content areas, learning styles, and learning goals that were state of the art during the time of search. Accordingly, this study should add new findings to the ongoing e-learning research in phoniatrics and SLP.

Methods

Protocol, Checklist, and Registration

Our study is an original, new, and exploratory investigation within the interdisciplinary field of medicine and SLP that targeted interactive learning tools on a niche topic. Therefore,

our study does not conform to the conventional framework of a systematic review. To find the tools, a novel approach was necessary. To adhere the tenets of good scientific practice, we developed structured protocols for the following processes: systematic web-based search, tool selection, and data abstraction and validation. For transparent, complete, and accurate reporting, we proceeded in accordance with PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 [41]. The checklist is provided in Multimedia Appendix 1. This study was registered at our institution Rheinisch-Westfälische Technische Hochschule (RWTH) Aachen University (no CTC-A 24 - 036) and has not been registered elsewhere.

Systematic Web-Based Search

In the summer of 2023, we conducted a systematic web-based search in "Google" (Google Search), Google Scholar, PubMed, peDOCS, the Apple App Store, and the Google Play Store using company owned devices. The use of nontraditional information sources such as "Google" and Google Scholar was necessary in the search for interactive tools. We combined general, academic, medical, and educational search engines to obtain optimal results. The combination of medical and educational databases was chosen since AAC is an interdisciplinary field. By searching App store, we targeted apps that teach about AAC. Gray literature search was not conducted. We used various search terms in English (American and British) and German ranging from specific to broad in the fields of education and medicine. Our IT center ensured that the IP address would not affect the search results. The same computer was used for all web-based searches. However, random checks with a different computer were done to ensure the results were the same. Multimedia Appendix 2 shows a protocol of our search including dates of search, search engines, search terms, number of records screened, and the final tools that met the inclusion and exclusion criteria.

Tool Selection Process

Records Identified, Removed, and Screened for Eligibility

Thousands of search results had been obtained, thus necessitating the implementation of limits for the screening process. We set a limit of the first 25 search results for each broad search term and a limit of the first 10 search results for specific terms. If the number of search results was less than 25 or 10, respectively, the records were limited to that number. In total, this strategy led to 1616 search results that were screened for eligibility. The screening was done by author [JB]. Automation tools were not used. We estimate that half of the records were discarded because they had no relation to AAC. The rationale behind this can be attributed to the polysemy of the acronym "AAC." For that matter, the originally planned search terms with the German abbreviation "UK" were dismissed. In addition, all tools pertaining to how AAC users can learn digitally, rather than how one can learn about AAC digitally, were removed, as well as duplicates.

Screening for Inclusion and Exclusion Criteria

After the initial filtering, approximately 700 - 800 tools remained to be screened for inclusion and exclusion criteria. Those results underwent a selection process based on access, content, and language. According to access, only tools that had a free, immediate, and full access were chosen for this study. Registration and email confirmation were tolerated. In terms of content, we defined a mandatory minimum. To be selected for this study, a tool had to contain a definition of AAC. In addition, a tool had to at least cover one of the following content areas: types of AAC, diagnostics, or therapy to ensure the tools were appropriate for the fields of phoniatrics and SLP. Tools, in this case apps, that only taught sign languages or functioned solely as a talker were excluded. Regarding language, only tools in English and German were included. However, social networks, video-sharing platforms, blogs, and forums were excluded. The search results obtained from the academic search engines yielded research papers. Since we targeted interactive tools rather than books and papers, we screened the results for links to free tools (eg, online courses). Reference lists were not reviewed. Using this technique, we found 3 websites that we added to our tool list. Nevertheless, these 3 tools had also previously been identified in the Google search. The tool list was supplemented by 2 websites added by the authors. Again, duplicates were removed. Finally, 131 tools were left for data abstraction.

Data Abstraction and Validation Process

To ensure a clear and concise method for data abstraction, we piloted a protocol (Multimedia Appendix 3). The data abstraction was done by author JB. Automation tools were not used. Author CNR checked for validity. Both authors followed the protocol and reported no bias. Uncertainties were solved in an interdisciplinary discussion between JB as a speech-language pathologist and CNR as a medical professor.

Tool Analysis

The tools were analyzed by basic and advanced level in terms of the following e-learning structures: formats (websites, online courses, apps, and podcasts), content areas (types of AAC, diagnostics, therapy, and other content areas), learning styles (visual, auditory, and audio-visual), and learning goals (receptive and performative).

Learner Level

We defined the criteria for basic-level and advanced-level tools based on our clinical and teaching experience. Basic level tools provided only general information. This level is appropriate for learners with no previous knowledge of AAC such as students of medicine and SLP. Advanced level tools exceeded general information and required either previous knowledge of AAC or clinical experience. Advanced-level tools are appropriate for students of medicine and SLP with previous knowledge as well as for professionals with clinical experience in AAC.

Formats

The tools of this study were either websites, online courses, apps, or podcasts. We analyzed websites of speech-language pathologists, clinics, consultation offices, self-help groups, institutions for special needs, and specific websites such as the website of the American Speech and Hearing Association and the German Society of AAC. We participated in online courses from universities and other teaching institutions. The online courses were recorded lectures, presentations, or modules on learning platforms. With regards to apps, we targeted those that

taught AAC. However, almost all apps functioned as a talker or trained the user in sign language while lacking a definition of AAC. Consequently, these apps were excluded from this study, leaving only 3 apps to our analysis. Finally, 2 podcasts were analyzed, although we had not explicitly searched for podcasts. In conclusion, the 4 formats in this study were websites, online courses, apps, and podcasts.

Content Areas

We analyzed the tools' content according to our previously defined 3 main competences that professionals in SLP and phoniatrics need: (1) knowledge about the types of AAC, (2) identification of potential AAC users, and (3) assignment of a type of AAC to a patient. Accordingly, the following 4 content areas were defined. The first content area was "types of AAC" for tools that provided a detailed explanation of at least 1 type of AAC or an overview of the types of AAC. The second content area was "diagnostics" for tools that identified at least 1 medical condition of AAC users. The third content area was "therapy" for tools that provided at least 1 example of a patient and their type of AAC. The fourth content area was "other content areas" for tools that provided other valuable information. This information could be downloads (eg, communication boards, sign language cards, collections of symbols, and other material), glossaries or descriptions of specific approaches in AAC. Thus, the 4 content areas in this study were "types of AAC," "diagnostics" "therapy," and "other content areas" to ensure the tools meet the needs of learners in phoniatrics and SLP.

Learning Styles

A total of 4 learning styles were identified, depending on whether the tools contained texts, pictures, diagrams, audio-files, or videos. When information was received via vision (eg, reading a text, interpreting diagrams, and looking at pictures), the tools were "visual (text)" or "visual (picture or diagram)." Auditory tools were audio-files where information was received via hearing. The audio-visual learning style was assigned to videos.

Learning Goals

Inspired by Lin and Neuschaefer-Rube [38], we defined the following learning goals for this study: "receptive" and "performative." A tool was "receptive" when information was only transmitted via reading or listening (ie, passive consumption). When a tool required action, it was "performative." Performative tools were further differentiated into "directive" and "guided discovery" [42]. A tool was "performative (directive)" when the learner had to fulfill directive tasks (eg, "fill-in-the-blank tests," multiple- and single choice tests, and assignment tasks). A tool was "performative (guided discovery)" when reasoning, thinking, and the integration of knowledge was required (eg, exploration of different chapters and submodules and decision making during the learning process) [42]. Performative tools could be both, "directive" and "guided discovery," while "performative," and "receptive" were mutually exclusive.

Statistical Analysis

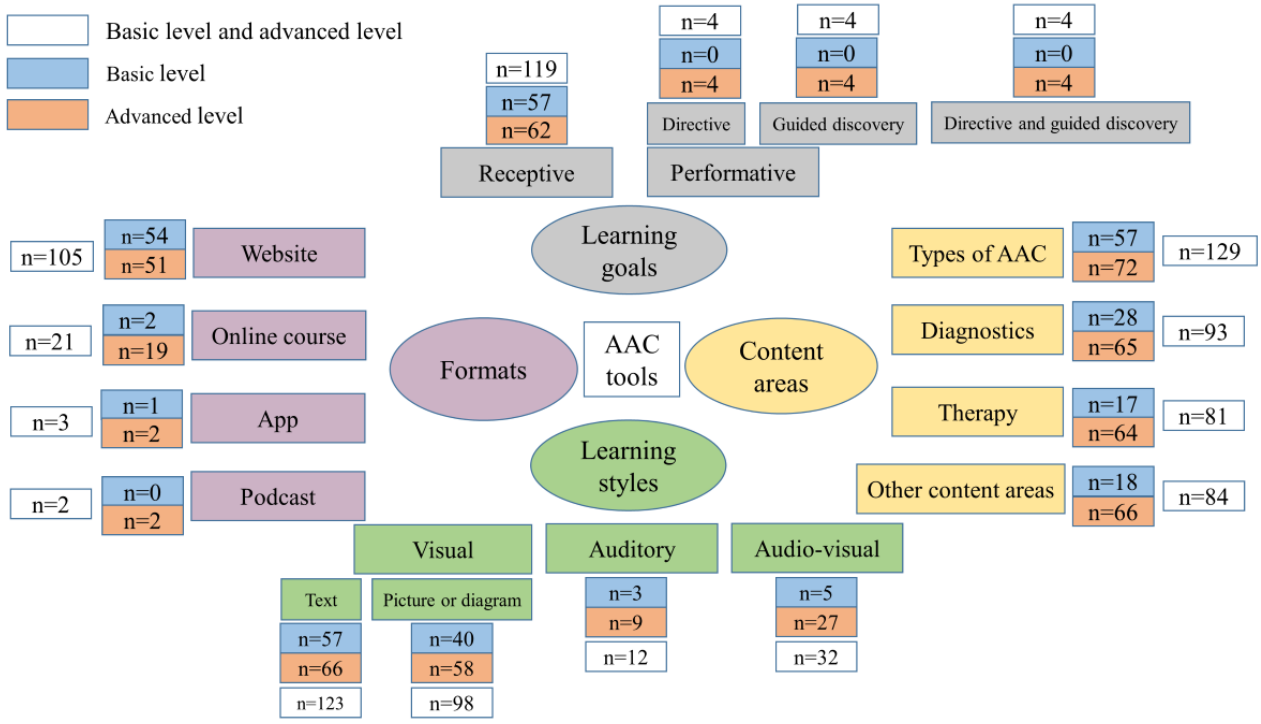
The results were analyzed using descriptive statistics. The attributes (formats, learning styles, content areas, and learning goals) of all tools were calculated, as well as inter- and intragroup differences regarding basic and advanced learner level. With respect to learning styles and content areas, overlaps were taken into consideration. The analysis was conducted using Microsoft Excel.

Results

Overview

We identified 131 tools that met the inclusion and exclusion criteria. Multimedia Appendix 4 shows a summary of the list of tools. Of all tools, 43.5% (57/131) were basic level and 56.5% (74/131) were advanced level. Figure 1 shows the number of basic-level tools and advanced-level tools according to formats, content areas, learning styles, and learning goals. The numbers of content areas and learning styles included overlaps (ie, a tool could cover multiple content areas).

Figure 1. Number of basic level tools and advanced level tools according to formats, content areas, learning styles, and learning goals; n=57 for basic level; n=74 for advanced level. AAC: augmentative and alternative communication.



e-Learning Structures Across All Tools

Of all 131 tools, 105 (80.2%) were websites, 21 (16%) were online courses, while only 3 (2.3%) were apps, and 2 (1.5%) were podcasts. In addition, it is worth mentioning that of the 21 online courses, only 2 (9.5%) were in German. Of the 387 content areas including overlaps, 129 (33.3%) pertained to “types of AAC,” 93 (24%) to “diagnostics,” 81 (20.9%) to “therapy,” and 84 (21.7%) to “other content areas.” Of the 265 learning styles including overlaps, “visual (text)” was predominant with 46.4% (123), followed by “visual (picture or diagram) with 37% (98). Only 12.1% (32) were “audio-visual” and only 4.5% (12) were “auditory.” For learning goals, 90.8% (119) of all 131 tools were “receptive,” while only 9.2% (12) were “performative.” Of these 12 performative tools, 4 (33.3%) pertained to “performative (directive),” 4 (33.3%) to “performative (guided discovery),” and 4 (33.3%) to “performative (directive and guided discovery).” Consequently, most of the tools were websites, taught about the types of AAC, showed the “visual (text)” learning style, and had a receptive learning goal. Multimedia Appendix 5 illustrates the distribution of formats, content areas, learning styles, and learning goals across all tools.

Within-Learner Level Analysis

Basic Level

Within all 57 basic level tools, 54 (94.7%) were websites, 2 (3.5%) were online courses, only 1 (1.8%) was an app and none was a podcast. The total number of content areas including overlaps was 120. Of these 120 counts, 57 (47.5%) pertained to “types of AAC,” 28 (23.3%) to “diagnostics,” 17 (14.2%) to “therapy,” and 18 (15%) to “other content areas.” The total number of learning styles including overlaps was 105. Of these 105 counts, 57 (54.3%) were “visual (text),” 40 (38.1%) were “visual (picture or diagram),” 5 (4.8%) were “audio-visual,” and 3 (2.9%) were “auditory.” As for learning goals, all basic-level tools were “receptive” and none were “performative. In conclusion, most of the basic level tools were websites, taught about the “types of AAC,” showed a “visual (text)” learning style, and all of them were "receptive" (Table 1).

Table . Distribution of augmentative and alternative communication e-learning tools in basic and advanced level according to formats, content areas, learning styles, and learning goals.

	Basic level	Advanced level
Formats		
Formats, n	57	74
Websites, n (%)	54 (94.7)	51 (68.9)
Online courses, n (%)	2 (3.5)	19 (25.7)
Apps, n (%)	1 (1.8)	2 (2.7)
Podcasts, n (%)	0 (0)	2 (2.7)
Content areas		
Content areas, n	120	267
Types of augmentative and alternative communication, n (%)	57 (47.5)	72 (27)
Diagnostics, n (%)	28 (23.3)	65 (24.3)
Therapy, n (%)	17 (14.2)	64 (24)
Other content areas, n (%)	18 (15)	66 (24.7)
Learning styles		
Learning styles, n	105	160
Visual (text), n (%)	57 (54.3)	66 (41.3)
Visual (picture or diagram), n (%)	40 (38.1)	58 (36.3)
Audio-visual, n (%)	5 (4.8)	27 (16.9)
Auditory, n (%)	3 (2.9)	9 (5.6)
Learning goals		
Learning goals, n	57	74
Receptive, n (%)	57 (100)	62 (83.8)
Performative (directive), n (%)	0 (0)	4 (5.4)
Performative (guided discovery), n (%)	0 (0)	4 (5.4)
Performative (directive and guided discovery), n (%)	0 (0)	4 (5.4)

Advanced Level

Of the 74 advanced-level tools, 51 (68.9%) were websites, 19 (25.7%) were online courses, 2 (2.7%) were apps, and 2 (2.7%) were podcasts. The total count of content areas including overlaps was 267. Of these 267 counts, 72 (27%) belonged to “types of AAC,” 65 (24.3%) to “diagnostics,” 64 (24%) to “therapy,” and 66 (24.7%) to “other content areas.” The total number of learning styles including overlaps was 160. Of these 160 counts, 66 (41.3%) were “visual (text),” 58 (36.3%) were “visual (picture or a diagram),” 27 (16.9%) were “audio-visual,” and only 9 (5.6%) were “auditory.” Within the 74 advanced-level tools, 62 (83.8%) were “receptive,” 4 (5.4%) required directive tasks, 4 (5.4%) offered a guided discovery, and 4 (5.4%) were directive and offered a guided discovery. To summarize, most of the advanced tools were websites, the content areas were evenly distributed, “visual (text)” was the predominant learning style, and most of the tools were “receptive” (Table 1)

Between-Learner Level Analysis

We investigated between-learner level differences in formats, content areas, learning styles, and learning goals. To compare basic and advanced level, the higher number of advanced level tools (respectively smaller number of basic level tools) had to be taken into account (57 basic and 74 advanced). Therefore, the absolute numbers cannot be compared. Instead, the percentages (of each component, eg, “website”) within learner levels were compared. In the following paragraphs, the results are presented as “basic level versus advanced level" when results for both learner levels appear in parentheses.

Formats

Websites were more common in basic tools (54/57, 94.7% vs 51/74, 68.9%), whereas online courses were more common in advanced tools (2/57, 3.5% vs 19/74, 25.7%). Apps were slightly more common in advanced tools (1/57, 1.8% vs 2/74, 2.7%). Podcasts only appeared in advanced tools (0/57, 0% vs 2/74, 2.7%). In conclusion, websites were more common in basic

level, whereas online courses, apps, and podcasts appeared more often in advanced level (Multimedia Appendix 6).

Content Areas

All basic level tools and almost all advanced level tools covered "types of AAC" (57/57, 100% vs 72/74, 97.3%). Diagnostic-related content was much more common in advanced level tools (28/57, 49.1% vs 65/74, 87.8%) as were "other content areas" (18/57, 31.6% vs 66/74, 89.2%) and therapeutic-related content (17/57, 29.8% vs 64/74, 86.5%). Advanced level tools were more likely to cover multiple content areas, while showing more diagnostic- and therapy-related content as well as more other content areas (Multimedia Appendix 7).

Learning Styles

All basic level tools and most of the advanced level tools had a learning style pertaining to "visual (text)" (57/57, 100% vs 66/74, 89.2%). "Visual (picture or diagram)" was more common in advanced level (40/57, 70.2% vs 58/74 78.4%), as were "audio-visual" (5/57, 8.8% vs 27/74, 36.5%) and "auditory" (3/57, 5.3% vs 9/74, 12.2%) learning styles. Consequently, "visual (picture or diagram)," "auditory," and "audio-visual" learning styles were more common in advanced level (Multimedia Appendix 8).

Learning Goals

All basic level tools and the majority of the advanced level tools were receptive (57/57, 100% vs 62/74, 83.8%). Respectively, basic level tools did not require performance whereas some of the advanced tools did (0/57, 0% vs 12/74, 16.2%) (Table 1).

Discussion

Principal Results

Our study set out to explore e-learning tools in AAC for phoniatrics and SLP, analyze their e-learning features and investigate differences in basic and advanced learner level. We identified 131 free access e-learning AAC tools and gained insight into their formats, content areas, learning styles, and learning goals. Most of the tools were websites, while apps and podcasts were rare. The predominant content area was "types of AAC" and "visual (text)" was the most common learning style. Most of the tools were "receptive." Within both learner levels, "website" was the predominant format. Within basic level, none of the tools were podcasts and the predominant content was "types of AAC." Within advanced level, the content areas were almost evenly distributed. "Visual (text)" was the predominant learning style in both learner levels. Most of the advanced tools and all basic level tools were "receptive." Websites were more common in basic level, whereas online courses, apps, and podcasts appeared more in advanced level. The content of advanced level tools was more diagnostic-related and therapy-related. "Visual (picture or diagram)," "auditory," and "audio-visual" learning styles were more common in advanced level. All basic-level tools and the majority of the advanced-level tools were receptive.

Interpretation

Number of Tools

The relatively large number of 131 tools is encouraging, given that AAC is not yet fully implemented in the curricula of phoniatrics and SLP (at least not in Germany). This number of tools was a good sample size for our analysis. In addition, the number of tools indicates that AAC appears to be a topic of interest in e-learning. This supports a recent study by Burgio [43], who claim that new AAC e-learning tools evolve constantly.

Formats

It is not surprising that websites were the predominant format for both learner levels due to the web-based nature of this study. However, we were surprised about the rare number of apps that met our inclusion and exclusion criteria. We found a fair amount of AAC-related apps. However, almost all apps functioned as a communication aid (eg, talker) while lacking a definition of AAC, therefore not being an e-learning tool. Other researchers dealt with these kinds of AAC apps [44]. Unanticipated was the relatively high number of online courses. This supports our claim that AAC is a topic of interest in e-learning. If we had not restricted the inclusion to "free access," we might have analyzed even more online courses. The fact that most of the online courses pertained to advanced level is in the nature of the thing. Since almost all online courses covered more than just general information, they were assigned to "advanced level." That made us question our learner level criteria. We could have set the boundary between basic level and advanced level differently (eg, only very detailed tools would be considered as advanced level) or formulated a third category (eg, "intermediate"). That only 2 of the 21 online courses were in German indicates that German-speaking countries lag behind English-speaking countries when it comes to e-learning in AAC. Nevertheless, it is obvious that English (as the predominant language worldwide) is used more often in teaching. The small number of podcasts can be explained by not having specifically searched for podcasts in the first place. Overall, AAC e-learning tools exist in various formats, which indicates that AAC is a topic of interest in e-learning, however, more so in English rather than in German.

Content Areas

It is not surprising that advanced level tools covered more content areas. The more detailed the content, the more likely was a tool assigned to advanced level. Again, this might question our learner-level criteria. However, it is interesting that the contents were almost evenly distributed in advanced level, whereas basic level tools covered more "types of AAC" content. Notwithstanding, it seems logical that general information is about the types of AAC. Therapeutic- and diagnostics-related content have a more "advanced" attribute. What do these results mean? We interpret that "types of AAC" seems to be a "basic" content or "general information" that is essential for learners with no previous knowledge of AAC. Information on the types of AAC seems to be an essential content of AAC e-learning tools and should therefore be considered in the development of future modules.

Learning Styles

As expected, the predominant learning style was "visual (text)," given that 80.2% of all tools were websites. The fact that online courses and podcasts were more likely in advanced level explains the prevalence of "auditory" and "audio-visual" learning styles in this category. These findings may be somewhat limited by the fact that we had not specifically searched for podcasts. However, it is still interesting to note that audio files and podcasts were more common in advanced level.

Learning Goals

In terms of learning goals, we found performative tools only in advanced level. In hindsight, this is obvious. Tools that exceed general information (respectively "advanced level"), more likely required active sensemaking and reasoning. However, we were surprised about the overall small number of performative tools. We therefore suggest that new AAC e-learning tools should offer performative tasks to fill this gap.

Comparison to Previous Work

This study appears to be the first to investigate the structures of AAC e-learning tools in English and German. In designing our study, we were inspired by Lin and Neuschaefer-Rube [38] who analyzed e-learning tools for SLP, phoniatrics, and otolaryngology by their e-learning structures. Although they set slightly different criteria, their overall idea of investigating e-learning tools can be compared with our study. In their study, for example, learner levels were classified into "academic level" and "clinical professional level." We chose a different classification for our study, since it cannot be assumed that all clinical professionals have previous knowledge in AAC. Most of our results reflect those of Lin and Neuschaefer-Rube [38] who also found that visual tools were predominant as well as most of the tools were receptive. The difference between the studies was, interestingly, that performative tools pertained more to academic level learners.

The comparison of our findings to yet another study turned out to be challenging. e-Learning studies increased since the COVID-19 pandemic resulting in numerous available research papers. However, other works in the fields of otolaryngology and SLP focused on surveys [45,46], the effectiveness of e-learning [47] or specific online programs [48,49], rather than exploring and analyzing the attributes of e-learning tools. Therefore, our study is a specific examination in the overall bewildering and evolving field of e-learning research.

Limitations

The findings of this study must be seen in light of 5 limitations. These are (1) incompleteness and difficult replication, (2) limited search terms and search engines, (3) strict inclusion criteria, (4) binary learner level categorization, and (5) quantitative assessment rather than quality rating.

First, e-learning tools are not static since the World Wide Web is an ongoing field of updates and changes [50]. Therefore, this study is only a snapshot of the available tools at the time of data collection undergoing certain inclusion and exclusion criteria. In addition, different IP addresses may lead to different search results. Accordingly, this study is not an investigation of all existing e-learning tools in AAC nor is it replicable. Nevertheless, with a total number of 131 tools, we managed to investigate a high amount, which appears to be a fair representation of the present AAC e-learning scope.

Second, in terms of our web-based search, we could have used more search engines to possibly find more tools [51]. Likewise, we could have added more search terms. Nonetheless, our chosen search engines and search terms led to many results that would not have been manageable without strict inclusion criteria.

Third, the inclusion criterium of a mandatory definition of AAC may have eliminated some advanced tools. In fact, the authors cannot recollect that this was the case during the process of finding the tools. Nevertheless, if we had included tools that did not provide a definition of AAC, probably almost every website within the search results would have entered our study.

Fourth, the binary categorization into either "basic level" or "advanced level" did not leave room for learners with knowledge in between those levels (eg, intermediate level). However, we think that this categorization allows students and clinical professionals to choose a tool according to their knowledge level rather than to their professional status.

Finally, this quantitative study cannot give quality evaluations of the 131 tools, nor do we have proof that the authors of the tools were professionals or well versed in AAC. However, since we investigated the content of the tools, we can conclude that all tools seemed to provide correct information. Despite of its limitations, our study certainly adds to the understanding of e-learning in AAC.

Future Directions

Our study lays the groundwork for future research in e-learning in AAC. More work needs to be done to gain further insight into e-learning in AAC, such as (1) expanding the analysis to other formats, (2) assessing the quality of the tools, (3) conducting a survey study on e-learning in AAC, and (4) developing new AAC e-learning tools.

First, the analysis of AAC e-learning tools on video sharing platforms, social networks, blogs, and forums would be interesting, since we excluded them in our study. Furthermore, an explicit search for a particular format (eg, podcasts) could lead to more results that might be worth investigating. In addition, an assessment of misinformation on AAC in social media would be informative.

Second, a qualitative analysis of AAC e-learning tools would be beneficial for the users of these tools. The best rated tools could be added to a possible future toolbox app [37], or to a German "AAC online learning platform" suggested by Burgio [43]. Furthermore, it would be interesting to investigate correlations between the e-learning structures and the quality of the tools.

Third, a survey study in phoniatrics and SLP could help identify the preferred structures of possible AAC e-learning tools. It would be worth investigating whether phoniatricians and speech-language pathologists have the same demands and would therefore benefit from the same tool.

Finally, we suggest practical applications for the development of new e-learning tools in AAC for students and professionals in phoniatrics and SLP. One example of a possible AAC e-learning tool could be a German, advanced level, audio-visual online course with a knowledge quiz that covers all content areas relevant to phoniatricians and speech-language pathologists. On the basis of this study, we developed such an online course which is currently being tested with students of medicine and SLP at RWTH Aachen University. Another example would be the development of apps that teach AAC (in both languages).

Conclusion

To the best of our knowledge, our exploratory study marks the beginning of the investigation of e-learning tools in AAC for phoniatrics and SLP. We found a fair number of tools and gained insight into their formats, content areas, learning styles, and learning goals. Our data indicate that e-learning in AAC is a topic of interest and needs to be further investigated. Overall, we established a basis for future research and suggested practical applications for e-learning in AAC.

Acknowledgments

This study had no funding or sponsors. JB conducted this study in her role as a scientific speech-language pathologist at the Clinic for Phoniatrics, Pedaudiology and Communication Disorders, University Hospital Rheinisch-Westfälische Technische Hochschule (RWTH) Aachen.

Authors' Contributions

JB conceived the study design, carried out the investigation, performed the analytical calculations, interpreted the data, created the figure, the table, and the multimedia appendices, wrote and revised the manuscript with input from CNR. CNR conceived the original idea, encouraged JB in the study design, suggested the investigation of learner level differences, supervised the conduction of the study, and revised the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist.
[DOCX File, 36 KB - mededu_v11i1e63392_app1.docx]

Multimedia Appendix 2
Search protocol.
[PDF File, 198 KB - mededu_v11i1e63392_app2.pdf]

Multimedia Appendix 3
Data abstraction protocol.
[PDF File, 131 KB - mededu_v11i1e63392_app3.pdf]

Multimedia Appendix 4
Summary list of tools.
[PDF File, 363 KB - mededu_v11i1e63392_app4.pdf]

Multimedia Appendix 5
Distribution of attributes across all tools.
[DOCX File, 167 KB - mededu_v11i1e63392_app5.docx]

Multimedia Appendix 6
Formats.
[DOCX File, 90 KB - mededu_v11i1e63392_app6.docx]

Multimedia Appendix 7
Content areas.
[DOCX File, 110 KB - mededu_v11i1e63392_app7.docx]

Multimedia Appendix 8

Learning styles.
[DOCX File, 154 KB - mededu_v11i1e63392_app8.docx]

References

1. American Speech-Language-Hearing Association. Augmentative and alternative communication (AAC). URL: https://www.asha.org/public/speech/disorders/aac/ [accessed 2025-03-11]
2. Zinkevich A, Uthoff SAK, Boenisch J, et al. Making a voice heard: evaluation of a new service delivery in augmentative and alternative communication through qualitative interviews with people without natural speech. BMC Res Notes 2023 Mar 29;16(1):42. [doi: 10.1186/s13104-023-06310-5] [Medline: 36991499]
3. Griffiths T, Clarke M, Price K. Augmentative and alternative communication for children with speech, language and communication needs. Paediatr Child Health (Oxford) 2022 Aug;32(8):277-281. [doi: 10.1016/j.paed.2022.05.001]
4. Lorang E, Maltman N, Venker C, Eith A, Sterling A. Speech-language pathologists' practices in augmentative and alternative communication during early intervention. Augment Altern Commun 2022 Mar;38(1):41-52. [doi: 10.1080/07434618.2022.2046853] [Medline: 35422176]
5. Vogel AP, Spencer C, Burke K, et al. Optimizing communication in ataxia: a multifaceted approach to alternative and augmentative communication (AAC). Cerebellum 2024 Oct;23(5):2142-2151. [doi: 10.1007/s12311-024-01675-0] [Medline: 38448793]
6. Richlin BC, Chow K, Cosetti MK. Augmentative and alternative communication (AAC) in pediatric cochlear implant recipients with complex needs: a scoping review. Int J Pediatr Otorhinolaryngol 2023 Aug;171:111610. [doi: 10.1016/j.ijporl.2023.111610] [Medline: 37329701]
7. Butt AK, Zubair R, Rathore FA. The role of augmentative and alternative communication in speech and language therapy: a mini review. J Pak Med Assoc 2022 Mar;72(3):581-584. [doi: 10.47391/JPMA.22-023] [Medline: 35320253]
8. Holyfield C, Drager K. Integrating familiar listeners and speech recognition technologies into augmentative and alternative communication intervention for adults with down syndrome: Descriptive exploration. Assist Technol 2022 Nov 2;34(6):734-744. [doi: 10.1080/10400435.2021.1934610] [Medline: 34033520]
9. Barbosa RTDA, de Oliveira ASB, de Lima Antão JYF, et al. Augmentative and alternative communication in children with Down's syndrome: a systematic review. BMC Pediatr 2018 Dec;18(1). [doi: 10.1186/s12887-018-1144-5]
10. Townend GS, Bartolotta TE, Urbanowicz A, Wandin H, Curfs LMG. Development of consensus-based guidelines for managing communication of individuals with Rett syndrome. Augment Altern Commun 2020 Jun;36(2):71-81. [doi: 10.1080/07434618.2020.1785009] [Medline: 32720526]
11. Calculator SN. Use and acceptance of AAC systems by children with Angelman syndrome. J Appl Res Intellect Disabil 2013 Nov;26(6):557-567. [doi: 10.1111/jar.12048] [Medline: 23606637]
12. Lillehaug HA, Klevberg GL, Stadskleiv K. Provision of augmentative and alternative communication interventions to Norwegian preschool children with cerebral palsy: are the right children receiving interventions? Augment Altern Commun 2023 Dec;39(4):219-229. [doi: 10.1080/07434618.2023.2212068] [Medline: 37212772]
13. Caron J, Light J. "Social Media has Opened a World of 'Open communication:'" experiences of adults with cerebral palsy who use augmentative and alternative communication and social media. Augment Altern Commun 2016;32(1):25-40. [doi: 10.3109/07434618.2015.1052887] [Medline: 26056722]
14. Creer S, Enderby P, Judge S, John A. Prevalence of people who could benefit from augmentative and alternative communication (AAC) in the UK: determining the need. Int J Lang Commun Disord 2016 Nov;51(6):639-653. [doi: 10.1111/1460-6984.12235] [Medline: 27113569]
15. Góral-Półrola J, Półrola P, Mirska N, Mirski A, Herman-Sucharska I, Pąchalska M. Augmentative and alternative communication (AAC) for a patient with a nonfluent/ agrammatic variant of PPA in the mutism stage. Ann Agric Environ Med 2016;23(1):182-192. [doi: 10.5604/12321966.1196877] [Medline: 27007540]
16. Dietz A, Wallace SE, Weissling K. Revisiting the role of augmentative and alternative communication in aphasia rehabilitation. Am J Speech Lang Pathol 2020 May 8;29(2):909-913. [doi: 10.1044/2019_AJSLP-19-00041] [Medline: 32109137]
17. Mackenzie L, Bhuta P, Rusten K, Devine J, Love A, Waterson P. Communications technology and motor neuron disease: an Australian survey of people with motor neuron disease. JMIR Rehabil Assist Technol 2016 Jan 25;3(1):e2. [doi: 10.2196/rehab.4017] [Medline: 28582251]
18. Ball LJ, Fager S, Fried-Oken M. Augmentative and alternative communication for people with progressive neuromuscular disease. Phys Med Rehabil Clin N Am 2012 Aug;23(3):689-699. [doi: 10.1016/j.pmr.2012.06.003] [Medline: 22938882]
19. Crowe B, Machalicek W, Wei Q, Drew C, Ganz JB. Augmentative and alternative communication for children with intellectual and developmental disability: a mega-review of the literature. J Dev Phys Disabil 2022;34(1):1-42. [doi: 10.1007/s10882-021-09790-0] [Medline: 33814873]
20. Šantić I, Bonetti L. Language intervention instead of speech intervention for children with cochlear implants. J Audiol Otol 2023 Apr;27(2):55-62. [doi: 10.7874/jao.2022.00584] [Medline: 37073450]

21. Meinzen-Derr J, Sheldon RM, Henry S, et al. Enhancing language in children who are deaf/hard-of-hearing using augmentative and alternative communication technology strategies. Int J Pediatr Otorhinolaryngol 2019 Oct;125:23-31. [doi: 10.1016/j.ijporl.2019.06.015] [Medline: 31238158]
22. Rasmus A, Góral-Półrola J, Orłowska E, Wiłkość-Dębczyńska M, Grzywniak C. Nonverbal communication of trauma patients in a state of minimal consciousness. Ann Agric Environ Med 2019 Jun 17;26(2):304-308. [doi: 10.26444/aaem/91911] [Medline: 31232063]
23. Waydhas C, Deffner T, Gaschler R, et al. Sedation, sleep-promotion, and non-verbal and verbal communication techniques in critically ill intubated or tracheostomized patients: results of a survey. BMC Anesthesiol 2022 Dec 12;22(1):384. [doi: 10.1186/s12871-022-01887-z] [Medline: 36503427]
24. Murtaugh B, Fager S, Sorenson T. Emergence from disorders of consciousness. Phys Med Rehabil Clin N Am 2024 Feb;35(1):175-191. [doi: 10.1016/j.pmr.2023.07.002]
25. Haring CT, Farlow JL, Leginza M, et al. Effect of augmentative technology on communication and quality of life after tracheostomy or total laryngectomy. Otolaryngol Head Neck Surg 2022 Dec;167(6):985-990. [doi: 10.1177/01945998211013778] [Medline: 34060949]
26. Lorah ER, Holyfield C, Kucharczyk S. Typical preschoolers' perceptions of augmentative and alternative communication modes of a preschooler with autism spectrum disorder. Augment Altern Commun 2021 Mar;37(1):52-63. [doi: 10.1080/07434618.2020.1864469] [Medline: 33583287]
27. Gevarter C, Groll M, Stone E. Dynamic assessment of augmentative and alternative communication application grid formats and communicative targets for children with autism spectrum disorder. Augment Altern Commun 2020 Dec;36(4):226-237. [doi: 10.1080/07434618.2020.1845236] [Medline: 33238754]
28. Wilkinson KM, Elko LR, Elko E, et al. An evidence-based approach to augmentative and alternative communication design for individuals with cortical visual impairment. Am J Speech Lang Pathol 2023 Sep 11;32(5):1939-1960. [doi: 10.1044/2023_AJSLP-22-00397] [Medline: 37594735]
29. Blackstone SW, Luo F, Canchola J, Wilkinson KM, Roman-Lantzy C. Children with cortical visual impairment and complex communication needs: identifying gaps between needs and current practice. Lang Speech Hear Serv Sch 2021 Apr 20;52(2):612-629. [doi: 10.1044/2020_LSHSS-20-00088] [Medline: 33592150]
30. Boster JB, Findlen UM, Pitt K, McCarthy JW. Design of aided augmentative and alternative communication systems for children with vision impairment: psychoacoustic perspectives. Augment Altern Commun 2024 Mar;40(1):57-67. [doi: 10.1080/07434618.2023.2262573] [Medline: 37811949]
31. Park SW, Yim YL, Yi SH, Kim HY, Jung SM. Augmentative and alternative communication training using eye blink switch for locked-in syndrome patient. Ann Rehabil Med 2012 Apr;36(2):268-272. [doi: 10.5535/arm.2012.36.2.268] [Medline: 22639753]
32. Pina S, Canellas M, Prazeres R, et al. Augmentative and alternative communication in ventilated patients: a scoping review. Rev Bras Enferm 2020;73(5):e20190562. [doi: 10.1590/0034-7167-2019-0562] [Medline: 32667397]
33. Zaylskie LE, Biggs EE, Minchin KJ, Abel ZK. Nurse perspectives on supporting children and youth who use augmentative and alternative communication (AAC) in the pediatric intensive care unit. Augment Altern Commun 2024 Dec;40(4):255-266. [doi: 10.1080/07434618.2023.2284269] [Medline: 38035596]
34. Light J, McNaughton D, Beukelman D, et al. Challenges and opportunities in augmentative and alternative communication: research and technology development to enhance communication and participation for individuals with complex communication needs. Augment Altern Commun 2019 Mar;35(1):1-12. [doi: 10.1080/07434618.2018.1556732] [Medline: 30648903]
35. Zinkevich A, Uthoff SAK, Boenisch J, Sachse SK, Bernasconi T, Ansmann L. Complex intervention in augmentative and alternative communication (AAC) care in Germany: a study protocol of an evaluation study with a controlled mixed-methods design. BMJ Open 2019 Aug 28;9(8):e029469. [doi: 10.1136/bmjopen-2019-029469] [Medline: 31467052]
36. Lin YC, Lemos M, Neuschaefer-Rube C. Digital health and digital learning experiences across speech-language pathology, phoniatrics, and otolaryngology: interdisciplinary survey study. JMIR Med Educ 2021 Nov 5;7(4):e30873. [doi: 10.2196/30873] [Medline: 34738911]
37. Lin YC, Lemos M, Neuschaefer-Rube C. Digital health and learning in speech-language pathology, phoniatrics, and otolaryngology: survey study for designing a digital learning toolbox app. JMIR Med Educ 2022 Apr 27;8(2):e34042. [doi: 10.2196/34042] [Medline: 35475980]
38. Lin YC, Neuschaefer-Rube C. Digital learning in speech-language pathology, phoniatrics, and otolaryngology: interdisciplinary and exploratory analysis of content, organizing structures, and formats. JMIR Med Educ 2021 Jul 27;7(3):e27901. [doi: 10.2196/27901] [Medline: 34313592]
39. Jadin T. Multimedia and memory. A cognitive-psychological perspective on learning using technologies. In: Martin E, Sandra S, editors. Lehrbuch für Lernen und Lehren mit Technologien, 2nd edition 2013. [doi: 10.25656/01:8346]
40. Mayer RE. Cognitive theory of multimedia learning. 2014 p. 43-71. [doi: 10.1017/CBO9781139547369.005]
41. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372(71):n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
42. Lavine RA. Guided discovery learning. . 2012 p. 1402-1403. [doi: 10.1007/978-1-4419-1428-6_526]

43. Burgio NM. E-learning in the field of further education and training in augmentative and alternative communication (AAC): recommendations for the design of an online platform for AAC and training to become an e-teacher for AAC. Behindertenpa¨dagogik 2020;59(4):394-403. [doi: 10.30820/0341-7301-2020-4-394]
44. Du Y, Choe S, Vega J, Liu Y, Trujillo A. Listening to stakeholders involved in speech-language therapy for children with communication disorders: content analysis of Apple app store reviews. JMIR Pediatr Parent 2022 Jan 21;5(1):e28661. [doi: 10.2196/28661] [Medline: 35060912]
45. Offergeld C, Ketterer M, Neudert M, et al. “Online from tomorrow on please”: comparison of digital framework conditions of curricular teaching at national university ENT clinics in times of COVID-19: Digital teaching at national university ENT clinics. HNO 2021 Mar;69(3):213-220. [doi: 10.1007/s00106-020-00939-5] [Medline: 32929523]
46. O’ Leary N, Brouder N, Bessell N, Frizelle P. An exploration of speech and language pathology student and facilitator perspectives on problem-based learning online. Clinical Linguistics & Phonetics 2023 Jul 3;37(7):599-617. [doi: 10.1080/02699206.2022.2061377]
47. Krauss F, Giesler M, Offergeld C. [On the effectiveness of digital teaching of practical skills in curricular ENT education]. HNO 2022 Apr;70(4):287-294. [doi: 10.1007/s00106-021-01107-z] [Medline: 34545415]
48. Lang F, Everad B, Knopf A, Kuhn S, Offergeld C. [Digitalization in curricular teaching: experiences with the Freiburg ENT Learning Program]. Laryngorhinootologie 2021 Dec;100(12):973-980. [doi: 10.1055/a-1334-4274] [Medline: 33352588]
49. Patel ST, Shah S, Sood RP, Siddiqui Z, McKay-Davies I. The implementation of virtual clinical skills teaching in improving procedural confidence in ENT trainees. Adv Med Educ Pract 2021;12:965-969. [doi: 10.2147/AMEP.S322965] [Medline: 34475794]
50. Jansen BJ, Pooch UW. A review of web searching studies and a framework for future research. J Am Soc Inf Sci 2001;52(3):235-246. [doi: 10.1002/1097-4571(2000)9999:9999<::AID-ASI1607>3.3.CO;2-6]
51. Bramer WM, Rethlefsen ML, Kleijnen J, Franco OH. Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. Syst Rev 2017 Dec 6;6(1):245. [doi: 10.1186/s13643-017-0644-y] [Medline: 29208034]

Abbreviations

AAC: augmentative and alternative communication
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RWTH: Rheinisch-Westfälische Technische Hochschule
SLP: speech-language pathology

Edited by B Lesselroth; submitted 18.06.24; peer-reviewed by D Tafiadis, M Vento-Wilson; revised version received 14.03.25; accepted 06.04.25; published 26.06.25.

Please cite as:
Büchs J, Neuschaefer-Rube C
e-Learning in Phoniatrics and Speech-Language Pathology: Exploratory Analysis of Free Access Tools in Augmentative and Alternative Communication
JMIR Med Educ 2025;11:e63392
URL: https://mededu.jmir.org/2025/1/e63392
doi:10.2196/63392

Evolution of Learning Styles in Surgery Comparing Residents and Teachers: Cross-Sectional Study

Gabriela Gouvea Silva, MD, MSc; Carlos Dario da Silva Costa, MD, MSc; Bruno Cardoso Gonçalves, MD, MSc; Luiz Vianney Saldanha Cidrão Nunes, MD; Emerson Roberto dos Santos, MSc; Natalia Almeida de Arnaldo Rodriguez Castro, MD, MSc; Alba Regina de Abreu Lima, MSc, PhD; Vânia Maria Sabadoto Brienze, MSc, PhD; Antônio Hélio Oliani, MD, MSc, PhD; Júlio César André, MD, MSc, PhD

Center for Studies and Development of Health Education, Faculdade de Medicina de São José do Rio Preto, Avenida Brigadeiro Faria Lima, 5416, São José do Rio Preto, Brazil

Corresponding Author:
Gabriela Gouvea Silva, MD, MSc
Center for Studies and Development of Health Education, Faculdade de Medicina de São José do Rio Preto, Avenida Brigadeiro Faria Lima, 5416, São José do Rio Preto, Brazil

Abstract

Background: Studies confirm a relationship between learning style and medical career choice in the learning style patterns observed in distinct types of residency programs. Such patterns can also be applied to general surgery, from medical school to the latest stages of training. Aligning teaching strategies with the predominant learning styles in surgical residency programs has the potential to make training more effective.

Objective: This study aimed to determine the learning styles of general surgery residents and professors in a Brazilian teaching hospital and compare the results with the existing literature.

Methods: This was a cross-sectional study conducted in a teaching hospital of a public university in Brazil. Thirty-four general surgery residents of any year of training and 30 professors participated in the study. Participants completed a sociodemographic survey and David Kolb's Learning Style Inventory. This was used to classify participants into one of four distinct types of learners: accommodating, diverging, assimilating, and converging. The relationship between sociodemographic data and learning styles was analyzed using the Fisher test, adjusted using the Bonferroni method, and the effect size was measured using the Cramer V test.

Results: The learning style distribution was similar in both groups, with 43,75% diverging, 42,18% accommodating, 10.93% assimilating, and 3.12% converging styles. A significant relationship was found between sex and learning style (P=.049) and between age and learning style for professors (P=.029). The effect sizes were strong (0.46) and very strong (0.506).

Conclusions: The prevalence of learning styles among general surgery residents and professors at this Brazilian hospital differs from that observed in previous studies, with more diverging and accommodating learners and fewer converging learners, suggesting a shift in learning styles. Understanding learning styles is important for effective surgical training programs. Further research with larger and more diverse populations is needed to confirm these results and explore the factors contributing to the observed differences in learning styles.

(JMIR Med Educ 2025;11:e64767) doi:10.2196/64767

KEYWORDS

learning; general surgery; medical education; internship and residency; surgeons; Brazil

Introduction

Background

The concept of learning styles was first developed at the beginning of 1960 as a result of the interest in individual differences while learning [1]. According to Dunn [2], everyone has a unique learning style, like a signature. In this prospect, adjusting the teaching to the different learning styles may help learners and improve educational outcomes. In the current literature, there are various models to determine the learning styles. There is a long and active discussion about whether learning styles are fixed or flexible, and to what extent the context can determine it [3]. Adapting learning styles can enhance student engagement, motivation, and academic performance [4]. Furthermore, the integration of technology and personalized learning approaches has shown promise in enhancing medical education [5].

To provide a more comprehensive understanding, learning styles can be defined as the cognitive, affective, and physiological traits that serve as relatively stable indicators of how learners

perceive, interact with, and respond to the learning environment [6]. Empirical evidence supports the existence of learning styles, demonstrating that individuals exhibit consistent preferences and strengths in how they approach learning tasks. For example, some learners may excel in visual tasks, while others thrive in auditory or kinesthetic activities [7].

The Kolb Experiential Learning Theory (ELT) is a prominent framework for understanding learning styles. The ELT posits that learning is a cyclical process involving four modes: concrete experience (CE), reflective observation (RO), abstract conceptualization (AC), and active experimentation (AE) [8]. Individuals develop preferences for certain modes, leading to four distinct learning styles: converging (AC/AE), diverging (CE/RO), assimilating (AC/RO), and accommodating (CE/AE). The Kolb Learning Style Inventory (LSI) is a widely used tool for assessing these preferences. The validity of Kolb's work in the context of medical education has been demonstrated in numerous studies [9-13], which have found that medical students and professionals exhibit distinct learning style preferences that can influence their academic and professional performance.

Knowledge is the main domain of medical education, but outcomes strongly depend on other domains such as attitude, lifelong learning, and empathy; in surgery, some domains are central including resilience, craftsmanship, and decision-making, among other domains [14].

Research Gap and Problem Statement

Despite the established importance of learning styles in education, limited research has specifically examined their prevalence and impact within general surgery residency programs, particularly in diverse cultural and geographical settings. The clinical and surgical environments present unique challenges for both trainees and educators, requiring the development of complex skills and behaviors [14]. Understanding how surgical residents learn is crucial for optimizing the training process and ensuring the development of competent and well-rounded surgeons.

Moreover, current surgical trainees come from diverse educational, cultural, ethnic, and gender backgrounds, and personal factors also influence their learning characteristics [15]. Little is known about the teaching and learning preferences among surgeons and how they influence the effectiveness of training [16]. Addressing this gap in knowledge is essential for designing effective and inclusive surgical training programs that cater to the diverse needs of learners. To this end, simulation-based surgical training has emerged as a valuable tool for enhancing technical skills and improving patient outcomes [17].

Research Aims and Objectives

Little is known about the teaching and learning preferences among surgeons and how they influence the effectiveness of training [16]. Despite its relevance, studies investigating learning styles in the context of general surgery residency are scarce, especially in countries outside North America and Europe.

Therefore, to address the gaps in understanding learning styles in general surgery, particularly in diverse cultural and geographical settings, this study aims to (1) determine the learning styles of general surgery residents and professors in a Brazilian teaching hospital; (2) compare these findings with existing literature on learning styles in surgery; and (3) discuss the implications of these findings for surgical training programs.

By providing data from a Brazilian teaching hospital, we aim to contribute to a more comprehensive understanding of learning styles in surgical training and inform the development of more effective and inclusive surgical education strategies. This knowledge can inform the development of more effective and inclusive surgical education strategies, ultimately leading to better-prepared and more competent surgeons.

Methods

Study Design and Setting

This cross-sectional study was conducted in 2022 at the Hospital de Base de São José do Rio Preto, a teaching hospital affiliated with Faculdade de Medicina de São José do Rio Preto (a public university in São Paulo, Brazil).

Participants and Recruitment

The study population consisted of general surgery residents in any year of training and hospital professors. All participants were over 18 years old and signed the free and informed consent form.

Data Collection

Data collection involved two instruments: a sociodemographic survey and David Kolb's LSI. The sociodemographic survey collected information on participants' age and sex, and years of residency (for residents) or teaching experience (for professors). The LSI is a validated tool that consists of 12 questions, each with four statements that the participants ranked from 1 to 4 according to their learning preferences. The LSI tool classifies the participants into one of four types of learners based on Kolb's learning cycle: (1) accommodating (learn primarily by experience), (2) diverging (learn by RO), (3) assimilating (learn by exploring associations and interrelationships), and (4) converging (learn by doing or trying things with practical results) [18].

The LSI test was administered in a controlled environment, with a researcher present to provide instructions and clarify any doubts. Participants had 30 minutes to complete it. The sociodemographic survey was completed immediately after completing the LSI test.

Statistical Analysis

Software

Data analysis was performed using the Statistical Package for the Social Sciences (SPSS), version 26.0 (IBM Corp.).

Normality Check

Due to the relatively small sample size and the nature of the data, the Shapiro-Wilk test was used to assess the normality of continuous variables (age and years of experience).

Statistical Tests

A *P* value <.05 was considered statistically significant. The relationship between data was calculated using the Fisher test, adjusted by the Bonferroni method [19]. The Fisher exact test was chosen due to the small sample size and the presence of categories with expected frequencies lower than 5 [20]. The size effect was measured using Cramer V test, which indicates the grade of association between variables: the result is stronger as it approaches the value of 1 [21].

Power

The sample size was calculated using the formula for finite populations, considering a confidence level of 95%, a margin of error of 5%, and an expected prevalence of 50% for each learning style. The minimum sample size was 67 participants, and the total number of residents and professors was 80 [22].

Data Exclusion

Questionnaires that were responded to incorrectly according to Kolb's rules were discarded.

Ethical Considerations

The study was approved by the Research Ethics Committee of Faculdade de Medicina de São José do Rio Preto (approval number: 12345/2022). All participants signed the free informed consent form. Data were anonymized.

Recruitment

This study is grounded in Kolb's ELT, which posits that learning is a cyclical process involving four modes: CE, RO, AC, and AE [23]. Individuals develop preferences for certain modes, leading to four distinct learning styles:

- Converging: Individuals with this learning style excel in AC and AE. They are practical, enjoy problem-solving, and are skilled at applying theories to real-world situations.
- Diverging: Individuals with this learning style excel in concrete CE and RO. They are imaginative, enjoy brainstorming, and are skilled at generating ideas.
- Assimilating: Individuals with this learning style excel in AC and RO. They are logical, enjoy analyzing data, and are skilled at creating models and theories.
- Accommodating: Individuals with this learning style excel in concrete CE and AE. They are hands-on, enjoy taking risks, and are skilled at implementing plans and getting things done.

Our logic model is based on the premise that aligning teaching strategies with the predominant learning styles of surgical residents and professors can enhance the effectiveness of surgical training. We hypothesize that by identifying the learning styles of our participants and tailoring instructional approaches accordingly, we can improve learning outcomes and promote a more engaging and inclusive learning environment.

Table 1 provides a more detailed overview of the four learning styles, including concrete examples of learning activities and instructional approaches that are best suited for each style.

Table . Learning styles, characteristics, and instructional approaches. Source: [8].

Learning style	Characteristics	Example learning activities	Example instructional approaches
Converging	Practical, problem-solver, applies theories	Simulation-based training, case studies	Problem-based learning, hands-on workshops
Diverging	Imaginative, brainstormer, generates ideas	Group discussions, reflective writing	Mentoring, collaborative projects
Assimilating	Logical, analytical, creates models	Literature reviews, data analysis	Lectures, seminars
Accommodating	Hands-on, risk-taker, implements plans	Surgical procedures, clinical rotations	Apprenticeship, on-the-job training

All general surgery residents were invited to answer printed free and informed consent form and the LSI's test, in person. The same was done with the faculty members. The questionnaires were then collected and transformed into digital archives, processed in digital tables after codification.

Statistical Analysis

Power

The sample size was calculated using the formula for finite populations, considering a confidence level of 95%, a margin of error of 5%, and an expected prevalence of 50% for each learning style. The minimum sample size was 67 participants, and the total number of residents and professors was 80 [22].

Results

A total of 64 participants (34 residents and 30 professors) were included in this study. The sociodemographic characteristics of the participants are presented in Table 2. Among the 34 residents, 18 (52.9%) were male and 16 (47.1%) were female. Most residents (91.2%, 31/34) were under 30 years of age. Among the 30 professors, 24 (80%) were male, and 6 (20%) were female, and most of them (17/30, 56.7%) were between 40 and 70 years of age. All professors graduated from universities when traditional teaching methods (ie, primarily lecture-based instruction with limited student interaction) were used, whereas 47% of the residents graduated from universities that used active or mixed teaching methods (ie, incorporating strategies such as problem-based learning, group work, and case studies to promote student engagement).

Table . Sociodemographic characteristics of participants.

Characteristics	Residents (n=34) N (%)	Professors (n=30) N (%)
Age (years)		
<30	31 (91.2)	2 (6.7)
30 - 39	3 (8.8)	11 (36.7)
40 - 70	0 (0)	17 (56.7)
Sex		
Male	18 (52.9)	24 (80)
Female	16 (47.1)	6 (20)
Teaching method used at the University of origin		
Traditional	18 (52.9)	30 (100)
Active or mixed	16 (47.1)	0 (0)

The distribution of Kolb's learning styles is presented in Table 3 and Multimedia Appendix 1. The most prevalent learning styles were diverging (18/34) in the residents' group and accommodating (17/30) in the professors' group.

Table . Learning styles among surgery groups.

Learning styles	Residents N (%)	Professors N (%)	Total N (%)
Converging	1/34 (2.94)	1/30 (3.33)	2/64 (3.12)
Assimilating	5/34 (14.7)	2/30 (6.7)	7/64 (10.93)
Accommodating	10/34 (29.4)	17/30 (56.7)	27/64 (42.18)
Diverging	18/34 (52.9)	10/30 (33.3)	28/64 (43.75)

The relationship between sociodemographic data and learning styles was analyzed using the Fisher exact test (Table 4). A significant association was found between sex and learning style (*P*=.049; Cramer V=0.46), indicating a strong effect size. However, determining which specific categories were significantly different using the Bonferroni post-hoc test was not possible. Among professors, a significant relationship was observed between age and learning style (*P*=.029; Cramer V=0.506), suggesting a very strong effect size. However, specific age groups that differed significantly could not be identified with the Bonferroni post-hoc test, possibly due to the small sample size.

Table . Relationship between sociodemographic data and learning styles.

Variables	*P* value (Fisher exact test)	Effect sizes (Cramer V)
Sex	0.049[a]	0.46 (strong)
Age (residents)	0.999	0.12 (weak)
Age (professors)	0.029[a]	0.506 (very strong)
Teaching method used at the university of origin (residents)	0.999	0.08 (weak)

[a]Statistically significant at P<.05

Discussion

Principal Findings

Our study, utilizing Kolb's LSI, identified the distribution of four learning styles among general surgery residents and professors at a Brazilian teaching hospital. These learning styles are:

- Diverging: Learners who excel in CE and RO, are imaginative, and generate ideas effectively.
- Accommodating: Learners who excel in CE and AE, are hands-on, and enjoy implementing plans.
- Assimilating: Learners who excel in AC and RO, are logical, and create models and theories.
- Converging: Learners who excel in AC and AE, are practical, and apply theories to real-world situations.

The most prevalent learning styles were diverging (52.9%) in the residents' group and accommodating (56.7%) in the professors' group (Table 2 and Multimedia Appendix 1). A significant association was found between sex and learning style

(*P*=.049; Cramer V=0.46), indicating a strong effect size. Among professors, a significant relationship was observed between age and learning style (*P*=0.029; Cramer V=0.506), suggesting a very strong effect size.

Table 2 and Multimedia Appendix 1 show that while diverging was the most common style among residents and accommodating was most common among professors, the overall learning style distribution was relatively similar between the two groups. This convergence, where both residents and professors exhibit a blend of diverging and accommodating tendencies, can potentially facilitate both teaching and learning [16]. The shared presence of these styles suggests that both groups may value CEs and RO (diverging) as well as hands-on activities and practical application (accommodating).

This alignment can be leveraged to support instruction in different ways. For diverging learners (both residents and some professors), emphasize group discussions, brainstorming sessions, and reflective writing assignments. Encourage the sharing of diverse perspectives and the exploration of different approaches to surgical problems. In contrast, for accommodating learners (both professors and some residents), Provide opportunities for hands-on practice, simulation-based training, and real-world clinical rotations. Encourage AE and problem-solving in practical settings. By incorporating these strategies, educators can create a learning environment that caters to the predominant learning styles of both residents and professors, fostering more effective communication, engagement, and knowledge acquisition.

However, it is important to acknowledge that the similarity in distribution does not guarantee a perfect match for all individuals. The relatively lower prevalence of converging and assimilating styles in both groups suggests that those learners might require more tailored support and learning opportunities to ensure their needs are met. This underscores the importance of mapping learning styles when designing a comprehensive residency program, as it provides a basis for guiding the learning needs of all residents and professors, not just the majority.

Implications of Findings

The findings of our study have important implications for surgical education. Understanding the predominant learning styles of residents and professors can help adapt teaching strategies and curriculum design to better meet their needs. For example, incorporating more RO and practical experiences can benefit diverging and accommodating learners while also providing opportunities for AC and AE to support assimilating and converging learners.

Furthermore, with the occurrence of the pandemic, the increased distances imposed by contact restrictions have further deepened these changes. The COVID-19 pandemic has also presented unique challenges to surgical training, with restrictions on in-person learning and clinical experiences. A pan-Romanian survey by Moldovan et al [24] highlighted the impact of the pandemic on orthopedic residents, including reduced surgical volume, limited access to educational resources, and increased stress and anxiety. These challenges may have further influenced the learning styles and preferences of surgical residents and professors during this period.

Comparison With Prior Work

Few studies on learning styles in surgery were found in the literature, but we can state that our results are different from previous results.

In the 1980s, Baker III et al [25] reported a prevalence of converging (46%), followed by accommodating (26%) and assimilating (20%) styles among surgeons. In the 2000s, this pattern was confirmed by Contessa et al [26]. They argued that surgical practice requires quick decision-making and problem-resolution, justifying the converging style and its more pragmatic view. In 2007, Mammen et al [27] published similar results obtained in the US population.

After Quillin [28] reduced his working hours in general surgery residency, he showed the results collected from 1999 to 2012. At that time, the proportion of accommodating learners was higher, especially after 2003, when the workload was reduced [28].

In 2017, for the first time, diverging learners became the majority in a study with 47 surgeons in the United Kingdom [29]. In 2018, also in the United Kingdom, a study with residents in various surgical areas found that converging, followed by accommodating styles were predominant [30]. In 2020, similar results were published in Scotland by Hopkins et al [15]. The most recent publication on the topic reported a predominance of assimilating followed by converging styles in Spain [31]. Table 5 and Multimedia Appendix 2 show the existing literature.

Table . Data of learning styles in surgery through time around the world.

Author	Publication year	Country	Population (n)	Diverging	Accomodator	Assimilating	Converging
Baker III	1985	USA	Surgeons (39)	8%	26%	20%	46%
Drew	1999	UK	Basic surgical trainees (52)	3.9%	27%	9.6%	59.5%
Mammen	2007	USA	General surgery residents (91)	10.6%	14.6%	17.2%	57.8%
Brown	2018	UK	Medical students (60)	20.8%	30.2%	17%	32%
Parra	2021	Spain	Surgical residents and staff (64)	14.1%	21.9%	39.1%	25%

The results of the present study were diverging (43,8%), accommodating (42,2%), assimilating (11,0%), and converging (3,12%) styles. These results amplify the existing literature, showing an increase in diverging and a decrease in converging styles over the years. These findings indicate a shift in the learning preferences of surgical residents and professors, which may have been influenced by various factors, such as changes in surgical education, technological advancements, and sociocultural aspects.

The geographical location may be a possible explanation for our results, as previous studies were conducted in North America and Europe (Multimedia Appendix 1). Cultural differences and variations in surgical training programs across countries may have contributed to the observed differences in learning styles. Another hypothesis may be the course of time: the last two decades have seen huge technological changes, when social media, smartphones, and laptops became widely available, greatly impacting the teaching-learning process [32]. Recent studies have further explored the impact of digital technologies on medical education, highlighting both the opportunities and challenges associated with their integration [5]. Furthermore, with the occurrence of the pandemic, the increased distances imposed by contact restrictions have further deepened these changes [33].

The differing proportions of female residents (47.1%) and professors (16.0%) highlight the ongoing evolution of gender representation in surgery. The historical underrepresentation of women in surgical fields may contribute to differences in observed learning styles between residents and professors [34]. Despite this, the increasing participation of women in surgery over recent decades is a positive trend [22].

Strengths and Limitations

The population included is a small sample of a larger Brazilian surgical group. More data can be further collected to compare the country with other nations, in America, Europe or even Asia. The medical reality in Brazil is diverse and worth a broader approach.

In addition to the small sample size, our study has several other limitations that should be acknowledged. The study had a sampling bias. Our sample was drawn from a single teaching hospital in Brazil, which may not be representative of all general surgery residents and professors in Brazil or other countries. This limits the generalizability of our findings. Additionally, the voluntary nature of participation may have introduced selection bias, as those who chose to participate may differ systematically from those who did not. In addition, there was measurement bias; the Kolb LSI is a self-report instrument, which is subject to social desirability bias and response bias. Participants may have answered the questions in a way that they perceived as more favorable or aligned with societal expectations, rather than reflecting their true learning preferences. Moreover, our study did not fully explore the potential influence of various sociodemographic factors, such as cultural background, socioeconomic status, and prior educational experiences, on learning styles. However, we did not collect data on other potentially relevant sociodemographic factors such as ethnicity, social class, migration background, or detailed information about prior educational experiences. These factors may interact with learning styles in complex ways and could have influenced our results.Finally, the cross-sectional design of our study limits our ability to draw causal inferences about the relationship between learning styles and other variables. A longitudinal study would be needed to examine how learning styles evolve over time and how they impact training outcomes. These limitations should be considered when interpreting our findings.

Further research is needed to explore the underlying factors that influence these learning styles, such as personality traits, prior educational experiences, and cultural background. Understanding these factors could allow for more tailored interventions to optimize learning. Moreover, future studies should investigate the potential impact of different learning styles on surgical performance metrics, such as technical skill acquisition, error rates, and patient outcomes. This would provide valuable insights into how learning style preferences translate into real-world surgical practice.

Conclusions

This study found that diverging and accommodating learning styles were more prevalent among general surgery residents and professors in a Brazilian university hospital, differing from previous North American and European studies. The decreased prevalence of the converging style is notable and may be due to changes in surgical education, technology, and cultural differences. Understanding these learning styles can guide more effective and inclusive teaching strategies in surgical residency programs. Further research with larger, diverse populations is needed to explore the relationships between learning styles, demographics, and training outcomes.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Learning styles prevalent in surgery groups.
[PNG File, 56 KB - mededu_v11i1e64767_app1.png]

Multimedia Appendix 2
Timeline of surgical learning styles according to the existent literature.

[PNG File, 72 KB - mededu_v11i1e64767_app2.png]

References

1. Curry L. An organization of learning styles theory and constructs. 1983 URL: https://files.eric.ed.gov/fulltext/ED235185.pdf [accessed 2025-05-05]
2. Dunn R. Understanding the dunn and dunn learning styles model and the need for individual diagnosis and prescription. Journal of Reading, Writing, and Learning Disabilities International 1990 Jan;6(3):223-247 [FREE Full text] [doi: 10.1080/0748763900060303]
3. Bernacki ML, Greene MJ, Lobczowski NG. A systematic review of research on personalized learning: personalized by whom, to what, how, and for what purpose(s)? Educ Psychol Rev 2021 Dec;33(4):1675-1715. [doi: 10.1007/s10648-021-09615-8]
4. El-Sabagh HA. Adaptive e-learning environment based on learning styles and its impact on development students' engagement. Int J Educ Technol High Educ 2021 Dec;18(1):53. [doi: 10.1186/s41239-021-00289-4]
5. Chowdhury PN, Vaish A, Puri B, Vaishya R. Medical education technology: past, present and future. Apollo Medicine 2024 Dec;21(4):374-380. [doi: 10.1177/09760016241256202]
6. (U.S.) NA of SSP. student learning styles: diagnosing and prescribing programs. National Association of Secondary School Principals. 1979. URL: https://books.google.com.br/books?id=PFYmAQAAIAAJ [accessed 2025-04-30]
7. Felder RM, Silverman LK. Learning and teaching styles in engineering education. 1988. URL: https://api.semanticscholar.org/CorpusID:140475379 [accessed 2025-04-30]
8. Kolb D. Experiential learning: experience as the source of learning and development. In: J Bus Ethics 1984, Vol. 1. URL: https://www.researchgate.net/publication/235701029_Experiential_Learning_Experience_As_The_Source_Of_Learning_And_Development [accessed 2025-05-05]
9. Villanueva F. The Use of Kolb's Learning Styles Inventory (LSI) in School Settings: BU Journal of Graduate Studies in Education, Vol. 12. URL: https://files.eric.ed.gov/fulltext/EJ1263067.pdf [accessed 2025-05-05]
10. Shakeri F, Ghazanfarpour M, MalaKoti N, Soleimani Houni M, Rajabzadeh Z, Saadat S. Learning Styles of Medical Students: A Systematic Review. Med Edu Bull 2022;3(2):441-456. [doi: 10.22034/MEB.2022.328652.1050]
11. Vemuri VR, Rao KA. Assessment of Learning Styles Using Kolb's Learning Style Inventory among Medical College Students: A Cross-sectional Study. J Clin Diagn Res . [doi: 10.7860/JCDR/2024/67440.18987]
12. Cortés B, Olaya G, Olaya M, Fabricio J. Estilos de aprendizaje en estudiantes de medicina [Article in Spanish]. Universitas Médica 2018;59(2):1-10.
13. Vemuri VR, Rao KA. J Clin of Diagn Res 2024;18(2):JC01-JC04 [FREE Full text]
14. Palejwala Z, Wallman KE, Maloney S, et al. Higher operating theatre temperature during burn surgery increases physiological heat strain, subjective workload, and fatigue of surgical staff. PLoS ONE 2023;18(6):e0286746. [doi: 10.1371/journal.pone.0286746] [Medline: 37267345]
15. Hopkins L, Robinson D, James O, Egan R, Lewis W, Bailey D. Surgical learning styles; a gender gap perspective. 2020. URL: https://www.webofscience.com/wos/woscc/full-record/WOS:000595958400223 [accessed 2025-05-05]
16. Dickinson KJ, Bass BL, Graviss EA, Nguyen DT, Pei KY. How learning preferences and teaching styles influence effectiveness of surgical educators. Am J Surg 2021 Feb;221(2):256-260. [doi: 10.1016/j.amjsurg.2020.08.028] [Medline: 32921405]
17. Foppiani J, Stanek K, Alvarez AH, et al. Merits of simulation-based education: a systematic review and meta-analysis. J Plast Reconstr Aesthet Surg 2024 Mar;90:227-239. [doi: 10.1016/j.bjps.2024.01.021] [Medline: 38387420]
18. Romanelli F, Bird E, Ryan M. Learning styles: a review of theory, application, and best practices. Am J Pharm Educ 2009 Feb 19;73(1):9. [doi: 10.5688/aj730109] [Medline: 19513146]
19. Argyrous G. The chi-square test for independence. In: Statistics for Social Research: Macmillan Education UK; 1997:257-284. [doi: 10.1007/978-1-349-14777-9_16]
20. Kim HY. Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test. Restor Dent Endod 2017 May;42(2):152-155. [doi: 10.5395/rde.2017.42.2.152] [Medline: 28503482]
21. Dias D, Silva JS, Bernardino A. The prediction of road-accident risk through data mining: a case study from Setubal, Portugal. Informatics (MDPI) 2023;10(1):17. [doi: 10.3390/informatics10010017]
22. Xepoleas MD, Munabi NCO, Auslander A, Magee WP, Yao CA. The experiences of female surgeons around the world: a scoping review. Hum Resour Health 2020 Oct 28;18(1):80. [doi: 10.1186/s12960-020-00526-3] [Medline: 33115509]
23. Kolb AY, Kolb DA. Learning styles and learning spaces: enhancing experiential learning in higher education. In: Academy of Management Learning and Education 2005, Vol. 4:193-212. [doi: 10.5465/amle.2005.17268566]
24. Moldovan F, Gligor A, Moldovan L, Bataga T. The impact of the COVID-19 pandemic on the orthopedic residents: a pan-romanian survey. Int J Environ Res Public Health 2022 Jul 27;19(15):9176. [doi: 10.3390/ijerph19159176] [Medline: 35954536]
25. Baker JD, Reines HD, Wallace CT. Learning style analysis in surgical training. Am Surg 1985 Sep;51(9):494-496. [doi: 10.5688/aj730109] [Medline: 4037546]

26. Contessa J, Ciardiello KA, Perlman S. Surgery resident learning styles and academic achievement. Curr Surg 2005;62(3):344-347. [doi: 10.1016/j.cursur.2004.09.012] [Medline: 15890222]
27. Mammen JMV, Fischer DR, Anderson A, et al. Learning styles vary among general surgery residents: analysis of 12 years of data. J Surg Educ 2007;64(6):386-389. [doi: 10.1016/j.jsurg.2007.08.005] [Medline: 18063274]
28. Quillin RC, Pritts TA, Young GB, Edwards MJ, Davis BR. Surgical resident learning styles have changed following the implementation of the 80-hour workweek. Journal of Surgical Research 2013 Feb;179(2):230. [doi: 10.1016/j.jss.2012.10.420] [Medline: 23721929]
29. Challapalli P. Evaluation of learning styles in a surgical unit. Abtracts from the Medical Research Symposium for Students and Foundation Doctors. 2017. URL: https://www.rcpe.ac.uk/sites/default/files/medical_research_symposium_abstracts_8_feb.pdf [accessed 2025-05-01]
30. Brown C, Luton OW, Robinson D, et al. Learning style variation: a British core surgical trainee perspective. J Am Coll Surg 2018;227(4):S221. [doi: 10.1016/j.jamcollsurg.2018.07.484]
31. Martín Parra JI, Toledo Martínez E, Martínez Pérez P, et al. Analysis of learning styles in a laparoscopic technical skills course. Implications for surgical training. Cirugía Española (English Edition) 2021 Dec;99(10):730-736. [doi: 10.1016/j.cireng.2021.10.011]
32. Frenk J, Chen LC, Chandran L, et al. Challenges and opportunities for educating health professionals after the COVID-19 pandemic. Lancet 2022 Oct 29;400(10362):1539-1556. [doi: 10.1016/S0140-6736(22)02092-X] [Medline: 36522209]
33. Trott M, Driscoll R, Irlado E, Pardhan S. Changes and correlates of screen time in adults and children during the COVID-19 pandemic: a systematic review and meta-analysis. EClinicalMedicine 2022 Jun;48:101452. [doi: 10.1016/j.eclinm.2022.101452] [Medline: 35615691]
34. Motter SB, Brandão GR, Iaroseski J, et al. Women representation in academic and leadership positions in surgery in Brazil. Am J Surg 2022 Jan;223(1):71-75. [doi: 10.1016/j.amjsurg.2021.07.023] [Medline: 34315578]

Abbreviations

AC: abstract conceptualization
AE: active experimentation
CE: concrete experience
ELT: Experiential Learning Theory
LSI: learning style inventory
RO: reflective observation

Edited by B Lesselroth; submitted 25.07.24; peer-reviewed by F Moldovan, J Olivier, M Jani; revised version received 11.03.25; accepted 06.04.25; published 08.05.25.

Please cite as:
Gouvea Silva G, da Silva Costa CD, Cardoso Gonçalves B, Vianney Saldanha Cidrão Nunes L, Roberto dos Santos E, Almeida de Arnaldo Rodriguez Castro N, de Abreu Lima AR, Sabadoto Brienze VM, Oliani AH, André JC
Evolution of Learning Styles in Surgery Comparing Residents and Teachers: Cross-Sectional Study
JMIR Med Educ 2025;11:e64767
URL: https://mededu.jmir.org/2025/1/e64767
doi:10.2196/64767

Exploring Connections Between Mental Health, Burnout, and Academic Factors Among Medical Students at an Iranian University: Cross-Sectional Questionnaire Study

Elham Faghihzadeh[1], PhD; Ali Eghtesad[2], MD; Muhammad Fawad[3], PhD; Xiaolin Xu[3], PhD

[1]Department of Biostatistics and Epidemiology, Zanjan University of Medical Sciences, Zanjan, Iran

[2]School of Medicine, Zanjan University of Medical Sciences, Zanjan, Iran

[3]School of Public Health, The Second Affiliated Hospital, Zhejiang University School of Medicine, 866 Yuhangtang Road, Hangzhou, Zhejiang, China

Corresponding Author:
Muhammad Fawad, PhD
School of Public Health, The Second Affiliated Hospital, Zhejiang University School of Medicine, 866 Yuhangtang Road, Hangzhou, Zhejiang, China

Abstract

Background: Medical students face high levels of burnout and mental health issues during training. Understanding associated factors can inform supportive interventions.

Objective: This study aimed to examine burnout, psychological well-being, and related demographics among Iranian medical students.

Methods: A cross-sectional survey was conducted among 131 medical students at an Iranian University. The instruments used included the Maslach Burnout Inventory-Student Survey and the Symptom Checklist-90-Revised. Descriptive statistics, multivariate regression, and tests for group differences were used to analyze the data.

Results: The MBI-SS subscale scores indicated moderate emotional exhaustion, mean 15.00 (SD 7.08) and academic efficacy, mean 14.98 (SD 6.29), with lower cynicism, mean 10.85 (SD 5.89). The most commonly reported mental health issues were depression and obsessive-compulsive disorder. Poor psychological well-being was associated with higher overall burnout, but no significant gender differences were found. Burnout levels varied by academic year across all Maslach Burnout Inventory-Student Survey domains.

Conclusions: Despite their health education, medical students in this study reported significant burnout and mental health distress, with strong associations between the two. These issues may impact student retention and post-graduation practice plans. Supporting well-being during training is critical for positive student and physician outcomes.

(JMIR Med Educ 2025;11:e58008) doi:10.2196/58008

KEYWORDS

emotional exhaustion; exhaustion; cynicism; academic efficacy; burnout; physician burnout; mental health; mental illness; mental disease; mental disorder; medical education; medical knowledge; medical training; medical student; resident physician; resident doctor; residency; residency training

Introduction

Kary and Pines [1] initially posited the concept of academic tedium and its impact on students. They suggested that this phenomenon is not confined to a specific educational level but can manifest at various stages of schooling, including both school and university environments [1]. Based on the author's viewpoint, students might be struggling with a condition marked by a fading enthusiasm for learning, noticeable lack of motivation, and an overwhelming sense of emotional exhaustion. Later, Maslach and Jackson [2,3] specified burnout as the experience of physical and emotional drain caused by chronic stress. Burnout is the state of physical and mental fatigue caused by work, study, or any caregiving activities. It can also be known as an adverse emotional, cognitive, and physical reaction to the study, work, and life pressures. Burnout was officially classified by the World Health Organization as an occupational phenomenon in 2019 and included in the International Classification of Diseases (*ICD-11*).

Educational burnout is a type of burnout experienced during studying. To better view educational burnout, it was expanded to three factors: emotional exhaustion, cynicism, and feelings of inefficacy [4,5]. Emotional exhaustion reflects feelings of exceeding emotional resources due to academic demands. Cynicism is a negative, unresponsive, or overly snapped response to a phenomenon. Feelings of inefficacy refer to a

reduction in academic effort, leading to a sense of incompetence and reduced academic achievement. Based on findings from a systematic review published in 2021, it was determined that educational burnout affected more than 40% of students [6]. This outcome implies a heightened susceptibility to burnout among medical students on a global scale. However, few studies have examined this problem, specifically among medical students in Iran. While 16% of Iranian medical students reported burnout in one study [7], assessing prevalence rates at individual universities could further inform supportive programs. Educational burnout has an essential role in medical students' overall health and could easily impact the quality of their learning [8].

A study on 14,000 students from different countries showed that approximately 35% of the students had been diagnosed with at least one mental health disorder, such as depression or anxiety [9]. Among students, university students showed a higher likelihood of mental health disorders, and among them, medical students' issues were significant. Medical schools pose multiple demands on students. First, enrollment in medical training coincides with adolescence and early adulthood, periods already associated with vulnerability to mental health disorders [10-12]. Second, the intense nature of medical education requires students to assimilate vast amounts of health information while coping with exposure to myriad diseases [11,12]. Consequently, studies report substantial rates of depression (11% - 37%), anxiety (7.4% - 30%), and other issues in this population internationally [13-15]. Evidence suggests that positive mental health aids coping [16], yet remains understudied in Iranian cultures.

Extensive evidence demonstrates intricate connections between burnout and mental health issues among medical students. Additional studies reveal substantially higher risks of depression, anxiety, suicidal ideation, concentration deficits, and physical symptoms compared to their peers [12,17-20]. Up to half of graduating students experience burnout, linking this syndrome to exacerbated mental health decline [18]. Ultimately, these concerning rates significantly exceed general population trends, underscoring the crisis of psychological well-being in medical education. Implementing supportive interventions requires further investigating specific student populations.

The aims of this study are twofold. Primarily, we assess the prevalence of mental health issues and burnout among native Iranian medical students at Zanjan University of Medical Sciences. Additionally, we delineate connections between mental health status and burnout risk by evaluating associated academic and personal factors. By understanding these relationships, targeted interventions can eventually be developed to promote the psychological well-being of Iran's future physicians during their demanding training period.

Methods

Study Design and Participants

This cross-sectional study was conducted at Zanjan University of Medical Sciences, Zanjan, Iran, focusing on the experiences of 1500 medical students. These trainees constituted the target of our research, with their perspectives and characteristics as students comprising the central subject of investigation. Participants were recruited using a convenience sampling method. Our research team directly contacted the students, explained the study's aims, invited their voluntary participation, and emphasized the confidentiality of their responses. We then sent an electronic survey link to consenting participants. Strict data quality control measures were implemented, with incomplete questionnaire submissions excluded from the analysis to uphold the integrity of the results. Based on a previous study [21], the minimum required sample size was 120 students; however, 140 surveys were distributed, and 131 fully completed questionnaires were returned. Those with missing data or students who indicated having a diagnosed mental health issue were excluded.

Measures

Demographics

The basic sociodemographic information included age, sex, residence, history of a positive COVID-19 test, underlying diseases, diagnosis of mental health issues, and level of education. The levels of education were categorized into three sections: the initial seven semesters, referred to as preclinical, followed by a two-semester externship, and finally, a three-semester internship. At the preclinical level, students learned about basic sciences and pathophysiology; in the externship phase, they would pass a short course in each hospital unit. The residential status comprises a parental home, independent home, or a dormitory. Students were asked directly about underlying diseases, including diabetes, hypertension, and chronic disease. Additionally, they were asked whether they had been diagnosed with any mental health condition and whether they had received any treatment.

Burnout Measurement

Burnout symptoms were measured by the Persian version of the MBI-SS [3,22,23]. It comprises 15 items, which are divided into three dimensions: emotional exhaustion, cynicism, and academic efficacy . Each item has been rated on a 7-pointed Likert scale. Academic efficacy scores were reverse-coded; therefore, it was scored oppositely. A high score in three dimensions indicated greater burnout. The maximum possible scores for emotional exhaustion, cynicism, and academic efficacy were 30, 24, and 36, respectively.

Mental Health Measurement

The Symptom Checklist 90 (SCL-90), developed by Derogatis was used to assess mental health. This scale consists of 90 items, each rated on a 5-point Likert scale, effectively measuring ten primary psychological symptoms [24]. The ten psychological symptoms measured by the Symptom Checklist-90-Revised (SCL-90-R) are somatization, obsessive-compulsive, interpersonal sensitivity, depression, anxiety, hostility, phobic anxiety, paranoid ideation, psychoticism, and sleep problems. If a person's average score for the questions related to these symptoms was greater than 2, it indicated potential psychological issues. The Global Severity Index (GSI) was calculated for our analysis, which measures the extent or depth of psychiatric disturbances. Specifically, the GSI is the average

score across all responded items and serves as an overall measure of psychiatric distress. Therefore, this study analyzed the positive rate of each subscale and the GSI. Notably, the validated Persian version of the SCL-90-R was used for this student population [25].

Statistical Analysis

We performed all statistical analyses using SPSS (version 20.0; IBM Corp) and Stata (version 12; StataCorp), and figures were drawn using R software (version 4.4.1; R Foundation for Statistical Computing) was used for visualization, including the ggplot2 package (version 3.5.1). Multivariable regression was used to assess factors associated with the three burnout subscales to examine the correlation between the response variables. The model included sociodemographic variables such as history of COVID-19, place of residence, and mental health status. The variables that were found to have a significant impact on the outcome were retained in the model.

Ethical Considerations

Ethics approval for this study was provided by the Ethics Committee of Zanjan University of Medical Sciences (IR.ZUMS.REC.1400.418). This committee approved all experimental protocols. The authors confirmed that relevant guidelines and regulations were used in all experiments. No participants were younger than 16 years. All students provided written informed consent.

Results

The study initially involved 140 students; after excluding those who dropped out due to missing answers or diagnosed mental illness, the remaining sample consisted of 131 participants. The average age of these remaining participants was approximately 24 years, with a mean of 23.95 (SD 3.69) years. Table 1 summarizes other sociodemographic characteristics of the student group. Approximately 66% were female students, while only 10% of the participants were married. An almost equal percentage of students across different academic levels completed the questionnaires. Additionally, 62% of the students had a history of a positive COVID-19 test, while 3.8% reported underlying diseases.

Table . Socio-demographic characteristics.

Sociodemographic variables	Participants (N=131), n (%)
Sex	
Male	44 (33.6)
Female	87 (66.4)
Marital status	
Single	118 (90.1)
Married	13 (9.9)
Residence	
Parental home	53 (40.5)
Own home	43 (32.8)
Dormitory	35 (26.7)
Positive COVID-19 test history	
No	50 (38.2)
Yes	81 (61.8)
Underlying diseases	
No	126 (96.2)
Yes	5 (3.8)
Academic level	
Preclinical	42 (32.1)
Externship	47 (35.9)
Internship	42 (32.1)

Table 2 shows the positive rates of SCL-90-R subscales by sex. Obsessive-compulsive disorder and depression showed the highest prevalence among symptoms. A χ^2 test examined the percentage differences between male and female students. The only symptom found to be statistically significant between the two sex was phobic anxiety. Among female students, paranoid ideation had the highest prevalence, whereas obsessive-compulsive disorder was more prevalent among male students.

Table . Comparison of SCL-90-R[a] subscales based on sex.

	SCL-90-R positive rates in male students n (%)	SCL-90-R positive rates in female students, n (%)	Total SCL-90-R positive rate, n (%)
Hostility	10 (22.7)	13 (14.9)	23 (17.6)
Anxiety	11 (25.0)	14 (16.1)	25 (19.1)
Obsessive-compulsive disorder	14 (31.8)	18 (20.7)	32 (24.4)
Interpersonal sensitivity	10 (22.7)	19 (21.8)	29 (22.1)
Somatization	6 (13.6)	12 (13.8)	18 (13.7)
Psychoticism	5 (11.4)	7 (8.0)	12 (9.2)
Paranoid ideation	8 (18.2)	21 (24.1)	29 (22.1)
Depression	12 (27.3)	20 (23.0)	32 (24.4)
Phobic anxiety[b]	9 (20.5)	7 (8.0)	16 (12.2)
Others	8 (18.2)	11 (12.6)	19 (14.5)

[a]SCL-90-R: Symptom Checklist-90-Revised.

[b]*P* value (χ^2 test)=.04 male versus female.

The boxplots in Figures 1 and 2 display the MBI-SS subscale scores across genders and academic levels. According to this figure, academic efficacy had the widest range of scores for male students. Additionally, female students exhibited lower mean scores compared to male students across all subscales. These boxplots indicate that interns had higher burnout scores overall. More detailed descriptive statistics can be found in Multimedia Appendix 1. Figure 3 shows a comparison of the total scores on the SCL-90-R between different levels, revealing that externs had the highest scores.

Figure 1. Comparison of Maslach Burnout Inventory-Student Survey subscale scores by sex.

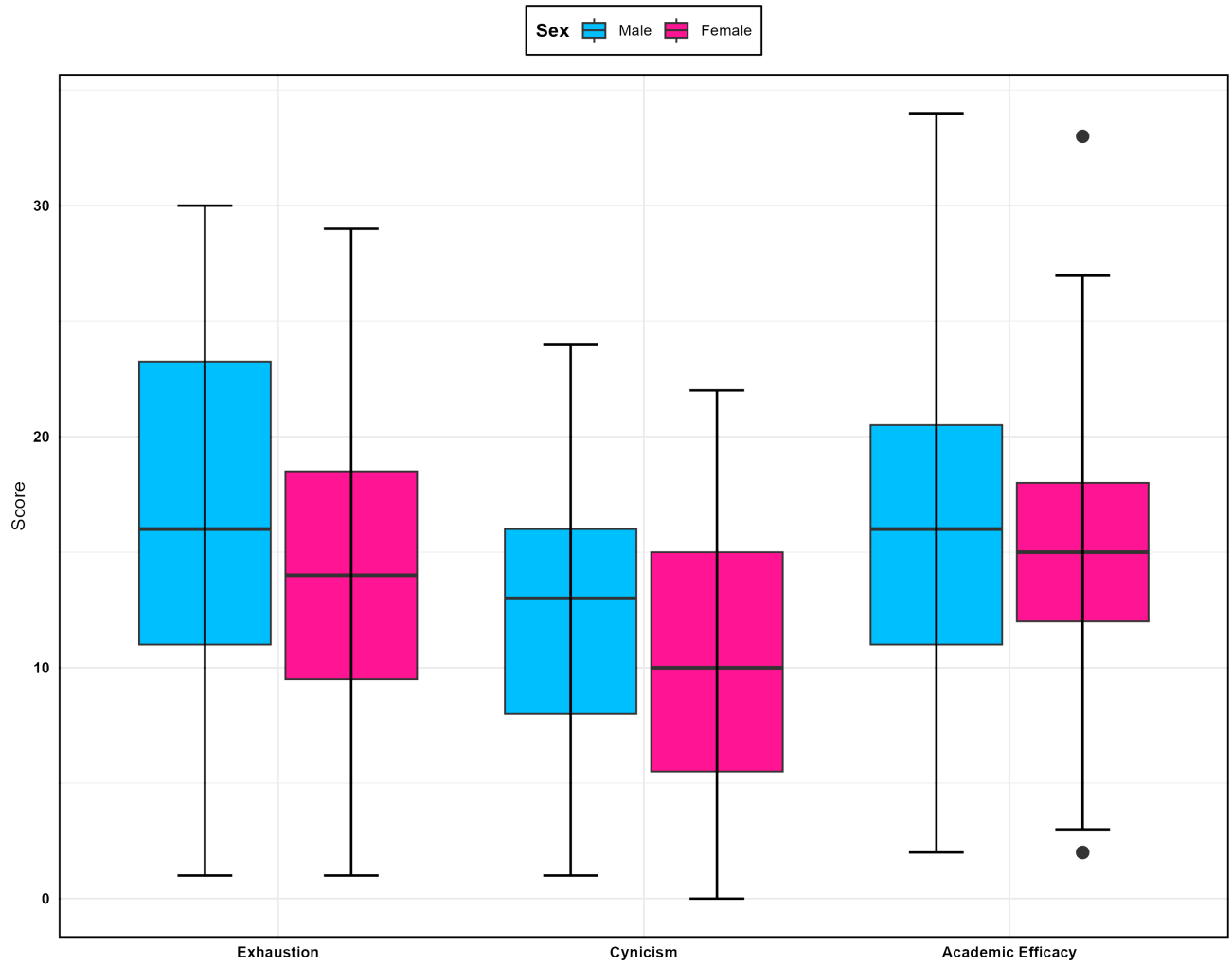


Figure 2. Comparison of Maslach Burnout Inventory-Student Survey subscale scores across academic levels.

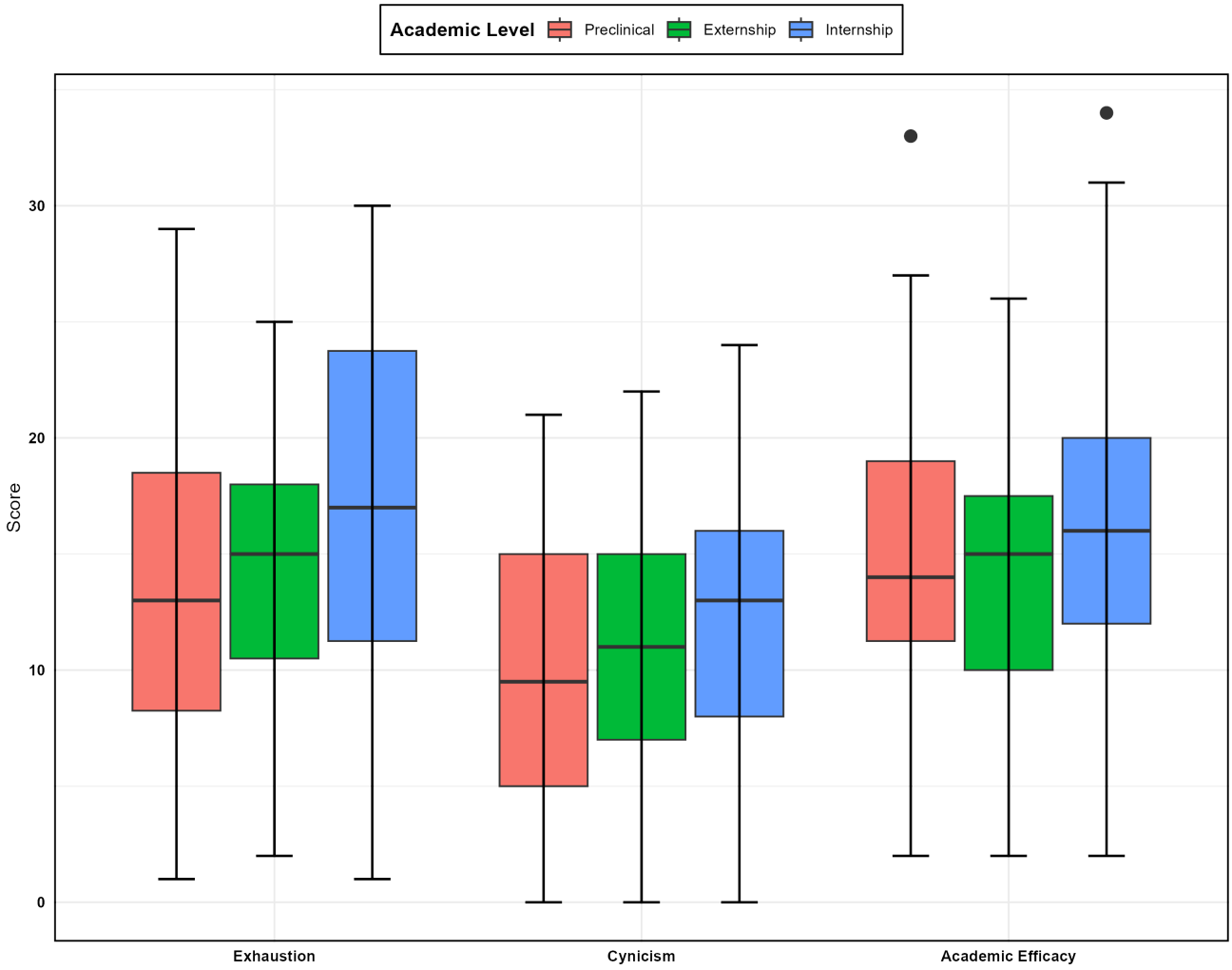
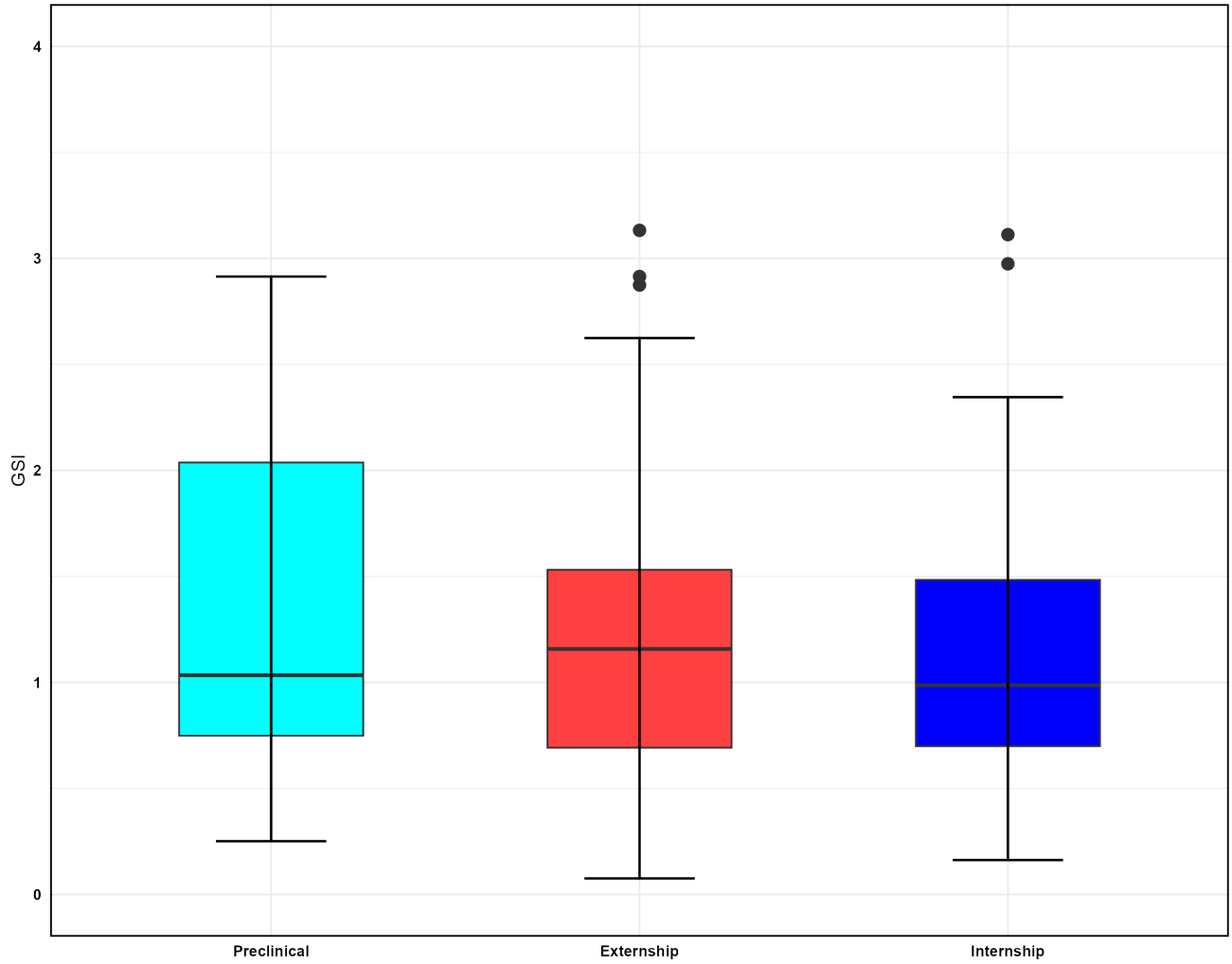


Figure 3. Comparison of Global Severity Index scores across academic levels.



Initially, the correlation between the three dimensions of burnout (academic efficacy, cynicism, and emotional exhaustion) was examined in the modeling data. The findings revealed that academic efficacy had a significant correlation with both cynicism (r=0.41, P<0.05) and emotional exhaustion (r=0.37, P<0.05). Additionally, emotional exhaustion positively correlated with cynicism (r=0.78, P<0.05). Given these significant correlations among the burnout dimensions, a multivariable regression analysis was deemed appropriate for further modeling. Table 3 presents the results of the multivariable regression analysis. An increase of one score in GSI corresponded to an increase of 5.67, 1.71, and 4.69 scores in emotional exhaustion, cynicism, and academic efficacy, respectively. Overall, the students in the internship phase has 4.19 and 3.02 scores higher than preclinical students in emotional exhaustion and academic efficacy, respectively, whereas they had only a 0.24-difference in cynicism. Furthermore, males scored 1.53 and 0.10 points lower than female students, respectively. A comparison of β coefficients shows that GSI and internship status had significantly different effects on the three dimensions of the MBI-SS.

Table . The association of educational burnout with mental well-being, academic level, and sex.

	Emotional exhaustion				Cynicism				Academic Efficacy				Equality of β coefficients	
	β coefficient	95% CI	*F* value	*P* value	β coefficient	95% CI	*F* value	*P* value	β coefficient	95% CI	*F* value	*P* value	*F* value	*P* value
GSI[a]	5.67	4.02-7.32	6.55	<.001	1.71	1.41-2.00	5.56	<.001	4.69	3.31-6.07	4.79	<.001	18.72	<.001
Academic levels														
Preclinical	-	-	-		-	-	-		-	-	-			
Externship	1.17	−1.83 to 4.18	0.68	.44	−0.32	−0.86 to 0.21	−1.00	.23	0.56	−1.95 to 3.08	0.52	.66	0.59	.49
Internship	4.19	0.97-7.41	2.41	.01	−0.24	−0.81 to 0.33	−1.98	.40	3.02	0.33-5.71	3.10	.03	2.01	.04
Sex														
Female	-	-	-		-	-	-		-	-	-			
Male	−1.53	−4.25 to 1.20	−1.29	.27	−0.10	−0.58 to 0.38	-1.40	.69	2.34	0.06-4.61	2.02	.04	0.99	.21
R^2	0.303				0.523				0.513					

[a]GSI: Global Severity Index.

Discussion

Principal Findings

Zanjan University, a prominent institution in Iran, attracts students from various cities. Therefore, studying its students' mental and physical well-being can provide valuable insights into the overall condition of Iranian students.

Our study examined crucial mental health issues such as hostility, obsessive-compulsive disorder, and interpersonal sensitivity among medical students. In brief, our study did not detect any statistically significant differences in overall mental health scores between students across different academic level or sex. However, phobic anxiety was the only mental health issue that was significantly different between genders. Students at the externship level had higher GSI scores. This increase in mental health problems among the students is understandable, as it occurred during their clinical rotations in hospitals, where they were exposed to diverse patient cases and experienced various illnesses for the first time in their academic careers. Facing such novel and potentially challenging situations can reasonably be expected to take an emotional toll.

In previous studies, depression, stress, and anxiety are the three most prevalent mental health issues among medical students [7,11-13,15,26,27]. Cuttilan et al [13], who reviewed studies from Asia in their meta-analysis, showed that 30% of Middle Eastern students experienced depression. While our study found a slightly lower prevalence, the difference could be caused by the university environment, sample size, and social or climatic differences. Nonetheless, the rate of depression remains considerable. Aghajani Liasi et al [7], who studied the prevalence of burnout and mental health at one of Tehran's universities, reported a 37% of depression among medical students. They used the Depression, Anxiety, and Stress Scale questionnaire to survey mental health. Therefore, despite being of the same nationality, the main reason for the variation between the findings of our study compared to those reported by Aghajani Liasi et al may be due to differences in the questionnaires.

Anxiety is one of the significant issues experienced by medical students. A systematic review revealed the wide range of anxiety prevalence across different countries (15.5%-70.0%) [27]. Our study found an anxiety disorder prevalence of 19% among students at Zanjan University of Medical Sciences; this places our sampled student population in the lower quartile of anxiety rates compared to the broader range reported by medical trainees across the world.

Unlike the questionnaire used in our study (ie, SCL-90-R), many previous studies on student stress have used the DASS scale and reported high rates of stress among students. For example, a meta-analysis found that 52.7% of medical students reported significant stress during training [13]. Additionally, studies by Aghajani Liasi et al [7] and Moutinho et al [28] reported stress rates of approximately 30% and 47%, respectively within their student samples, despite the differences in study populations. While these percentages vary, these studies collectively highlight that clinically significant stress is a widespread and impactful issue for many students across educational contexts [7,28]. However, our study did not directly measure student stress, which is a limitation compared to previous existing research.

Our study showed that although there was no statistically significant difference in burnout scores between male and female students, female students reported lower burnout levels in the three burnout subscales compared to male students. Additionally, students in later years of medical education reported higher burnout levels than students in the initial phases. This indicates

that the interns about to graduate showed higher burnout levels, especially feeling emotionally drained.

The relationship between years of medical education and burnout levels is interesting. While some studies have suggested that burnout levels may increase with advancing years of medical education due to prolonged exposure to stressors, the evidence remains inconclusive [29,30]. This suggests that the intense pressures of medical school take an cumulative toll.

Prior studies have found that medical students experience some of the highest rates of burnout compared to other populations [4,6,11,20]. However, findings regarding the relationship between gender and burnout have been mixed [29,31-33]. There was more evidence suggesting that male students are more likely to face burnout than female students [6]. Therefore, it can be concluded that the relationship between gender and burnout in medical students may be influenced by various factors such as the specific population, sample sizes, and the definition of burnout used in the research.

Our study explored several potential influencing factors on the three burnout dimensions in medical students, including mental health status, gender, and academic level. These variables significantly impacted emotional exhaustion, cynicism, and academic efficacy scores. To date, no study had directly examined the linkage between mental health disorders and burnout in this population, representing a gap in understanding. However, related research by Dyrbye et al [16] showed associations between positive well-being and professionalism, which burnout may undermine. Additionally, psychologists have suggested that students with psychiatric conditions demonstrate greater emotional exhaustion [17,18,34]. Notably, in our analysis, mental health had a much more significant effect on emotional exhaustion compared to the other burnout facets. Other studies found students with higher burnout reported more suicidal thoughts and behaviors [6,17,18,26,27,34,35]. Integrating those findings with our results suggests that mental health could play an intermediary role between burnout and suicidal risks. These interrelationships between wellness, distress, and functioning highlight the need for more holistic support to promote student resilience.

Limitations

This study has some limitations. The cross-sectional design cannot determine causal relationships between variables. Additionally, the convenience sampling and voluntary participation could indicate that students with psychological issues may have been less inclined to take part or answer honestly. While different variables were recorded, others such as physical activity, social support, and economic status, should be investigated in future studies. Longitudinal follow-up studies warrant a better understanding of mental health's impact on burnout trajectories.

Another limitation was the rate of female participation compared to male participants for two reasons. According to the university's annual statistics, about 55% of students are women, increasing the female sample rate in the convenience sampling method. However, among the Iranian population, women are more interested in psychological issues and experience exhaustion about improving mental health, resulting in more women participation in our study.

An additional limitation is the potential link between financial issues, mental health, and burnout. Our survey did not include detailed questions about participants' financial situations, which could have influenced their responses to other questions. To partially address this, we included a question about place of residence, which could indirectly reflect financial circumstances and their possible effects on other survey responses.

Conclusion

The demanding nature of academic work and personal lives faced by medical students can take a severe mental toll, leading to burnout. Despite being educated on physical and psychological health, students often neglect their own well-being. This research confirms that mental health issues directly contribute to students' emotional exhaustion, cynicism, and reduced feelings of academic self-efficacy. Both burnout and psychological problems increase the risk of students dropping out or deciding against careers as general practitioners after graduation, resulting in wasted resources invested in their training. Most alarmingly, if the society cannot ensure the mental well-being of its future doctors, the overall population's health will suffer consequences. There is an undeniable connection between medical trainees' health and the communities they will serve. Fostering resilience and coping abilities in students must be a key priority, as their personal health and capacity to provide quality patient care in the future hinges on it. The findings of this study highlight the prevalence of burnout and mental health issues among medical students, underscoring the profound importance of addressing this problem for the well-being of the general population, who will rely on these future physicians for care.

Acknowledgments

The authors express their gratitude to Prof. Xiaolin Xu for his contributions as a senior author in guiding the research and to Dr. Omid Saed for his suggestion for the questionnaires, Department of Clinical Psychology, Zanjan University of Medical Sciences.

Data Availability

The datasets used and analyzed during the current study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: EF

Data curation: AE
Formal analysis: EF, AE
Funding acquisition: XX, MF
Methodology: EF, MF
Project administration: XX
Resources: EF
Supervision: MF, XX
Visualization: EF, MF
Writing – original draft: EF
Writing – review & editing: EF, MF, XX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Descriptive statistics of exhaustion, cynicism, academic efficacy across sex and academic level.
[DOCX File, 16 KB - mededu_v11i1e58008_app1.docx]

References

1. Kafry D, Pines A. The experience of Tedium in life and work. Human Relations 1980 Jul;33(7):477-503. [doi: 10.1177/001872678003300703]
2. Maslach C, Jackson SE. The measurement of experienced burnout. J Organ Behavior 1981 Apr;2(2):99-113. [doi: 10.1002/job.4030020205]
3. Maslach C. Burnout: a multidimensional perspective. In: Professional Burnout: Routledge; 2018:19-32. [doi: 10.4324/9781315227979-3]
4. Abreu Alves S, Sinval J, Lucas Neto L, Marôco J, Gonçalves Ferreira A, Oliveira P. Burnout and dropout intention in medical students: the protective role of academic engagement. BMC Med Educ 2022 Feb 7;22(1):83. [doi: 10.1186/s12909-021-03094-9] [Medline: 35130892]
5. Schaufeli WB, Martínez IM, Pinto AM, Salanova M, Barker AB. Burnout and engagement in university students a cross-national study. J Cross Cult Psychol 2002;33:464-481. [doi: 10.1177/0022022102033005003]
6. Rosales-Ricardo Y, Rizzo-Chunga F, Mocha-Bonilla J, Ferreira JP. Prevalence of burnout syndrome in university students: a systematic review. Salud Ment 2021 Apr 9;44(2):91-102. [doi: 10.17711/SM.0185-3325.2021.013]
7. Aghajani Liasi G, Mahdi Nejad S, Sami N, Khakpour S, Ghorbani Yekta B. The prevalence of educational burnout, depression, anxiety, and stress among medical students of the Islamic Azad University in Tehran, Iran. BMC Med Educ 2021 Sep 5;21(1):471. [doi: 10.1186/s12909-021-02874-7] [Medline: 34482821]
8. Zis P, Artemiadis A, Bargiotas P, Nteveros A, Hadjigeorgiou GM. Medical studies during the COVID-19 pandemic: the impact of digital learning on medical students' burnout and mental health. Int J Environ Res Public Health 2021 Jan 5;18(1):349. [doi: 10.3390/ijerph18010349] [Medline: 33466459]
9. Auerbach RP, Alonso J, Axinn WG, et al. Mental disorders among college students in the World Health Organization World Mental Health Surveys - CORRIGENDUM. Psychol Med 2017 Nov;47(15):2737. [doi: 10.1017/S0033291717001039] [Medline: 28462760]
10. Gerber M, Lang C, Feldmeth AK, et al. Burnout and mental health in Swiss vocational students: the moderating role of physical activity. J of Research on Adolesc 2015 Mar;25(1):63-74. [doi: 10.1111/jora.12097]
11. Wassif GO, Gamal-Eldin DA, Boulos DNK. Stress and burnout among medical students. Journal of High Institute of Public Health 2019 Dec 5;0:189-197. [doi: 10.21608/jhiph.2019.63794]
12. Rotenstein LS, Ramos MA, Torre M, et al. Prevalence of depression, depressive symptoms, and suicidal ideation among medical students: a systematic review and meta-analysis. JAMA 2016 Dec 6;316(21):2214-2236. [doi: 10.1001/jama.2016.17324] [Medline: 27923088]
13. Cuttilan AN, Sayampanathan AA, Ho RCM. Mental health issues amongst medical students in Asia: a systematic review [2000-2015]. Ann Transl Med 2016 Feb;4(4):72. [doi: 10.3978/j.issn.2305-5839.2016.02.07] [Medline: 27004219]
14. Healy C, Ryan Á, Moran CN, Harkin DW, Doyle F, Hickey A. Medical students, mental health and the role of resilience - a cross-sectional study. Med Teach 2023 Jan;45(1):40-48. [doi: 10.1080/0142159X.2022.2128735] [Medline: 36214365]
15. Zeng W, Chen R, Wang X, Zhang Q, Deng W. Prevalence of mental health problems among medical students in China. Medicine (Abingdon) 2019;98(18):e15337. [doi: 10.1097/MD.0000000000015337] [Medline: 31045774]
16. Dyrbye LN, Harper W, Moutier C, et al. A multi-institutional study exploring the impact of positive mental health on medical students' professionalism in an era of high burnout. Acad Med 2012 Aug;87(8):1024-1031. [doi: 10.1097/ACM.0b013e31825cfa35] [Medline: 22722352]

17. Morcos G, Awan OA. Burnout in medical school: a medical student's perspective. Acad Radiol 2023 Jun;30(6):1223-1225. [doi: 10.1016/j.acra.2022.11.023] [Medline: 36586757]
18. Klein HJ, McCarthy SM. Student wellness trends and interventions in medical education: a narrative review. Humanit Soc Sci Commun 2022;9(1):92. [doi: 10.1057/s41599-022-01105-8] [Medline: 36254165]
19. Hu W, Yu Z, Liang X, Abulaiti A, Aini X, Kelimu A. A cross-sectional study on the analysis of the current situation of depression and anxiety among primary and secondary school students in Urumqi City in 2021: a case study of S district. J Affect Disord 2024 Feb 15;347:210-219. [doi: 10.1016/j.jad.2023.11.079] [Medline: 37995925]
20. Dyrbye LN, West CP, Satele D, et al. Burnout among U.S. medical students, residents, and early career physicians relative to the general U.S. population. Acad Med 2014 Mar;89(3):443-451. [doi: 10.1097/ACM.0000000000000134] [Medline: 24448053]
21. Koutsimani P, Montgomery A, Georganta K. The relationship between burnout, depression, and anxiety: a systematic review and meta-analysis. Front Psychol 2019;10:284. [doi: 10.3389/fpsyg.2019.00284] [Medline: 30918490]
22. Rostami Z, Abedi MR, Schuffli VB. Standardization of Maslash burnout inventory among female students at University of Isfahan. New Educational Approaches 2011;6:21-38 [FREE Full text]
23. Maslach C, Jackson SE, Leiter MP. Evaluating stress: a book of resources. In: Maslach Burnout Inventory, th edition: Scarecrow Education; 1997.
24. Derogatis LR, Melisaratos N. The Brief Symptom Inventory: an introductory report. Psychol Med 1983 Aug;13(3):595-605. [Medline: 6622612]
25. Akhavan Abiri F, Shir Mohammadi R. Validity and reliability of Symptom Checklist-90-revised (SCL-90-R) and Brief Symptom Inventory-53 (BSI-53). Clinical Psychology and Personality 2019;17:169-195 [FREE Full text]
26. Fares J, Al Tabosh H, Saadeddin Z, El Mouhayyar C, Aridi H. Stress, burnout and coping strategies in preclinical medical students. N Am J Med Sci 2016 Feb;8(2):75-81. [doi: 10.4103/1947-2714.177299] [Medline: 27042604]
27. Mirza AA, Baig M, Beyari GM, Halawani MA, Mirza AA. Depression and anxiety among medical students: a brief overview. Adv Med Educ Pract 2021;12:393-398. [doi: 10.2147/AMEP.S302897] [Medline: 33911913]
28. Moutinho ILD, Maddalena NDC, Roland RK, et al. Depression, stress and anxiety in medical students: a cross-sectional comparison between students from different semesters. Rev Assoc Med Bras (1992) 2017 Jan 1;63(1):21-28. [doi: 10.1590/1806-9282.63.01.21] [Medline: 28225885]
29. Almutairi H, Alsubaiei A, Abduljawad S, et al. Prevalence of burnout in medical students: a systematic review and meta-analysis. Int J Soc Psychiatry 2022 Sep;68(6):1157-1170. [doi: 10.1177/00207640221106691] [Medline: 35775726]
30. Grech M. The effect of the educational environment on the rate of burnout among postgraduate medical trainees - a narrative literature review. J Med Educ Curric Dev 2021;8:23821205211018700. [doi: 10.1177/23821205211018700] [Medline: 34104789]
31. Galán F, Sanmartín A, Polo J, Giner L. Burnout risk in medical students in Spain using the Maslach Burnout Inventory-Student Survey. Int Arch Occup Environ Health 2011 Apr;84(4):453-459. [doi: 10.1007/s00420-011-0623-x] [Medline: 21373879]
32. Backović DV, Zivojinović JI, Maksimović J, Maksimović M. Gender differences in academic stress and burnout among medical students in final years of education. Psychiatr Danub 2012 Jun;24(2):175-181 [FREE Full text] [Medline: 22706416]
33. Briggs LG, Riew GJ, Kim NH, et al. Racial and gender differences in medical student burnout: a 2021 national survey. Mayo Clin Proc 2023 May;98(5):723-735. [doi: 10.1016/j.mayocp.2022.11.003] [Medline: 37137644]
34. Ishak W, Nikravesh R, Lederer S, Perry R, Ogunyemi D, Bernstein C. Burnout in medical students: a systematic review. Clin Teach 2013 Aug;10(4):242-245. [doi: 10.1111/tct.12014] [Medline: 23834570]
35. Frajerman A, Morvan Y, Krebs MO, Gorwood P, Chaumette B. Burnout in medical students before residency: a systematic review and meta-analysis. Eur Psychiatry 2019 Jan;55:36-42. [doi: 10.1016/j.eurpsy.2018.08.006] [Medline: 30384110]

Abbreviations

GSI: Global Severity Index
MBI-SS: Maslach Burnout Inventory-Student Survey
SCL-90: Symptom Checklist-90
SCL-90-R: Symptom Checklist-90-Revised

Edited by B Lesselroth; submitted 02.03.24; peer-reviewed by A Hassan, CI Sartorão Filho, S Kale; revised version received 22.06.24; accepted 23.11.24; published 15.05.25.

Please cite as:
Faghihzadeh E, Eghtesad A, Fawad M, Xu X
Exploring Connections Between Mental Health, Burnout, and Academic Factors Among Medical Students at an Iranian University: Cross-Sectional Questionnaire Study
JMIR Med Educ 2025;11:e58008
URL: https://mededu.jmir.org/2025/1/e58008
doi:10.2196/58008

Barriers to and Facilitators of Implementing Team-Based Extracorporeal Membrane Oxygenation Simulation Study: Exploratory Analysis

Joan Brown[1], EdD, MBA, CCE; Sophia De-Oliveira[2], MPH; Christopher Mitchell[1,3], BSN, RN, CCRN; Rachel Carmen Cesar[4], PhD; Li Ding[4], MD; Melissa Fix[1,3], BSN, RN, CCRN-CMC; Daniel Stemen[5], MSRS, RRT-ACCS, E-AEC; Krisda Yacharn[1,3], BSN, RN, CNML; Se Fum Wong[6], MD; Anahat Dhillon[7], MD

[1]Department of Surgery, Keck School of Medicine, University of South California, Los Angeles, CA, United States

[2]Office of Performance & Transformation, Keck Hospital of USC, Los Angeles, CA, United StatesUnited States

[3]Department of Nursing, Keck Hospital of USC, Los Angeles, CA, United StatesUnited States

[4]Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA, United StatesUnited States

[5]Deparment of Respiratory Therapy, Keck Hospital of USC, Los Angeles, CA, United StatesUnited States

[6]Department of Anesthesia, Kaiser Permanante, Los Angeles, CA, United States

[7]Department of Anesthesia Critical Care Medicine, Keck School of Medicine, University of South California, Los Angeles, CA, United StatesUnited States

Corresponding Author:
Joan Brown, EdD, MBA, CCE
Department of Surgery, Keck School of Medicine, University of South California, Los Angeles, CA, United States

Abstract

Introduction: Extracorporeal membrane oxygenation (ECMO) is a critical tool in the care of severe cardiorespiratory dysfunction. Simulation training for ECMO has become standard practice. Therefore, Keck Medicine of the University of California (USC) holds simulation-training sessions to reinforce and improve providers knowledge.

Objective: This study aimed to understand the impact of simulation training approaches on interprofessional collaboration. We believed simulation-based ECMO training would improve interprofessional collaboration through increased communication and enhance teamwork.

Methods: This was a single-center, mixed methods study of the Cardiac and Vascular Institute Intensive Care Unit at Keck Medicine of USC conducted from September 2021 to April 2023. Simulation training was offered for 1 hour monthly to the clinical team focused on the collaboration and decision-making needed to evaluate the initiation of ECMO therapy. Electronic surveys were distributed before, after, and 3 months post training. The survey evaluated teamwork and the effectiveness of training, and focus groups were held to understand social environment factors. Additionally, trainee and peer evaluation focus groups were held to understand socioenvironmental factors.

Results: In total, 37 trainees attended the training simulation from August 2021 to August 2022. Using 27 records for exploratory factor analysis, the standardized Cronbach α was 0.717. The survey results descriptively demonstrated a positive shift in teamwork ability. Qualitative themes identified improved confidence and decision-making.

Conclusions: The study design was flawed, indicating improvement opportunities for future research on simulation training in the clinical setting. The paper outlines what to avoid when designing and implementing studies that assess an educational intervention in a complex clinical setting. The hypothesis deserves further exploration and is supported by the results of this study.

(JMIR Med Educ 2025;11:e57424) doi:10.2196/57424

KEYWORDS

intensive care unit; ICU; teamwork in the ICU; team dynamics; collaboration; interprofessional collaboration; simulation; simulation training; ECMO; extracorporeal membrane oxygenation; life support; cardiorespiratory dysfunction; cardiorespiratory; cardiology; respiratory; heart; lungs

Introduction

Simulation training for extracorporeal membrane oxygenation (ECMO) has become standard practice for reinforcing technical skills, facilitating troubleshooting, and building teamwork [1]. ECMO is a critical tool in the care of severe cardiorespiratory dysfunction among patients of all ages [1]. Within the intensive care unit (ICU), ECMO is one of the most complicated therapies, requiring not only extensive knowledge of cardiopulmonary physiology and expertise with intricate circuit components but also skills to rapidly respond to emergent situations [2]. Therefore, high-fidelity simulation trainings are critical to practice skills and work through different emergency scenarios, such as the blood pump falling from the drive unit [3]. A randomized control study concluded that exposure to high-fidelity simulated ECMO emergencies leads to significant improvements in technical and behavioral skills among clinicians. This study demonstrated that simulation training creates a learning environment that replicates the clinical setting and fosters acquisition of cognitive, technical, and behavioral skills [4].

The Extracorporeal Life Support Organization, an international nonprofit association of health care institutions focused on ECMO research and education, recommends simulation training didactic sessions, water drills, animal sessions, and bedside training [5]. However, a randomized controlled trial published in *Critical Care Medicine* compared traditional water drill with simulation and found that simulation-based training is more effective than traditional training [6]. Water-based drills do not offer the same hands-on experience of real-time troubleshooting, and the use of animals is expensive and complex [6]. Nevertheless, traditional and simulation-based training are both beneficial to ECMO education. The benefits of simulation training on reinforcing skills have been noted in the literature [3,7-9]. Therefore, Keck Medicine of the University of California (USC) has held ECMO simulation-training sessions since 2013 for nursing education and 2021 for interprofessional simulation to reinforce and improve providers knowledge and hands-on skills in high-risk, low-frequency scenarios at no risk to patients [6].

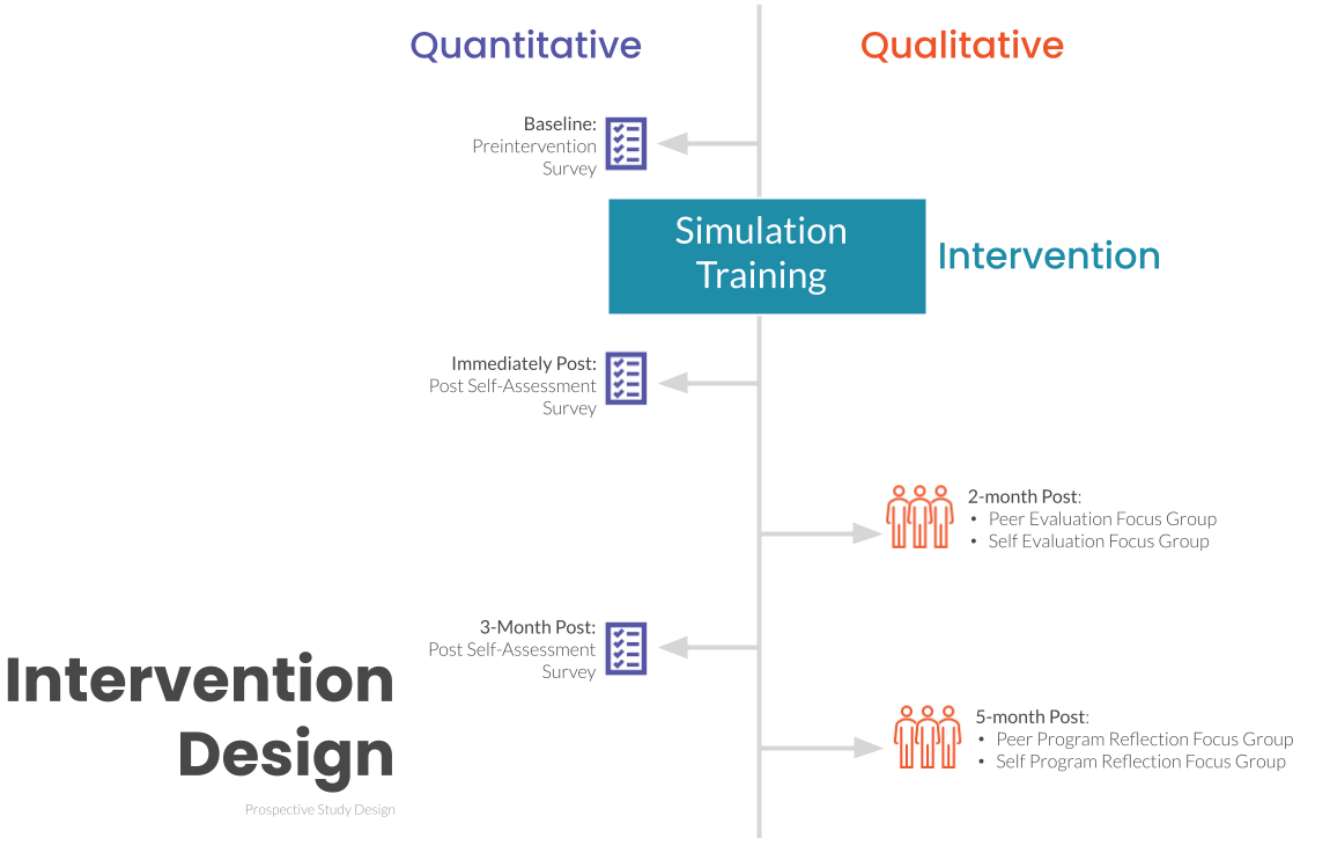
We implemented simulation-based ECMO training to improve interprofessional collaboration through increased communication and enhanced teamwork. Moreover, the intention of the simulation training was to strengthen collaboration skills and increase confidence in providers to work through emergency scenarios. The specific aim of the study was to understand the impact of our simulation training approach on interprofessional collaboration. However, ICU staffing models impacted the ability to execute the study design as intended. This paper outlines the original study design, the challenges the research team faced during the study, and the lessons learned to ensure future studies mitigate the challenges posed by real-world ICU operations. Our primary outcome shifted to the development and validation of measurement tools and offers recommendations for the evaluation of simulation approaches in future studies.

Methods

Overview

This was a single-center, mixed methods study of the Cardiac and Vascular Institute (CVI) ICU at Keck Medicine of USC conducted from September 2021 to April 2023. The study was designed to elicit quantitative feedback through an electronic survey before and post a voluntary training simulation exercise, and qualitative feedback from a participants via a series of focus groups that incorporated self- and peer evaluations (Figure 1).

Figure 1. Study design.



Participants

A census sampling strategy was used to recruit participants; in other words, all trainees that participated in the training were offered participation in the study. Participants included: (1) trainee physician fellows and residents in the CVI ICU who were offered attendance to the simulation training by their program director, and (2) peer evaluators, including the CVI ICU's Medical Director, nurse manager, nurse clinical educator, and lead respiratory therapist. Participant trainees attended a single 1-hour simulation training with roles played by clinical staff members from the CVI ICU, including an intensivist, clinical nurse educator, and respiratory therapist. To be eligible for the study, participants needed to be in a fellowship on rotation at Keck Medicine of USC. Fellows were recruited by intensivist leaders of the CVI from departments that rotated through or interacted with the CVI ICU (Pulmonary Critical Care Medicine, Cardiology, Anesthesia, Surgical Critical Care, and Cardiac Surgery). All study recruitment took place via email by the CVI Medical Director and Program Director to physicians in fellowship based on department and rotation schedule from pulmonary critical care medicine, surgical critical care medicine, cardiac surgery, and cardiology.

Simulation Training

Simulation training was designed as part of the continuing clinical education offered to the clinical team for 1-hour monthly, where participants attended a single session. Simulations were designed to focus on the interprofessional collaboration and decision-making needed to evaluate a patient for the initiation of ECMO therapy (Table S1 in Multimedia Appendix 1). Initially, low-fidelity simulations were held in a conference room using (1) a resuscitation training mannequin, (2) simulated vital signs via a hospital patient monitor connected to a rhythm simulator, (3) simulated intravenous access, (4) simulated medications, and (5) emergency equipment. In January 2022, collaboration with the Keck School of Medicine Simulation department allowed for training to be held in a simulation lab with a high-fidelity simulation mannequin and integrated simulation software LLEAP, version 8.5 from Laerdal. The availability of a higher fidelity training environment was meant to improve the training experience of the learners.

Each training session began with an orientation to the simulation environment and assigned roles. The scenario (Table S1 in Multimedia Appendix 1) was created to include relative contraindications to ECMO therapy and a potentially reversible condition that led to a cardiac arrest requiring resuscitation. Participants were assigned into roles of primary physician, code blue response provider, and cardiac surgeon prior to entering the simulation and entered the scenario when prompted by the facilitator or requested during the simulation by another participant. The patient was introduced to the learners as a 65-year-old female in-patient on a hospital cardiac telemetry unit with a past medical history of coronary artery disease, congestive heart failure, and peripheral vascular disease. The simulation began when a facilitator in the role of the patient's nurse requested help from a participant. The simulated patient was initially responsive with complaints of palpitations and shortness of breath with intermittent ventricular tachycardia displayed on the cardiac monitor. The simulated patient then became unresponsive in persistent ventricular tachycardia, and the imbedded facilitator activated the resuscitation team. When

the simulated patient's cardiac rhythm changes, the participants performed the roles of a code blue response, including coordinating the resuscitation, performing a simulated echocardiograph, and performing simulated invasive procedures including endotracheal intubation, arterial line insertion, and central line insertion. The participants collaborated to identify the candidacy of the simulated patient for ECMO therapy and proceeded to participate in a moderate-fidelity mock cannulation with ECMO training equipment. The participant in the surgeon role chose a method and site of cannulation for the simulated patient, and a practice ECMO circuit was connected to the simulator. Participants proceeded to respond and troubleshoot as the patient was set to be initially unstable during the transition to ECMO support. The simulated patient remained in ventricular tachycardia, and the participants were required to decide whether to continue attempting interventions, including, for example, chest compressions, medication, and defibrillation once the patient was placed on ECMO. The simulation ended when the patient was stabilized on ECMO and the participants decided to transfer the patient to the ICU. Areas of safety concern (Figure S1 in Multimedia Appendix 2) were emphasized in the training as points for communication to consider the decision to initiate ECMO with an unstable patient. A postsimulation debriefing session was facilitated by the simulation faculty.

Qualitative Approach

To understand the social environment factors the simulation training impacted, a total of 12 qualitative focus groups were planned (Figure 1). The 12 interviews were divided into 6 focus groups with the trainee attendees of the simulation training as a self-evaluation and 6 focus groups with colleague participants as a peer evaluation (Figure S2 in Multimedia Appendix 3). Each focus group was designed to have 2 - 4 participants. The study was designed to use the same peer evaluators for each peer evaluation focus group for the study duration. Each peer evaluation focus group was meant to target the evaluation of the individuals in the 6 simulation cohorts with a total of 4 peer participants. The focus groups were designed to be a duration of 30 minutes. Questions were developed to assess how the simulation training impacted the trainees' practice related to collaboration and teamwork (Table S3 in Multimedia Appendix 4). Questions were reviewed by the study expert in mixed methods study application.

Interviews were conducted by the simulation facilitators experienced in ECMO therapy and simulation education. Sessions were recorded using the Voice Memo application (Apple, Inc.). Once the focus groups were completed, the simulation facilitators sent the audio file to the data management author for transcription. The transcribed focus group sessions were de-identified, then uploaded and stored to a HIPAA (Health Insurance Portability and Accountability Act)-compliant Microsoft OneDrive. All audio and video files containing identifiers were deleted following transcription. Transcription documents were reviewed and coded for key themes using grounded theory methodology, an iterative process that will identify conceptual categories emerging from the comparative analyses of the data.

Quantitative Approach

An electronic survey was distributed with a QR code in person and electronically via email using Qualtrics XM software version December 2019. A total of 49 questions were posed to trainees across the pretraining, posttraining, and 3-month posttraining questionnaires (Table S2 in Multimedia Appendix 5). Teamwork-focused questions were obtained from the validated Mayo High Performance Teamwork scale (16 questions) [10]. The remaining questions regarding the effectiveness of training were devised using the Kirkpatrick Training Evaluation Framework as a basis for query design [11]. The study biostatistician performed a psychometric review to assess the validity and reliability of the survey questions. The use of existing validated tools ensured high reliability and validity of the teamwork elements of the survey tool [10]. Additionally, field tests (1 MD, 2 RNs, and 1 RT) of the survey tool showed an average survey duration of 7 minutes or less and promoted consistent comprehension of the study questions across individuals.

Exploratory factor analysis was conducted to assess questions reflecting underlying factors. The number of factors included in the final model was determined by eigenvalue and Scree plot. In the final factor pattern table, questions with a value>0.4 were considered well loaded for the factor. The Validated Mayo High Performance Teamwork scale (16 questions) was used as a sum score as recommended by the study [10]. Secondary outcomes analysis included department-level patient mortality, average device days on ECMO, decannulation percentage, and percentage of staff that had simulation training.

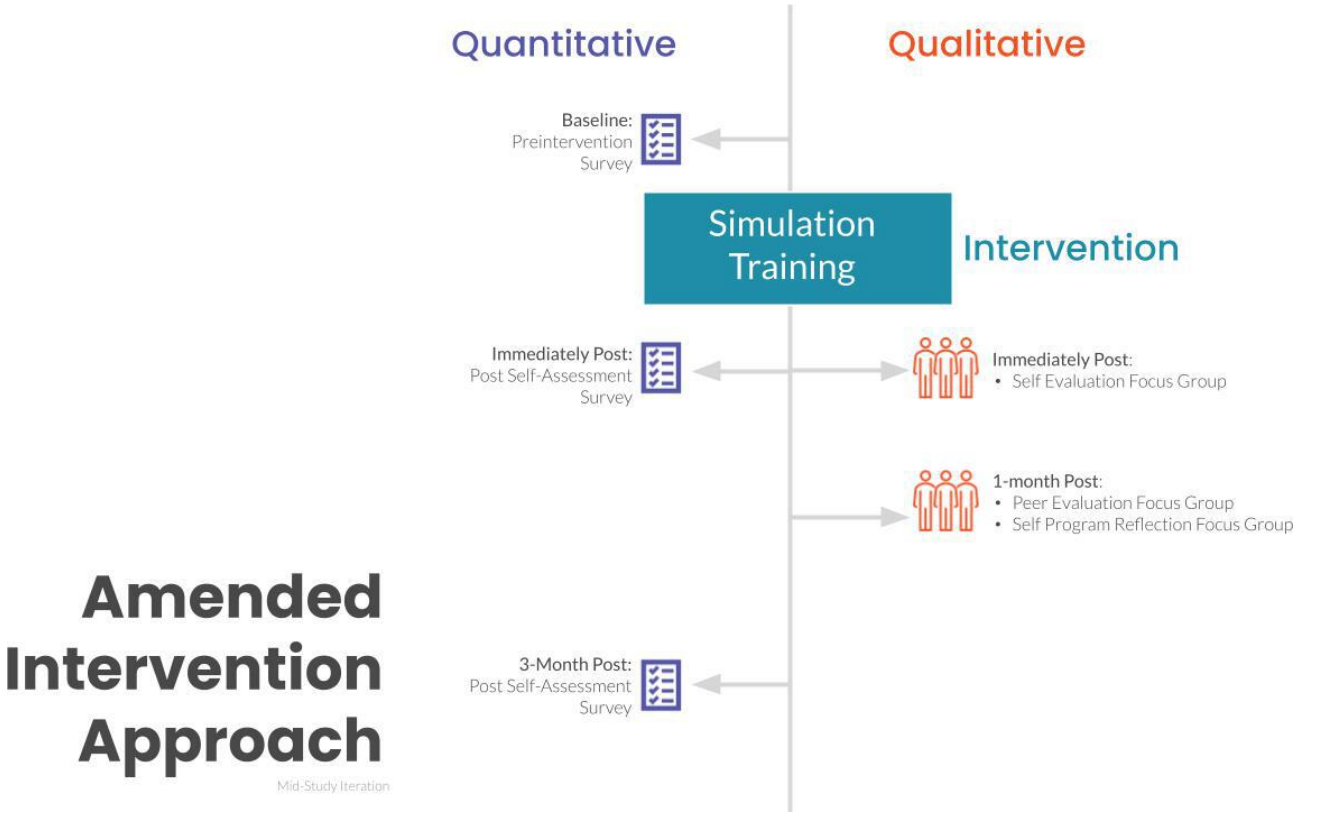
Ethical Considerations

The study was approved by the University of Southern California institutional review board (UP-21-01021). Prior to participation, all study participants were required to sign an informed consent form, thereby confirming their voluntary engagement in the survey process. The study data were anonymous.

Results

A total of 37 trainees attended the training simulation from August 2021 to August 2022. However, only 7 trainees opted to participate in the qualitative portion of the study. Due to lack of participant engagement, mid-study the study design was amended to increase study participation (Figure 2). The quantitative approach remained as originally designed; however, the analysis approach was done descriptively due to the inability to compare pre- and postsurvey results on an individual basis. In other words, survey analysis aggregated all preresponses and then postresponses to compare the pregroup and postgroup responses. The qualitative approach shifted to trainee participants attending a total of 2 focus groups, an initial self-evaluation immediately following simulation and a 1-month post-program reflection if they were available (Figure S3 in Multimedia Appendix 6).

Figure 2. Amended study design.



Qualitative Approach

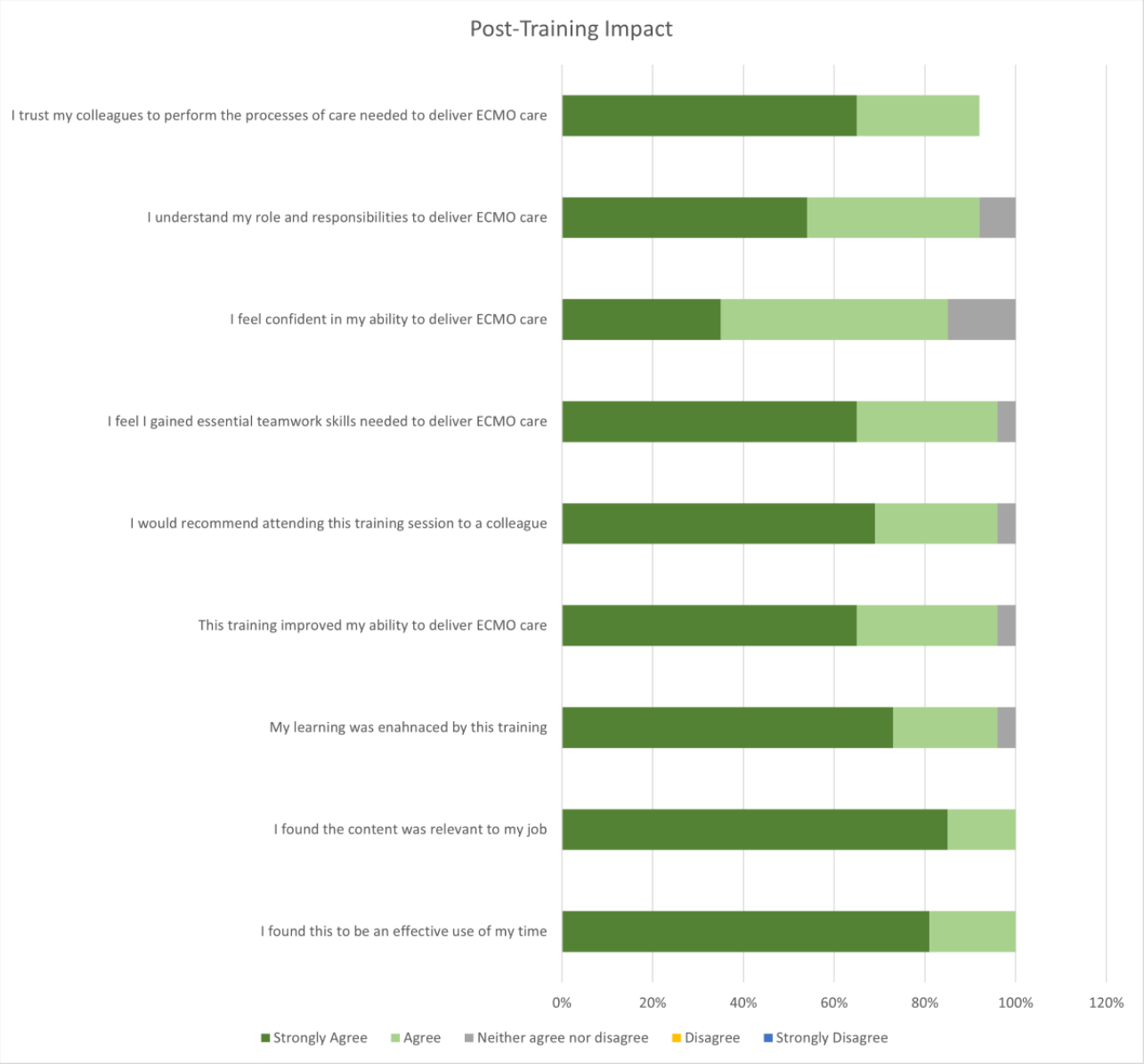
A total of 4 focus groups were conducted between January 2022 and August 2022 including 2 trainee self-evaluations (n=7) immediately post the training simulation and 2 peer evaluations 2-months post-training evaluation with the group of 4 peers. Focus groups for participants between August 2021 and December 2021 were not coordinated due to lack of trainee engagement in study participation. The 2 peer focus groups held highlighted an issue in trainee exposure with the peer team. Peer participants noted that they had limited clinical working exposure to the trainees being evaluated due to the nature of the fellow's rotation in the ICU. Meaning, peers did not have any recollection of working with the trainees prior to and following the simulation training to provide an appropriate evaluation of their skills in teamwork during ECMO therapy decision-making. In other words, peers remembered a trainee prior to or following the simulation, but not both. The 2 trainee self-evaluations that occurred highlighted themes that showed the simulation training benefited the trainees, including that the simulation training resulted in (1) working together as a stronger and more confident team because of simulation and (2) the creation of a space to improve communications, decision-making, and express concerns (Table S5 in Multimedia Appendix 7).

Quantitative Approach

All trainees were asked to complete the pre-, immediately post-, and 3-month surveys as part of the simulation training experience. There were a total of 37 entries recorded for the pre-survey, yielding a 100% response rate. Of those entries, 9 records were excluded due to lack of record ID or mismatching question numbers, leaving 28 entries for analysis. There was a total of 35 entries recorded for the postsurvey, yielding a 95% response rate. Following data cleaning, 9 records were excluded, leaving 26 postsurvey entries for analysis. There were a total of 2 entries for the 3-month posttraining survey, yielding a 5.4% response rate. The 3-month postsurvey results were excluded from analysis due to the low response rate. Table S4 in Multimedia Appendix 8 details the survey results of each question.

Questions posed in the post-survey focused on assessing levels 1 (reaction) and 2 (learning) of the Kirkpatrick Training Evaluation Framework demonstrated a high level of agreement for positive postsimulation training impact (Figure 3). Additionally, an increase in knowledge and understanding were noted descriptively when comparing the pre- and postresponses of the survey (Table S4 in Multimedia Appendix 8). For example, question "I understand the mechanism for activating ECMO at Keck Hospital" in the presurvey 46% of respondents agreed or strongly agreed to the statement compared to the postsimulation training survey where 100% of participants responded with a level of agreement. Levels 3 (impact) and 4 (results) of Kirkpatrick's framework were unable to be assessed due to the low response rate for the 3-month postsimulation survey.

Figure 3. Posttraining impact. ECMO: extracorporeal membrane oxygenation.



The Mayo Teamwork Scale was used to understand changes in the trainee's perspective on teamwork before and after the simulation training. The 16 focused teamwork questions demonstrated a positive shift in teamwork ability; that is, in the presurvey participants had a 71% average response on performing each question consistently, postsurvey showed an increase of the average to 95% consistently.

Exploratory Factor Analysis

An exploratory factor analysis was conducted to validate the use of the survey in evaluating the effectiveness of team training in ECMO simulation (Table 1). Entries with missing questions were excluded from the factor analysis; that is, only records with all questions answered were used. A total of 28 data points were available in the presurvey and 26 for the postsurvey. For the factor analysis, 0.4 was used as the cutoff. For the Mayo High-Performance Teamwork scale, 23 records were collected with a mean sum score of 26.87 (SD 6.41) at the presurvey and 21 records at the postsurvey with a mean sum score of 31.1 (SD 1.88). For the presurvey, we asked six 5-likelihood questions. Using 27 records for exploratory factor analysis, only 1 question, "Q5," did not reflect the underlying factor. Standardized Cronbach α was 0.686 when using all 6 questions. After excluding Q5, the standardized Cronbach α is 0.717. For the postsurvey, 26 records were used for analysis with 3 factors. Standardized Cronbach α was 0.919 when using all 15 questions. For questions from the Mayo High-Performance Teamwork scale, 21 records were used for analysis with a mean sum score of 31.1 and a SD of 1.88.

Table . Exploratory factor analysis.

		Factor 1	Factor 2	Factor 3
Presurvey				
	1. Please rate the following statements—I understand the mechanism for activating ECMO[a] at Keck Hospital	0.556	—[b]	—
	2. Please rate the following statements—I understand my role in a bedside cannulation	0.755	—	—
	3. Please rate the following statements—I feel comfortable using the ECMO equipment specific to my role	0.723	—	—
	4. Please rate the following statements—I feel comfortable using the 2-challenge rule	0.478	—	—
	5. Please rate the following statements—I feel confident to voice concerns to leadership during a critical situation	0.201	—	—
	6. Please rate the following statements—I trust my colleagues to perform the processes of care needed to deliver ECMO care	0.417	—	—
Postsurvey				
	1. Please rate the following statements—I understand the mechanism for activating ECMO at Keck Hospital	—	0.880	—
	2. Please rate the following statements—I understand my role in a bedside cannulation	—	0.602	—
	3. Please rate the following statements—I feel comfortable using the ECMO equipment specific to my role	—	—	0.412
	4. Please rate the following statements—The initiating team communicates efficiently during a bedside cannulation	0.543	0.456	—
	5. Please rate the following statements—I feel comfortable using the 2-challenge rule	—	—	0.659
	6. Please rate the following statements—I feel confident to voice concerns to leadership during a critical situation	—	0.705	—

		Factor 1	Factor 2	Factor 3
	7. Please rate the following statements—I found this to be an effective use of my time	0.851	—	—
	8. Please rate the following statements—I found the content was relevant to my job	0.813	—	—
	9. Please rate the following statements—My learning was enhanced by this training	0.927	—	—
	10. Please rate the following statements—This training improved my ability to deliver ECMO care	0.891	—	—
	11. Please rate the following statements—I would recommend attending this training session to a colleague	0.844	—	—
	12. Please rate the following statements—I feel I gained essential teamwork skills needed to deliver ECMO care	0.916	—	—
	13. Please rate the following statements—I feel confident in my ability to deliver ECMO care	0.532	—	0.670
	14. Please rate the following statements—I understand my role and responsibilities to deliver ECMO care	0.695	—	—
	15. Please rate the following statements—I trust my colleagues to perform the processes of care needed to deliver ECMO care	0.788	—	—

[a]ECMO: extracorporeal membrane oxygenation.

[b]Not applicable.

Triangulation of Quantitative and Qualitative Results

Applying procedures of convergent mixed methods design, we converged the quantitative and qualitative results that were obtained separately to obtain a nuanced understanding of the core research aims. The themes identified of teamwork and improved communication in the qualitative analysis were supported by the quantitative survey results (Tables 2 and 3). Qualitative subthemes were supported by the positive shift observed descriptively from the pre- compared to the post-simulation training survey results.

Table . Triangulation of quantitative and qualitative results between frequently endorsed survey items and themes emerging from postsimulation focus groups (part).

Question	Agree level (Strongly Agree + Agree), %	Neither agree nor disagree, %	Disagree level (Strongly Disagree + Disagree),	Qualitative theme
I understand the mechanism for activating ECMO [a] at Keck Hospital				Working together as a stronger and more confident team because of simulation
Prestimulation	46	18	36	
Poststimulation	100	0	0	
I understand my role in a bedside cannulation				Working together as a stronger and more confident team because of simulation
Prestimulation	23	48	30	
Poststimulation	100	0	0	
I feel comfortable using the 2- challenge rule				Working together as a stronger and more confident team because of simulation
Prestimulation	4	29	68	
Poststimulation	69	15	15	
I feel confident to voice concerns to leadership during a critical situation				Working together as a stronger and more confident team because of simulation
Prestimulation	78	11	11	
Poststimulation	96	4	0	
I found this to be an effective use of my time (poststimulation)	100	0	0	Creating a space to improve communications, decision-making, and express concerns via simulation
I found the content was relevant to my job (poststimulation)	100	0	0	Creating a space to improve communications, decision-making, and express concerns via simulation
My learning was enhanced by this training (poststimulation)	96	4	0	Creating a space to improve communications, decision-making, and express concerns via simulation
This training improved my ability to deliver ECMO care (poststimulation)	96	4	0	Creating a space to improve communications, decision-making, and express concerns via simulation
I feel I gained essential teamwork skills needed to deliver ECMO care (poststimulation)	96	4	0	Creating a space to improve communications, decision-making, and express concerns via simulation

Question	Agree level (Strongly Agree + Agree), %	Neither agree nor disagree, %	Disagree level (Strongly Disagree + Disagree),	Qualitative theme
I feel confident in my ability to deliver ECMO care (poststimulation)	85	15	0	Creating a space to improve communications, decision-making, and express concerns via simulation
I understand my role and responsibilities to deliver ECMO care (poststimulation)	92	8	0	Creating a space to improve communications, decision-making, and express concerns via simulation
I trust my colleagues to perform the processes of care needed to deliver ECMO care (poststimulation)	92	0	0	Working together as a stronger and more confident team because of simulation

[a]ECMO: extracorporeal membrane oxygenation.

Table . Triangulation of quantitative and qualitative results triangulation between frequently endorsed survey items and themes emerging from post-simulation focus groups (part 2).

Question	% Never or rarely	% Inconsistently	% Consistently	Qualitative theme
A leader is clearly recognized by all team members				Working together as a stronger and more confident team because of simulation
Prestimulation	0	41	59	
Poststimulation	0	17	83	
Each team member demonstrates a clear understanding of his or her role				Working together as a stronger and more confident team because of simulation
Prestimulation	0	37	63	
Poststimulation	0	8	92	
The team prompts each other to attend to all significant clinical indicators throughout the procedure or intervention				Working together as a stronger and more confident team because of simulation
Prestimulation	0	26	74	
Poststimulation	0	8	92	
Disagreements or conflicts among team members are addressed without a loss of situation awareness				Working together as a stronger and more confident team because of simulation
Prestimulation	0	30	70	
Poststimulation	0	4	96	
Poststimulation	0	4	96	
Poststimulation	0	0	100	

Discussion

Principal Findings

This study was designed to evaluate the impact of ECMO therapy simulation training, specifically focused on enhancing teamwork and communication. The study was successful in validating the survey for future use in assessing the effectiveness of ECMO simulation training in improving teamwork and communication. However, while rigorous and well thought out in design, clear flaws were identified that need to be addressed in future attempts to study this type of simulation exercise. We outline the limitations of the study with recommendations for research with the intention to share what to avoid when

designing and implementing studies that assess a clinical education approach in a complex clinical setting. We provide a unique validated tool to assess teamwork and collaboration across clinical disciplines during ECMO therapy, where existing evidence assesses the impact of simulation approaches on knowledge.

Strengths and Limitations

First, the original focus of the study targeted physician fellows and residents from various clinical teams that practice in the CVI unit. An assumption was made in the study design that peer evaluators would have enough interaction with trainees before and after the simulation training to evaluate changes in their behavior. Due to the nature of the rotation of this participant population, the peers were unable to assess any impact. Additionally, the rotation of the trainees contributed to difficulty in follow-up for study participation in both the quantitative and qualitative aspects of the study. Only 2 responses were received for the 3-month postsimulation training survey, and the qualitative study was altered midstudy to garner more participation in study focus groups. The team was unable to obtain commitment from trainees for the 2-month and 5-month planned focus groups and amended the study for a trainee self-evaluation focus group immediately following the simulation training and 1 month post. The study team was unable to coordinate the 1-month post–focus group due to a lack of availability of the fellow and resident trainees. The lack of participation led to the inability to assess levels 3 and 4 of the Kirkpatrick Training Evaluation Framework [11]. Additionally, the lack of participation reduced the validity of the qualitative data obtained in the focus groups. To generalize the qualitative results of the study, the original target of 6 simulation cohorts with a total of 4 peer participants each would be necessary. We suggest future studies alter the study design to broaden study participants to the entire interprofessional team to ensure the target participant enrollment and focus groups are reached. Second, the trainee rotation also did not guarantee exposure of the trainees to ECMO cannulation postsimulation training to practice the technical skills gained from the simulation training. Third, quantitative results demonstrated there is merit to this training simulation approach. Where there were positive shifts from pre- compared to postsimulation training survey results. However, we were unable to calculate statistical significance in pre- and postresponses due to survey collection methods. Survey participation was anonymous and a routine part of the simulation training program. We were unable to align individual pre- and postsurvey responses to apply this statistical strategy or follow up with specific trainees that missed questions. Fourth, although the trainees had a qualitatively and quantitatively favorable response in ECMO initiation following the simulation exercise per survey results, the study did not conclusively demonstrate their ability to actively use that attained knowledge beyond the original simulation date given the lack of actual cannulations and, again, being observed by staff who could claim that teamwork was significantly improved in future interactions. Lastly, the study team anticipated a larger sample of participants, but the recruitment challenges, focus on physician fellows and residents, and staff shortages due to the impact of the COVID-19 pandemic were severe limitations of the study.

Despite these limitations, the quantitative survey results descriptively highlighted the positive impact on the trainees. Level 1 questions of the Kirkpatrick Training evaluation [11] were met with a strong level of agreement, with no level of disagreement responses (Figure 1). Additionally, each of the level 2 questions shifted to a higher level of agreement post the simulation training. The same was true for the responses to the Mayo Teamwork Scale, where each response shifted to more consistent teamwork behavior pre- and postsimulation training (Table S4 in Multimedia Appendix 8).

We know team-based interprofessional care has historically demonstrated gains in positive patient outcomes in the ICU and is seen as the solution to reduce medical errors and poor quality [12-15]. Moreover, a key component of ECMO care is interprofessional collaboration, as it requires a large and multifaceted team of providers collaborating to carry out complementary tasks to one another [16]. Simulation-based team training can cultivate and preserve interprofessional teamwork and communication [16]. However, collaboration across the care team is not a standard topic covered in clinical curriculum [13,15]. We believe our survey results support the merit of our teamwork-focused simulation training approach and its ability to foster a higher level of collaboration when the clinical team is faced with deciding to initiate ECMO therapy in the cardiac and vascular patient population. These findings highlight the importance of simulation training from other innovative ways of ECMO skills training, such as game-based mobile apps, which might not cultivate a teamwork approach to the same extent [17]. This approach could be applied to supplement the lack of practical teamwork focus in today's clinical curriculum.

Despite the identified limitations, the study underscored several positive aspects of ECMO simulation training. The quantitative survey results notably revealed a significant positive impact on the trainees. Level 1 questions of the Kirkpatrick Training evaluation demonstrated a strong level of agreement without any disagreement responses, indicating a high degree of satisfaction with the training (Figure 1). Furthermore, each of the level 2 questions exhibited a shift towards higher levels of agreement postsimulation training. Similarly, responses to the Mayo Teamwork Scale demonstrated a consistent improvement in teamwork behavior before and after simulation training (Table S4 in Multimedia Appendix 8). This reaffirms the notion that team-based interprofessional care, a cornerstone in ICU settings, can lead to enhanced patient outcomes and reduced medical errors. The study's focus on cultivating interprofessional collaboration through simulation-based training aligns with the demands of ECMO care, which relies heavily on coordinated efforts among various health care professionals. These findings highlight the effectiveness of the teamwork-focused simulation training approach in preparing clinical teams to make critical decisions regarding ECMO therapy in the cardiac and vascular patient population. Moreover, they emphasize the importance of incorporating such training into clinical curricula to ensure a holistic approach to health care education. The study's insights pave the way for future research endeavors to further explore

and refine the application of simulation training in improving teamwork and patient outcomes in complex clinical settings. By addressing the outlined recommendations and leveraging innovative approaches, such as virtual reality simulation, the medical community can continue to advance ECMO care delivery and interprofessional collaboration, ultimately enhancing patient care outcomes.

Future research may build upon the learning of this study to strengthen the understanding of a teamwork-focused simulation approach. We would encourage implementation of the following and plan for our future studies to include (1) continuing the study with the entire interprofessional team, using the survey to build on exploratory factor analysis that validated the survey questions and provide a confirmatory factor analysis to validate results; (2) emphasize established continuity with the learners and the peer evaluators in the study design to mitigate the limited interactions with the study participants outside of the actual simulation and during their clinical rotation; (3) training operational staff participating in gathering data on best practices of data collection for operations and research; (4) the team would encourage incorporating and evaluating the impact of the results on patient outcomes. Answering if patient outcomes improved with increased teamwork and collaboration of the interprofessional team. This would require a larger sample size of trainees involved in simulation training.

Conclusions

We were challenged with the reality of executing a research protocol in a highly complex health care environment, for example, clinician availability, time, response, ability for follow-up, change in protocol, data collection from clinical staff, etc. While these difficulties altered our study approach, the study team believes the design attempted in this study had merit in understanding the impact of a teamwork-focused ECMO simulation approach. We would encourage the medical community to build on the strengths of the design, fortify the weaknesses, and continue to emphasize the need for simulation training to improve ECMO care delivery and teamwork in the clinical setting. Especially as the field of simulation training continues to expand into new mediums like virtual reality [18].

Acknowledgments

Keck Medicine of USC (University of Southern California) Cardiac and Vascular Institute. Li Ding is supported by a grant (UL1TR001855) from the National Center for Advancing Translational Science (NCATS) of the US National Institutes of Health.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Training curriculum.
[DOCX File, 36 KB - mededu_v11i1e57424_app1.docx]

Multimedia Appendix 2
Extracorporeal membrane oxygenation circuit.
[DOCX File, 95 KB - mededu_v11i1e57424_app2.docx]

Multimedia Appendix 3
Study timeline.
[PDF File, 58 KB - mededu_v11i1e57424_app3.pdf]

Multimedia Appendix 4
Qualitative interview guide.
[DOCX File, 26 KB - mededu_v11i1e57424_app4.docx]

Multimedia Appendix 5
Quantitative survey tool.
[DOCX File, 39 KB - mededu_v11i1e57424_app5.docx]

Multimedia Appendix 6
Amended study timeline.
[PDF File, 58 KB - mededu_v11i1e57424_app6.pdf]

Multimedia Appendix 7
Qualitative focus group themes.

[DOCX File, 16 KB - mededu_v11i1e57424_app7.docx]

Multimedia Appendix 8
Quantitative survey results.
[DOCX File, 24 KB - mededu_v11i1e57424_app8.docx]

References

1. Betit P. Technical advances in the field of ECMO. Respir Care 2018 Sep;63(9):1162-1173. [doi: 10.4187/respcare.06320] [Medline: 30166411]
2. Banfi C, Bendjelid K, Giraud R. High-fidelity simulation for extracorporeal membrane oxygenation training, utile or futile? J Thorac Dis 2017 Nov;9(11):4283-4285. [doi: 10.21037/jtd.2017.10.54] [Medline: 29268492]
3. Fouilloux V, Gran C, Guervilly C, Breaud J, El Louali F, Rostini P. Impact of education and training course for ECMO patients based on high-fidelity simulation: a pilot study dedicated to ICU nurses. Perfusion 2019 Jan;34(1):29-34. [doi: 10.1177/0267659118789824] [Medline: 30014779]
4. Burton KS, Pendergrass TL, Byczkowski TL, et al. Impact of simulation-based extracorporeal membrane oxygenation training in the simulation laboratory and clinical environment. Simul Healthc 2011;6(5):284-291. [doi: 10.1097/SIH.0b013e31821dfcea]
5. ELSO guidelines for training and continuing education of ECMO specialists. 2010. URL: http://www.elso.org/Portals/0/IGD/Archive/FileManager/97000963d6cusersshyerdocumentselsoguidelinesfortrainingandcontinuingeducationofecmospecialists.pdf [accessed 2021-08-21]
6. Zakhary BM, Kam LM, Kaufman BS, Felner KJ. The utility of high-fidelity simulation for training critical care fellows in the management of extracorporeal membrane oxygenation emergencies: a randomized controlled trial. Crit Care Med 2017 Aug;45(8):1367-1373. [doi: 10.1097/CCM.0000000000002437] [Medline: 28422779]
7. Glass KM. Research in ECMO simulation: a review of the literature. In: Johnston LC, Su L, editors. Comprehensive Healthcare Simulation: ECMO Simulation: Springer International Publishing; 2021:147-152. [doi: 10.1007/978-3-030-53844-6_17]
8. Fehr JJ, Shepard M, McBride ME, et al. Simulation-based assessment of ECMO clinical specialists. Simul Healthcare 2016;11(3):194-199. [doi: 10.1097/SIH.0000000000000153]
9. Loeb D, Shoemaker J, Parsons A, Schumacher D, Zackoff M. How augmenting reality changes the reality of simulation: ethnographic analysis. JMIR Med Educ 2023 Jun 30;9:e45538. [doi: 10.2196/45538] [Medline: 37389920]
10. Malec JF, Torsher LC, Dunn WF, et al. The Mayo High Performance Teamwork Scale: reliability and validity for evaluating key crew resource management skills. Simul Healthcare 2007 ;2(1):4-10. [doi: 10.1097/SIH.0b013e31802b68ee]
11. Kirkpatrick JD, Kirkpatrick WK. Kirkpatrick's Four Levels of Training Evaluation: ATD Press; 2016.
12. Donovan AL, Aldrich JM, Gross AK, et al. Interprofessional care and teamwork in the ICU. Crit Care Med 2018 Jun;46(6):980-990. [doi: 10.1097/CCM.0000000000003067] [Medline: 29521716]
13. Ervin JN, Kahn JM, Cohen TR, Weingart LR. Teamwork in the intensive care unit. Am Psychol 2018;73(4):468-477. [doi: 10.1037/amp0000247] [Medline: 29792461]
14. Institute of Medicine (US) Committee on Quality of Health Care in America. Crossing the Quality Chasm: A New Health System for the 21st Century: National Academies Press; 2001. [Medline: http://www.ncbi.nlm.nih.gov/books/NBK222274]
15. Manthous CA, Hollingshead AB. Team science and critical care. Am J Respir Crit Care Med 2011 Jul 1;184(1):17-25. [doi: 10.1164/rccm.201101-0185CI] [Medline: 21471081]
16. Elshenawy S, Franciscovich CD, Williams SB, French HM. Interprofessional education and ECMO simulation. In: Johnston LC, Su L, editors. Comprehensive Healthcare Simulation: ECMO Simulation: Springer International Publishing; 2021:89-98. [doi: 10.1007/978-3-030-53844-6_10]
17. Wang Z, Gu R, Wang J, et al. Effectiveness of a game-based mobile app for educating intensive critical care specialist nurses in extracorporeal membrane oxygenation pipeline preflushing: quasi-experimental trial. JMIR Serious Games 2023 Dec 7;11:e43181. [doi: 10.2196/43181] [Medline: 38062643]
18. Gupta S, Wilcocks K, Matava C, Wiegelmann J, Kaustov L, Alam F. Creating a successful virtual reality-based medical simulation environment: tutorial. JMIR Med Educ 2023 Feb 14;9:e41090. [doi: 10.2196/41090] [Medline: 36787169]

Abbreviations

CVI: Cardiac and Vascular Institute
ECMO: extracorporeal membrane oxygenation
HIPAA: Health Insurance Portability and Accountability Act
ICU: intensive care unit
USC: University of California

Edited by B Lesselroth; submitted 23.02.24; peer-reviewed by A Hutchinson, FM Carr, M Capoccia; revised version received 16.05.24; accepted 03.12.24; published 24.01.25.

Please cite as:
Brown J, De-Oliveira S, Mitchell C, Cesar RC, Ding L, Fix M, Stemen D, Yacharn K, Wong SF, Dhillon A
Barriers to and Facilitators of Implementing Team-Based Extracorporeal Membrane Oxygenation Simulation Study: Exploratory Analysis
JMIR Med Educ 2025;11:e57424
URL: https://mededu.jmir.org/2025/1/e57424
doi:10.2196/57424

Game-Based Assessment of Cognitive Abilities and Personality Characteristics for Surgical Resident Selection: A Preliminary Validation Study

Noa Gazit[1,2], PhD; Gilad Ben-Gal[1*], PhD; Ron Eliashar[2*], MD

[1]Department of Prosthodontics, Faculty of Dental Medicine, Hebrew University of Jerusalem, Hadassah Medical Center, Kalman Ya'akov Man 1, Jerusalem, Israel

[2]Department of Otolaryngology/HNS, Faculty of Medicine, Hebrew University of Jerusalem, Hadassah Medical Center, Jerusalem, Israel

[*]these authors contributed equally

Corresponding Author:
Noa Gazit, PhD
Department of Prosthodontics, Faculty of Dental Medicine, Hebrew University of Jerusalem, Hadassah Medical Center, Kalman Ya'akov Man 1, Jerusalem, Israel

Abstract

Background: Assessment of nontechnical attributes is important in selecting candidates for surgical training. Currently, these assessments are typically made based on ineffective methods, which have been shown to be poorly correlated with later performance.

Objective: The study aimed to examine preliminary evidence regarding the use of game-based assessment (GBA) for assessing cognitive abilities and personality characteristics in candidates for surgical residencies.

Methods: The study had 2 phases. In the first phase, a gamified test was developed to assess competencies relevant for surgical residents. Three games were chosen, assessing 14 competencies: planning, problem-solving, ingenuity, goal orientation, self-reflection, endurance, analytical thinking, learning ability, flexibility, concentration, conformity, multitasking, working memory, and precision. In the second phase, we collected data from 152 medical interns and 30 expert surgeons to evaluate the test's feasibility, acceptability, and validity for candidate selection.

Results: Feedback from the interns and surgeons supported the relevance of the test for selection of surgical residents. In addition, analyses of the interns' performance data supported the appropriateness of the score calculation process and the internal structure of the test. Based on this data, the test showed good psychometric properties, including reliability (α=0.76) and discrimination (mean discrimination 0.39, SD 0.18). Correlations between test scores and background variables indicated significant correlations with gender, video game experience, and technical aptitude test scores (all P<.001).

Conclusions: This study presents an innovative GBA testing cognitive abilities and personality characteristics. Preliminary evidence supports the validity, feasibility, and acceptability of the test for the selection of surgical residents. However, evidence for test-criterion relationships, particularly the GBA's ability to predict future surgical performance, remains to be established. Future longitudinal studies are necessary to confirm its utility as a selection tool.

(JMIR Med Educ 2025;11:e72264) doi:10.2196/72264

KEYWORDS

resident selection; assessment; surgical training; cognitive abilities; personality characteristics; gamification; game-based assessment

Introduction

Selection of surgical training residents is an essential process aimed at ensuring that only the most capable candidates are chosen to undergo the rigorous training required to become qualified surgeons. Alongside technical skills, there is broad consensus that it is also crucial to assess nontechnical attributes, including cognitive abilities (eg, deductive reasoning, learning ability, and concentration) and personality characteristics (eg, decision-making, stress tolerance, and communication skills), in potential surgical residents [1-7]. Indeed, some even consider nontechnical attributes to be more relevant for selecting surgical trainees than technical aptitude [7-9]. In a recent study [7], 19 nontechnical competencies were identified as relevant to surgeons in the 21st century (6 cognitive abilities and 13 personality characteristics).

Traditionally, surgical training programs have assessed nontechnical attributes almost exclusively through proxies such as academic achievement, curricula vitae, letters of recommendation, and unstructured interviews [10,11]. However, studies suggest that these methods are poorly correlated with

later performance during residency [11-16]. In light of such findings, some studies have examined the use of self-report measures as a potential alternative. For example, studies have explored the potential of self-report questionnaires for assessing personality, emotional intelligence, and grit. But there is as yet no consistent evidence that these methods improve the selection of surgical residents [5,17]; and these tools are subject to all the potential problems and biases of self-reports, from poor introspective ability to outright dishonesty [18,19]. Hence, better ways of assessing surgical residency candidates are needed.

One promising new approach is to analyze behavior itself using simulated tasks, where examinees are exposed to controlled situations designed to elicit behaviors relevant to the assessment of specific competencies. This method is expected to have higher predictive value than either traditional methods or self-reports.

A simulation test can be conducted in the real world by evaluators or actors, or on a computer using emerging technologies such as virtual reality and gamification. Gamification refers to the incorporation of game elements into nongaming activities, and its application to personnel selection has led to the development of game-based assessments (GBAs). GBAs use gameplay behaviors to assess job-related skills, abilities, and characteristics, and they have many advantages over traditional assessments and noncomputerized simulation tests for predicting job performance [20-23]. First, GBAs promote a more positive assessment experience that reduces examinees' stress levels and increases their engagement and motivation. Second, GBAs are based on an automated scoring system, which eliminates the bias often associated with human assessments. Finally, GBAs can collect rich high-resolution spatiotemporal data capturing examinees' behavior throughout the test, allowing the entire solving process to be examined rather than just the final result or answer. These advantages may lead to a more reliable and valid assessment of examinees' skills and abilities.

As GBAs are still relatively new, only a limited number of studies have examined their use in hiring and recruitment [24-26], and to the best of our knowledge, no study has evaluated GBAs as a tool for selecting medical residents. The current study examines the use of GBA for assessing cognitive abilities and personality characteristics identified as relevant for surgical residents in an initial phase of job analysis [7]. This study is the first in a planned series of studies aimed at establishing the validity of the GBA. Here, we present preliminary evidence of its feasibility, acceptability, and validity in the context of surgical resident selection, based on feedback and behavioral data from potential candidates and expert surgeons. Further research linking the GBA scores to future surgical performance will be necessary to complete the validation process.

Methods

We developed a gamified assessment test relevant for appraising the cognitive abilities and personality characteristics of potential surgical residents and examined preliminary evidence for its validity, feasibility, and acceptability. In accordance with the contemporary understanding of validity as a unified concept, we collected and evaluated evidence related to 4 sources of validity: content, internal structure, response process, and relationships with other variables [27,28], although the evidence for relationships with other variables was limited and did not include test-criterion relationships. The evidence collected is based on both the procedures used in the development and revision of the test and the empirical data collected during the study.

Ethical Considerations

The study was approved by the ethics committee of the Hebrew University of Jerusalem (approval no. 13032023), and all participants provided informed consent. Participant data were stored using a unique fake identifier; the key linking these identifiers to real identities was kept in a password-protected file stored offline, ensuring that no identifying information was accessible online. Interns received US $75 for participating in the study, as well as feedback regarding their performance in both tests relative to the rest of the sample (the percentile rankings of their total scores).

Test Development

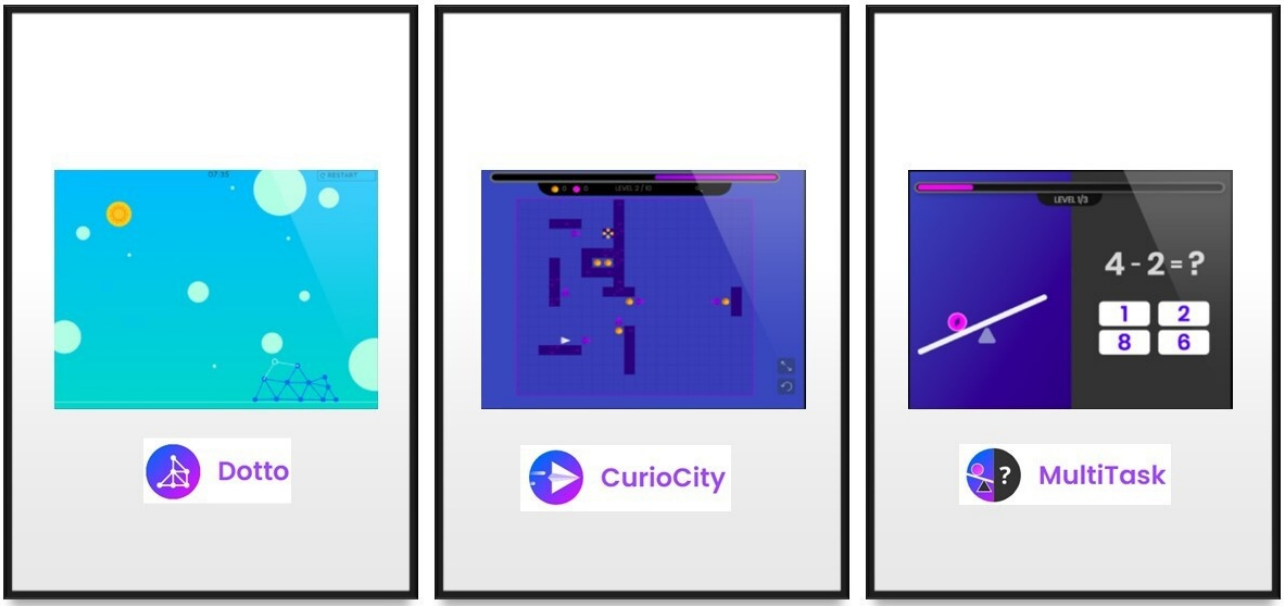
The GBA

The GBA used in this study was developed in cooperation with Benchmark.games LTD (Hungary), a company that produces GBAs for use in organizational hiring and recruitment. Tests are tailored to the organization's needs, based on video games developed specifically for the assessment of various competencies (eg, analytical thinking, planning, or multitasking). Each test is administered on a standard computer and requires only a stable internet connection and a mouse.

The test developed for this study is based on three video games adapted to capture competencies needed by surgical residents: (1) Dotto, (2) CurioCity, and (3) MultiTask (refer to Figure 1). In the Dotto game, the goal is to build a structure by inserting and manipulating points and lines to reach a target while overcoming physics-based challenges. The game confronts examinees with a problem-solving situation that is not clearly defined, requiring them to discover the rules for solving the problem on their own. In CurioCity, examinees are tasked with finding their way through a maze to reach the target area. The game consists of 16 levels with varying requirements and levels of difficulty. Once again, some of the rules must be discovered by examinees, and some rules change as the game proceeds, to test the adaptability and flexibility of the examinees. Finally, in the MultiTask game, examinees are asked to perform 2 nonverbal tasks simultaneously (eg, a swing balancing task and a simple arithmetic task). The game has three levels, each using a different combination of 2 tasks. The initial versions of the games were developed by psychometricians and psychologists employed by Benchmark.games, and the games were validated using data from hundreds of employees by Benchmark.games for general personnel selection. For this study, all 3 games were modified based on feedback from the research team in 3 ways: levels that were insufficiently challenging for candidates with high abilities were excluded; tasks that assessed irrelevant competencies (eg, typing speed and accuracy) were replaced with tasks assessing competencies relevant for surgical trainees

(eg, concentration and working memory); and, to ensure that the assessment would be objective and standardized, the instructions and demonstrations for each game were revised and improved. Instructions were provided in English and included both written instructions and video demonstrations. Furthermore, to ensure that the instructions were understood correctly, each game was preceded by a few minutes of practice. The initial version of the test was then pilot-tested with 8 medical students. Based on their feedback, changes were made in the instructions and in the test interface. The entire test takes about 45 - 60 minutes to complete, with each game taking 15 - 20 minutes.

Figure 1. Illustrations of the video game assessments selected for the test. The games are shown in the order in which they appeared in the test.



The video games were selected to assess 14 relevant competencies: planning, problem-solving, ingenuity, goal orientation, self-reflection, endurance, analytical thinking, learning ability, flexibility, concentration, conformity, multitasking, working memory, and precision. Definitions of the competencies are provided in Table 1.

The competencies were drawn from a set of cognitive abilities and personality characteristics identified as relevant for selection of surgical residents in a previous phase of job analysis conducted by the research team [7]. However, the GBA does not assess some of the competencies which were identified as relevant to selection for surgical training (in particular, “soft skills” such as interpersonal skills, teamwork, leadership, and integrity). These competencies were not assessed in the present research because they are not susceptible to measurement using computerized and automated systems. The GBA was designed such that each game would elicit specific behaviors relevant to 2 or more of the 14 competencies, with each competency assessed using information obtained from one game (except for precision, which was assessed in all 3 games).

Table . Competencies assessed in the game-based assessment (GBA) test.

Competency	Description	Video game used to assess the competency
Planning	Ability to plan the steps required to solve the task, and to implement the plan in order to achieve the goal.	Dotto
Problem-solving	Ability to work through unexpected obstacles and challenges that arise during the task.	Dotto
Ingenuity	Ability to test the boundaries of a problem and to seek unique and creative solutions.	Dotto
Goal orientation	Ability to translate an intention into action (ie, to stay focused on achieving the goal).	Dotto
Self-reflection	Ability to learn from failure and to adopt a new approach.	Dotto
Endurance	Ability to invest effort for an extended period of time.	Dotto
Analytical thinking	Ability to collect, organize, and implement the information needed to solve the problem.	CurioCity
Learning ability	Ability to recognize "rules" quickly and effectively and apply them in the relevant situation.	CurioCity
Flexibility	Ability to adapt to changes in the situation.	CurioCity
Concentration	Ability to stay focused and to maintain high performance even in monotonous repetitive tasks.	CurioCity
Conformity	Ability and willingness to follow rules and instructions.	CurioCity
Multitasking	Ability to split attention between two tasks without harming performance.	MultiTask
Working memory	Ability to store and retrieve information in short-term memory.	MultiTask
Precision	Ability to perform the task in an accurate manner, with few errors.	All games

Scoring

The gamified tasks provide the stimuli by which the program measures candidates' behavior. In each game, all actions of examinees (eg, mouse movements and key presses) are recorded and logged. Approximately 2000 data points are recorded for each 15-minute gameplay session. These raw data are then transformed into higher-level variables that describe a set of meaningful measurements (eg, time to first response, time between actions, accuracy, number of steps, and learning curve). Then, competency scores are calculated using an aggregation (ie, linear combination) of the relevant variables, with higher weight given to variables characterized by larger variance between candidates.

The initial mapping between different variables and competencies was determined by a team of psychologists and psychometricians employed by the company following a theory-driven approach [20]. This mapping was tested and improved based on empirical data from hundreds of employees, and variables that did not converge with the expected pattern were excluded from consideration. The mapping was then further validated based on correlations with other measures of cognitive abilities and personality (eg, Raven's Progressive Matrices, the Stroop test, scales of the International Personality Item Pool, and the Bar-On Emotional Quotient Inventory; refer to Table S1 in Multimedia Appendix 1).

Competency scores are computed and standardized based on a norm created using a database of over 5000 observations. Scores are presented on a scale of 1 - 10. For this study, we also calculated a total test score for each examinee by averaging the individual competency scores (with equal weight for each competency). To facilitate interpretation of the results, the total scores were then scaled to have a mean of 100 and a SD of 20.

Validation

Sample and Procedure

To evaluate the test's validity, feasibility, and acceptability, we recruited 30 experienced surgeons from 3 hospitals and 152 medical interns from 10 hospitals in Israel. The surgeons were asked to review the test and then complete a feedback questionnaire (see below). The interns were asked to complete the test, and their test data was collected and analyzed to evaluate the internal structure and psychometric characteristics of the test (discrimination, reliability, and correlations between competency scores). The interns also completed a feedback questionnaire similar to that filled in by the surgeons.

The expert surgeons were recruited using an email invitation. Email addresses of potential participants were obtained from hospital websites or from the Israeli medical association database. Recruitment continued until we had 30 participants. Surgeons who were willing to participate in the study were invited to review the gamified test and to complete the feedback questionnaire.

The interns were recruited using an invitation posted in relevant Facebook and WhatsApp groups. Recruitment continued until at least 150 participants were enrolled. Participants were invited to attend a session in which we administered the gamified assessment test and a separate technical aptitude test developed by Gazit et al [29]. The technical aptitude test included 10 basic tasks performed on the Lap-X VR laparoscopic simulator [30] and was designed to assess technical skills relevant for surgery such as dexterity, visuospatial perception, coordination, and arm-hand steadiness. The order of the tests varied, such that some participants started with the GBA and others with the technical aptitude test, with a short break between the two. The interns were told that each game in the GBA should take around 15 - 20 minutes to complete.

Questionnaire

The questionnaires filled in by the surgeons and interns were nearly identical. Participants in both samples were asked to provide four main ratings for each game: (1) its relevance for selecting candidates for surgical training (on a 5-point Likert scale, 1=not relevant, 5=extremely relevant); (2) its difficulty (also on a 5-point Likert scale, 1=very easy, 5=extremely difficult); (3) whether the time limit was sufficient (yes or no); and (4) whether the instructions were clear (yes or no). In addition, participants provided 2 ratings for the test as a whole: the relevance of the entire test and the comfort of the test platform (both on 5-point Likert scales, 1=not relevant or not comfortable, 5=extremely relevant or comfortable). Participants were also invited to share general comments and suggestions for improving each game and the whole test using free text. Finally, each participant provided demographic information (for interns: age, gender, dominant hand, desired training field [surgical or nonsurgical], and previous experience with video games; for the surgeons: age, gender, surgical specialty, and number of years working in the field). Previous experience with video games was reported on a 5-point scale (1=no experience, 5=very extensive experience).

Analyses

Some validity evidence is encompassed in the procedures used in the development of the test described above (selection of games and tasks based on job analysis; development of the games and scoring method by psychometricians and psychologists; and calculation of scores based on a norm sample). Further evidence of validity is derived from the empirical data collected in this study. In particular, internal structure evidence, response process evidence, and relationships with other variables were obtained from analysis of the interns' test performance data. Content evidence, feasibility, and acceptability were obtained from the feedback questionnaires completed by both the interns and surgeons.

To analyze the performance data of the interns, we first examined the distribution of the competency scores and calculated Pearson correlations between them to support computation of a composite score for each participant, representing that participant's total performance in the test (response process evidence of validity). We then conducted an item analysis to assess the discrimination of each competency and the reliability of the whole test, and a factor analysis to assess whether the structure of the test variables accords with what is theoretically expected (together these provide internal structure evidence for validity). Finally, we calculated correlations between participants' scores in the gamified test and other variables: their demographic characteristics (age, gender, dominant hand, desired training field, and previous experience with video games) and their technical aptitude test scores (evidence of relationship to other variables).

To analyze the data from the feedback questionnaires of the interns and surgeons, we first calculated, for each sample, mean relevance and difficulty ratings for each game. We then analyzed the data on the time limits and clarity of instructions for each game, as described above, and calculated the mean relevance and comfort ratings for the whole test. Finally, we analyzed the general comments obtained from participants in the open-ended question to identify common remarks and suggestions. All statistical analyses were performed using R, version 4.2.2 (R Foundation for Statistical Computing, Vienna, Austria).

Results

Overview

In total, 152 interns (71 females, 46%) from 10 academic hospitals in Israel and 30 expert surgeons (4 females, 13%) from three academic hospitals in Israel participated in the study. Demographic characteristics of the participants are presented in Table 2.

Table . Demographic characteristics of study participants.

Group and characteristic	Values
Interns (n=152)	
Age in years, mean (SD)	28.3 (3.8)
Gender (female), n (%)	71 (46)
Dominant hand (left), n (%)	13 (9)
Desired training field, n (%)	
Surgical training	100 (65)
Nonsurgical training	36 (24)
Not decided	17 (11)
Experience with video games, n (%)	
No experience	22 (14)
Little experience	45 (29)
Moderate experience	46 (30)
Considerable experience	20 (13)
Very extensive experience	20 (13)
Expert surgeons (n=30)	
Age in years, mean (SD)	53.8 (8.4)
Gender (female), n (%)	4 (13)
Years of experience, mean (SD)	13.5 (7.9)
Surgical specialty, n (%)	
General surgery	8 (27)
Gynecology	5 (17)
Orthopedics	10 (33)
Otorhinolaryngology–head and neck surgery	4 (13)
Urology	3 (10)

Performance Data of Interns

First, competency scores and total test scores were calculated for each of the interns. The means and SDs of the competency scores and total scores are presented in Table 3. The total test scores ranged from 44 to 142 (a range of 98). Figure 2 displays the distribution of the total scores for the 152 interns (The distributions of the individual competency scores can be found in Figure S1 in Multimedia Appendix 1).

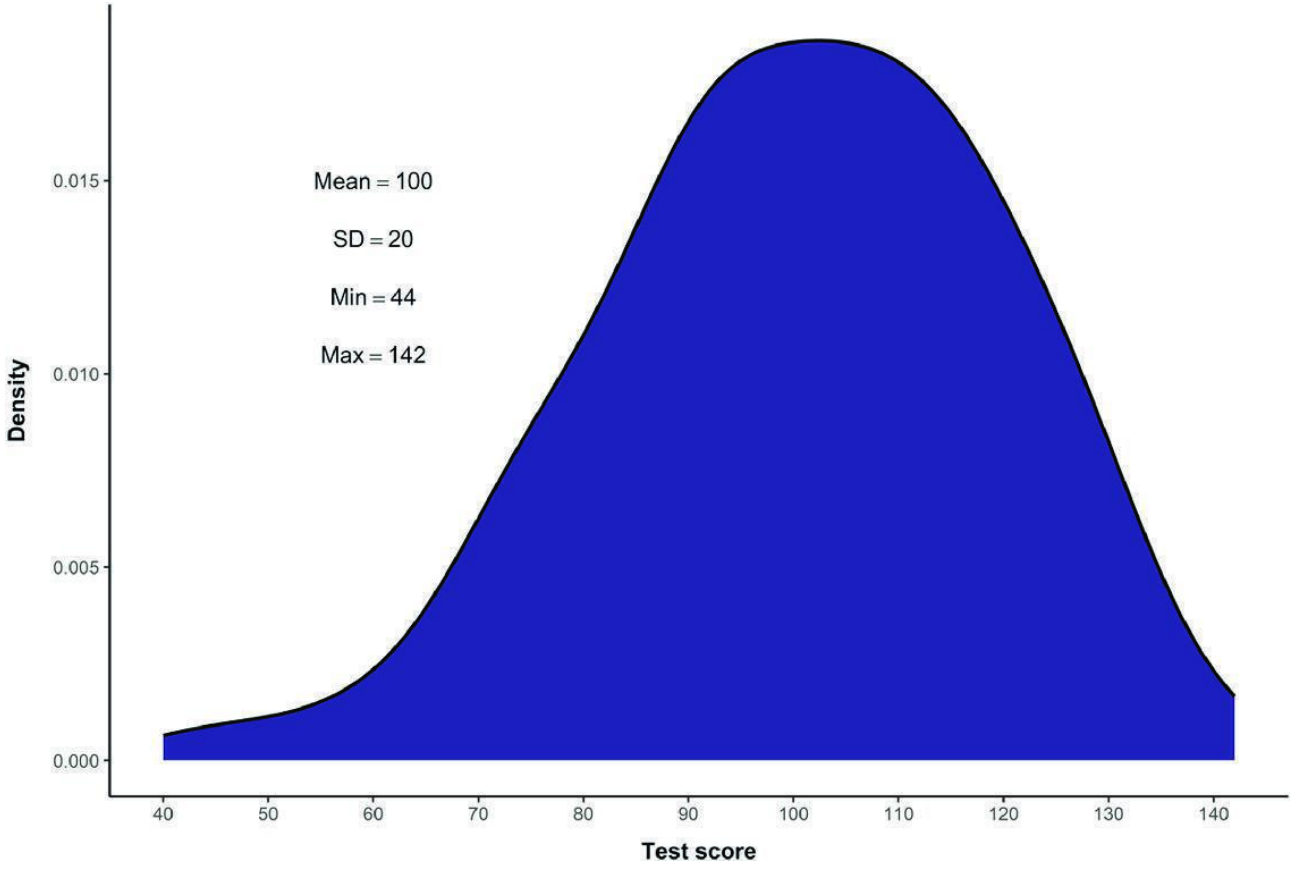
Table . Descriptive statistics and item analysis of the game-based assessment (GBA) test.

Competency	Mean	SD	Skew	Competency discrimination	Cronbach α if deleted
Planning	6.30	2.24	−0.29	0.64	0.72
Problem-solving	5.75	2.51	−0.08	0.47	0.73
Ingenuity	4.22	2.19	0.13	0.34	0.74
Goal orientation	4.79	2.38	−0.29	0.21	0.76
Self-reflection	4.87	3.09	0.14	0.23	0.76
Endurance	3.77	2.50	0.31	0.06	0.77
Analytical thinking	7.79	1.87	−1.13	0.46	0.74
Learning ability	7.23	1.97	−0.70	0.40	0.74
Flexibility	6.15	2.71	−0.25	0.36	0.74
Concentration	7.98	1.91	−1.30	0.34	0.75
Conformity	4.63	2.33	−0.10	0.17	0.76
Multitasking	7.39	2.38	−1.13	0.56	0.72
Working memory	6.22	3.47	−0.43	0.46	0.73
Precision	7.24	1.82	−0.71	0.71	0.72
Total test score[a]	100.00	20.00	−0.55	—[b]	—

[a]Cronbach α=0.76.

[b]Not applicable.

Figure 2. Distribution of total GBA test scores in the intern sample (n=152). GBA: game-based assessment.



To support the calculation of a total test score based on the competency scores, we examined the Pearson correlations between the competency scores. Most of the correlations were high (refer to Table S2 in Multimedia Appendix 1). To support the internal structure of the test, an item analysis was then

conducted to assess the discrimination of each competency and the reliability of the whole test (see Table 3).

The results showed good psychometric properties: the discrimination was satisfactory for all competencies (mean 0.39, SD 0.18), and the test's internal reliability was high (α=0.76). In addition, we conducted exploratory factor analysis with Promax rotation on the 14 competencies. The Kaiser–Meyer–Olkin measure of sampling adequacy suggested that the data was factorable (Kaiser–Meyer–Olkin=0.69). The factor analysis yielded a 2-factor solution, such that seven competencies (analytical thinking, learning, flexibility, concentration, working memory, multitasking, and precision) loaded on one factor, and 6 competencies (planning, problem-solving, ingenuity, goal orientation, self-reflection, and endurance) loaded on the second factor. The only exception was conformity, which did not load on either of the factors. Based on our previous job analysis [7], we defined the first group as cognitive abilities and the second group as personality characteristics. The correlation between the two factors was 0.5. Detailed results for the factor loadings can be found in Table S3 in Multimedia Appendix 1.

We next calculated correlations between the total test scores in the GBA and external variables, including participants' demographic characteristics and their scores in the separate technical aptitude test described earlier. No significant correlations were found between age, dominant hand, or desired training field and the total GBA scores. However, a significant difference emerged with respect to gender, such that males (mean 104.6, SD 16.8) scored significantly higher than females (mean 94.3, SD 21.9) on the gamified test (mean difference 10.9, 95% CI 3.1-17.6, t_{150}=2.8, P=.002, Cohen d=0.52). This represents a small-to-medium effect size. In addition, we found a significant low positive correlation between the total GBA scores and reported amount of previous experience with video games (r_{150}=0.26, P<.001). Interestingly, when we controlled for video game experience, the difference between the genders was no longer significant, suggesting that this difference is mainly due to different levels of video game experience.

Finally, we also calculated the correlation between the total GBA scores and scores in the technical aptitude test. We found a significant correlation between the 2 sets of scores (r_{150}=0.46, P<.001). When controlling for video game experience, the correlation remained significant, though slightly reduced (semipartial r_{152}=0.38, P<.001), suggesting that while gaming experience contributes to the association, the majority of the shared variance likely reflects underlying competencies relevant to both assessments. Supporting this interpretation, we found significant correlations between technical aptitude test scores and several nontechnical competencies measured by the GBA: planning, r_{150}=0.28; problem-solving, r_{150}=0.28; analytical thinking, r_{150}=0.27; learning ability, r_{150}=0.30; flexibility, r_{150}=0.50; and precision, r_{150}=0.30; all P<.001. In the absence of these 6 competencies, the total GBA scores showed no significant correlation with the technical aptitude test (r_{150}=0.11, P=.17). These findings suggest that shared cognitive and behavioral attributes may play an important role in performance on both tests.

Questionnaire Data

Table 4 presents the main results for the questionnaire data, including mean relevance and difficulty ratings for each game, and the rates at which participants judged the time limits as sufficient and the instructions as clear.

Table . Feedback of interns and expert surgeons on the relevance,[a] difficulty,[b] time limit,[c] and clarity of instructions[d] for each game in the game-based assessment (GBA) test.

Game and group		Relevance rating, mean (SD)	Difficulty rating, mean (SD)	Time limit, n (%)	Clarity of instructions, n (%)
Dotto					
	Interns	3.5 (0.8)	4.5 (0.4)	95 (62)	94 (61)
	Surgeons	3.8 (0.6)	4.2 (0.7)	22 (73)	21 (70)
CurioCity					
	Interns	3.8 (0.6)	2.9 (0.8)	151 (99)	144 (94)
	Surgeons	3.7 (0.7)	3.7 (0.6)	27 (90)	24 (80)
MultiTask					
	Interns	3.7 (0.6)	2.9 (0.7)	142 (93)	147 (96)
	Surgeons	3.6 (0.7)	2.5 (0.5)	29 (97)	27 (90)

[a]The relevance rating scale ranged from 1 to 5, with higher scores indicating greater relevance for selection of surgical residents (1="not relevant", 2="slightly relevant", 3="moderately relevant", 4="very relevant", 5="extremely relevant").

[b]The difficulty rating scale ranged from 1 to 5, with higher scores indicating greater difficulty (1="very easy", 2="easy", 3="moderately difficult", 4="very difficult", 5="extremely difficult").

[c]Participants were asked whether the time limit was sufficient for the task. The number in the table represents the number of interns and surgeons who responded "yes."

[d]Participants were asked whether the instructions for the task were clear. The number in the table represents the number of interns and surgeons who responded "yes." The instructions were modified slightly based on the surgeons' feedback before the test was administered to the interns.

Addressing the latter first, overall, both the interns and expert surgeons regarded the time limits as sufficient (the lowest time limit approval rating was 62% of the interns for the Dotto game; for CurioCity and MultiTask, all ratings were 90% or above). Both samples also considered the instructions to be generally clear (again, the lowest approval rating was by the interns for the Dotto game, at 61%; see Table 4). Before the test was administered to the interns, some of the instructions were modified slightly and improved based on feedback provided by the expert surgeons either verbally or in writing.

The difficulty ratings varied between games, with the CurioCity and MultiTask games perceived overall as being moderately difficult, and the Dotto game largely perceived as very difficult to extremely difficult. The mean difficulty rating across the games and samples was 3.5 (SD 0.7), meaning that the test as a whole was perceived as moderately to very difficult. All games were considered by both the expert surgeons and the interns as relevant for assessing cognitive abilities and personality characteristics in the selection of candidates for surgical training (manifested in average ratings of 3.5 or above; see Table 4). The mean relevance rating across the games and samples was 3.6 (SD 0.1). Looking at the whole-test ratings, the mean relevance ratings were relatively high (interns: mean 3.6, SD 0.7; expert surgeons: mean 3.7, SD 0.6). In addition, the test platform was perceived as comfortable to use (interns: mean 4.2, SD 0.2; expert surgeons: mean 4.0, SD 0.3).

As noted, we also analyzed participants' written feedback (in the free-text portion of the questionnaire), as well as feedback provided orally by the expert surgeons. Some of the surgeons indicated that their relevance ratings would have been higher if the tasks in the GBA were more directly related to surgical tasks and scenarios. Some participants also suggested that the test would be more relevant if it assessed other important competencies not covered in the current version, such as interpersonal skills, teamwork, leadership, and integrity. Finally, participants also expressed concern that prior experience with video games could affect performance on the test.

Discussion

Study Overview and Significance

This paper presents an innovative gamified test designed to assess cognitive abilities and personality characteristics relevant to the selection of surgical residents. While several studies have evaluated the use of GBAs in assessing applicants for employment, this is, to our knowledge, the first to evaluate their use in selecting surgical residents. As part of a broader program of validation research, this initial study provides preliminary evidence supporting the tool's feasibility, acceptability, and validity.

Evidence for Validity

Overview

On the basis of feedback from surgeons and interns regarding the test's relevance, difficulty, and administration, the results of this study support the feasibility and acceptability of the test. We also present preliminary evidence concerning 4 of the 5 main components of construct validity: content, response process, internal structure, and relationships with other variables (the fifth component, consequences, could not be examined in this study) [27,28]. In some cases, the evidence is based on procedures used in the development and adaptation of the test; in others, it is based on empirical data collected during the study.

Content

In terms of content, the games used in the GBA were selected to assess relevant cognitive abilities and personality characteristics based on competencies identified in a previous job analysis [7]. The games were developed and validated by psychometricians and psychologists to evaluate these specific competencies, and both the interns and surgeons participating in the study rated the games as relevant for selecting candidates for surgical training. Some of the expert surgeons indicated that their relevance ratings would have been higher if the content of the games were more directly related to surgery or medicine. This weakens somewhat the content evidence for validity. However, the literature on gamification suggests that GBAs can effectively assess relevant competencies even when the game scenario seems unrelated to the profession [26]. Future studies should examine whether GBAs that more directly mimic job-related situations are more valid for selecting qualified candidates.

Response Process Evidence

Response process evidence of validity has 2 components. The first is the elimination of sources of error associated with test administration [28]. Toward this end, we provided detailed and thorough instructions for each game. The instructions were revised based on feedback provided by the expert surgeons before the test was administered to the interns. The ratings of both the expert surgeons and interns indicate that on the whole, the instructions were perceived as clear.

The second component of response process evidence is the appropriateness of the methods used to combine different performance parameters to produce a composite score. To support the calculation of a total test score based on the competency scores, we examined the correlations between the competency scores. Strong correlations were obtained, supporting the calculation of a composite performance score.

Internal Structure Evidence

Internal structure, as a source of validity, relates to the statistical or psychometric characteristics of the test. The item analysis conducted on the test data of the interns showed good psychometric properties, supporting the internal structure of the test. In addition, the factor analysis yielded two groups of competencies, one reflecting cognitive abilities and the other personality characteristics. This result is consistent with previous classifications of these competencies [4,7,31], and therefore also in keeping with the test's expected internal structure.

Relationships With Other Variables

This source of evidence relates to the "degree to which these relationships are consistent with the construct underlying the proposed test score interpretation" [32]. Most commonly, this evidence is assessed based on correlations of assessment scores

with a criterion measure of future workplace performance. While this type of evidence is indeed crucial for the validation of the current test, it was not available in this initial study.

Instead, the present analysis relies on a different methodology, namely, examining whether the relationships found in this study between test scores and external variables are consistent with what is known from the literature regarding the relationship between nontechnical competencies and those variables. Based on the data of interns, we calculated the correlations between participants' performance on the gamified test and other variables.

As expected, no correlations were found with age, dominant hand, or the intern's desired training field. We found relatively small but statistically significant correlations with both gender and self-reported video game experience, with males and frequent gamers obtaining higher GBA scores. Notably, the gender difference was largely accounted for by differences in video game experience, suggesting that the observed gender effect is explained by greater familiarity with video games among males. These findings are in line with other studies showing that gamers and males may potentially have advantages over nongamers and females in the context of GBAs [33,34], and they raise questions regarding the fairness of these tests. Since there is evidence that playing video games improves cognitive and mental abilities [35,36], it is unclear whether the correlation between video game experience and the gamified test scores found in this study reflects a genuine positive influence of video games on gamers' abilities, or whether it is simply an artifact of the test format that may bias the selection process. Future research should examine whether changes in instructions, allowing more practice time before the test, or changes in GBA features and measures may eliminate these advantages [33]. In addition, further studies should examine whether increasing women's exposure to video games in general would help to minimize this gender gap. However, it is important to note that the observed gender effect was small to medium in size, and the effect of video game experience was small. Thus, while caution is warranted, these differences should not be overstated. Until further evidence is available, the use of adjusted cutoffs or gender-specific norms may help avoid exacerbating the underrepresentation of women in surgical fields.

In addition, it is important to acknowledge that the GBA examined in this study does not encompass the full range of cognitive abilities and personality characteristics relevant for selecting surgical residents. Notably, key nontechnical competencies such as interpersonal skills, teamwork, leadership, and integrity were not addressed in the current assessment. Furthermore, the tasks included were primarily procedural and did not involve verbal abilities. As previous research has shown that males and females may excel in different domains—with females often demonstrating strengths in tasks that require verbal abilities [37] and interpersonal skills [38,39]—it is plausible that a more comprehensive assessment approach could mitigate the small gender differences observed in this study. For example, incorporating tools that evaluate verbal and interpersonal competencies might balance the overall selection outcomes. Future research should investigate whether expanding the assessment battery to include gamified situational judgment tests [21,40] or other instruments targeting these nontechnical domains could enhance fairness and reduce gender disparities in selection.

Moreover, we found a medium correlation between the gamified test scores and scores on a technical aptitude test performed using a virtual reality laparoscopic simulator. Since video game experience has been shown to correlate with initial performance on laparoscopic simulators [41], we considered the possibility that this shared factor may contribute to the observed association, that is, that previous video game experience might positively influence performance on both assessments. However, the correlation remained significant even after controlling for video game experience, suggesting that gaming experience only partially explains the relationship between the 2 tests.

In addition to this shared factor, our findings suggest that common underlying competencies may also play a role. Specifically, scores on the technical aptitude test were significantly associated with nontechnical competencies measured by the GBA, such as planning, problem-solving, analytical thinking, learning ability, flexibility, and precision. These results indicate that both assessments may tap into similar cognitive processes or behavioral tendencies. This interpretation is supported by prior research demonstrating meaningful correlations between nontechnical skills and performance on laparoscopic simulators [42-44].

To further disentangle the effects of gaming experience from shared competencies, future research should examine whether the correlation between GBA and laparoscopic simulator performance persists among individuals with previous laparoscopic experience. Alternatively, exploring the relationship between GBA scores and performance on open surgery tasks—which are not influenced by video game experience—could help clarify whether the observed correlation is driven by familiarity with gaming or by genuine overlap in nontechnical competencies.

Finally, as only 21% of the variance in GBA scores is explained by the technical aptitude test, it is clear that the GBA primarily measures competencies beyond those assessed by the laparoscopic simulator. This finding supports both the convergent and divergent validity of the GBA and aligns with its intended construct interpretation [32].

Implications

Nontechnical skills are important for surgeons no less, and perhaps even more, than technical skills [7]. Indeed, many underlying causes of error within and outside the operating room originate from nontechnical aspects of performance [8]. Hence, training programs recognize the importance of assessing candidates' cognitive abilities and personality characteristics when selecting each year's cohort of surgical residents. Yet traditional assessment methods (academic achievement, curricula vitae, letters of recommendation, and interviews) are poorly correlated with later performance; and self-report measures, a potential alternative, are subject to bias and dishonesty.

The present study introduces an innovative solution for assessing relevant competencies: game-based assessment [21,22,25]. Building on existing GBAs developed for hiring and recruitment contexts, we implemented a systematic process to develop a gamified test tailored for surgical resident selection and conducted an initial investigation into its validity. Gamified assessment tests offer numerous advantages over other assessment approaches. First, they examine the entire solving process, as opposed to traditional tests which only examine the final product, allowing for a deeper understanding of the candidate's competencies and work style. Compared to self-report measures, GBAs measure candidates' actual behavior, which is harder to fake. Finally, gamified tests are based on automated scoring, thus minimizing the influence of bias in the selection process.

The present findings provide preliminary support for the feasibility, acceptability, and validity of the gamified test, suggesting that it may contribute to improving the selection of surgical residents by offering a potentially more reliable assessment of candidates' abilities and attributes. It follows that implementing this test—or a similar tool—may assist program directors in identifying candidates with strong potential for success in surgical training. This improved selection process should, in turn, result in more capable surgical residents and surgeons, ultimately leading to better surgical outcomes and increased patient safety. Our findings may be relevant to nonsurgical training programs as well, since some of the competencies assessed in the gamified test developed in this study apply to residents in all medical fields.

The gamified test presented in this study does not assess all cognitive abilities and personality characteristics relevant for selecting surgical residents. As mentioned by the participants in this study, competencies missing in the present work include interpersonal skills, teamwork, leadership, and integrity. Future studies should examine whether other types of GBAs, such as gamified situational judgment tests [21], or other assessment methods may be useful in improving this area.

Strengths and Limitations

This is the first study to examine the use of GBAs in selecting surgical residents, or indeed medical residents in any field. As such, one of its key strengths is use of a systematic process to develop a novel test for assessing candidates' cognitive abilities and personality characteristics and to evaluate its validity, feasibility, and acceptability. Another strength is the large sample of expert surgeons (30) and interns (152) from various hospitals who provided data for statistical analysis (the interns) and feedback (both samples).

The study has some limitations. First, our participants came from a single country, thereby restricting the generalizability of our findings. However, it seems unlikely that the competencies we assessed are distributed differently among candidates from other nations. In addition, since the interns in our study were volunteers, it is possible that our sample does not represent the population of candidates for surgical training. Future studies should aim to recruit a more randomized and representative sample to ensure the findings are generalizable to the broader population of surgical trainees. However, the large variance in competency and test scores observed in our sample suggests that our sample was likely sufficiently representative of candidates with different qualifications. Finally, an important limitation of this study is the absence of evidence for test-criterion relationships. While we present data supporting various sources of validity, we have not yet assessed whether the GBA scores predict future performance in surgical residency. Given the high-stakes nature of surgical selection, establishing evidence for test-criterion relationships is critical before the tool can be adopted for widespread use. Longitudinal studies that track residents' real-world performance over time are planned to address this essential aspect.

Conclusions

The use of GBAs holds potential for contributing to improvements in resident selection. The present study presents an innovative gamified test designed to assess cognitive abilities and personality characteristics relevant to the selection of surgical residents. Preliminary evidence supports the feasibility, acceptability, and validity of the gamified test. However, further research is needed, particularly to assess evidence for test-criterion relationships, before the tool can be fully recommended for surgical resident selection.

Acknowledgments

This research was supported by the Israel Science Foundation (grant No. 1830/20). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability

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

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary material.
[PDF File, 339 KB - mededu_v11i1e72264_app1.pdf]

References

1. Cuschieri A, Francis N, Crosby J, Hanna GB. What do master surgeons think of surgical competence and revalidation? Am J Surg 2001 Aug;182(2):110-116. [doi: 10.1016/s0002-9610(01)00667-5] [Medline: 11574079]
2. Baldwin PJ, Paisley AM, Brown SP. Consultant surgeons' opinion of the skills required of basic surgical trainees. Br J Surg 1999 Aug;86(8):1078-1082. [doi: 10.1046/j.1365-2168.1999.01169.x] [Medline: 10460649]
3. Dean B, Jones L, Garfjeld Roberts P, Rees J. What is known about the attributes of a successful surgical trainer? A systematic review. J Surg Educ 2017;74(5):843-850. [doi: 10.1016/j.jsurg.2017.01.010] [Medline: 28392267]
4. Gardner AK, Cavanaugh KJ, Willis RE, et al. Great expectations? Future competency requirements among candidates entering surgery training. J Surg Educ 2020;77(2):267-272. [doi: 10.1016/j.jsurg.2019.09.001] [Medline: 31606376]
5. Bann S, Darzi A. Selection of individuals for training in surgery. Am J Surg 2005 Jul;190(1):98-102. [doi: 10.1016/j.amjsurg.2005.04.002] [Medline: 15972179]
6. Grantcharov TP, Reznick RK. Training tomorrow's surgeons: what are we looking for and how can we achieve it? ANZ J Surg 2009 Mar;79(3):104-107. [doi: 10.1111/j.1445-2197.2008.04823.x] [Medline: 19317771]
7. Gazit N, Ben-Gal G, Eliashar R. Using job analysis for identifying the desired competencies of 21st-century surgeons for improving trainees selection. J Surg Educ 2023 Jan;80(1):81-92. [doi: 10.1016/j.jsurg.2022.08.015] [Medline: 36175291]
8. Yule S, Flin R, Paterson-Brown S, Maran N. Non-technical skills for surgeons in the operating room: a review of the literature. Surgery 2006 Feb;139(2):140-149. [doi: 10.1016/j.surg.2005.06.017] [Medline: 16455321]
9. Flin R, Yule S, Paterson-Brown S, Maran N, Rowley D, Youngson G. Teaching surgeons about non-technical skills. Surgeon 2007 Apr;5(2):86-89. [doi: 10.1016/S1479-666X(07)80059-X]
10. Schaverien MV. Selection for surgical training: an evidence-based review. J Surg Educ 2016;73(4):721-729. [doi: 10.1016/j.jsurg.2016.02.007] [Medline: 27133583]
11. Lipman JM, Colbert CY, Ashton R, et al. A systematic review of metrics utilized in the selection and prediction of future performance of residents in the United States. J Grad Med Educ 2023 Dec;15(6):652-668. [doi: 10.4300/JGME-D-22-00955.1] [Medline: 38045930]
12. Bowe SN, Laury AM, Gray ST. Associations between otolaryngology applicant characteristics and future performance in residency or practice: a systematic review. Otolaryngol Head Neck Surg 2017 Jun;156(6):1011-1017. [doi: 10.1177/0194599817698430]
13. Harfmann KL, Zirwas MJ. Can performance in medical school predict performance in residency? A compilation and review of correlative studies. J Am Acad Dermatol 2011 Nov;65(5):1010-1022. [doi: 10.1016/j.jaad.2010.07.034] [Medline: 21612841]
14. Kenny S, McInnes M, Singh V. Associations between residency selection strategies and doctor performance: a meta-analysis. Med Educ 2013 Aug;47(8):790-800. [doi: 10.1111/medu.12234] [Medline: 23837425]
15. Oldfield Z, Beasley SW, Smith J, Anthony A, Watt A. Correlation of selection scores with subsequent assessment scores during surgical training. ANZ J Surg 2013 Jun;83(6):412-416. [doi: 10.1111/ans.12176] [Medline: 23647783]
16. Stephenson-Famy A, Houmard BS, Oberoi S, Manyak A, Chiang S, Kim S. Use of the interview in resident candidate selection: a review of the literature. J Grad Med Educ 2015 Dec;7(4):539-548. [doi: 10.4300/JGME-D-14-00236.1] [Medline: 26692964]
17. Gardner AK, Dunkin BJ. Evaluation of validity evidence for personality, emotional intelligence, and situational judgment tests to identify successful residents. JAMA Surg 2018 May 1;153(5):409-416. [doi: 10.1001/jamasurg.2017.5013] [Medline: 29282462]
18. Niessen ASM, Meijer RR, Tendeiro JN. Measuring non-cognitive predictors in high-stakes contexts: the effect of self-presentation on self-report instruments used in admission to higher education. Pers Individ Dif 2017 Feb;106:183-189. [doi: 10.1016/j.paid.2016.11.014]
19. Griffin B, Wilson IG. Faking good: self-enhancement in medical school applicants. Med Educ 2012 May;46(5):485-490. [doi: 10.1111/j.1365-2923.2011.04208.x] [Medline: 22515756]
20. Landers RN, Sanchez DR. Game‐based, gamified, and gamefully designed assessments for employee selection: definitions, distinctions, design, and validation. Int J Selection Assessment 2022 Mar;30(1):1-13. [doi: 10.1111/ijsa.12376]
21. Georgiou K, Gouras A, Nikolaou I. Gamification in employee selection: the development of a gamified assessment. Int J Selection Assessment 2019 Jun;27(2):91-103. [doi: 10.1111/ijsa.12240]
22. Gomez MJ, Ruipérez-Valiente JA, Clemente FJG. A systematic literature review of game-based assessment studies: trends and challenges. IEEE Trans Learning Technol 2023;16(4):500-515. [doi: 10.1109/TLT.2022.3226661]
23. Ramos-Villagrasa PJ, Fernández-del-Río E, Castro Á. Game-related assessments for personnel selection: a systematic review. Front Psychol 2022;13:952002. [doi: 10.3389/FPSYG.2022.952002/BIBTEX]
24. Simons A, Wohlgenannt I, Zelt S, Weinmann M, Schneider J, vom Brocke J. Intelligence at play: game-based assessment using a virtual-reality application. Virtual Real 2023 Sep;27(3):1827-1843. [doi: 10.1007/s10055-023-00752-9]
25. Wiernik BM, Raghavan M, Caretta TR, Coovert MD. Developing and validating a serious game‐based assessment for cyber occupations in the US Air Force. Int J Selection Assessment 2022 Mar;30(1):27-47. [doi: 10.1111/ijsa.12378]

26. Landers RN, Armstrong MB, Collmus AB, Mujcic S, Blaik J. Theory-driven game-based assessment of general cognitive ability: design theory, measurement, prediction of performance, and test fairness. J Appl Psychol 2022 Oct;107(10):1655-1677. [doi: 10.1037/apl0000954] [Medline: 34672652]
27. Cook DA, Hatala R. Validation of educational assessments: a primer for simulation and beyond. Adv Simul 2016 Jan;1(1):1-12. [doi: 10.1186/s41077-016-0033-y]
28. Downing SM. Validity: on meaningful interpretation of assessment data. Med Educ 2003 Sep;37(9):830-837. [doi: 10.1046/j.1365-2923.2003.01594.x] [Medline: 14506816]
29. Gazit N, Ben-Gal G, Eliashar R. Development and validation of an objective virtual reality tool for assessing technical aptitude among potential candidates for surgical training. BMC Med Educ 2024 Mar 14;24(1):286. [doi: 10.1186/s12909-024-05228-1] [Medline: 38486166]
30. Kawaguchi K, Egi H, Hattori M, Sawada H, Suzuki T, Ohdan H. Validation of a novel basic virtual reality simulator, the LAP-X, for training basic laparoscopic skills. Minim Invasive Ther Allied Technol 2014 Oct;23(5):287-293. [doi: 10.3109/13645706.2014.903853] [Medline: 24773373]
31. Patterson F, Ferguson E, Thomas S. Using job analysis to identify core and specific competencies: implications for selection and recruitment. Med Educ 2008 Dec;42(12):1195-1204. [doi: 10.1111/j.1365-2923.2008.03174.x] [Medline: 19120950]
32. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for Educational and Psychological Testing: American Educational Research Association; 2014.
33. Kim YJ, Shute VJ. The interplay of game elements with psychometric qualities, learning, and enjoyment in game-based assessment. Computers & Education 2015 Sep;87:340-356. [doi: 10.1016/j.compedu.2015.07.009]
34. Ventura M, Shute V. The validity of a game-based assessment of persistence. Comput Human Behav 2013 Nov;29(6):2568-2572. [doi: 10.1016/j.chb.2013.06.033]
35. Granic I, Lobel A, Engels RCME. The benefits of playing video games. Am Psychol 2014 Jan;69(1):66-78. [doi: 10.1037/a0034857] [Medline: 24295515]
36. Reynaldo C, Christian R, Hosea H, Gunawan AAS. Using video games to improve capabilities in decision making and cognitive skill: a literature review. Procedia Comput Sci 2021;179:211-221. [doi: 10.1016/j.procs.2020.12.027]
37. Kheloui S, Jacmin-Park S, Larocque O, et al. Sex/gender differences in cognitive abilities. Neurosci Biobehav Rev 2023 Sep;152:105333. [doi: 10.1016/j.neubiorev.2023.105333] [Medline: 37517542]
38. Sugawara A, Ishikawa K, Motoya R, Kobayashi G, Moroi Y, Fukushima T. Characteristics and gender differences in the medical interview skills of Japanese medical students. Intern Med 2017;56(12):1507-1513. [doi: 10.2169/internalmedicine.56.8135] [Medline: 28626175]
39. Graf J, Smolka R, Simoes E, et al. Communication skills of medical students during the OSCE: gender-specific differences in a longitudinal trend study. BMC Med Educ 2017 Dec;17(1):1-9. [doi: 10.1186/s12909-017-0913-4]
40. Gardner AK, Costa P. Predicting surgical resident performance with situational judgment tests. Acad Med 2024 Aug 1;99(8):884-888. [doi: 10.1097/ACM.0000000000005680] [Medline: 38412475]
41. Lynch J, Aughwane P, Hammond TM. Video games and surgical ability: a literature review. J Surg Educ 2010;67(3):184-189. [doi: 10.1016/j.jsurg.2010.02.010] [Medline: 20630431]
42. Kengen B, IJgosse WM, van Goor H, Luursema JM. Fast or safe? The role of impulsiveness in laparoscopic simulator performance. Am J Surg 2020 Oct;220(4):914-919. [doi: 10.1016/j.amjsurg.2020.02.056] [Medline: 32145917]
43. Wetzel CM, Black SA, Hanna GB, et al. The effects of stress and coping on surgical performance during simulations. Ann Surg 2010 Jan;251(1):171-176. [doi: 10.1097/SLA.0b013e3181b3b2be] [Medline: 20032721]
44. Rosendal AA, Sloth SB, Rölfing JD, Bie M, Jensen RD. Technical, non-technical, or both? A scoping review of skills in simulation-based surgical training. J Surg Educ 2023 May;80(5):731-749. [doi: 10.1016/j.jsurg.2023.02.011] [Medline: 36906398]

Abbreviations

GBA: game-based assessment

Edited by LT Car; submitted 06.02.25; peer-reviewed by DA O'Keeffe, EM Doherty; revised version received 23.05.25; accepted 31.05.25; published 15.08.25.

Please cite as:
Gazit N, Ben-Gal G, Eliashar R
Game-Based Assessment of Cognitive Abilities and Personality Characteristics for Surgical Resident Selection: A Preliminary Validation Study
JMIR Med Educ 2025;11:e72264
URL: https://mededu.jmir.org/2025/1/e72264
doi:10.2196/72264

Original Paper

Trends in the Japanese National Medical Licensing Examination: Cross-Sectional Study

Yuki Morimoto[1*], MD; Kiyoshi Shikino[2*], MD, MHPE, PhD; Yukihiro Nomura[3], PhD; Shoichi Ito[2,4], MD, PhD

[1]Department of Medical Engineering, Faculty of Engineering, Chiba University, Chiba, Japan

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

[3]Center for Frontier Medical Engineering, Chiba University, Chiba, Japan

[4]Department of Medical Education, Graduate School of Medicine, Chiba University, Chiba, Japan

[*]these authors contributed equally

Corresponding Author:
Kiyoshi Shikino, MD, MHPE, PhD
Department of Community-Oriented Medical Education
Graduate School of Medicine
Chiba University
1-8-1, Inohana
Chu-ou-ku
Chiba, 260-8670
Japan
Phone: 1 43 222 7171
Email: kshikino@gmail.com

Abstract

Background: The Japanese National Medical Licensing Examination (NMLE) is mandatory for all medical graduates seeking to become licensed physicians in Japan. Given the cultural emphasis on summative assessment, the NMLE has had a significant impact on Japanese medical education. Although the NMLE Content Guidelines have been revised approximately every five years over the last 2 decades, objective literature analyzing how the examination itself has evolved is absent.

Objective: To provide a holistic view of the trends of the actual examination over time, this study used a combined rule-based and data-driven approach. We primarily focused on classifying the items according to the perspectives outlined in the NMLE Content Guidelines, complementing this approach with a natural language processing technique called topic modeling to identify latent topics.

Methods: We collected publicly available NMLE data for 2001-2024. Six examination iterations (2880 items) were manually classified from 3 perspectives (level, content, and taxonomy) based on pre-established rules derived from the guidelines. Temporal trends within each classification were evaluated using the Cochran-Armitage test. Additionally, we conducted topic modeling for all 24 examination iterations (11,540 items) using the bidirectional encoder representations from transformers–based topic modeling framework. Temporal trends were traced using linear regression models of topic frequencies to identify topics growing in prominence.

Results: In the level classification, the proportion of items addressing common or emergent diseases increased from 60% (115/193) to 76% (111/147; $P<.001$). In the content classification, the proportion of items assessing knowledge of pathophysiology decreased from 52% (237/459) to 33% (98/293; $P<.001$), whereas the proportion assessing practical knowledge of primary emergency care increased from 21% (95/459) to 29% (84/293; $P<.001$). In the taxonomy classification, the proportion of items that could be answered solely through simple recall of knowledge decreased from 51% (279/550) to 30% (118/400; $P<.001$), while the proportion assessing advanced analytical skills, such as interpreting and evaluating the meaning of each answer choice according to the given context, increased from 4% (21/550) to 19% (75/400; $P<.001$). Topic modeling identified 25 distinct topics, of which 10 exhibited an increasing trend. Non–organ-specific topics with notable increases included “comprehensive clinical items,” “accountability in medical practice and patients’ rights,” “care, daily living support, and community health care,” and “infection control and safety management in basic clinical procedures.”

Conclusions: This study identified significant shifts in the Japanese NMLE over the past 2 decades, suggesting that Japanese undergraduate medical education is evolving to place greater importance on practical problem-solving abilities than on rote

memorization. This study also identified latent topics that showed increased prominence, possibly reflecting underlying social conditions.

(JMIR Med Educ 2025;11:e78214) doi:10.2196/78214

KEYWORDS

BERTopic; bidirectional encoder representations from transformers–based topic modeling framework; exam content; Japan; National Medical Licensing Examination; natural language processing; NMLE; taxonomy; topic modeling; trend analysis

Introduction

Background

The National Medical Licensing Examination (NMLE) is a mandatory, high-pressure test that all medical school graduates in Japan must pass before becoming licensed physicians. Japan has 82 medical schools with an annual enrollment capacity of more than 9000 students [1]. Upon completing the 6-year medical curriculum, students are eligible to participate in the NMLE program [1]. Unlike licensing examinations in countries such as the United Kingdom, Australia, and Singapore, which primarily assess foreign medical graduates [2], passing the Japanese NMLE is compulsory for all domestic and international graduates who wish to practice medicine in Japan. In 2024, approximately 10,000 candidates took the examination [3]. The NMLE comprises multiple-choice items similar in format to examinations in South Korea, China, Thailand, and Germany [4]. Although there have been discussions regarding transitioning to a computer-based format [1], the examination remains paper-based, with all the candidates in a given year responding to identical items [5]. It is held annually, and all examinees take the examination on the same day. The items and answers are disclosed after the examination, and the item set is completely altered each year. Unlike the United States Medical Licensing Examination (USMLE), repeated attempts are not permitted in the same year, making it a one-shot opportunity. A partially norm-referenced grading system is applied to the pass-or-fail decision, ensuring that approximately 10% of candidates fail each year, thereby maintaining the examination's competitive nature [1]. Consequently, the pressure on examinees is considerable because failing the examination delays medical licensure by a full year.

Cultural factors contribute to the heightened significance of the NMLE in Japan. The belief that academic achievement determines future stability is deeply rooted in Japanese society, a mindset often associated with Confucian-influenced cultures [1]. This perspective is reinforced by the highly competitive nature of Japanese university entrance examinations, with medical school admissions being particularly competitive and reportedly having an acceptance rate of approximately 8.3% (12 applicants per spot). This fosters a learning culture that prioritizes summative assessments over formative feedback. In this context, students are accustomed to receiving model answers from instructors rather than engaging in critical discussions [6]. These educational values shape how medical students approach learning and preparation for high-stakes examinations such as the NMLE.

The scope and content of the NMLE are defined by the "NMLE Content Guidelines," published by the Ministry of Health, Labour and Welfare (Japan) [7] (an English summary of which is provided in Multimedia Appendix 1, as no official English version exists). The original Japanese document is titled "*Ishi Kokka Shiken Shutsudai Naiyō Shishin (Reiwa 6 Edition)*." They specify the fundamental knowledge and skills physicians are expected to possess when taking their first steps into clinical practice. The NMLE Content Guidelines categorize these competencies by topic and outline the performance expectations for each one. To streamline the examination's administration, the NMLE was reduced from a 3-day, 500-question format to a 2-day, 400-question format [1,7,8]. In addition, the question content was regulated to avoid excessive specialization, and competency levels were clearly defined to ensure clarity regarding the knowledge expected of medical graduates [1]. These guidelines are revised approximately every five years to reflect the evolving requirements for medical professionals.

Another key framework shaping Japanese medical education is the "Medical Education Model Core Curriculum (MCC)," issued by the Ministry of Education, Culture, Sports, Science, and Technology (Japan) [9]. The official Japanese title is "*Igaku Kyouiku Moderu Koa Kyarikyuramu (Reiwa 4 Edition)*." The MCC guidelines oversee approximately two-thirds of the medical education curriculum and are revised approximately every five years. Since its introduction in 2001, the MCC has evolved significantly; the 2016 revision adopted a competency-based medical education approach and shifted from traditional discipline-based learning to an integrated curriculum [10], aligned with international standards followed in the United States, Singapore, Canada, the Netherlands, and the United Kingdom [11]. The most recent 2022 MCC further elaborates on good practice and offers recommendations for medical education strategy and assessment [12].

Despite these comprehensive educational guidelines, the NMLE continues to exert considerable influence on medical education in Japan. Some universities underemphasize the MCC, and students often prioritize knowledge acquisition for NMLE preparation over more holistic learning experiences [13,14]. To enhance medical education in Japan, it is essential to consider the influence of the NMLE on learning behaviors and curriculum design, particularly given the strong preference for summative assessments in Japanese educational culture. As both the NMLE Content Guidelines and the MCC have evolved, it is increasingly important to examine how the actual content of the NMLE reflects these changes. However, there is a lack of objective longitudinal analysis of the examination itself, highlighting the need for systematic research in this area.

Study Goal

The objective of this study was to analyze trends in the content of the NMLE across the 21st century. A priori dimensions of interest were defined by the recent NMLE Content Guidelines and included 3 perspectives: level, content, and taxonomy classifications of examination items. Accordingly, the specific research questions were as follows: (1) How has the distribution of examination items changed across these predefined dimensions since 2001? (2) What latent themes, which cannot be fully captured by these dimensions, can be identified through topic modeling, and how have they evolved over time?

Based on prior revisions of the MCC and national examination policy reports, the following a priori hypotheses were formulated: (1) the proportion of items requiring highly specialized knowledge beyond the scope of generalist training has decreased; (2) the overall proportion of practical questions has increased; (3) among the items with higher-order cognitive demands, the proportion of more complex, context-dependent problem-solving items has increased; and (4) in addition to the above dimensions, topic modeling can identify latent themes whose prominence has increased over time, reflecting evolving social priorities and educational expectations. To address these questions, we used a hybrid methodology that integrates rule-based manual classification with exploratory natural language processing (NLP) techniques, particularly topic modeling [15]. Manual classification allowed us to directly reflect the emphases outlined in official medical education guidelines, whereas topic modeling enabled the discovery of previously unknown patterns beyond predefined frameworks.

By integrating these complementary approaches, this study aims to not only provide an evidence-based understanding of how the NMLE has evolved in terms of content emphasis and cognitive skill requirements but also offer practical insights for key stakeholders. For policymakers, our longitudinal data offers insights into the current medical training system's adaptive flexibility and helps identify remaining areas for further adjustment. For curriculum designers and educators, the trend data serve as an empirical basis for the timely update of educational content to align with contemporary medical education needs. Finally, for medical education researchers, this study's approach to quantitatively capturing the evolution of examination content offers a new analytical model for education evaluation research, both domestically and internationally.

Methods

Study Design

This cross-sectional study comprehensively analyzed NMLE examination items from 2001 to 2024, using a hybrid methodology that integrates rule-based and data-driven approaches in a complementary manner. To elucidate trends in the NMLE, we combined systematic analysis through manual classification with exploratory analysis using artificial intelligence (AI).

In the rule-based component, we focused on 6 specific examination sessions—2001, 2005, 2009, 2013, 2018, and 2024—that each represented the first examination administered after a major revision of the NMLE Content Guidelines. These datasets, comprising a total of 2880 items, were manually classified using 3 classification systems developed based on the guidelines: level, content, and taxonomy.

In the data-driven component, we performed topic modeling across all 24 NMLE sessions conducted between 2001 and 2024, encompassing a total of 11,540 items. This AI-based analysis provided an objective perspective independent of the guidelines, offering valuable complementary insights. Specifically, we used bidirectional encoder representations from transformers–based topic modeling framework (BERTopic) [16,17], a state-of-the-art NLP framework.

Setting

The NMLE Content Guidelines accompanying the 2024 edition outlined the revisions to be made to the NMLE. The 2024 NMLE was the first to be administered following the revisions [7]. The examination comprises 3 sections: essential, general clinical, and specialized clinical [7]. Each section includes items covering a wide range of medical specialties and public health domains (Figure S1 in Multimedia Appendix 2).

Classification Procedures

Level Classification

The revision process of the NMLE Content Guidelines emphasized the careful selection of diseases to be tested, clarification of the required depth of knowledge for each disease, and the exclusion of topics requiring only basic factual recall that should be mastered before clinical training [5]. To define the breadth and depth of knowledge appropriate for new graduates entering clinical residency, a level classification system was introduced [5]. The 2024 edition of the NMLE Content Guidelines explicitly assigns a level classification to each subcategory within a specialized clinical section [7].

In this study, the level of each question was determined by identifying the disease or condition being tested, either as the main topic or as the correct answer. These were matched to corresponding items in the 2024 NMLE Content Guidelines, and the preassigned level classification was used to categorize each question (Table 1). For items involving multiple key diseases or answer choices, the highest assigned level among the relevant classifications was used to represent the level of the question, with the ranking order as follows: "A > B > C > not classifiable." Detailed classification criteria are provided in Multimedia Appendix 3.

Table 1. Disease categorization. English translations of terms are based on official sources where available. For terms without official translations, translations were made by the authors and validated against published English-language literature.

Level	Disease types	Competencies required for initial treatment	Competencies required for subsequent treatment	Knowledge to be tested
A	Common diseases in primary care settings and acute diseases requiring emergency treatment	Possess enough knowledge to diagnose and manage patients under supervision, while appropriately consulting attending physicians as necessary	Possess enough knowledge to solve problems arising in subsequent treatment	• Knowledge of pathophysiology • Clinical reasoning skills • Knowledge of primary emergency care • Knowledge of continued care
B	Diseases that should be learned during postgraduate clinical training	Possess enough fundamental knowledge to manage patients under supervision	Possess enough knowledge to recognize when and how to present concerns to supervisors	• Knowledge of pathophysiology • Clinical reasoning skills • Knowledge of primary care
C	Diseases requiring a high level of clinical experience (beyond the postgraduate clinical training level)	Ability to integrate understanding of the outline of illnesses and clinical reasoning to reach a differential diagnosis	N/A[a]	• Ability to recall the names of diseases

[a]N/A: not applicable.

Content Classification

The NMLE Content Guidelines impose certain restrictions on test content based on the level classification of each question. As seen in Table 1, the guidelines classify required knowledge such as pathophysiology, clinical reasoning, primary emergency care, and continued care, to help define the focus of examination items [7]. Although explicit definitions of these terms are not provided, they generally align with the cognitive processes expected of a practicing clinician.

This study categorized the items based on the cognitive processes required for clinical practice. The following 4 content categories were defined: pathophysiology, clinical reasoning, primary emergency care, and continued care. Pathophysiology items require reasoning based on a given diagnosis, typically presenting a disease or pathological condition and assessing knowledge of its mechanisms, associated symptoms, and potential complications. Items requiring an approach to making a diagnosis were classified as clinical reasoning. While clinical reasoning is broadly defined and encompasses multiple components—including data gathering, hypothesis generation, problem representation, differential diagnosis, provisional diagnosis, diagnostic justification, management, and treatment [18]—this study applies the term in a narrower sense. Specifically, it is limited to the diagnostic process, as described in the **Accreditation Council for Graduate Medical Education** Internal Medicine Milestones [19], excluding elements related to management and treatment to avoid overlap with other categories. Primary emergency care items focus on urgent decision-making, which is commonly applied in emergency settings and requires immediate problem-solving. Continued care items address long-term management and preventive strategies, often in the context of scheduled outpatient visits or inpatient care. Detailed classification criteria are provided in Multimedia Appendix 4.

Notably, in the official Ministry of Health, Labour and Welfare guidelines, both “primary emergency care” and “continued care” are conventionally categorized under “examination & treatment.” In this study, these were purposefully separated into distinct categories to enable a more process-oriented analysis and to track the evolving focus of the NMLE more accurately.

Taxonomy Classification

Internationally, examination items in medical education are often evaluated using Bloom’s taxonomy, which categorizes cognitive complexity levels [20,21]. In Japan, a traditional classification system derived from Bloom’s taxonomy is used to categorize items based on the cognitive skills required to answer them. This system classifies items into 3 levels, as outlined in Figures 1 and 2 [5].

Figure 1. Schematic representation of the cognitive processes required to solve examination questions. Traditional 3-tier taxonomy commonly used in Japan. A potential gap exists between Type II and Type III in terms of cognitive progression.

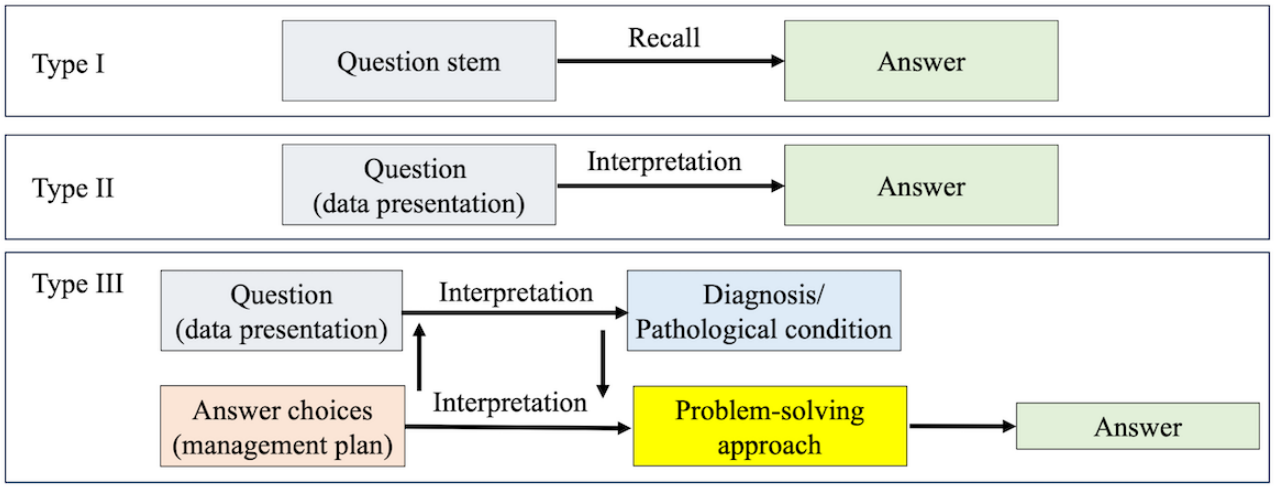
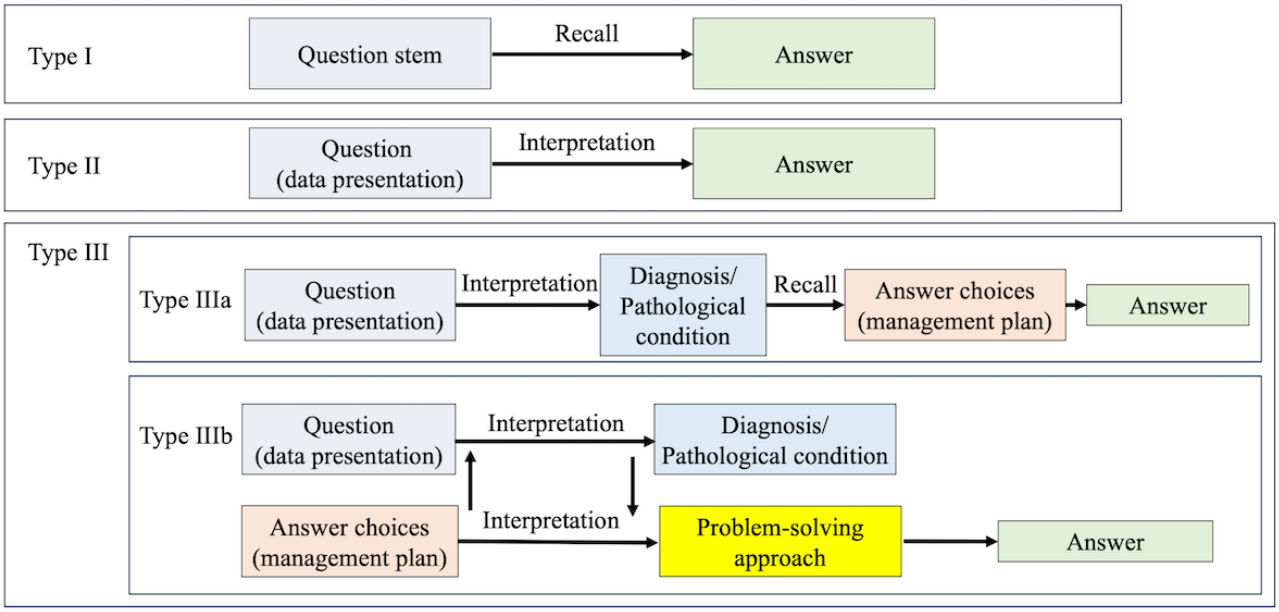


Figure 2. Schematic representation of the cognitive processes required to solve examination questions. Modified 4-tier taxonomy that introduces Type IIIa and Type IIIb to improve continuity between cognitive categories. Type III is defined as a combination of Type IIIa and Type IIIb in this system.



- Type I (recall-based items) can be answered purely by recalling factual knowledge.
- Type II (interpretation-based items) require understanding and interpreting the given information to derive an answer.
- Type III (problem-solving items) involve comprehending the scenario, analyzing the meaning of each answer choice, and solving a clinical problem.

Past revisions of the NMLE Content Guidelines have strongly emphasized Type II and Type III items, which test higher-order cognitive skills rather than simple recall.

Although the traditional classification system offers a straightforward framework, it does not guarantee an exhaustive or mutually exclusive categorization of NMLE items. An undefined region between Type II and Type III items may exist. We noted a substantial number of items wherein once the stem was correctly interpreted to reach a diagnostic conclusion, the subsequent answering process required little more than rote recall. Such items often arise when medical students memorize a direct correspondence between diagnoses and their associated treatments. Although prevalent, these items are not adequately defined within conventional classification frameworks.

Therefore, we introduced a novel classification system to address this limitation and enhance exclusivity and comprehensiveness. While retaining the traditional definitions for Types I and II, we refined the definition of Type III and subdivided it into 2 distinct categories: Types IIIa and IIIb. The revised classification system is as follows (Figures 1 and 2):

- Type I (recall-based): Same as the conventional Type I.
- Type II (interpretation-based): Same as the conventional Type II.
- Type III (problem-solving–based): A collective term encompassing Types IIIa and IIIb.
- Type IIIa (interpretation + recall-based): Items in which once the stem is correctly interpreted, the answer can be derived through a simple recall process. This type

corresponds to the previously undefined region between Types II and III.

- Type IIIb (interpretation + interpretation-based): Items requiring interpretation of both the stem and the meaning of the answer choices before responding. Unlike Type IIIa, the mere recognition of a diagnosis is insufficient; the examinee must also consider the context and apply higher-order reasoning. This includes scenarios where appropriate decision making is contingent upon factors such as disease severity, stage, comorbidities, allergies, or other patient-specific conditions that influence the selection of an optimal intervention.

The introduction of Type IIIa is particularly relevant, as it addresses a previously unclassified domain wherein interpretation and recall are intertwined. However, the cognitive demand is less rigorous than that in Type IIIb. The distinction between Types IIIa and IIIb is intended to capture the complexity of clinical problem-solving more accurately, particularly in cases where decision making requires contextual understanding beyond simple diagnosis. Detailed classification criteria are provided in Multimedia Appendix 5. To enhance the validity of the revised taxonomy, we aligned the definitions of Types IIIa and IIIb with established educational frameworks, including Revised Bloom's Taxonomy and the Accreditation Council for Graduate Medical Education Internal Medicine Milestones, thereby ensuring content validity. Furthermore, the distinction between Types IIIa and IIIb reflects whether the clinical reasoning process involves context-sensitive decision making, which supports its construct validity. Reliability was ensured through the use of a detailed classification manual, structured assessor training, calibration exercises, and interrater reliability testing, with discrepancies adjudicated by a senior physician.

Topic Modeling

Various topic-modeling techniques exist, such as Non-Negative Matrix Factorization (NMF) and Latent Dirichlet Allocation (LDA). However, with rapid advancements in NLP following the emergence of transformer models, more sophisticated analytical methods have been developed [22]. One such method is BERTopic, an advanced topic-modeling approach that leverages BERT (bidirectional encoder representations from transformers) embeddings, which are context-aware representations of text. Unlike traditional models that assign a fixed meaning to each word, BERT learns word meanings by considering the words that come before and after it. This bidirectional contextual understanding enables the model to capture nuanced relationships between terms—particularly important for medical texts wherein similar concepts can be expressed in diverse ways. By transforming each examination item into a BERT embedding—a numerical vector reflecting its contextual meaning—BERTopic can more accurately cluster semantically related items and identify coherent topics [16]. For Japanese texts, additional preprocessing was required because, unlike English, words are not separated by spaces. We therefore applied morphological analysis using MeCab—the most widely used tool for Japanese language processing—to segment sentences into words. To ensure accurate recognition of medical terminology, a specialized dictionary (MANBYO) was incorporated into MeCab. We then used a Japanese Sentence-BERT model ("sonoisa/sentence-bert-base-ja-mean-tokens-v2") to generate embeddings, which provided context-aware numerical representations of each examination item (stem + answer choices). These embeddings were subsequently processed through dimensionality reduction and clustering—core steps of the BERTopic topic modeling pipeline. Owing to its high customizability, BERTopic is well suited for applications beyond English-language texts and particularly effective for time-series analyses [23]. It has gained attention for its ability to uncover novel patterns and provide insights into textual data [17,23].

In this study, we applied BERTopic by inputting textual data that combined the item stems and answer choices for each examination item. The topic-modeling process was conducted in six steps using the BERTopic Python package:

- Preprocessing: For the tokenization of the Japanese text, we used the morphological analyzer "MeCab" [24,25]. To ensure appropriate analysis of texts containing symptoms and disease names, we applied a specialized dictionary for medical terms, "MANBYO" [26].
- Embedding: As a Japanese Sentence Transformer model, we used "sonoisa/sentence-bert-base-ja-mean-tokens-v2," available on Hugging Face [27].
- Dimensionality Reduction: Uniform Manifold Approximation and Projection (UMAP) was used.
- Clustering: Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) was used.
- Tokenization: CountVectorizer was used.
- Weighting Scheme: Class-based term frequency–inverse document frequency (c-TF-IDF) scoring was used.

After analyzing the extracted keywords and various representative items, the authors determined the topic labels through mutual agreement.

To enhance transparency, we also describe the process of parameter selection in BERTopic. While default settings were maintained whenever possible, 3 key parameters were adjusted through systematic trial-and-error to optimize interpretability and stability. For UMAP, n_neighbors was set to 3 to emphasize local structures and uncover fine-grained latent topics, and n_components was set to 10 to preserve information while ensuring clustering stability. For HDBSCAN, min_cluster_size was set to 20, which balanced interpretability and granularity and resulted in 25 coherent topics. We also confirmed that the main conclusions of this study, such as the increase in items under "comprehensive clinical items" and "accountability in healthcare and patient rights," remained stable across reasonable parameter variations.

Each examination item was assigned to a single topic only, reflecting the multiclass nature of the clustering process. While some items (eg, questions combining "renal disease" and "research") could plausibly relate to more than one topic, the dimensionality reduction and clustering steps ultimately placed each item in one cluster. This design choice facilitates clearer identification of longitudinal shifts in topic prevalence, whereas

a multilabel approach would have assigned multiple simultaneous labels and obscured such temporal trends.

In addition, the process of topic interpretation and labeling was refined. For each topic, the top 20 representative keywords with the highest c-TF-IDF values and the list of examination items assigned to that topic were extracted. The research team carefully reviewed these materials, designating clearly irrelevant words (eg, laboratory units, particles, auxiliary verbs) as stop words to improve the quality of the keyword lists without altering the BERTopic analysis. Final topic labels were established through repeated focus group discussions, with reference to representative items and prior literature, until consensus was reached. For transparency, the representative keyword lists for each topic are provided in Table S2 in Multimedia Appendix 6.

Following the automatic clustering, several closely related topics were manually integrated through consensus discussions (eg, merging "male" and "female" patient clusters into "comprehensive clinical items") to enhance thematic coherence and prevent over-splitting.

Selection Criteria for Analysis

For each rule-based analysis, we examined question datasets from the NMLEs for 2001, 2005, 2009, 2013, 2018, and 2024. The 2001 NMLE was included because of its coincidence with the initial publication of the MCC. For the level classification analysis, only items from the specialized clinical section were considered, as the 2024 NMLE Content Guidelines have defined level classifications for this section. Items without an explicitly assigned level were excluded from the statistical analysis. For the content classification analysis, items related to public health knowledge (eg, health policies, legal frameworks, and statistical concepts) were excluded because of their distinct nature. The analysis focused on items from other medical specialties, whereas items that could not be assigned to a predefined content category were excluded from statistical analysis. For the Taxonomy classification analysis, all items were included without exception.

For the data-driven analysis, we used complete question datasets from the 2001-2024 NMLEs. All items were included in the topic-modeling analysis without exception.

Between 2001 and 2006, each examination included 30 or 50 pilot items that were not subject to scoring. However, as the specific pilot items were not disclosed, the dataset was analyzed with all items included.

Study Variables

The primary outcome variable was the distribution of items across the 3 classification systems in each examination year. The primary predictor variable was the examination year, which was treated as representing time progression, allowing for the assessment of trends in the item distribution. Secondary outcome variables included topic distribution, measured as the proportion of items assigned to each topic, and the slope coefficient from the linear regression, which represented the annual rate of change in the proportion of each topic over time.

Data Sources and Measurement

Each classification system was evaluated by 10-12 assessors, including clerkship students and licensed physicians. All assessors were provided with a detailed classification manual developed iteratively through expert discussion (Multimedia Appendices 3-5). Prior to the formal classification, assessors participated in structured training sessions, which included video-based instructions and multiple practice rounds with sample items. These sessions were followed by calibration exercises wherein assessors independently classified a pilot set of items, then discussed discrepancies with senior physicians to ensure consistent interpretation of the manual. Clerkship students were included only if they had completed at least one year of clinical clerkship. In addition, when required, their classifications were cross-checked against those of licensed physicians to confirm reliability. The initial draft of the manual was created by a general internist (YM) with reference to the 2024 (Reiwa 6) NMLE Content Guidelines and the official report of the Ministry of Health, Labour and Welfare's Working Group on Improvements to the NMLE. The draft was reviewed by the study team and refined through expert focus group discussions involving KS, an experienced physician and medical educator. During these discussions, operational rules were established for each system. For the level classification, each item was mapped to the most relevant blueprint subitem, and when multiple disease themes or answer keys were plausible, assessors selected the theme most central to the item's core; if options were equally plausible, priority was given to the condition with higher frequency or urgency (A > B > C). For the content classification, we adopted 4 mutually exclusive categories—pathophysiology, clinical reasoning, primary emergency care, and continued care—derived from the official guidelines, and refined definitions to avoid overlap. For the Taxonomy classification, we clarified the boundaries among Types II, IIIa, and IIIb through iterative examination of numerous examination items and consensus-based discussions among the research team. In particular, the cognitive distinction between "recall" and "interpretation" was deliberated extensively to define the boundary between Types IIIa and IIIb. Disagreements encountered during manual development were resolved through consensus, supported by additional sample questions when necessary. When discrepancies occurred, a third-party adjudicator (a general internist) made the final decision. We believe that this structured process ensured that all assessors, including students, were adequately trained and able to apply the classification framework reliably.

In this study, clerkship students are defined as medical students who have passed common achievement tests (computer-based tests and objective structured clinical examinations), and been certified to possess the necessary knowledge, skills, and professionalism to participate in clinical clerkships. Given the importance of clinical experience in accurate classification, content and taxonomy classifications were conducted exclusively by licensed physicians or clerkship students with at least two years of clinical clerkship experience.

This ensured that classifications requiring higher clinical expertise were performed only by adequately trained and experienced assessors. All translations of Japanese terms were

based on official bilingual documents whenever available. When no official translation was available, translations were performed by the study team and cross-checked against existing English-language literature to ensure consistency and accuracy.

Statistical Analysis

Interrater reliability of the classification system was assessed using Fleiss' κ coefficient [28] (0.8-1.0=almost perfect; 0.6-0.8=substantial; 0.4-0.6=moderate; 0.2-0.4=fair). Chi-square tests were conducted to evaluate the relationship between item distribution and examination year. Cochran-Armitage trend tests were conducted to assess time-series trends in question distribution. Statistical analyses were performed using R (version 4.3.3; The R Project for Statistical Computing). The trends of each topic over time were traced using linear regression models based on topic frequencies using NumPy (version 1.26.4; NumPy Developers) in Python (version 3.12.4; Python Software Foundation).

Ethical Statement

This study did not collect any human or patient information and was thus exempt from ethical approval, informed consent requirements, and institutional review board approval. Additionally, as no identifying information was included, the data did not need to be anonymized or deidentified. No compensation was offered because there were no human participants in this study.

Results

Overview

Fleiss' κ coefficients were calculated to assess the reliability of all 3 classification systems (level, content, and taxonomy). Most coefficients indicated substantial agreement (κ>0.6), although the content classification in the 2013 examination showed a moderate κ coefficient (κ=0.54). Detailed values for each classification across 6 representative examination years are provided in Table S3 in Multimedia Appendix 7, which also describes the procedures for calculating κ coefficients. For the level, content, and taxonomy classifications, a chi-square test revealed a significant association between examination year and the distribution of question categories (P=.02, P<.001, and P<.001, respectively). Based on these results, Cochran-Armitage trend tests were conducted for all 3 classifications to evaluate temporal trends.

Level Classification

Among the 1100 items in the specialized clinical section, a total of 1073 (97.5%) were successfully classified into one of 3 predefined levels (Table 2). The remaining 27 (2.5%) items were classified as "not classifiable" and excluded from further statistical analysis. Level A items accounted for 743 (69.2%) of classified specialized clinical items, making them the most common. Items at levels B and C accounted for 244 (22.7%) and 86 (8%), respectively. Over 23 years, the proportion of level A items increased significantly from 59.6% (115/193) in 2001 to 75.5% (111/147) in 2024 (P<.001). By contrast, the proportion of level B and C items declined significantly, from 29.5% (57/193) to 19.7% (29/147; P=.008) and 10.9% (21/193) to 4.8% (7/147; P=.009), respectively (Figure 3A). These trends suggest a shift toward assessing higher-priority knowledge over time.

Figure 3. Temporal trends in question distribution across classification systems. (A) Level classification, (B) content classification, and (C) taxonomy classification (Types I, II, and III), and (D) taxonomy classification (Types IIIa and IIIb). Points represent observed data. Solid lines indicate fitted trends estimated using linear regression, with shaded areas representing the 95% CIs. Cochran-Armitage trend tests were performed for each category, and the associated *P* values are displayed. The significant increase in Type IIIb questions observed in subpart D primarily accounts for the overall increase in Type III questions shown in subpart C, whereas no significant trend was found for Type IIIa.

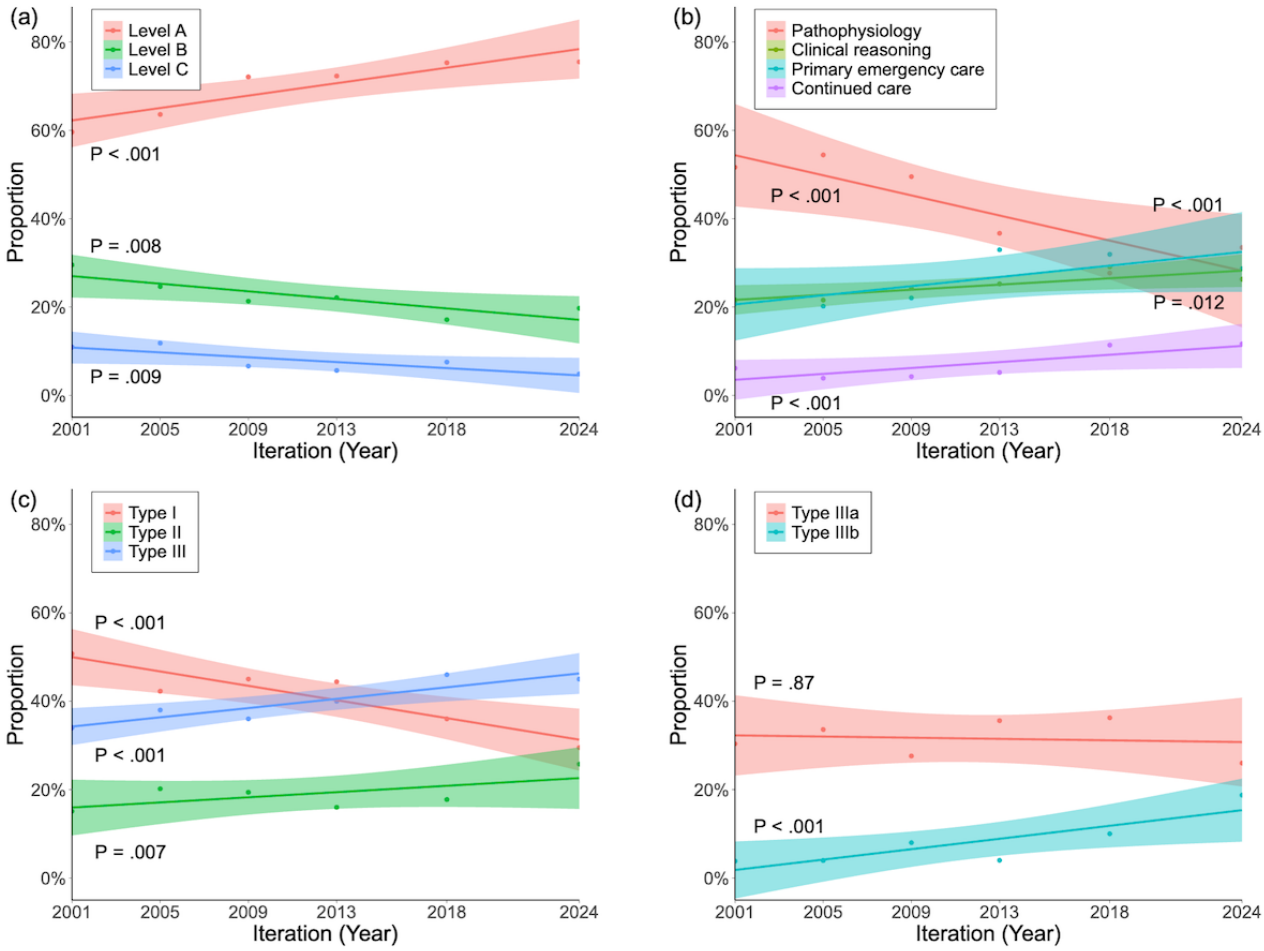


Table 2. Level classification.

Year	2001 (n=200), n (%)	2005 (n=200), n (%)	2009 (n=200), n (%)	2013 (n=200), n (%)	2018 (n=150), n (%)	2024 (n=150), n (%)	Total (n=1100), n (%)
Level A	115 (57.5)	124 (62)	142 (71)	141 (70.5)	110 (73.3)	111 (74)	743 (67.6)
Level B	57 (28.5)	48 (24)	42 (21)	43 (21.5)	25 (16.7)	29 (19.3)	244 (22.2)
Level C	21 (10.5)	23 (11.5)	13 (6.5)	11 (5.5)	11 (7.3)	7 (4.7)	86 (7.8)
Not classifiable	7 (3.5)	5 (2.5)	3 (1.5)	5 (2.5)	4 (2.7)	3 (2)	27 (2.5)

Content Classification

Of the 2518 items (excluding public health–related items), a total of 2228 (88.5%) were assigned to one of 4 content categories (Table 3). The remaining 290 (11.5%) items were classified as "not classifiable" and excluded from further statistical analysis. Pathophysiology was the most frequently tested category, appearing in 981 (44%) of all classified items, followed by primary emergency care (562, 25.2%), clinical reasoning (539, 24.2%), and continued care was the least common (146, 6.6%). However, the proportion of pathophysiology items exhibited substantial fluctuations, ranging from 54.4% (240/441) in 2005 to 27.7% (78/282) in 2018. Over 23 years, the overall proportion of pathophysiology items declined significantly, from 51.6% (237/459) to 33.4% (98/293; *P*<.001). Conversely, the proportions of clinical reasoning, primary emergency care, and continued care items showed significant upward trends. From 2001 to 2024, the proportions increased from 21.6% (99/459) to 26.3% (77/293) for clinical reasoning (*P*=.01), 20.7% (95/459) to 28.7% (84/293) for primary emergency care (*P*<.001), and 6.1% (28/459) to 11.6% (34/293) for continued care (*P*<.001; Figure 3B). These trends indicate a progressive shift toward assessing clinical decision-making and patient management skills.

Table 3. Content classification.

Year	2001 (n=488), n (%)	2005 (n=473), n (5)	2009 (n=442), n (%)	2013 (n=436), n (%)	2018 (n=343), n (%)	2024 (n=336), n (%)	Total (n=2518), n (%)
Pathophysiology	237 (48.6)	240 (50.7)	200 (45.2)	128 (29.4)	78 (22.7)	98 (29.2)	981 (39)
Clinical reasoning	99 (20.3)	95 (20.1)	98 (22.2)	88 (20.2)	82 (23.9)	77 (22.9)	539 (21.4)
Primary emergency care	95 (19.5)	89 (18.8)	89 (20.1)	115 (26.4)	90 (26.2)	84 (25)	562 (22.3)
Continued care	28 (5.7)	17 (3.6)	17 (3.8)	18 (4.1)	32 (9.3)	34 (10.1)	146 (5.8)
Not classifiable	29 (5.9)	32 (6.8)	38 (8.6)	87 (20)	61 (17.8)	43 (12.8)	290 (11.5)

Taxonomy Classification

All 2880 items were successfully classified into one of the 4 taxonomy types (Table 4). Type I items were the most common, accounting for 1212 (42.1%) of all items, followed by Type IIIa (n=910, 31.6%), Type II (n=541, 18.8%), and Type IIIb (n=217, 7.5%). Over 23 years, the proportion of Type I items declined significantly, from 50.7% (279/550) in 2001 to 29.5% (118/400) in 2024 (P<.001). Conversely, the proportions of Types II and IIIb increased significantly, from 15.1% (83/550) to 25.8% (103/400; P=.007) and from 3.8% (21/550) to 18.8% (75/400; P<.001), respectively (Figures 3C and D). Notably, the combined proportion of Types IIIa and IIIb surpassed Type I in 2018 and 2024, reflecting a progressive shift toward assessing higher-order problem-solving skills.

Table 4. Taxonomy classification.

Year	2001 (n=550), n (%)	2005 (n=530), n (%)	2009 (n=500), n (%)	2013 (n=500), n (%)	2018 (n=400), n (%)	2024 (n=400), n (%)	Total (n=2880), n (%)
Type I	279 (50.7)	224 (42.3)	225 (45.0)	222 (44.4)	144 (36)	118 (29.5)	1212 (42.1)
Type II	83 (15.1)	107 (20.2)	97 (19.4)	80 (16)	71 (17.8)	103 (25.8)	541 (18.8)
Type III	188 (34.2)	199 (37.6)	178 (35.6)	198 (39.6)	185 (46.3)	179 (44.8)	1127 (39.1)
Type IIIa	167 (30.4)	178 (33.6)	138 (27.6)	178 (35.6)	145 (36.3)	104 (26)	910 (31.6)
Type IIIb	21 (3.8)	21 (4)	40 (8)	20 (4)	40 (10)	75 (18.8)	217 (7.5)

Topic Modeling

Topic modeling effectively classified 10,129 of the 11,540 (88%) items into 25 topics. The remaining 1411 (12%) items were treated as outliers and could not be assigned to any topic. A summary of these topics is presented in Table 5. Further details on these topics are presented in Multimedia Appendix 6. The most frequently appearing topics included “comprehensive clinical items,” “pediatrics,” “accountability in medical practice and patients’ rights,” “cardiology,” and “metabolic and endocrinology.” A visualization of the popular topics is shown in Figure 4. Ten topics exhibited an increasing trend, among which topics not limited to specific organs were identified, including “comprehensive clinical items,” “accountability in medical practice and patients’ rights,” “care, daily living support, and community health care,” “intensive care,” and “infection control and safety management in basic clinical procedures.”

Figure 4. Hot and cold topics. Bar chart showing topic trends from 2001 to 2024, based on the slope of linear regression lines fitted to topic proportions over time. Topics are sorted in descending order of slope, with “hot” (increasing) topics at the top and “cold” (decreasing) topics at the bottom.

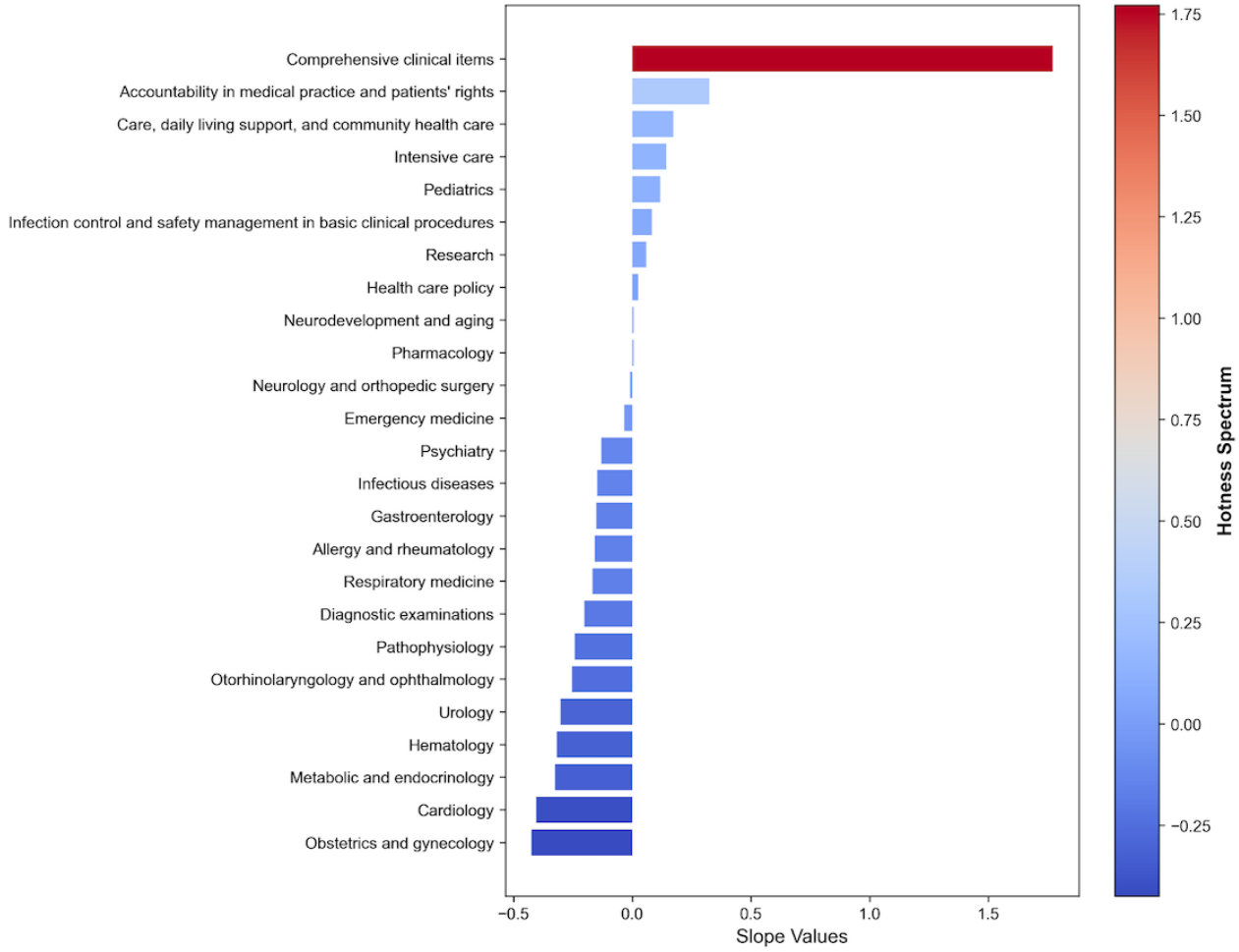


Table 5. Summary of the topics.

No	Topic name	Number of items, n (%)
1	Comprehensive clinical items	4515 (39.12)
2	Pediatrics	752 (6.52)
3	Accountability in medical practice and patients' rights	412 (3.57)
4	Cardiology	381 (3.30)
5	Metabolic and endocrinology	324 (2.81)
6	Obstetrics and gynecology	305 (2.64)
7	Diagnostic examinations	303 (2.63)
8	Otorhinolaryngology and ophthalmology	288 (2.50)
9	Emergency medicine	271 (2.35)
10	Hematology	239 (2.07)
11	Pathophysiology	237 (2.05)
12	Neurology and orthopedic surgery	208 (1.80)
13	Respiratory medicine	208 (1.80)
14	Urology	190 (1.65)
15	Care, daily living support, and community health care	183 (1.59)
16	Infectious diseases	171 (1.48)
17	Research	171 (1.48)
18	Neurodevelopment and aging	169 (1.46)
19	Psychiatry	154 (1.33)
20	Gastroenterology	147 (1.27)
21	Allergy and rheumatology	139 (1.20)
22	Pharmacology	128 (1.11)
23	Intensive care	122 (1.06)
24	Health care policy	60 (0.52)
25	Infection control and safety management in basic clinical procedures	52 (0.45)

Discussion

Principal Results

This study revealed significant changes in the Japanese NMLE over the past 2 decades, highlighting a shift in content and cognitive complexity. The increasing emphasis on common diseases (level A), with a corresponding reduction in highly specialized or rare conditions, aligns with the NMLE Content Guidelines' focus on core medical knowledge relevant to general clinical practice [9]. Additionally, the dominance of pathophysiology-related items decreased. Concurrently, clinical reasoning, primary emergency care, and continued care have become more prominent, reflecting a transition from knowledge-based assessment to practical clinical skills and decision-making. Furthermore, the decline in factual recall items (Type I) and rise in higher-order cognitive skill items (Type IIIb), which require data interpretation and problem-solving, indicate a broader educational shift toward competency-based medical education, consistent with the MCC's emphasis on clinical reasoning and problem-solving abilities. These findings suggest that the NMLE is evolving into a more practice-oriented examination, better aligned with the skills required for real-world clinical settings, including clinical clerkship.

The level classification analysis indicates that common diseases (level A) have been increasingly emphasized in the NMLE, with their proportion rising significantly in recent years. This suggests a greater focus on common diseases, while highly specialized or less frequently encountered conditions have been increasingly restricted in the examination content. This shift aligns with the policies outlined in the NMLE Content Guidelines, which emphasize core medical knowledge applicable to general clinical practice. Although the level classification was first formally introduced in the 2024 edition of the NMLE Content Guidelines, efforts to minimize discrepancies between the MCC and NMLE Content Guidelines by aligning lists of examinable diseases have already been reported [13]. To fully integrate this shift in medical education, students must be sufficiently exposed to common diseases during clinical clerkships [29,30]. Since university hospitals often manage complex or rare cases [31], structured rotations in regional or community hospitals, where students can encounter a broader range of common conditions, may enhance their preparedness

for the NMLE and future clinical practice. In this context, future analyses of the NMLE based on level classification will be essential for monitoring the alignment between national medical education policies, clinical clerkship experiences, and licensing examinations.

The results of the content classification analysis indicated a decreasing dominance of pathophysiology-related items, with a concurrent increase in items assessing clinical reasoning, primary emergency care, and continued care. This suggests a transition from a knowledge-based focus to a more practice-oriented approach in the NMLE, emphasizing clinical decision-making and patient management skills. Compared with pathophysiology, primary emergency care and continued care require more practical competence, as they involve real-time decision making, prioritization, and the ability to manage ongoing patient care [32,33]. Such changes align with the directive of the NMLE Content Guidelines to assess the practical knowledge and skills acquired through clinical clerkship. Moreover, this shift is consistent with the ongoing transition in Japanese medical education, from a discipline-based to a competency-based framework, as promoted by the MCC. Notably, the NMLE blueprint also specifies organ system-based distributions (eg, nephrology, urology, and reproductive medicine: approximately 12%), and these proportions have remained relatively stable across the years. Because such distributions are explicitly monitored and reported in the blueprint, further analysis of organ system–based proportions would provide limited additional value. Conversely, the categorization into pathophysiology, clinical reasoning, primary emergency care, and continued care was newly introduced in the 2024 (Reiwa 6) NMLE Content Guidelines, without any specified proportions. Therefore, retrospectively applying this framework to past examinations enabled us to conduct the first longitudinal analysis of how the examination's process-oriented emphases have evolved. This approach provides educators and policymakers with unique insights that complement the existing organ system-based blueprint.

Taxonomy classification analysis highlights a declining emphasis on simple factual recall items (Type I), with a corresponding increase in items requiring data interpretation and a critical analysis of answer choices (Types II and III). Notably, the proportion of Type IIIb items requiring interpretation of both the stem and answer choices increased. This suggests that the NMLE's emphasis on higher-order cognitive skills extends beyond surface-level modifications in question format, reflecting a genuine shift toward assessing problem-solving and clinical reasoning abilities. This trend can be understood from the perspective of Revised Bloom's Taxonomy, which categorizes cognitive processes into 6 levels: remembering, understanding, applying, analyzing, evaluating, and creating [20,21,34]. While traditional medical assessments have focused primarily on recall (remembering) and comprehension (understanding), the increasing prevalence of Type IIIb items suggests a shift toward the "analyzing" level, where examinees must break down complex clinical information, assess relationships between data points, and apply their knowledge in a more integrative manner. These changes align with the MCC's competency framework, particularly in the problem-solving domain, which aims to develop medical graduates capable of integrating evidence-based medicine, experiential learning, and clinical reasoning to address complex patient issues. Consequently, the NMLE increasingly demands the ability to apply knowledge and critically evaluate and interpret clinical scenarios, requiring a higher degree of practical competence.

Educational and Assessment Implications of the Distinction Between Types IIIa and IIIb

Our introduction of Type IIIa (interpretation + recall-based) and Type IIIb (interpretation + patient-specific contextualization) further clarifies the trend toward practical competence. The increasing proportion of Type IIIb items highlights the NMLE's emphasis on context-sensitive clinical decision-making, where even after reaching a diagnosis, learners must make management choices based on severity, comorbidities, contraindications, or patient-specific factors. From an educational perspective, this has several implications. First, students can be guided away from "one disease–one answer" memorization and toward comparative reasoning among multiple plausible options. The level classification can help prioritize high-yield diseases for such practice. Second, curricula should incorporate structured case-based or simulation-based exercises that vary clinical modifiers (eg, renal function, allergies, and social context), requiring explicit justification of choices. Third, the use of visual or video-based materials may further reinforce the reasoning processes captured by Type IIIb, as they allow learners to engage with patient cues and procedural nuances beyond what text alone can convey [35]. For examination development, the IIIa-IIIb framework offers guidance for constructing items that embed contextual modifiers, use distractors requiring applied reasoning, or leverage multimedia formats. Collectively, these strategies can foster more authentic assessment and deeper learner preparation.

The following section discusses selected topics that show an increasing trend in the topic-modeling analysis.

- Comprehensive clinical items: Approximately 40% of the items are categorized under this topic, which includes questions that ask examinees to select appropriate diagnoses, tests, or treatments based on a patient's history, physical findings, and test results, excluding discipline- or organ-specific cases. As many items from various departments share a common format, they appear to have been grouped into this overarching category. For examinees, these items require a cross-disciplinary perspective because the relevant organ system is not immediately apparent at first glance. The increasing prevalence of such items likely reflects the shift in medical education toward a competency-based approach, moving away from traditional discipline-based outcome assessments [36].
- Accountability and patients' rights: This topic pertains to the knowledge essential for the prevention and management of medical accidents, as well as physicians' accountability to patients and society, and how they confront such responsibilities. This overlaps with 2 of the 4 pillars of professionalism described by Arnold and Stern's model: accountability and altruism [37]. These elements are

considered fundamental components of medical professionalism. According to the MCC, professionalism is to "acknowledge the professional responsibility of physicians to be deeply involved in people's lives and to protect health, respect diversity and humanity, and take an altruistic approach to medical practice throughout one's career" [9]. Each revision of the MCC has increasingly emphasized the importance of professionalism. The current version lists it as the foremost quality and competency required by physicians. This heightened emphasis likely contributed to the observed increase in this topic.

- Care, daily living support, and community health care: This topic encompasses items related to long-term care insurance policies, the roles of long-term care hospitals and welfare facilities, and housing with daily living support. In Japan, where the population is rapidly aging alongside a declining birthrate, the comprehensive and seamless provision of health care, long-term care, and social support services within the community has become increasingly important [38]. This growing need is reflected in the increased attention given to this topic following the government's introduction of a community-based integrated care system in 2006.
- Infection control and safety management in basic clinical procedures: This topic encompasses items that assess knowledge related to safety management and infection control during basic clinical procedures, such as the safe implementation of blood collection and intravenous access, appropriate handling of collected specimens, hand hygiene, and precautions against needlestick injuries. While such clinical skills are also evaluated in the objective structured clinical examination (OSCE), concerns have been raised regarding the lack of sufficiently validated OSCE stations for assessing patient safety competencies [39]. Medical incidents began to attract public attention in the 2000s, culminating in the establishment of a Medical Accident Investigation System in 2014. Reflecting the growing emphasis on patient safety, the 2018 revision of the examination guidelines incorporated terminology such as "medical accident investigation system," "medical accident prevention manual," and "medical safety management departments and risk managers." The increase in items from this topic area on the NMLE has likely been in response to this heightened awareness.

The observed trends in the NMLE reflect the broader expectations of undergraduate medical education in Japan. There is a growing emphasis on acquiring practical knowledge that cannot be gained solely through lectures, developing problem-solving skills for real-world clinical challenges, and fostering deeper analytical thinking beyond rote memorization [40,41]. Given these trends, clinical clerkships play a crucial role in medical education [42-45]. Japanese medical schools allocate an average of 67 weeks to clinical training [1]. While efforts have been made to extend clerkship duration, the focus must shift toward ensuring that these experiences are participatory rather than observational. The MCC and Japan Accreditation Council for Medical Education advocate participatory clinical clerkship models, emphasizing active involvement in patient care [46]. Furthermore, clinical clerkships at community hospitals could be strategically utilized if the goal is to enhance exposure to common diseases [47-49]. In addition, the topic-modeling results suggest that themes showing an increasing trend, such as "care, daily living support, and community health care," can be particularly well demonstrated and understood through community hospital clerkships, wherein students are more likely to encounter patients requiring long-term care, home-based support, and interprofessional collaboration. Recent studies have highlighted the educational value of such experiences for developing patient-centered and community-oriented care competencies [50,51]. A well-balanced integration of university-affiliated hospital rotations and regional hospital training will foster more practical hands-on learning experiences in undergraduate medical education.

In addition, the major topics highlighted in this study, such as "comprehensive clinical questions" and "accountability in medical practice and patients' rights," were robust to reasonable variations in parameter settings (eg, *n_neighbors*, *n_components*, and *min_cluster_size*). The same topics were consistently extracted across these sensitivity checks and continued to demonstrate increasing trends, suggesting that this study's main conclusions are stable and not merely artifacts of specific parameter configurations.

Practical and Policy Implications

The findings of this study provide practical insights for both policymakers and educators. For policymakers, longitudinal analysis of the NMLE offers an evidence-based foundation for examination reform, helping ensure a balance between theoretical knowledge and practical competencies, and between common and specialized conditions. Meanwhile, educators can apply the hybrid methodology demonstrated in this study to evaluate whether undergraduate curricula are aligned with national examination priorities and to analyze internal institutional examinations for continuous improvement of test content balance. From a curricular perspective, structuring clerkships to include rotations at community hospitals provides valuable opportunities for students to directly encounter common diseases, continuity of care, and outpatient management in real-world contexts. Such experiences complement the predominantly specialty-oriented training at university hospitals and strengthen the acquisition of the competencies emphasized by recent trends in the NMLE.

Limitations

This study has several limitations that should be considered when interpreting the findings. First, the classification framework used in this study was designed specifically for the Japanese NMLE, based on the perspectives outlined in the NMLE Content Guidelines. Therefore, applying this methodology to other medical licensing examinations would require modifications to reflect country-specific guidelines and educational frameworks. In particular, regional variations in disease prevalence, differences in the expected competencies of first-year physicians, and the complexity of taxonomy classifications based on examination formats must be considered when adapting this approach to other national examinations. Second, while interrater reliability was generally acceptable, the κ coefficient for the content classification in specific

examination sessions was lower than ideal. In particular, the 2013 examination (107th NMLE) showed a moderate κ coefficient (0.54), which was lower than that of other years. Closer examination revealed that many disagreements arose between the categories "pathophysiology" and "not classifiable." The 2013 examination also represented a transitional phase in test design, as the proportion of "pathophysiology" items decreased markedly compared to earlier years (eg, 2009), while the proportion of "not classifiable" items increased. These shifts suggest that evolving examination policy and the emergence of intermediate item styles between pathophysiology and other categories contributed to assessor disagreement and the lower κ coefficient. To improve the classification reliability in future studies, it may be beneficial to use experienced physicians instead of clerkship students as assessors, particularly for the content classification, which requires clinical judgment. Third, approximately 11.5% of items in the content classification were categorized as "not classifiable." This group was heterogeneous and included items such as specific cautions for diagnostic procedures, drugs, and surgical techniques; calculation problems (eg, A–aDO_2); and procedural tasks such as donning and doffing personal protective equipment. Although these are clinically relevant, they do not align neatly with disease- or patient-centered cognitive processes and, therefore, were grouped under "not classifiable" in this study. Representative examples are provided in Multimedia Appendix 4. However, because of their diversity, designing a more detailed classification system would be necessary to identify meaningful patterns, and future systematic analyses of this group may provide valuable insights for medical education and examination design. Fourth, public health–related items were excluded from the content classification analysis. While this decision was made to maintain consistency with the focus on clinical competencies, public health items can vary in their emphasis on factual knowledge versus practical applications. Future research should explore how public health–related items align with competency-based medical education frameworks, especially as global health and preventive medicine are increasingly emphasized in clinical training. Fifth, this study did not assess item difficulty levels; instead, it focused on the examination content and cognitive complexity. However, item difficulty plays a crucial role in examinations wherein the absolute grading criteria determine the pass–fail outcomes. If these findings are to be applied to other examinations, the potential confounding effect of difficulty adjustments must be considered, as difficulty regulation could influence the trends observed in this study. Sixth, this study did not consider the impact of examination preparation on the taxonomy classification. Because Japanese medical students frequently study past NMLE items, previously used or closely resembling past items may be answered at a lower cognitive level than initially intended. This effect is particularly relevant for examinations with a high proportion of reused or recycled items, wherein students may rely on pattern recognition rather than problem-solving skills. If this classification system is to be applied to such examinations, additional considerations in taxonomy categorization may be necessary. Finally, while BERTopic provided valuable insights by capturing contextual information more effectively than traditional topic-modeling methods such as LDA, it also introduced specific challenges when attempting a more rigorous validation. We adjusted several model parameters (Multimedia Appendix 8) to obtain a reasonable number of topics; however, the relative ranking of topics and the slopes of their temporal trends in the linear regression were sensitive to these settings. Therefore, it should be noted that the statistical robustness of these observations is not guaranteed, although our discussion focused on topics that consistently appeared at the top or showed increasing trends over time. Moreover, a key characteristic—both a potential strength and limitation—of this method is its flexibility: entirely different parameter configurations may yield novel and potentially insightful topics. Nevertheless, because the interpretation of topics ultimately relies on human judgment, there is an increased risk that certain outputs may lack clear interpretability or thematic coherence.

Conclusions

This study identified significant shifts in the content and cognitive complexity of the Japanese NMLE over the past 2 decades. The findings indicated a greater emphasis on common diseases (level A) and a decline in highly specialized topics, suggesting a prioritization of core medical knowledge applicable to general clinical practice. Additionally, the transition from content heavy on pathophysiology to a greater focus on clinical reasoning, primary emergency care, and continued care reflects a broader shift in assessment priorities, aligning with competency-based medical education reforms in Japan.

Moreover, the reduction in factual recall items (Type I) and increase in problem-solving items (Type III), particularly those requiring dual interpretation (Type IIIb), indicate a growing emphasis on higher-order cognitive skills. These trends correspond with the MCC's competency framework, particularly its problem-solving domain, which aims to develop physicians capable of integrating evidence-based medicine, clinical reasoning, and patient-centered care.

Topic modeling using NLP suggested an increasing emphasis on clinical items considered from a cross-organ perspective. In addition, topics not limited to specific organ systems that are increasingly represented in the NMLE, such as "accountability in medical practice and patients' rights," "care, daily living support, and community health care," and "infection control and safety management in basic clinical procedures," were identified. These trends are considered to reflect broader societal developments.

The observed changes in the NMLE suggest that Japanese undergraduate medical education is evolving to place greater importance on practical problem-solving abilities than on rote memorization. Given this trend, enhancing participatory clinical clerkship and leveraging regional hospital rotations will be crucial in further aligning medical education with evolving expectations. Future research should continue to monitor the alignment between national medical education policies and licensing examinations to ensure that assessments reflect the skills and knowledge necessary for competent medical practice.

Acknowledgments

We thank M Ueda for kindly facilitating the initial connections between the first and corresponding authors. We also thank S Sasaki for organizing and managing the dataset, including the examination items, answer sheets, and classification results. We thank T Watanabe for managing the task allocation among the physician staff and clerkship students. Finally, we thank the clerkship students for performing the classification according to the manual.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The datasets generated and analyzed in this study are available from the corresponding author upon reasonable request. However, the original datasets are owned by Medic Media Corp, Tokyo, Japan, and any use of the data requires permission from the company.

Authors' Contributions

YM, KS, NY, and SI planned, designed, and conceived the study. YM drafted the manuscript. YM gathered data resources. YM, KS, NY, and SI analyzed the qualitative and quantitative data. YM, KS, NY, and SI have read and approved the final version of the manuscript.

Conflicts of Interest

Medic Media Corp, Tokyo, Japan, was involved in collecting public data related to the National Medical Licensure Examination. However, it did not participate in the design or conduct of the study, data analysis or interpretation, manuscript review or approval, or the decision to submit the manuscript for publication.

Multimedia Appendix 1
English Summary of National Medical Licensing Examination (NMLE) Content Guidelines.
[DOCX File , 23 KB - mededu_v11i1e78214_app1.docx]

Multimedia Appendix 2
Mapping of the sections to subjects.
[DOCX File , 650 KB - mededu_v11i1e78214_app2.docx]

Multimedia Appendix 3
Manual for level classification.
[DOCX File , 17 KB - mededu_v11i1e78214_app3.docx]

Multimedia Appendix 4
Manual for content classification.
[DOCX File , 21403 KB - mededu_v11i1e78214_app4.docx]

Multimedia Appendix 5
Manual for taxonomy classification.
[DOCX File , 18 KB - mededu_v11i1e78214_app5.docx]

Multimedia Appendix 6
Detailed topic information extracted by the topic modeling.
[DOCX File , 22 KB - mededu_v11i1e78214_app6.docx]

Multimedia Appendix 7
Interrater reliability of the classifications.
[DOCX File , 16 KB - mededu_v11i1e78214_app7.docx]

Multimedia Appendix 8

Model parameters.
[DOCX File , 17 KB - mededu_v11i1e78214_app8.docx]

References

1. Nishigori H. Medical education in Japan. Med Teach 2024;46(sup1):S4-S10. [doi: 10.1080/0142159x.2024.2372108]
2. Price T, Lynn N, Coombes L, Roberts M, Gale T, de Bere SR, et al. The international landscape of medical licensing examinations: a typology derived from a systematic review. Int J Health Policy Manag 2018;7(9):782-790 [FREE Full text] [doi: 10.15171/ijhpm.2018.32] [Medline: 30316226]
3. Announcement of the results for the 118th national medical licensing examination. Ministry of Health, Labour and Welfare, Japan. URL: https://www.mhlw.go.jp/general/sikaku/successlist/2024/siken01/about.html [accessed 2025-11-07]
4. Kuwabara N, Yamashita M, Yee K, Kurahara D. The evolution of the Japanese medical education system: a historical perspective. Hawaii J Med Public Health 2015;74(3):96-100 [FREE Full text] [Medline: 25821652]
5. Ban N, Monden M. Study for revision recommendations of the guidelines for the national examination for medical practitioners 2022 report. Med Educ (Japan) 2022;53(3):207-213 [FREE Full text]
6. Kozato A, Shikino K, Matsuyama Y, Hayashi M, Kondo S, Uchida S, et al. A qualitative study examining the critical differences in the experience of and response to formative feedback by undergraduate medical students in Japan and the UK. BMC Med Educ 2023;23(1):408 [FREE Full text] [doi: 10.1186/s12909-023-04257-6] [Medline: 37277728]
7. Regarding the 2024 edition of the national medical licensing examination content guidelines. Ministry of Health, Labour and Welfare, Japan. URL: https://www.mhlw.go.jp/stf/shingi2/0000128981_00001.html [accessed 2025-11-07]
8. Kozu T. Medical education in Japan. Acad Med 2006;81(12):1069-1075. [doi: 10.1097/01.ACM.0000246682.45610.dd] [Medline: 17122471]
9. The model core curriculum for medical education in Japan 2022. Medical Education Model Core Curriculum Expert Research Committee. 2022. URL: https://www.mext.go.jp/content/20250411-mxt_igaku-000028108_00003-2.pdf [accessed 2025-11-07]
10. Nishigori H, Haruta J, Urushibara-Miyachi Y. What can Japanese medical education contribute to the world? Med Teach 2024;46(sup1):S1-S3. [doi: 10.1080/0142159X.2024.2320456] [Medline: 39545498]
11. Matsuyama Y, Nomura O, Oikawa S, Kikukawa M, Shimizu I, Gomi H. Competency-based medical education guidelines are context-based: lessons from national guidelines in five countries. Med Teach 2024;46(sup1):S38-S45. [doi: 10.1080/0142159X.2024.2351215] [Medline: 39545497]
12. Bond MP, Vaillant JS. An empirical study of the relationship between diagnosis and defense style. Arch Gen Psychiatry 1986;43(3):285-288. [doi: 10.1001/archpsyc.1986.01800030103012] [Medline: 3954550]
13. Nomura O, Komatsu H, Matsuyama Y, Onoue T, Ikusaka M, Okazaki H, et al. Development of medical knowledge content for problem-solving competencies through dialogue with the undergraduate medical education community in Japan. Med Teach 2024;46(sup1):S61-S66 [FREE Full text] [doi: 10.1080/0142159X.2024.2385707] [Medline: 39545494]
14. Tokuda Y, Goto E, Otaki J, Jacobs J, Omata F, Obara H, et al. Undergraduate educational environment, perceived preparedness for postgraduate clinical training, and pass rate on the national medical licensure examination in Japan. BMC Med Educ 2010;10:35 [FREE Full text] [doi: 10.1186/1472-6920-10-35] [Medline: 20487536]
15. Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, et al. Latent dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimed Tools Appl 2018;78(11):15169-15211. [doi: 10.1007/s11042-018-6894-4]
16. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. Preprint posted online on March 11, 2022 2022 [FREE Full text]
17. Raman R, Pattnaik D, Hughes L, Nedungadi P. Unveiling the dynamics of AI applications: a review of reviews using scientometrics and BERTopic modeling. Journal of Innovation & Knowledge 2024;9(3):100517. [doi: 10.1016/j.jik.2024.100517]
18. Daniel M, Rencic J, Durning SJ, Holmboe E, Santen SA, Lang V, et al. Clinical reasoning assessment methods: a scoping review and practical guidance. Acad Med 2019;94(6):902-912. [doi: 10.1097/ACM.0000000000002618] [Medline: 30720527]
19. Internal medicine milestones 2.0: supplemental guide. Accreditation Council for Graduate Medical Education (ACGME). 2020. URL: https://www.acgme.org/globalassets/pdfs/milestones/internalmedicinesupplementalguide.pdf [accessed 2025-11-07]
20. Anderson LW, Krathwohl DR, Airasian PW. A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. New York: Longman; 2001.
21. Bloom BS, Englehart MD, Furst EJ, Hill WH, Krathwohl DR. Taxonomy of Educational Objectives: The Classification of Educational Goals, Handbook I: Cognitive Domain. New York: Longmans, Green, and Co; 1956.
22. Ogunleye B, Lancho Barrantes BS, Zakariyyah KI. Topic modelling through the bibliometrics lens and its technique. Artif Intell Rev 2025;58(3):74. [doi: 10.1007/s10462-024-11011-x]
23. Egger R, Yu J. A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. Front Sociol 2022;7:886498 [FREE Full text] [doi: 10.3389/fsoc.2022.886498] [Medline: 35602001]
24. Applying Conditional Random Fields to Japanese Morphological Analysis. 2025. URL: https://aclanthology.org/W04-3230.pdf [accessed 2025-11-14]

25. Yet Another Part-Of-Speech and Morphological Analyzer. 2025. URL: https://taku910.github.io/mecab/ [accessed 2025-11-07]
26. Large-scale disease name dictionary for tabulating and analyzing disease names actually used in clinical settings. "MANBYO" Dictionary. 2025. URL: https://sociocom.jp/~data/2018-manbyo/index.html [accessed 2025-11-07]
27. Hugging face. Sonoisa. 2025. URL: https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2 [accessed 2025-03-30]
28. Fleiss JL. Measuring nominal scale agreement among many raters. Psychological Bulletin 1971;76(5):378-382. [doi: 10.1037/h0031619]
29. Goldman D, Smith JP. The increasing value of education to health. Soc Sci Med 2011;72(10):1728-1737 [FREE Full text] [doi: 10.1016/j.socscimed.2011.02.047] [Medline: 21555176]
30. Nichols L. Medical education needs typical cases of common diseases. Autops Case Rep 2019;9(1):e2018080. [doi: 10.4322/acr.2018.080] [Medline: 30881928]
31. Walter A, Baty F, Rassouli F, Bilz S, Brutsche MH. Diagnostic precision and identification of rare diseases is dependent on distance of residence relative to tertiary medical facilities. Orphanet J Rare Dis 2021;16(1):131 [FREE Full text] [doi: 10.1186/s13023-021-01769-6] [Medline: 33745447]
32. Choi A, Lee K, Hyun H, Kim KJ, Ahn B, Lee KH, et al. A novel deep learning algorithm for real-time prediction of clinical deterioration in the emergency department for a multimodal clinical decision support system. Sci Rep 2024;14(1):30116 [FREE Full text] [doi: 10.1038/s41598-024-80268-7] [Medline: 39627310]
33. Leo RJ. Competency and the capacity to make treatment decisions: a primer for primary care physicians. Prim Care Companion J Clin Psychiatry 1999;1(5):131-141 [FREE Full text] [doi: 10.4088/pcc.v01n0501] [Medline: 15014674]
34. Shikino K, Rosu CA, Yokokawa D, Suzuki S, Hirota Y, Nishiya K, et al. Flexible e-learning video approach to improve fundus examination skills for medical students: a mixed-methods study. BMC Med Educ 2021;21(1):428 [FREE Full text] [doi: 10.1186/s12909-021-02857-8] [Medline: 34389012]
35. Shikino K, Nishizaki Y, Fukui S, Yokokawa D, Yamamoto Y, Kobayashi H, et al. Development of a clinical simulation video to evaluate multiple domains of clinical competence: cross-sectional study. JMIR Med Educ 2024;10:e54401 [FREE Full text] [doi: 10.2196/54401] [Medline: 38421691]
36. Frank JR, Snell LS, Cate OT, Holmboe ES, Carraccio C, Swing SR, et al. Competency-based medical education: theory to practice. Med Teach 2010;32(8):638-645. [doi: 10.3109/0142159X.2010.501190]
37. Arnold L, Stern DT. What is medical professionalism? In: Measuring Medical Professionalism. New York, NY: Oxford Academic; 2005:15-37.
38. Tsutsui T. Implementation process and challenges for the community-based integrated care system in Japan. Int J Integr Care 2014;14:e002. [doi: 10.5334/ijic.988] [Medline: 24478614]
39. Shimizu I, Tanaka K, Mori J, Yamauchi A, Kato S, Masuda Y, et al. Objective structured clinical examination to assess patient safety competencies of Japanese Medical Students: development and validation argument. Cureus 2024;16(11):e73969. [doi: 10.7759/cureus.73969] [Medline: 39563687]
40. McEvoy MD, Fowler LC, Robertson A, Gelfand BJ, Fleming GM, Miller B, et al. Comparison of two learning modalities on continuing medical education consumption and knowledge acquisition: a pilot randomized controlled trial. J Educ Perioper Med 2021;23(3):E668 [FREE Full text] [doi: 10.46374/volxxiii_issue3_mcevoy] [Medline: 34631966]
41. Norman GR. Problem-solving skills, solving problems and problem-based learning. Med Educ 1988;22(4):279-286. [doi: 10.1111/j.1365-2923.1988.tb00754.x] [Medline: 3050382]
42. Cho H, Jeong H, Yu J, Lee J, Jung HJ. Becoming a doctor: using social constructivism and situated learning to understand the clinical clerkship experiences of undergraduate medical students. BMC Med Educ 2024;24(1):236 [FREE Full text] [doi: 10.1186/s12909-024-05113-x] [Medline: 38443907]
43. Lee HJ, Kim D, Kang YJ. Understanding medical students' transition to and development in clerkship education: a qualitative study using grounded theory. BMC Med Educ 2024;24(1):910 [FREE Full text] [doi: 10.1186/s12909-024-05778-4] [Medline: 39223489]
44. Taylor JS, George PF, MacNamara MMC, Zink D, Patel NK, Gainor J, et al. A new clinical skills clerkship for medical students. Fam Med 2014 Jun;46(6):433-439. [Medline: 24911298]
45. Wimmers PF, Schmidt HG, Splinter TAW. Influence of clerkship experiences on clinical competence. Med Educ 2006;40(5):450-458. [doi: 10.1111/j.1365-2929.2006.02447.x] [Medline: 16635125]
46. The Japan Accreditation Council for Medical Education. URL: https://www.jacme.or.jp/en/index.php [accessed 2025-11-07]
47. Hiramine S, Nagata M, Kainuma M. The impact of an undergraduate community-based medical education program in a Japanese urban city. Cureus 2024;16(2):e54204 [FREE Full text] [doi: 10.7759/cureus.54204] [Medline: 38496076]
48. Kato D, Wakabayashi H, Takamura A, Takemura YC. Identifying the learning objectives of clinical clerkship in community health in Japan: focus group. J Gen Fam Med 2020;21(2):3-8 [FREE Full text] [doi: 10.1002/jgf2.289] [Medline: 32161694]
49. Ohta R, Ryu Y, Katsube T, Moriwaki Y, Otani J. Students' perceptions of general medicine following community-based medical education in rural Japan. J Gen Fam Med 2019;20(6):236-243 [FREE Full text] [doi: 10.1002/jgf2.274] [Medline: 31788401]

50. Ohta R, Sano C. Utilizing learn-to-rank systems for more effective diagnosis in rural family medicine. Cureus 2023;15(10):e47219. [doi: 10.7759/cureus.47219] [Medline: 38022090]
51. Yoshimura M, Saiki T, Imafuku R, Fujisaki K, Suzuki Y. Experiential learning of overnight home care by medical trainees for professional development: an exploratory study. Int J Med Educ 2020;11:146-154 [FREE Full text] [doi: 10.5116/ijme.5f01.c78f] [Medline: 32712596]

Abbreviations

AI: artificial intelligence
BERT: bidirectional encoder representations from transformers
BERTopic: bidirectional encoder representations from transformers–based topic modeling framework
c-TF-IDF: class-based term frequency–inverse document frequency
HDBSCAN: Hierarchical Density-Based Spatial Clustering of Applications with Noise
LDA: Latent Dirichlet Allocation
MCC: Model Core Curriculum
NLP: natural language processing
NMF: Non-Negative Matrix Factorization
NMLE: National Medical Licensing Examination
OSCE: objective structured clinical examination
UMAP: Uniform Manifold Approximation and Projection
USMLE: United States Medical Licensing Examination

Edited by J Moen; submitted 28.05.25; peer-reviewed by K Sakaguchi, S Huh, K Tsunekawa; comments to author 31.08.25; revised version received 13.11.25; accepted 05.12.25; published 23.12.25.

Please cite as:
Morimoto Y, Shikino K, Nomura Y, Ito S
Trends in the Japanese National Medical Licensing Examination: Cross-Sectional Study
JMIR Med Educ 2025;11:e78214
URL: https://mededu.jmir.org/2025/1/e78214
doi:10.2196/78214
PMID:

Evaluation of the Inverted Classroom Approach in a Case-Study Course on Antithrombotic Drug Use in a PharmD Curriculum: French Monocentric Randomized Study

Georges Jourdi[1,2], Prof Dr; Mayssa Selmi[2], PharmD; Pascale Gaussem[3,4], Prof Dr; Jennifer Truchot[5,6,7], Prof Dr Med; Isabelle Margaill[1], Prof Dr; Virginie Siguret[1,2], Prof Dr

[1]Optimisation thérapeutique en neuropharmacologie, INSERM U1144, Université Paris Cité, Paris, France

[2]Service d'Hématologie Biologique, Hôpital Lariboisière, AP-HP.Nord, Paris, France

[3]INSERM UMR-S970, Paris Cardiovascular Research Center, Université Paris Cité, Paris, France

[4]Service d'Hématologie Biologique, Hôpital Européen Georges Pompidou, AP-HP.Centre, Paris, France

[5]Emergency Department, Hôpital Cochin, AP-HP.Centre, Université Paris Cité, Paris, France

[6]iLumens, Université Paris Cité, Paris, France

[7]US UPPERS, Faculte´ de Sante´, Unite´ de Service Pour la Pe´dagogie, l'Enseignement et la Recherche, Université Paris Cité, Paris, France

Corresponding Author:
Georges Jourdi, Prof Dr
Optimisation thérapeutique en neuropharmacologie, INSERM U1144, Université Paris Cité, Paris, France

Abstract

Background: Appropriate antithrombotic drug use is crucial knowledge for pharmacy students.

Objective: We sought to compare the inverted classroom (IC) approach to a traditional question-and-answer educational approach with the aim of enhancing pharmacy students' engagement with a case-study course on antithrombotic drug use.

Methods: Third-year PharmD (Doctor of Pharmacy) students from Paris Cité University were randomly assigned to control (n=171) and IC (n=175) groups. The latter were instructed to read and prepare the preprovided course material 1 week before the in-class session to assume the instructor role on the target day, whereas students of the control group attended a traditional case-study course carried out by the same instructor. All students completed pre- and posttest multiple-choice questions surveys assessing their knowledge levels as well as stress, empathy, and satisfaction questionnaires.

Results: A significantly higher participation rate was observed in the control group (93/171, 54%) compared to the IC group (65/175, 37%; *P*=.002). Women (110/213, 52%) participated more than men (48/133, 36%; *P*=.002) whatever the group was. Students' knowledge scores from both groups had similar results with no difference neither in the prescore (1.17, SD 0.66 and 1.24, SD 0.72 of 5, respectively) nor in the short-term knowledge retention (2.45, SD 0.61 and 2.35, SD 0.73, respectively). The IC approach did not increase student stress or enhance their empathy for the instructor. It increased the preclass workload (*P*=.02) and was not well received among students.

Conclusions: This study showed that the traditional educational approach remains an efficient method for case-study courses in the early stages (ie, third-year) of the 6-year PharmD curriculum, yet dynamic methods improving the active role of students in the learning process are still needed.

(JMIR Med Educ 2025;11:e67419) doi:10.2196/67419

KEYWORDS

antithrombotic drugs; case-study course; inverted classroom; pharmacy students; traditional educational approach; medical education

Introduction

The French Regional Centers of Pharmacovigilance recently revealed that 8.5% of hospital admissions of 141 short-stay specialist medical wards randomly selected in 69 public hospitals, were related to adverse drug reactions [1]. Antithrombotic drugs were involved in 11.6% of the cases, placing them in second place right behind antineoplastic drugs. Some of the adverse drug reactions (mainly bleeding) were considered to be preventable because the drugs had not been used per the summary of product characteristics or guidelines [2]. In the French health care system, antithrombotics are mainly available in pharmacies, and for some also on websites but under the responsibility of a pharmacist. Pharmacists, particularly community-based pharmacists, are easily accessible health care professionals [3]. They are experts in drug therapy use, assessing

each patient through observation, dialogue, and consideration of clinical indicators. They are involved in monitoring the patient's compliance with treatment as well as their response to drug therapy through regular follow-up. This allows for the early detection of adverse effects or drug misuse. Therefore, pharmacist intervention should have a positive impact on the management of patients on antithrombotic therapy. All the above reasons makes the optimal use of antithrombotic drugs of utmost importance to learn for pharmacy students to prevent iatrogenesis.

In France, the PharmD (Doctor of Pharmacy) curriculum consists of a 6-year course. Three teaching models are used mainly: the lecture-based classroom, the case-study class, and the practical session. Case-study class is a hands-on approach to learning that involves presenting realistic scenarios and helps students to apply theoretical knowledge in clinic-like settings and attain a high-order cognitive level per Bloom's taxonomy [4]. The instructor asks students to participate in the case study analysis and discussion. This method favors the development of a deep understanding of the subject and avoids passive note-taking in students. However, this objective is not necessarily always reached. Since its introduction in 2000, the inverted classroom (IC) approach, switching away from the traditional educational approach for the lecture-based classroom, literally inverts the focus per Bloom's taxonomy: the bottom parts of the taxonomy (ie, understand and memorize basic concepts) are reserved for student self-instruction through readings, short recorded video, audio lectures, etc, while the class time is focused on the upper parts of the taxonomy (ie, analyze, justify a stand, and create original work). The IC approach has been increasingly studied in health professions students' education including pharmacy school [5-17]. These studies reported a positive impact of this approach on students' knowledge and skills in most of the cases in comparison to lecture-based courses [8-11,13,14,18-24]. It has been introduced into various courses including pharmacotherapy, pharmacokinetics, pharmaceutical calculations, pharmacy practice, and others [7,12,25-32]. That said, the added value of the IC approach has rarely, if at all, been tested for case-study courses in pharmacy education. Hence, we sought to conduct a monocentric study investigating the added value of an IC approach in a case-study course during the PharmD curriculum at Paris Cité University. In this IC approach, students received the course material before the in-class session and were asked to prepare it and assume the role of the instructor on the target day. We aimed to assess knowledge acquisition, preclass workload and students' self-assessment of their stress, empathy, and global satisfaction. It was hypothesized that the IC approach would lead to improved outcomes compared to the traditional question-and-answer educational approach.

Methods

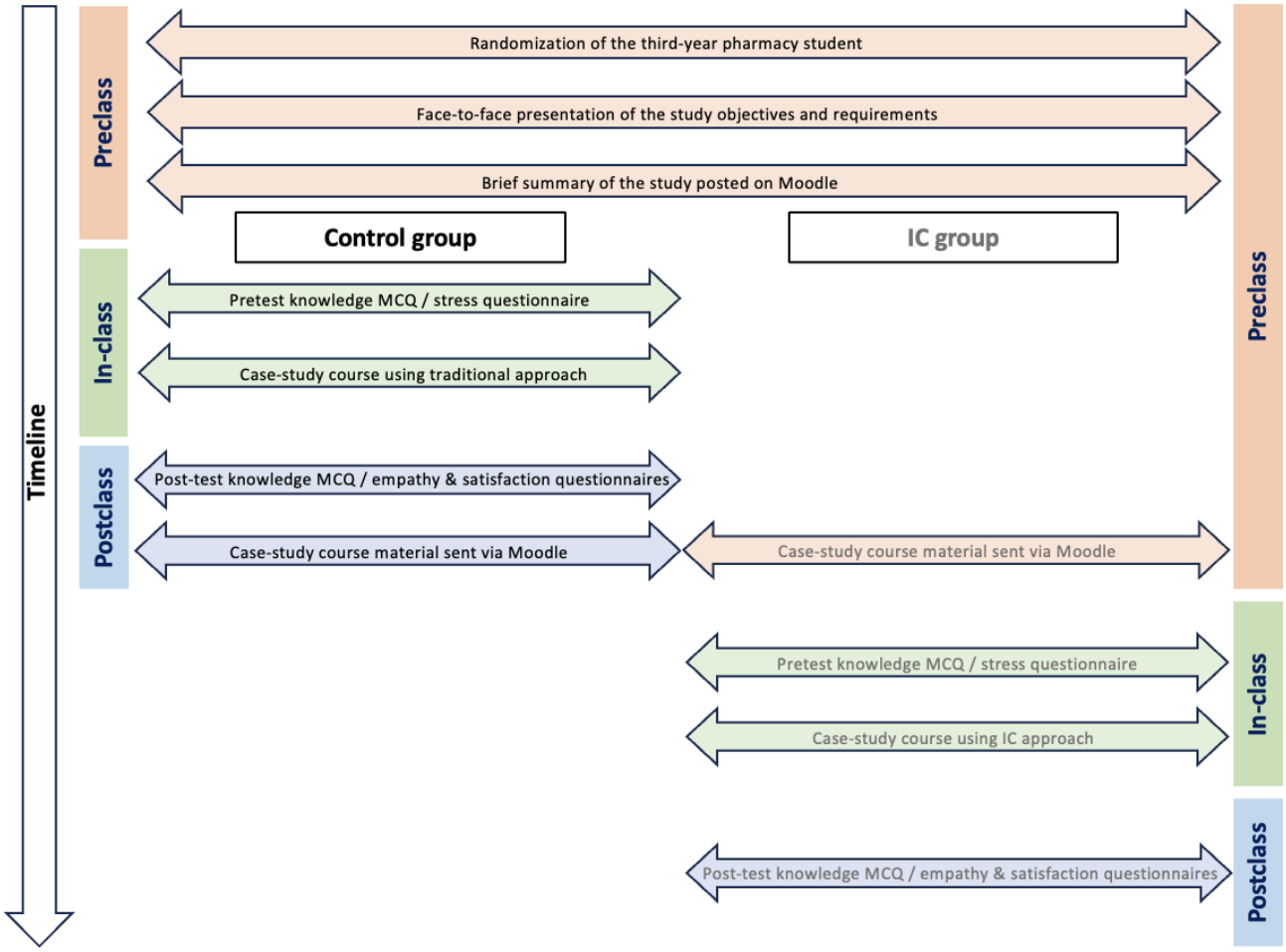
Ethical Considerations

This study was approved by the Institutional Review Board and Ethics Committee of Paris Cité University (00012023 - 20) and all procedures were performed per the Helsinki Declaration. Students were completely free to participate or not in this study. They were informed that neither participation nor nonparticipation in this study would influence their passing of this course or their grades. The participants were also informed about the option to withdraw from this study at any time. Informed consent was obtained from all the students who participated in this study. They were not paid for their participation. Collected data were anonymously analyzed.

Participants

We conducted this study at the faculty of Pharmacy of Paris Cité University in March 2023 (2022/2023 academic year). The participants were in the spring semester of the third year of the PharmD curriculum. At the beginning of the semester, students had basic courses on hemostasis (physiology and pathology) and thrombosis. During the month preceding the case-study course, they also had 3 lecture-based courses for a total of 4 hours on antithrombotic drugs. The recordings of the lectures were then available on a teaching platform ("Moodle"). The 90-minute case-study course on antithrombotic drug use entailed 16 groups consisting of about 20 to 22 students each, which were randomly assigned to the groups by the office of student affairs of the faculty of Pharmacy without any influence from the instructor. Therefore, a randomized assignment can be assumed. Students were highly required to respect their group assignment. Each time, 2 groups assisted simultaneously in one course, thus the course was repeated 8 times by the same instructor. Participants were assigned to 2 major groups: the control group and the IC group (Figure 1). Attendance was not mandatory.

Figure 1. Overview of this study's design. IC: inverted classroom; MCQ: multiple-choice question.



Study Design and Procedure

Three weeks before the case-study course, the instructor informed all the students about the concept of the IC approach and this study's process in addition to the importance of their engagement in the learning process. A summary of the process was also posted on Moodle. Students were informed that half of them would be assigned to the IC group where they would receive the course material for preparation 1 week before the in-class session to take the role of the instructor on the target day. The course material was delivered as a PDF document of a PowerPoint (Microsoft Corp) presentation (encompassing 63 slides) file via Moodle. The same course material was used in both approaches. The courses took place on Monday and Tuesday. One week separated the courses of the control from the IC groups.

Six cases concerning various clinically relevant scenarios were included in this course in addition to some incorporated slides to review and emphasize selected foundational concepts. They were centered on the most relevant aspects of antithrombotic drug use in different clinical settings: (1) treatment of pulmonary embolism associated with proximal deep vein thrombosis in a posttrauma patient aged 63 years; (2) prevention of thrombotic events postelective knee replacement surgery in patient aged 78 years; (3) treatment of pulmonary embolism that occurred under combined pill contraception in a woman aged 20 years; (4) prevention of thrombotic events in an acutely ill medical patient aged 80 years; (5) prevention of stroke and systemic embolism in a patient aged 82 years with atrial fibrillation and renal insufficiency; and (6) antithrombotic treatment of myocardial infarction in a patient aged 54 years. Each case was accompanied by a set of standardized questions about antithrombotic treatment decisions and adequate monitoring. Students in the control group attended a case-study course carried out by the instructor with a traditional question-and-answer approach, whereas students in the IC group took on the role of the instructor and were able to recall basic concepts or add details to the course material freely. Two to three students were randomly asked to present and discuss 1 of the 6 cases on the target day. By doing so, students were able to draw on each others' knowledge and understanding. The instructor added details, provided guidance, clarity and feedback whenever required during the progress of the students' presentation and summarized the main features at the end of each case. The classes in the control and IC groups were conducted by the same experienced instructor who is familiar with the content and organization of the case-study course to guarantee the consistency of the teaching content and objectives in the 2 educational approaches. This study's design and progress are illustrated in Figure 1.

Data Collection

On the target day, students in both control and IC groups were asked to complete a pretest (ie, at the beginning) and a posttest (at the end) survey (Multimedia Appendix 1): it consisted of

the same 5 multiple-choice questions (MCQs) to be completed within 5 minutes, then collected in an identified way (ie, including the student name, and the date and the hour of the questionnaire completion) to pair pre- and posttest scripts. Questionnaires were anonymized by a secretary and corrected afterward by an independent assistant instructor. Another 5-minute survey assessing the stress in the week preceding the in-class course (thus assessing how they were affected from the moment they knew which group they belonged to till the target day) was also completed by all the students at the beginning of the course. At the end of the course, students were asked to complete 2 additional surveys, 1 assessing their empathy for the instructor and the other their global satisfaction. The first consisted of rating 3 items by a 7-point Likert scale (1=strongly disagree, 2=disagree, 3=somewhat disagree, 4=neutral, 5=somewhat agree, 6=agree, and 7=strongly agree). The second included 3 questions regarding the preclass workload associated with the educational approach and the related students' perception, and 6 others linked to the students' satisfaction with the course objectives, course material, in-class progress, and educational approach. Surveys were built based on previously validated assessment tools [33-35]. Stress, empathy, and satisfaction surveys (Multimedia Appendices 2-4) were filled out anonymously. Students were not allowed to keep a copy of the different surveys nor to take a photo of these documents.

Statistical Analysis

The distribution of the data was evaluated using the Shapiro-Wilk test. The percentage of participation was compared between groups and sexes using a 2-way ANOVA followed by the 2-stage setup method of Benjamini, Krieger, and Yekutieli [36] for multiple comparisons. The results of the pre- and posttest MCQs were compared for statistically significant differences using the nonparametric Wilcoxon matched-pairs signed rank test. The preclass workload was compared between the 2 groups using the Mann Whitney test. The data relative to the empathy and stress self-assessment were compared between both groups using chi-square test whereas those relative to satisfaction were analyzed using the Fisher exact test. Error probability with a *P* value less than .05 was considered significant. Statistical analysis and graphical representation were performed using GraphPad Prism (version 10.0.2, GraphPad Software, Inc).

Results

Participants Characteristics

In this study, 346 third-year adult students (women: n=213, 62%; men: n=133, 38%) were randomized. As attendance is not mandatory, only 46% (n=158) attended the in-class session. All of them took part in this study. Ninety-three (women: n=70, 75%; men: n=23, 25%) were in the control group whereas 65 (women: n=40, 62% women; men: n=25, 38%) were in the IC group (Table 1). A significantly higher participation rate was observed in the control group (93/171, 54%) compared to the IC group (65/175, 37%; *P*=.002). Women (110/213, 52%) participated more than men (48/133, 36%; *P*=.002) in the case-study course no matter the allocation group. No other demographic information regarding the students was collected.

Table . Characteristics of the participants. Absolute numbers with the percentages concerning the corresponding randomized participants are reported. Women participated more than men in the case-study course whatever the group was (*P*=.002) and a significantly higher participation was observed in the control group compared to the IC[a] group (*P*=.002).

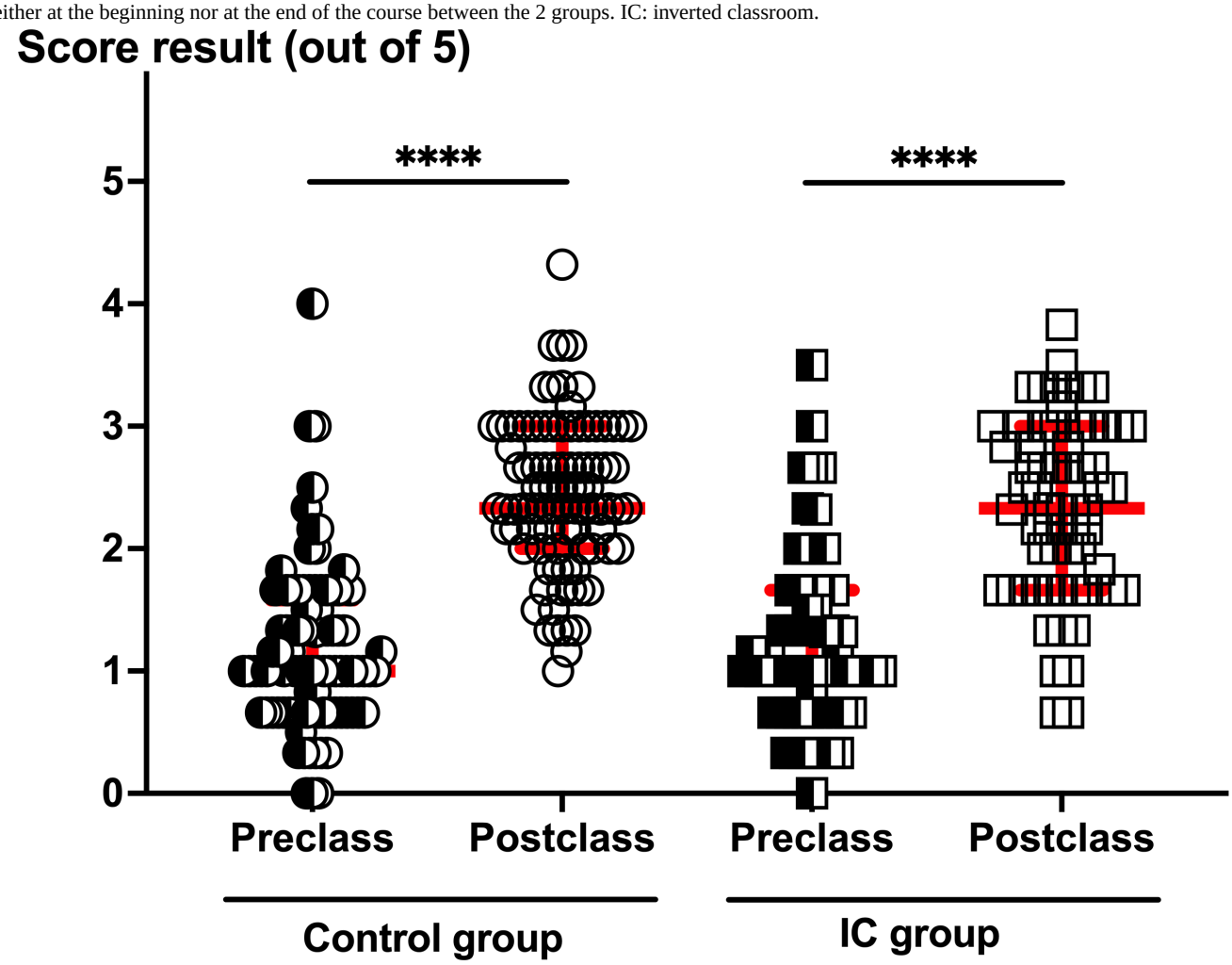
	Sex	Control group	IC group	Total
Randomized participants				
	Men	55	78	133
	Women	116	97	213
	Total	171	175	346
Effective participants, n (%)				
	Men	23 (42)	25 (32)	48 (36)
	Women	70 (60)	40 (41)	110 (52)
	Total	93 (54)	65 (37)	158 (46)

[a]IC: inverted classroom.

Pre- and Postclass Knowledge Survey

On the target day, the in-class session started for both groups with a 5 MCQ survey for 5 minutes to assess the students' readiness to discuss cases and stimulate the recall of knowledge learned before the case-study course. Questions were on the "take-home messages" related to antithrombotic drug use in real-life clinical settings. These MCQs also help students to identify their possible misconceptions at the beginning of the course. The mean and SD of the prescore (out of 5) did not differ between the control group (1.17, SD 0.66) and the IC group (1.24, SD 0.72; Figure 2). To evaluate the students' short-term knowledge retention, the same survey was completed immediately after the class. The mean score improved from the prescore (*P*<.001), in both the control group 2.45 (SD 0.61) and the IC group 2.35 (SD 0.73; Figure 2). Knowledge improvement was comparable in students from both groups.

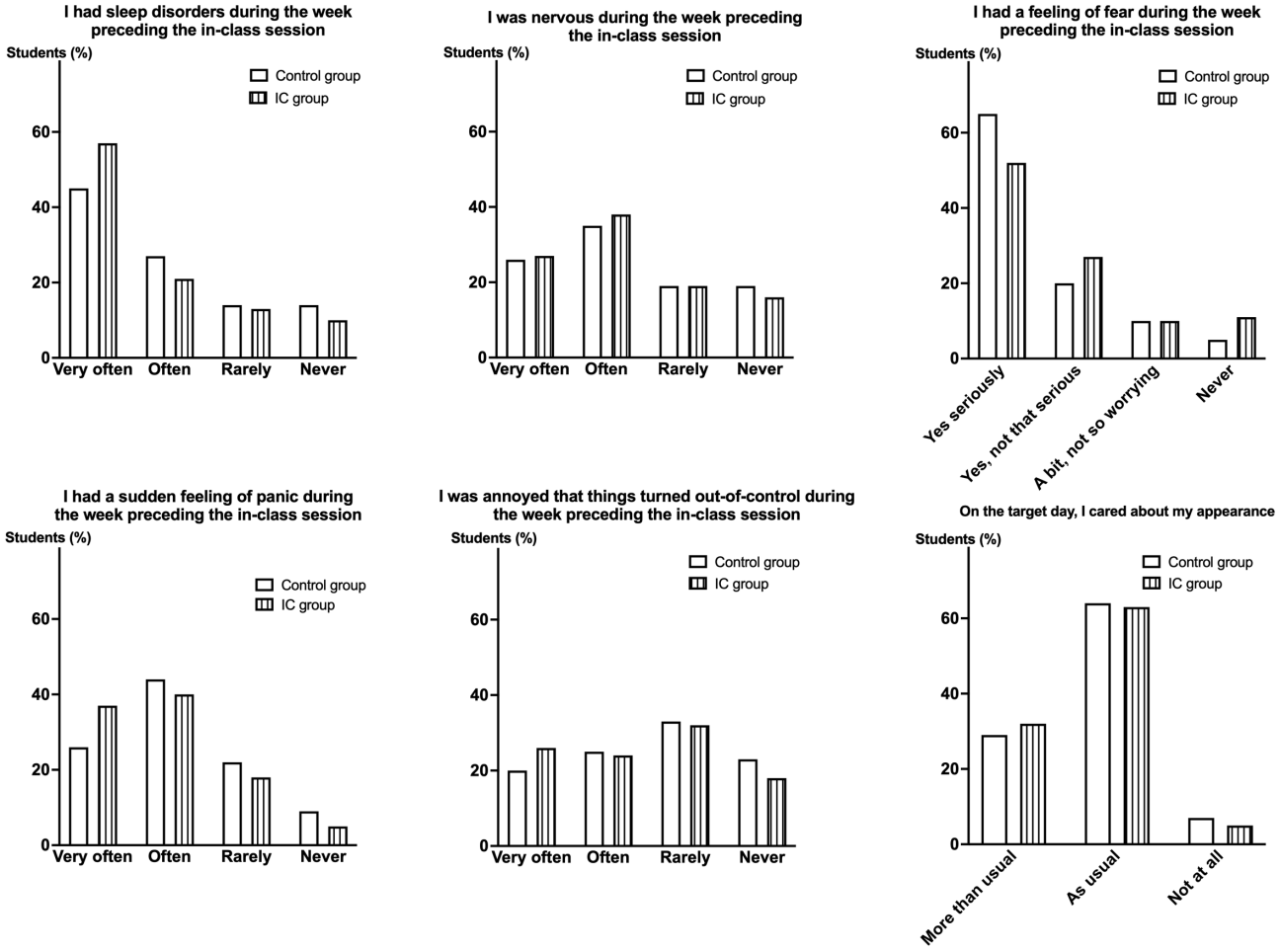
Figure 2. Pre- and postclass knowledge assessment survey per group. Five multiple-choice questions were completed by the control (circles, n=93) and IC (squares, n=65) groups before (semiclosed symbols) and after (open symbols) the completion of the case-study course. Red bars reflect median values with IQRs. Students' scores significantly increased (P<.001) at the end of the course in both groups. No difference in the scores was observed neither at the beginning nor at the end of the course between the 2 groups. IC: inverted classroom.



Stress Self-Assessment

Apart from the preclass knowledge assessment survey, students in both groups completed a stress survey at the beginning of the class during 5 minutes. Questions related to sleep disorders, nervousness, fear, panic, and annoyance during the week preceding the in-class course were completed (Figure 3). Although approximately 50% of the students reported sleep disorders and a feeling of fear during the week preceding the in-class session, no significant difference was observed between the 2 groups for any of the above-mentioned parameters. Likewise, students care for their appearance on the target day did not differ between the control and the IC groups.

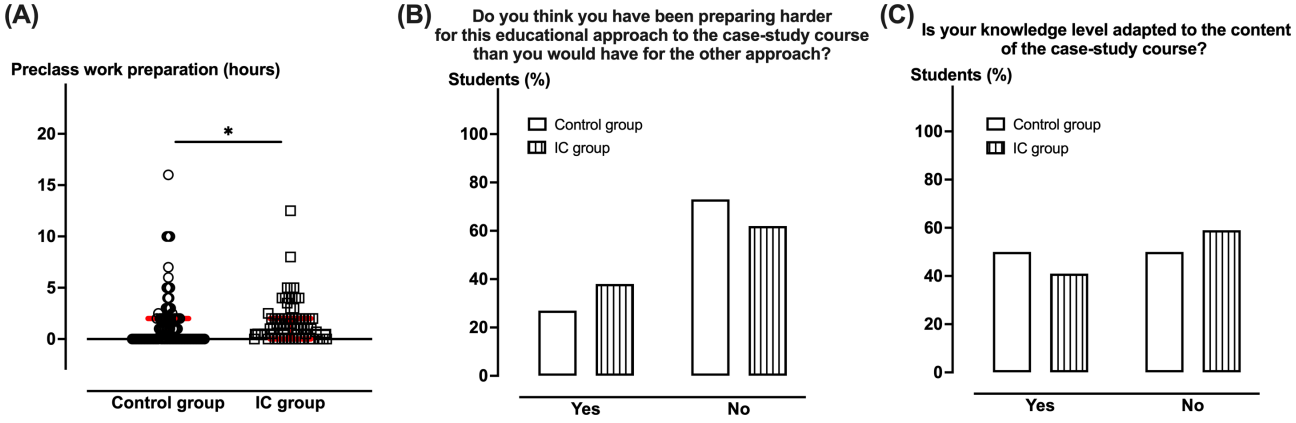
Figure 3. Stress self-assessment questionnaire. A 6-question stress self-assessment survey was completed by both the control (n=93) and IC (n=65) groups at the beginning of the case-study course. No significant difference for any of the 6 questions was observed between the 2 groups. IC: inverted classroom.



Preclass Workload

Students of both groups were asked to provide the number of hours spent preparing for the course, their perception of the preclass preparation difficulty in comparison to the other educational approach as well as the self-assessment of the required skill level. Preclass workload was estimated at 1 (IQR 0-2) hour in the IC group as a median, which was significantly higher (P=.02) than in the control group, 0 (IQR 0-2) hour (Figure 4A). While 38% of the students in the IC group considered that they had been preparing harder for the case-study course than it would have been if they were in the control group, 27% of the latter considered that as such it resulted in no significant difference between both groups (Figure 4B). Approximately half of the students in each group considered having a knowledge level adapted to the case-study course content. No difference was observed between both groups (Figure 4C).

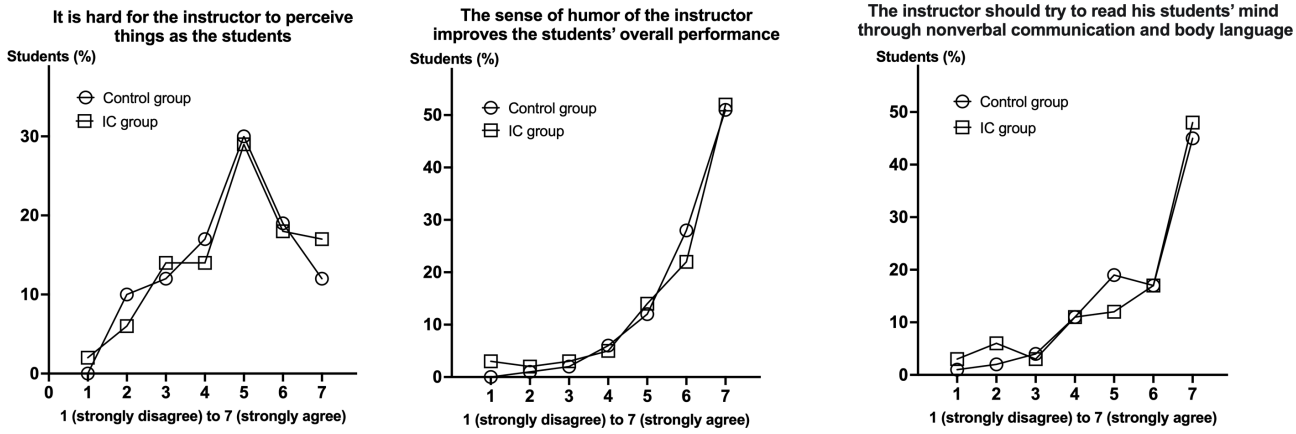
Figure 4. Self-assessment of the preclass work per group. (**A**) Preclass preparation requirements concerning working hours, (**B**) student perception, and (**C**) knowledge level self-assessment were compared between the control and IC groups. Red bars reflect median values with IQRs. While preclass work preparation necessitated more time for the IC compared to the control groups (P=.02), no difference was observed between the 2 groups in terms of work difficulty perception and the self-assessment of the knowledge level adapted to the course. IC: inverted classroom.



Empathy Self-Assessment

Student empathy for their instructor favors engagement and learning behavior in class, hence 3 related questions rated by a 7-point Likert scale were completed by students of both groups at the end of the course (Figure 5). Although only students in the IC group assumed the role of the instructor within the class, students' opinions were quite similar between both groups. Approximately 30% of the students considered that it is hard for the instructor to perceive things as the students while 50% confirmed the importance of the sense of humor of the instructor in enhancing students' performance and of reading the students' minds through nonverbal communication and body language. These results did not differ between both groups.

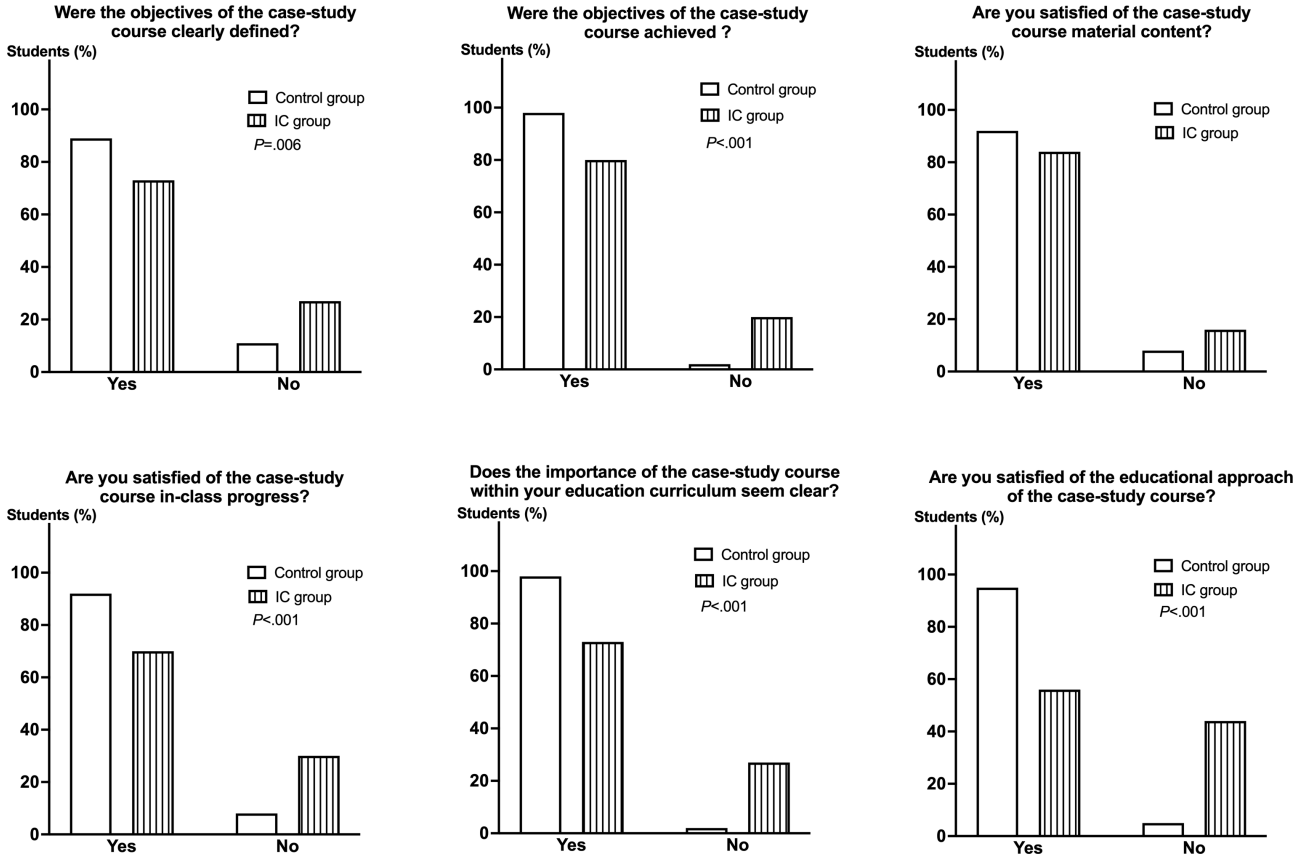
Figure 5. Empathy assessment questionnaire. The empathy of students for the instructor was assessed using 3 questions completed at the end of the case-study course. No difference was observed between the control and the IC groups for any of the 3 raised issues. IC: inverted classroom.



Global Satisfaction Assessment

The students' global satisfaction was evaluated at the end of the course using a 6-item questionnaire (Figure 6). While more than 80% of the students were satisfied with the content of the course material in both groups, the satisfaction survey revealed that the IC approach was not well received among the students. Indeed, 18/65 (28%) of the students found that the objectives of the case-study course not clearly defined and 13/65 (20%) considered that they were not achieved, in comparison to 10/93 (11%*; P*=.006) and 2/93 (2%*; P*<.001) in the control group, respectively. A total of 20/65 (30%) of the students in the IC group were not satisfied with the in-class progress of the course versus 7/93 (8%) in the control group (*P*<.001). While 91/93 (98%) of the students in the control group considered this case-study course on antithrombotic drug use important for pharmacy education, only 47/65 (72%) did so in the IC group (*P*<.001). When it came to the question of the overall satisfaction of the educational approach, 88/93 (95%) of the students in the control group were satisfied versus (36/65) 55% in the IC group (*P*<.001).

Figure 6. Satisfaction survey. A 6-item questionnaire relative to the students' satisfaction with the pedagogical strategy (ie, traditional [n=93] vs IC [n=65]) was completed at the end of the case-study course. Overall, students in the control group were significantly more satisfied than those in the IC group. IC: inverted classroom.



Discussion

Principal Findings

Education methodology is increasingly shifting from a teacher- to student-centered learning approach [37]. In this context, we sought to assess whether applying an IC approach for a case-study course has an added value for conveying optimal antithrombotic drug use skills to pharmacy students. Such an educational approach would increase the communication, critical-thinking, problem-solving, and self-learning skills of pharmacy students. Our study revealed that while the IC approach did not increase student stress, it did not enhance their short-term knowledge retention or their empathy for the instructor. It increased the preclass workload and was not well received among the students. Of note, the case-study course on antithrombotic use was the first on this topic in the PharmD curriculum.

For any educational method to be considered successful, there must be evidence that student learning is enhanced. Many small-scale studies showed enhanced student learning following flipping lecture-based pharmacy courses [5-11,24]. Here, the IC approach applied to the case-study course was not associated with an enhanced preclass knowledge level and did not improve the students' short-term knowledge retention. This was also the case in the Everly and Cochran study in which no significant differences in examination question performance between students in the lecture-based section and the flipped format section were observed [9]. That said, MCQ and examination scores are often used as a proxy measure for learning; however, the ultimate goal is for students to be able to apply classroom learning to real-life situations, which is more difficult to assess. Perhaps we need to examine other nonquantitative student characteristics outcomes following the IC approach in case-study courses (such as intellectual curiosity, personal responsibility, reasoning skills, etc) to identify pharmacy students most likely to succeed.

A recent systematic review including 45 studies with a total of 8426 students from various health professional pathways showed that implementing flipped classes may improve academic performance, and may support student satisfaction, yet the certainty of the evidence is low [37]. A second systematic review and meta-analysis including 11 randomized controlled Chinese studies enrolling 1200 participants suggested that flipped classroom pedagogy enhances students' learning enthusiasm, self-learning ability, thinking and communication skills as well as cooperative ability [38]. Although, another systematic review focusing on students in pharmacy education, incorporated 6 observational studies with 1395 participants. No overall significant difference in final academic performance between the 2 educational models was reported [39]. An important heterogeneity of student perspectives from flipped classrooms has emerged in the literature, ranging from positive [10,24,40-42] to negative [21,39,43,44] and mixed [45] perceptions. These differences are potentially due to the different contexts in which these studies were carried out as well as different student populations, backgrounds, sample sizes, and outcome measures. Apart from that, some instructors may be

more effective teachers than others regardless of the teaching modality. Research evaluating which elements contribute to the efficacy of an IC approach in pharmacy education is still needed. Moreover, it is still unclear if there is a particular area or topic that is better suited for the use of the IC approach in the PharmD curriculum. Besides, with the increased use of the IC method, it is important to consider the impact on students when this approach is incorporated into multiple concurrent courses. Determination of the ideal amount of preclass preparation time across the curriculum would provide helpful guidance to pharmacy faculties implementing such teaching methods.

One of the major limitations of the IC approach is its high dependence on the attendance to class time and on student engagement and discipline for reading and preparing the preclass material as previously emphasized [26]. As the class time attendance is not obligatory in our faculty, only 46% of the students were present in the antithrombotic drug use case-study course. The percentage of attendance was more important in the control group than in the IC group. A more elaborate strategy for students' motivation should thus be implemented to obtain a higher engagement and adherence to our case-study courses in general, and to such IC experience if it shall be repeated. Providing clear expectations to students, keeping the preparation tasks focused, and explicitly linking preparation activities to in-class active learning could be some key methods for instructors to increase the proportion of students who prepare for classes. Future research should also be devoted to assessing the potential effect of sex, gender, socioeconomic background, and age on the outcomes of such an approach in the PharmD curriculum.

Preclass preparation could be considered as a considerable "extra" work [10,29,30,46]. Indeed, students in the IC group of our study reported an increased preclass workload which might take up an amount of their spare time leading to negative feedback on this approach. It is to be mentioned that the course materials made available beforehand should not be too complex to be understood by the students on their own. We did our best to include 6 uncomplicated cases issued from real-life settings and incorporate few slides to recall knowledge learned in the three lecture-based courses during the month preceding the case-study course.

Case-study courses with an IC approach are probably important in pharmacy education as they would help students promote higher-level critical-thinking skills, foster analyzing, and improve their communication skills therefore improving their motivation and attitudes [47]. Assisting them in learning how to think and communicate like a graduated pharmacist will prepare them for their future beyond pharmacy school. Communication skills are required to ensure patient understanding and compliance [48,49]. If a patient does not understand the purpose of the antithrombotic treatment, adherence will likely be low. Communication training in pharmacy students is thus mandatory to improve their later effectiveness as future health professionals. Although third-year students were overall not satisfied with this experience, such an approach is worth being retested with "older students," from the fourth to the sixth year of the PharmD curriculum. Successful pharmacy students are expected to have the ability to manage their learning and adequately communicate their knowledge. Self-learning skills are particularly crucial to achieving effective lifelong learning in pharmacy, where scientific knowledge is continually evolving. Consequently, pharmacy students should be trained to be effective self-learners. Antithrombotic drug use assessed using a case-study course is an application-based activity, while the material taught previously is mostly a knowledge-based presentation. Therefore, third-year students may have significant difficulty in providing an application-based activity when their skill level may have been still on a much lower level of Bloom's taxonomy [4] which may explain the low postclass knowledge scores in both groups. They will complete 2 additional learning years where it is anticipated they will gain more knowledge, clinical experience, and communication skills through their traineeships in community-based and hospital pharmacies as well as clinical laboratories. That said, many students were very interested in participating in this pedagogical study and found the experience to be innovative and enjoyable.

From the instructor's perspective, the IC approach might be more challenging than a traditional session with a question-and-answer approach due to the risk that students' presentation and case discussion activities create an unsettled classroom, thus a chaotic environment in which students may feel lost, and the fear that students may be unable to deliver the course adequately. However, the IC approach makes student engagement in the course easier and empowers them as active participants in their learning in comparison to the traditional educational approach. It allows the instructor to guide students in deeper learning processes as previously shown [38,46]. A more interactive teaching strategy may be more attractive than the IC approach, such as the adventure game recently developed by Perrin et al [50]. It is a video game in which the player assumes the role of a protagonist in an interactive story driven by exploration and problem-solving tests. Briefly, the pharmacy student assumes the role of a hematology superhero named SUPER HEMO. SUPER HEMO can meet 5 unwell characters in 5 different steps. The player must answer their questions and find the best way to diagnose and cure them. At the final hematology evaluation, students who played SUPER HEMO had a slightly better (but not statistically significant) median knowledge score than those who did not [50]. The value of such an innovative strategy for case-study courses on antithrombotic drug use in the PharmD curriculum remains to be established. Team-based learning is another educational approach that provides structure, defined timeframes, and formative assessment opportunities. It was previously shown to develop students learning enthusiasm, self-study, and thinking abilities as well as communication skills [51]. Therefore, it might be considered as an alternative approach for improving case-study courses in the PharmD curriculum, and thus is worth being assessed.

One possible limitation of our study is that the long-term effect of our IC approach on knowledge retention and skill application was not assessed. The question relative to this course was deliberately not included in the final exam in order not to create any lack of equity among the students of both groups. We did not complete any postclass knowledge survey 3 to 6 months

later, nor did we assess the acquired skills through, for instance, an objective structured clinical examination. This remains to be specifically investigated. Second, we did not collect data on how many students effectively accessed the course material before the in-class session, although it might be feasible via the information technology service. As the preclass results of the students in the IC group were not better than those of the control group, we hypothesized that a lot of students had not read the course material before the target day. Noteworthily, it is hard to control whether students had effectively read and prepared the course material before the in-class session. Students might only click on the material folder without reading it or read parts of it. Also, several students might have accessed the material via 1 student user ID. We also did not ask students in the presurvey questionnaire whether they had read the assignment. However, they would, most probably, have not told the truth. Third, ideally, students in the IC group should have been asked to prepare the course material by themselves, yet this was not the case to avoid a high level of absenteeism as in-class attendance is not mandatory according to Paris Cité University policy. Despite this, a relatively small number of students effectively participated in this study. Fourth, in-class sessions with the traditional approach were completed 1 week before those with the IC approach to prevent students in the former group from having access to the material course from those of the latter group before the in-class session. Consequently, we cannot rule out the possibility that some students initially assigned to the IC group had changed their group assignment to get the case-study course at an early date. Finally, our findings cannot be generalized to other contexts, particularly to students in other year cohorts or to other specialties. This remains to be specifically investigated.

Conclusions

Our study showed that an IC approach does not appear to be suited to the case-study course on antithrombotic drug use in the third-year PharmD curriculum. While no additional gain in short-term knowledge was observed using this approach in comparison to the traditional educational approach, we perceived significantly lower student satisfaction. However, the increased instructor-student and student-student interactions are still convincing arguments to try this pedagogical approach. Hence, additional research in this field is still needed to implement innovative educational approaches aiming at improving the knowledge and skills of our future pharmacists.

Acknowledgments

The authors thank the members of the Institutional Review Board and Ethics Committee of Paris Cité University for the rapid review of the project, Farah Berrich, MSc, for her administrative support, and the 2022/2023 third-year pharmacy students of Paris Cité University for their participation in this study.

Data Availability

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

Authors' Contributions

GJ, VS, JT, and IM conceptualized this study. GJ worked on investigation. GJ and MS did the data analysis. GJ wrote the original draft. VS, MS, PG, JT, and IM reviewed and edited the writing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Pre- and postclass multiple-choice questions.
[DOCX File, 16 KB - mededu_v11i1e67419_app1.docx]

Multimedia Appendix 2
Stress self-assessment survey
[DOCX File, 16 KB - mededu_v11i1e67419_app2.docx]

Multimedia Appendix 3
Empathy assessment questionnaire
[DOCX File, 14 KB - mededu_v11i1e67419_app3.docx]

Multimedia Appendix 4
Preclass workload self-assessment and satisfaction questionnaire
[DOCX File, 16 KB - mededu_v11i1e67419_app4.docx]

Checklist 1
CONSORT 2010 checklist. CONSORT: Consolidated Standards of Reporting Trials.
[PDF File, 126 KB - mededu_v11i1e67419_app5.pdf]

References

1. Laroche ML, Gautier S, Polard E, et al. Incidence and preventability of hospital admissions for adverse drug reactions in France: a prospective observational study (IATROSTAT). Br J Clin Pharmacol 2023 Jan;89(1):390-400. [doi: 10.1111/bcp.15510] [Medline: 36002314]
2. Réseau Français des Centres Régionaux de Pharmacovigilance [Website in French]. URL: https://www.rfcrpv.fr/ [accessed 2025-03-25]
3. Kehrer JP, Eberhart G, Wing M, Horon K. Pharmacy's role in a modern health continuum. Can Pharm J (Ott) 2013 Nov;146(6):321-324. [doi: 10.1177/1715163513506370] [Medline: 24228046]
4. Krathwohl DR. A revision of Bloom's taxonomy: an overview. THEORY Pract 2002 Nov 1;41(4):212-218. [doi: 10.1207/s15430421tip4104_2]
5. Roth MT, Mumper RJ, Singleton SF, et al. A renaissance in pharmacy education at the University of North Carolina at Chapel Hill. N C Med J 2014;75(1):48-52. [doi: 10.18043/ncm.75.1.48] [Medline: 24487762]
6. Stewart DW, Brown SD, Clavier CW, Wyatt J. Active-learning processes used in US pharmacy education. Am J Pharm Educ 2011 May 10;75(4):68. [doi: 10.5688/ajpe75468] [Medline: 21769144]
7. Bossaer JB, Panus P, Stewart DW, Hagemeier NE, George J. Student performance in a pharmacotherapy oncology module before and after flipping the classroom. Am J Pharm Educ 2016 Mar 25;80(2):31. [doi: 10.5688/ajpe80231] [Medline: 27073284]
8. Ferreri SP, O'Connor SK. Redesign of a large lecture course into a small-group learning course. Am J Pharm Educ 2013 Feb 12;77(1):13. [doi: 10.5688/ajpe77113] [Medline: 23459199]
9. 115th Annual Meeting of the American Associaton of Colleges of Pharmacy, Grapevine, TX, July 26-30, 2014. Am J Pharm Educ 2014 Jun;78(5):111. [doi: 10.5688/ajpe785111] [Medline: 4064488]
10. Wong TH, Ip EJ, Lopes I, Rajagopalan V. Pharmacy students' performance and perceptions in a flipped teaching pilot on cardiac arrhythmias. Am J Pharm Educ 2014 Dec 15;78(10):185. [doi: 10.5688/ajpe7810185] [Medline: 25657372]
11. Persky AM, Dupuis RE. An eight-year retrospective study in "flipped" pharmacokinetics courses. Am J Pharm Educ 2014 Dec 15;78(10):190. [doi: 10.5688/ajpe7810190] [Medline: 25657377]
12. McLaughlin JE, Gharkholonarehe N, Khanova J, Deyo ZM, Rodgers JE. The impact of blended learning on student performance in a cardiovascular pharmacotherapy course. Am J Pharm Educ 2015 Mar 25;79(2):24. [doi: 10.5688/ajpe79224] [Medline: 25861105]
13. Edginton A, Holbrook J. A blended learning approach to teaching basic pharmacokinetics and the significance of face-to-face interaction. Am J Pharm Educ 2010 Jun 15;74(5):88. [doi: 10.5688/aj740588] [Medline: 20798797]
14. Crouch MA. An advanced cardiovascular pharmacotherapy course blending online and face-to-face instruction. Am J Pharm Educ 2009 May 27;73(3):51. [doi: 10.5688/aj730351] [Medline: 19564994]
15. Wilson JA, Waghel RC, Dinkins MM. Flipped classroom versus a didactic method with active learning in a modified team-based learning self-care pharmacotherapy course. Curr Pharm Teach Learn 2019 Dec;11(12):1287-1295. [doi: 10.1016/j.cptl.2019.09.017] [Medline: 31836155]
16. Wong WJ, Lee SWH, White PJ, Efendie B, Lee RFS. Perspectives on opportunities and challenges in a predominantly flipped classroom-based pharmacy curriculum: a qualitative study. Curr Pharm Teach Learn 2023 Mar;15(3):242-251. [doi: 10.1016/j.cptl.2023.03.004] [Medline: 37055316]
17. Youhasan P, Chen Y, Lyndon M, Henning MA. Exploring the pedagogical design features of the flipped classroom in undergraduate nursing education: a systematic review. BMC Nurs 2021 Mar 22;20(1):50. [doi: 10.1186/s12912-021-00555-w] [Medline: 33752654]
18. Hess R, Hagemeier NE, Blackwelder R, Rose D, Ansari N, Branham T. Teaching communication skills to medical and pharmacy students through a blended learning course. Am J Pharm Educ 2016 May 25;80(4):64. [doi: 10.5688/ajpe80464] [Medline: 27293231]
19. Bergmann J, Sams A. Flip Your Classroom: Reach Every Student in Every Class Every Day: International Society for Technology in Education; 2012.
20. Cai L, Li YL, Hu XY, Li R. Implementation of flipped classroom combined with case-based learning: a promising and effective teaching modality in undergraduate pathology education. Medicine (Baltimore) 2022 Feb 4;101(5):e28782. [doi: 10.1097/MD.0000000000028782] [Medline: 35119043]
21. Chen KS, Monrouxe L, Lu YH, et al. Academic outcomes of flipped classroom learning: a meta-analysis. Med Educ 2018 Jun 25;52(9):910-924. [doi: 10.1111/medu.13616] [Medline: 29943399]
22. Crome M, Adam K, Flohr M, Rahman A, Staufenbiel I. Application of the inverted classroom model in the teaching module "new classification of periodontal and peri-implant diseases and conditions" during the COVID-19 pandemic. GMS J Med Educ 2021;38(5):Doc89. [doi: 10.3205/zma001485] [Medline: 34286069]

23. Malik PRV, Nakhla N. Instructor-blinded study of pharmacy student learning when a flipped online classroom was implemented during the COVID-19 pandemic. Pharmacy (Basel) 2022 May 11;10(3):53. [doi: 10.3390/pharmacy10030053] [Medline: 35645332]
24. McLaughlin JE, Roth MT, Glatt DM, et al. The flipped classroom: a course redesign to foster learning and engagement in a health professions school. Acad Med 2014 Feb;89(2):236-243. [doi: 10.1097/ACM.0000000000000086] [Medline: 24270916]
25. McCabe C, Smith MG, Ferreri SP. Comparison of flipped model to traditional classroom learning in a professional pharmacy course. Educ Sci 2017;7(3):73. [doi: 10.3390/educsci7030073]
26. White PJ, Naidu S, Yuriev E, Short JL, McLaughlin JE, Larson IC. Student engagement with a flipped classroom teaching design affects pharmacology examination performance in a manner dependent on question type. Am J Pharm Educ 2017 Nov;81(9):5931. [doi: 10.5688/ajpe5931] [Medline: 29302082]
27. Saseen JJ, Linnebur SA, Borgelt LM, Trujillo J, Fish DN, Mueller S. A pharmacotherapy capstone course to target student learning and programmatic curricular assessment. Am J Pharm Educ 2017 Apr;81(3):45. [doi: 10.5688/ajpe81345] [Medline: 28496265]
28. Koo CL, Demps EL, Farris C, Bowman JD, Panahi L, Boyle P. Impact of flipped classroom design on student performance and perceptions in a pharmacotherapy course. Am J Pharm Educ 2016 Mar 25;80(2):33. [doi: 10.5688/ajpe80233] [Medline: 27073286]
29. Gloudeman MW, Shah-Manek B, Wong TH, Vo C, Ip EJ. Use of condensed videos in a flipped classroom for pharmaceutical calculations: student perceptions and academic performance. Curr Pharm Teach Learn 2018 Feb;10(2):206-210. [doi: 10.1016/j.cptl.2017.10.001] [Medline: 29706277]
30. Cotta KI, Shah S, Almgren MM, Macías-Moriarity LZ, Mody V. Effectiveness of flipped classroom instructional model in teaching pharmaceutical calculations. Curr Pharm Teach Learn 2016 Sep;8(5):646-653. [doi: 10.1016/j.cptl.2016.06.011]
31. Patanwala AE, Erstad BL, Murphy JE. Student use of flipped classroom videos in a therapeutics course. Curr Pharm Teach Learn 2017;9(1):50-54. [doi: 10.1016/j.cptl.2016.08.043] [Medline: 29180154]
32. See S, Conry JM. Flip my class! A faculty development demonstration of a flipped-classroom. Curr Pharm Teach Learn 2014 Jul;6(4):585-588. [doi: 10.1016/j.cptl.2014.03.003]
33. Bjelland I, Dahl AA, Haug TT, Neckelmann D. The validity of the Hospital Anxiety and Depression Scale. J Psychosom Res 2002 Feb;52(2):69-77. [doi: 10.1016/S0022-3999(01)00296-3]
34. Cohen S, Sherrod DR, Clark MS. Social skills and the stress-protective role of social support. J Pers Soc Psychol 1986 May;50(5):963-973. [doi: 10.1037//0022-3514.50.5.963] [Medline: 3486973]
35. Mirani SH, Shaikh NA, Tahir A. Assessment of clinical empathy among medical students using the Jefferson scale of empathy-student version. Cureus 2019 Feb 28;11(2):e4160. [doi: 10.7759/cureus.4160] [Medline: 31058043]
36. Benjamini Y, Krieger AM, Yekutieli D. Adaptive linear step-up procedures that control the false discovery rate. Biometrika 2006 Sep 1;93(3):491-507 [FREE Full text] [doi: 10.1093/biomet/93.3.491]
37. Naing C, Whittaker MA, Aung HH, Chellappan DK, Riegelman A. The effects of flipped classrooms to improve learning outcomes in undergraduate health professional education: a systematic review. Campbell Syst Rev 2023 Sep;19(3):e1339. [doi: 10.1002/cl2.1339] [Medline: 37425620]
38. Peng W, Xiong Y, Wei J, et al. Flipped classroom improves student learning outcome in Chinese pharmacy education: a systematic review and meta-analysis. Front Pharmacol 2022;13:936899. [doi: 10.3389/fphar.2022.936899] [Medline: 36110553]
39. Gillette C, Rudolph M, Kimble C, Rockich-Winston N, Smith L, Broedel-Zaugg K. A meta-analysis of outcomes comparing flipped classroom and lecture. Am J Pharm Educ 2018 Jun;82(5):6898. [doi: 10.5688/ajpe6898] [Medline: 30013248]
40. White PJ, Larson I, Styles K, et al. Adopting an active learning approach to teaching in a research-intensive higher education context transformed staff teaching attitudes and behaviours. HERD 2016 May 3;35(3):619-633. [doi: 10.1080/07294360.2015.1107887]
41. He Y, Lu J, Huang H, et al. The effects of flipped classrooms on undergraduate pharmaceutical marketing learning: a clustered randomized controlled study. PLoS ONE 2019;14(4):e0214624. [doi: 10.1371/journal.pone.0214624] [Medline: 30969976]
42. Anderson HG Jr, Frazier L, Anderson SL, et al. Comparison of pharmaceutical calculations learning outcomes achieved within a traditional lecture or flipped classroom andragogy. Am J Pharm Educ 2017 May;81(4):70. [doi: 10.5688/ajpe81470] [Medline: 28630511]
43. Missildine K, Fountain R, Summers L, Gosselin K. Flipping the classroom to improve student performance and satisfaction. J Nurs Educ 2013 Oct;52(10):597-599. [doi: 10.3928/01484834-20130919-03] [Medline: 24044386]
44. Kugler AJ, Gogineni HP, Garavalia LS. Learning outcomes and student preferences with flipped vs lecture/case teaching model in a block curriculum. Am J Pharm Educ 2019 Oct;83(8):7044. [doi: 10.5688/ajpe7044] [Medline: 31831896]
45. McLaughlin JE, White PJ, Khanova J, Yuriev E. Flipped classroom implementation: a case report of two higher education institutions in the United States and Australia. Comput Sch 2016 Jan 2;33(1):24-37. [doi: 10.1080/07380569.2016.1137734]
46. Akçayır G, Akçayır M. The flipped classroom: a review of its advantages and challenges. Comput Educ 2018 Nov;126:334-345. [doi: 10.1016/j.compedu.2018.07.021]

47. Freeman S, O'Connor E, Parks JW, et al. Prescribed active learning increases performance in introductory biology. CBE Life Sci Educ 2007;6(2):132-139. [doi: 10.1187/cbe.06-09-0194] [Medline: 17548875]
48. Ranjan P, Kumari A, Chakrawarty A. How can doctors improve their communication skills? J Clin Diagn Res 2015 Mar;9(3):JE01-JE04. [doi: 10.7860/JCDR/2015/12072.5712] [Medline: 25954636]
49. Maguire P, Pitceathly C. Key communication skills and how to acquire them. BMJ 2002 Sep 28;325(7366):697-700. [doi: 10.1136/bmj.325.7366.697] [Medline: 12351365]
50. Perrin J, Meeus A, Broseus J, et al. A serious game about hematology for health care workers (SUPER HEMO): development and validation study. JMIR Serious Games 2023 Feb 13;11:e40350. [doi: 10.2196/40350] [Medline: 36780215]
51. Lang B, Zhang L, Lin Y, Han L, Zhang C, Liu Y. Team-based learning pedagogy enhances the quality of Chinese pharmacy education: a systematic review and meta-analysis. BMC Med Educ 2019 Jul 29;19(1):286. [doi: 10.1186/s12909-019-1724-6] [Medline: 31357986]

Abbreviations

IC: inverted classroom
MCQ: multiple-choice question
PharmD: Doctor of Pharmacy

Edited by TDA Cardoso; submitted 11.10.24; peer-reviewed by D Schwarz, MM Vasquez, S Arsić; revised version received 25.02.25; accepted 12.03.25; published 10.04.25.

Please cite as:
Jourdi G, Selmi M, Gaussem P, Truchot J, Margaill I, Siguret V
Evaluation of the Inverted Classroom Approach in a Case-Study Course on Antithrombotic Drug Use in a PharmD Curriculum: French Monocentric Randomized Study
JMIR Med Educ 2025;11:e67419
URL: https://mededu.jmir.org/2025/1/e67419
doi:10.2196/67419

Balancing Academics and Life: Qualitative Study of Health Professions Students' Perceptions of a Four-Day Academic Week in the United Arab Emirates

Ashokan Arumugam[1,2,3,4*], MPT, PhD; Jacqueline Maria Dias[5*], PhD; Sangeetha Narasimhan[6], PhD; Raneen Mohammed Qadah[1], MSc; Reime Jamal Shalash[1], PhD; Taif A M Omran[1], BSc; Bashair M Mussa[7], PhD; Basema Saddik[8,9], PhD; Nadia Rashed Al Mazrouei[10], PhD; Sivapriya Ramakrishnan[1], MA, MPT

[1]Department of Physiotherapy, College of Health Sciences, University of Sharjah, PO Box 27272, Sharjah, United Arab Emirates

[10]Department of Pharmacy Practice and Pharmacotherapeutics, College of Pharmacy, University of Sharjah, Sharjah, United Arab Emirates

[2]Neuromusculoskeletal Rehabilitation Research Group, RIMHS – Research Institute of Medical and Health Sciences, University of Sharjah, Sharjah, United Arab Emirates

[3]Sustainable Engineering Asset Management Research Group, RISE - Research Institute of Sciences and Engineering, University of Sharjah, Sharjah, United Arab Emirates

[4]Department of Physiotherapy, Manipal College of Health Professions, Manipal Academy of Higher Education, Manipal, Karnataka, India

[5]Department of Nursing, College of Health Sciences, University of Sharjah, Sharjah, United Arab Emirates

[6]Department of Oral and Craniofacial Health Sciences, College of Dental Medicine, University of Sharjah, Sharjah, United Arab Emirates

[7]Department of Basic Medical Sciences, College of Medicine, University of Sharjah, Sharjah, United Arab Emirates

[8]Department of Public Health and Epidemiology, College of Medicine and Health Sciences, Khalifa University, Abu Dhabi, United Arab Emirates

[9]School of Population Health, Faculty of Medicine and Health, University of New South Wales, Sydney, Australia

[*]these authors contributed equally

Corresponding Author:
Ashokan Arumugam, MPT, PhD
Department of Physiotherapy, College of Health Sciences, University of Sharjah, PO Box 27272, Sharjah, United Arab Emirates

Abstract

Background: Since January 2022, Sharjah in the United Arab Emirates has implemented the 4-day work week model for the first time in the public and private sectors, including universities. While this framework may enhance productivity and work-life balance for many professionals, the current study specifically explores the perceptions of students in medicine, dentistry, and health sciences programs regarding the impact of transitioning from a 5-day to a 4-day work week on their academic performance.

Objective: The objective of this study was to explore and analyze the perceptions of students in medicine, dentistry, and health sciences regarding the implementation of a 4-day academic week in the Emirate of Sharjah, United Arab Emirates.

Methods: Twenty-four university students (mean age 20.95, SD 1.30 years; 12 men) studying in medicine, dentistry, physiotherapy, nursing, or medical diagnostic imaging programs, who experienced a transition from a 5-day week to a 4-day week, participated in semistructured interviews lasting approximately 20 - 30 minutes. All interviews were recorded and transcribed verbatim. The Braun and Clarke 6-phase framework of thematic analysis was used.

Results: We identified 5 themes: academic journey, academic work-life balance, support systems, classroom dynamics, and common stressors of the 4-day academic week. Overall, most students reported increased motivation, engagement, and academic achievement following the transition from a 5-day to a 4-day week. In addition, participants described a positive academic work-life balance, improved physical and mental well-being, optimal use of time for both academic and personal commitments, favorable support from faculty and family members, and maintained or even improved attendance levels. Nevertheless, some students expressed concerns about condensed schedules and longer days, increased stress, disrupted work-life balance, and inadequate support systems to cope with this new framework.

Conclusions: Overall, the 4-day academic week enhanced motivation, academic performance, work-life balance, and the physical and mental well-being of medicine, dental, and health science students. However, some students experienced challenges related to condensed schedules and increased stress. These mixed outcomes highlight that while the 4-day work week offers notable advantages, careful planning and support are essential to mitigate the potential drawbacks and ensure all students can succeed within this new academic framework. Future research could explore strategies to address these challenges and further improve the 4-day week experience for all students.

(JMIR Med Educ 2025;11:e67775) doi:10.2196/67775

KEYWORDS

four-day academic week; students; academic performance; work-life balance; support systems; United Arab Emirates

Introduction

The concept of a 4-day work week has gained significant attention, particularly in the aftermath of the global pandemic [1]. This shift has sparked a movement toward reducing the traditional 5-day work week to 4 days across various industries, including higher education. Typically, universities operate on a 5-day schedule, but with the adoption of a 4-day week, classes are now condensed into 4 days, which allows a long (3-day) weekend. Numerous studies have highlighted the advantages of this model, such as enhanced productivity, improved job satisfaction, better academic work-life balance, lower stress levels, and reduced personal expenses, such as commuting costs [2,3].

The 4-day school week was introduced in a South Dakota school in 1936. In 1973, many schools across the northeastern United States adopted this model to address economic challenges, particularly those related to energy and operational costs [4,5]. In more recent years, the number of school districts transitioning from the traditional 5-day week to a 4-day week with extended hours has increased [6]. Hewitt and Denny [4] reported minimal differences in the academic performance of students attending schools on the 5-day schedule compared to those following the 4-day system. Furthermore, this study recommended considering multiple factors beyond academic performance when evaluating the effects of the 4-day work week model [4].

According to Sagness et al [7], the 4-day school week offers potential advantages for both rural and suburban schools, particularly in terms of student achievement, behavior, attendance, as well as job satisfaction among teachers and staff. Thompson and Ward [6] revealed that students in nonrural 4-day schools had a reduction in on-time grade progression and absenteeism with the implementation of 4-day week. However, a few United States district schools reverted to a 5-day week due to reduced academic achievement [8].

A recent study found that school schedules during a 4-day school week have had detrimental impacts on student achievement, with a decline in math and reading scores [9]. On the contrary, another study conducted on students found that there is a positive correlation between implementing a 4-day school week and academic performance in reading and mathematics [10].

Since January 2022, Sharjah, one of the 7 Emirates in the United Arab Emirates, has adopted the 4-day (work) week model across both the public and private sectors. This happened after the United Arab Emirates moved its weekend from Friday-Saturday to Saturday-Sunday, mainly for governmental institutions, to better align its business economy with global markets. A government study, presented to the Sharjah Executive Council in 2023, revealed significant improvements in job performance, happiness, and mental well-being among the government employees [11]. A survey conducted in 2023 among 40,000 parents and employees of schools by the Sharjah Private Education Authority revealed that school children's absenteeism decreased, motivation increased, and behavioral problems reduced while teachers' work-life balance improved significantly [12]. A recent survey analyzing the impact of a 4-day academic week on dental students in the United Arab Emirates revealed high levels of satisfaction with both their academic and clinical training. Students expressed contentment with the extended weekend, which they found beneficial for reducing stress, improving exam preparation, and allowing them to spend more time with family and friends [13]. However, only dental students with or without the experience of both 4-day week and 5-day week were included in this survey. These findings cannot be generalized to all medicine and health sciences students, as their coursework, clinical practice, and laboratory skills requirements are relatively different. In addition, those students who have not experienced a transition from a 5-day week to a 4-day week may not be able to appreciate the benefits and challenges of this changeover. Although electronic surveys are easy to administer, cost-effective, and quick to deploy, they present challenges such as difficulty in identifying the characteristics of nonrespondents and maintaining confidentiality, among others. They do not provide the same in-depth understanding of participants' perceptions, actions, thoughts, and experiences as (semistructured) interviews do [14].

Although numerous studies have investigated the effects of a 4-day academic or work week in schools and various sectors, no research has specifically examined its impact on university students in medicine, dental, and health sciences programs who transitioned from a 5-day to a 4-day academic week. Therefore, it is crucial to assess how this shift affects the academic performance of students in these programs. To the best of our knowledge, this is the first qualitative study of the perceptions of medicine and health sciences students in Sharjah regarding the 4-day work week model using semistructured interviews. If the 4-day work week is found to improve academic outcomes, the results could have broader implications for educational institutions still operating on a 5-day work week, prompting them to consider adopting a 4-day work week schedule to support student performance in the other 6 Emirates. The themes and insights emerging from this study will provide valuable perspectives on how the 4-day work week impacts the academic performance of students in medicine, dental, and health sciences programs.

Methods

Study Design, Setting, and Participants

In this study, we adopted an exploratory qualitative research design using semistructured interviews. A qualitative approach enabled the understanding of the perceptions of students who had used both the 5-day week and were not exposed to the 4-day week. We adhered to the COREQ (Consolidated Criteria for Reporting Qualitative Studies) guidelines for reporting this qualitative research [15].

Study Population

All students enrolled in the medical and health sciences colleges at the University of Sharjah, United Arab Emirates, were invited to participate in the study through advertisements posted on university notice boards, emails, and word of mouth. The participants included students who had experienced the transition from a 5-day week to a 4-day week, with the study taking place during the Fall semester of 2023. These students had been exposed to 3 semesters since the introduction of the 4-day work week. Students were selected based on their willingness and availability to participate. Purposive sampling was used to ensure representation from the 4 different colleges on the medical campus, including medicine, dental medicine, pharmacy, and health sciences. Recruitment of participants continued until data saturation was reached. The interviews were conducted in one of the classrooms on the medical campus, free from noise and distractions.

Instrument

An electronic search of PubMed, Scopus, and Google Scholar was conducted to inform the development of interview questions. We used the following search terms: ("student" OR "learner" OR "pupil" OR "scholar") AND ("perception" OR "attitude" OR "view" OR "opinion") AND ("transition" OR "change" OR "shift" OR "move") AND ("four-day" OR "4-day" OR "four day") AND ("five-day" OR "5-day" OR "five day" OR "traditional") AND ("academic week" OR "school week" OR "educational week" OR "work week"). We could not find qualitative studies on university students' perceptions of transitioning from a 4-day to a 5-day academic week. Only one survey published in 2024 has reported the impact of a 4-day week on academic performance and study–life balance of university students studying dental medicine [13]. To inform the development of the interview guide, we reviewed existing literature on the transitions in education, school and university students, and alternative workweek models in other sectors [6,13,16-19]. Although some studies were based on survey designs rather than qualitative studies, they provided relevant prompts, clues, and contextual insights that shaped the structure and content of our questions tailored to answer our aims of the study.

Our interview questions and probes were explicitly designed to align with the study's specific research aims, allowing a focused exploration of participants' perspectives that are consistent with our objectives. Two independent external content experts with qualitative research experience—one a physiotherapy academic from India and the other a nursing academic from the United Arab Emirates, both not involved in the study—reviewed the interview guide and suggested minor amendments. The interview guide was piloted during mock interviews prior to data collection to ensure clarity. The final interview questions are presented in Textbox 1.

Textbox 1. Semistructured interview questions.

Questions

1. In your opinion, would you tell me about the impact of a 4-day workday week on your lectures and lab sessions?
   a. How would you describe your current attendance levels in lectures and labs compared to a 5-day week?
   b. In your opinion, are there any specific changes in the method of lectures and lab taught?
   c. Would you like to describe any specific consequences or concerns over the 4-day work week in your academic life?
2. In what ways has the 4-day work week impacted your assignments and coursework?
   a. In what ways has the 4-day week impacted individual and group assignments compared to a 5-day week?
   b. In your opinion, were there any specific changes in the type of assignments and course work compared to a 5-day week?
3. In what ways has the 4-day work week impacted on learning theoretical information and honing practical (lab or clinical) skills?
4. In what ways did the 4-day work week affect your performance in mid-term and final exams?
   a. How do you feel that the 4-day work week has affected your ability to prepare for exams or complete assignments in a timely manner?
5. How did the 4-day work week affect your clinical placements and gaining clinical skills?
6. Would you share with us some strategies which you adopted to maintain your academic performance during a 4-day week compared to a 5-day week?
7. Could you describe the support or help you would have received from your faculty, department, and college to support your academic performance following the transition to a 4-day weekday week?
   a. What external help or support have you received?
   b. How effective was the help or support? In what ways did you benefit from the support?
8. Can you tell me how you feel about the 4-day work week in terms of your academic performance? Has it improved or gone down?
9. In your opinion, what are the benefits and drawbacks of having a 4-day work week on your academic performance?
10. How would you describe any changes in your motivation or level of engagement with your coursework or lab sessions since the implementation of the 4-day work week?
11. How do you think the 4-day work week has affected your ability to balance your academic workload with other commitments and responsibilities?
12. Do you think there's anything else I can do to understand your academic performance during a 4-day week better, but I have not asked?

Collection of Data

This qualitative research relied on collecting data directly from students using an exploratory approach, which enabled an in-depth understanding of their perception of the 4-day work week and its impact on their academic performance [20]. Semistructured interviews are a well-established method in qualitative research. Each interview lasted between 20 and 30 minutes. The interviews were recorded after obtaining written informed consent from all participants. Two female team members (RMQ and RJS) who are graduate research assistants with master's degrees of science in physiotherapy received training in qualitative interview techniques from an expert researcher in the research team (JMD) who was experienced in conducting qualitative interviews. The first 4 interviews were conducted in the presence of JMD, after which RMQ and RJS conducted the remaining interviews independently. They wrote field notes to facilitate data analysis. They recorded all interviews and transcribed them verbatim in English. Throughout the interviews, participants were encouraged to ask questions or express doubts to ensure clarity and reduce ambiguity. The research team members were always available to answer any questions from students and provide clarifications where required. All transcripts were sent back to the students to ensure accuracy of the data captured and to ensure member checking. A range of viewpoints was captured, and data saturation was achieved.

Data Analysis

Thematic analysis was used to analyze the data. The interviews were transcribed verbatim by two members of our team (RMQ and RJS). The 6 phases of thematic analysis as defined by Braun and Clarke were followed [21] : (1) familiarizing oneself with the data, (2) generating codes, (3) constructing themes, (4) reviewing potential themes, (5) defining and naming themes, and (6) producing the report. Two investigators (AA and JMD) initially read the transcripts to generate codes and to document where and how patterns occurred. They performed data reduction, and they collapsed data into labels to create categories to facilitate efficient analysis. Six team members (AA, JMS, RMQ, RJS, SN, and SR) were involved in an iterative and analytical process to develop themes from the initial codes. We identified patterns (differences and similarities) in the data by repeatedly reading the transcripts and memos. The use of memos throughout the data analysis promoted reflexivity, precluded a priori biases, and ensured the credibility of our analysis and associated findings. Themes were revised through constant comparison across transcripts, and an audit log was maintained

to document data-driven decisions and ensure transparency. All team members reviewed the data to ensure consistency and coherence between the raw data, the generated codes, and the resulting themes. Representative quotations were used to support the themes, ensuring transparency, and justification of interpretations in the data.

Team Reflexivity

Our research team included members from diverse disciplinary backgrounds, including physiotherapy, nursing, medicine, dental medicine, pharmacy, public health, medical education, and qualitative research. Moreover, the team's experience levels ranged from research assistants and lecturers to senior academics and professors. Our multidisciplinary approach with variations in roles and expertise allowed for a rich interpretation of the data while promoting reflexivity throughout the study. All investigators acknowledged and discussed their own professional perspectives, assumptions, and potential biases prior to data collection and during the thematic analysis process.

To enhance reflexivity, memos were used throughout data analysis to document emerging thoughts, evolving decisions, and reflections. Regular team meetings were held to critically examine how our positionalities and prior experiences might influence coding and theme development. Two research assistants not involved in teaching responsibilities conducted the interviews to mitigate the influence of faculty on students and social desirability bias. We involved two external content experts in reviewing the interview guide, further improving the objectivity of our semistructured interview guide.

Ethical Considerations

The study was performed in line with the principles of the Declaration of Helsinki. Ethical approval was granted by the University of Sharjah Ethics Committee (reference number: REC - 23 - 06 - 13 - 01 - F). Informed consent was obtained from all the participants who were assured of the anonymity of their responses and that their participation in this research would not affect their course grades. Each participant was assigned a random, alphanumeric code to ensure confidentiality. Interview content was analyzed by the research team, and data were securely stored with the primary investigator (AA) on a password-protected computer. The data will be retained for 5 years, after which it will be destroyed.

Results

Overview

In total, there were 24 participants with equal male and female representation in the health professions. All 4 colleges at the medical campus of our university were represented, including medicine, dentistry, pharmacy, and health sciences. The demographics of the study participants are listed in Table 1.

The findings from this study offer a wealth of valuable data. Students' perspectives were grouped into 5 main themes. Each theme was further divided into one or more subthemes. Figure 1 and Multimedia Appendix 1 summarize the themes and subthemes derived from the students' responses.

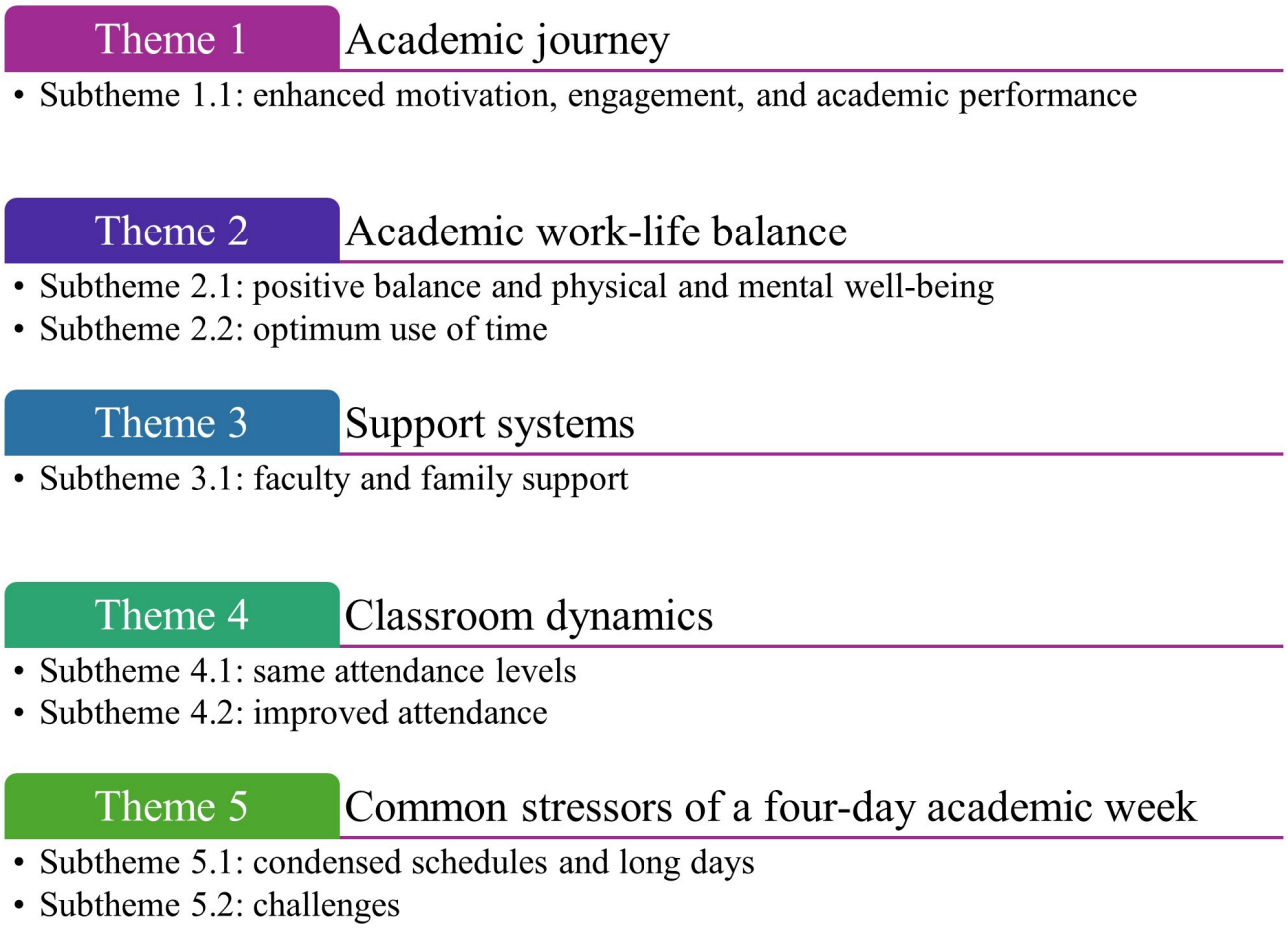
Table . Demographic characteristics of participants.

Participant ID	Age (years)	Sex	College	Program of study
P1	20	F[a]	Health sciences	Medical Diagnostic Imaging
P2	21	F	Health sciences	Nursing
P3	22	M[b]	Health sciences	Nursing
P4	20	M	Health sciences	Physiotherapy
P5	20	M	Health sciences	Physiotherapy
P6	21	F	Health sciences	Physiotherapy
P7	20	M	Medicine	Medicine
P8	20	F	Medicine	Medicine
P9	19	F	Medicine	Medicine
P10	19	F	Medicine	Medicine
P11	20	M	Medicine	Medicine
P12	19	M	Medicine	Medicine
P13	23	M	Dentistry	Dentistry
P14	22	M	Dentistry	Dentistry
P15	23	M	Dentistry	Dentistry
P16	23	M	Dentistry	Dentistry
P17	22	M	Dentistry	Dentistry
P18	22	M	Dentistry	Dentistry
P19	22	F	Dentistry	Dentistry
P20	22	F	Dentistry	Dentistry
P21	22	F	Dentistry	Dentistry
P22	20	F	Pharmacy	Pharmacy
P23	20	F	Pharmacy	Pharmacy
P24	21	F	Pharmacy	Pharmacy

[a]F: female.

[b]M: male.

Figure 1. A summary of themes and subthemes reflecting medicine, dentistry, and health sciences students' perspectives on the impact of the transition from a 5-day to a 4-day week on their academic performance.



Theme 1: Academic Journey

Subtheme 1.1: Enhanced Motivation, Engagement, and Academic Performance

The implementation of a 4-day academic week has notably enhanced students' motivation, engagement, and academic performance, as expressed by many health sciences students. Many students reported a marked increase in motivation, attributing it to having more time to prepare for exams, complete assignments, and organize their studies.

> *Since the implementation of the four-day week, my motivation levels increased. Having an extra weekend day will help you have more time to prepare for upcoming exams and complete assignments on time.* [P6]

One student added that additional time during the weekend improved his grades.

> *My grades got better because I had more time to study.* [P13]

Some students observed that with more time to rest during the extended weekend, they returned to their studies feeling more focused and refreshed. The extra day off also encouraged greater self-reliance for one's own learning, with students feeling more motivated to study independently rather than relying solely on instructors. Furthermore, the 4-day work week not only helped them stay engaged but also made them more productive, providing a sense of balance and convenience that positively impacted their academic efforts.

Theme 2: Academic Work-Life Balance

Subtheme 2.1: Positive Balance and Physical and Mental Well-Being

A significant number of study participants reported that the transition to a 4-day academic week markedly improved their academic work-life balance. Many appreciated that the additional day off allowed them to allocate more time to their studies, extracurricular activities, and personal commitments. Students highlighted that the extra time enabled them to better manage academic workload and responsibilities, while also dedicating time for rest, exercise, family gatherings, and social interactions.

> *As I mentioned, because now I have one extra day, I can enjoy during the weekend, I can distribute the work better, so I have better actual balance between both my studies and extracurricular (activities).* [P5]

The resident students in the dormitory found the extended weekend particularly beneficial for traveling back to their families.

> *So, it's easier to travel to your family outside the country.* [P21]

Some students noted that having an additional day to themselves contributed to better mental health, allowing for reflection,

relaxation, and a reprieve from the daily academic pressures. While the 4-day schedule condensed the lab and lecture sessions, students acknowledged that the 3 days off helped them recharge and maintain a healthy balance between their academic and personal lives. Overall, the shift to the 4-day academic week was seen as highly beneficial in promoting both academic productivity and overall well-being.

Subtheme 2.2: Optimum Usage of Time

The shift to a 4-day academic week had a positive impact on the students' ability to manage and optimize their time. Many expressed that under the previous 5-day schedule, there was insufficient time to complete coursework, particularly clinical documentation, which often had to be completed at home. With the additional day off, students found they had more time to focus on assignments, manage their academic workload, and balance extracurricular activities. An extra weekend day allowed for better organization of tasks, leading to improved time management and more efficient studying.

> *Having the three-day weekend helped me organize my time better.* [P7]
>
> *Having an extra weekend day allowed me to balance both workload and commitments, helping me manage my time better.* [P11]

In addition, students noted that the 3-day weekend gave them more energy and focus during the week, as they had more time to rest and recharge. Some also emphasized that the extra day helped them achieve a better balance between academic responsibilities and personal commitments, including spending time with family, which had been challenging with the previous schedule. In summary, a considerable proportion of the university students noted a major improvement in terms of time management and productivity after shifting to the 4-day academic week.

Theme 3: Support Systems

Subtheme 3.1: Faculty and Family Support

Health care students expressed a strong sense of support from both faculty and family, which played a pivotal role in their academic success and well-being. Many students highlighted the encouragement they received from specific faculty members who were not only understanding but also flexible in accommodating their needs. Whether it was adjusting lab timings, rescheduling exams, or providing additional resources such as opening labs on weekends, faculty demonstrated a commitment to helping students balance their academic and personal responsibilities. One student mentioned how faculty members worked around their professional commitments, enabling them to complete a movie shoot while staying on track academically.

> *Yes, from Doctor xxxx (a faculty member's name) and Doctor xxxx (another faculty member's name) as well; basically, in one of the semesters, I had Netflix shoot, and I was going to be absent from university for two months in a row. I talked with Doctor xxxx (a faculty member's name) as she's my advisor and she helped me out, also Doctor xxxx (another faculty member's name) helped me and they made it possible for me to do my shooting, to do my work and to attend my makeup (exams) better.* [P3]

Moreover, the emotional and motivational support from family and friends further empowered students to stay focused and motivated, allowing them to better organize their priorities and excel in their studies. Based on the perceptions of the 4-day academic week, it can be inferred that the combined support from faculty, family, and peers significantly enhanced the students' academic experience, making the 4-day academic week more manageable and effective whilst attending to personal responsibility outside of academia.

Theme 4: Classroom Dynamics

Subtheme 4.1: Same Attendance Levels

Nearly 70% of the participants reported that their attendance levels did not alter in the shift from 5-day week to 4-day academic week.

Subtheme 4.2: Improved Attendance

Around 25% of the study respondents recorded that their attendance got better, and they have started attending all the classes with the 4-day academic week, while before they used to stay absent or attend the maximum number of required classes and clinical practicums.

Theme 5: Common Stressors of a 4-Day Academic Week

Subtheme 5.1: Condensed Schedules and Long Days

The extended working hours and condensed schedules brought about by the shift to a 4-day academic week were one of the major concerns raised by the health care students. Many found that the teaching content was compressed into fewer days, leaving minimal time for breaks or adequate reflection.

> *The negatives of having four days are that the materials are compressed into these four days, leaving minimal time.* [P1]
>
> *But on the other side, the drawbacks for having 4 days is (are) that we usually have a lot of materials that's compressed on us in these four days that we have minimal time.* [P5]

Some students noted that their weekdays became more stressful, with longer hours and more back-to-back lectures, making it difficult to manage their workload effectively. The compact schedule, though allowing for a longer weekend, resulted in longer days, leading to increased fatigue and diminished focus during lectures.

> *The days became longer and compact. It became more stressful because the days became longer.* [P4]

Despite these challenges, some students accepted the trade-off, stating that they valued the extra day off on weekends, even though the compressed weekdays required more intense time management.

> *The drawback is for example, nowadays the classes are many; let's say for example today is Monday, so I started at 8:00 (am) and I finished at 4.00 (pm), but*

> *before I used to start at 10 and then I would finish that too (at 4.00 pm). We have more time to divide the classes, but for me personally, and I think most people will agree with me, it's fine to have everything compacted on the weekdays as long as on the weekend we have an extra day off.* [P3]

Therefore, the main drawback of the 4-day academic week appeared to be the increased workload on each individual day, which students found challenging to manage, despite the perceived benefits of extended weekends for relaxation and study.

Subtheme 5.2: Challenges

While the 4-day academic week was beneficial to the health care students, many of them also voiced several challenges associated with this model. Some noted that the longer weekend led to a tendency to neglect studies during the break, resulting in lapses in focus or forgotten material.

> *…. but the drawbacks that we may forget things we may like that we will not study in these three days and sometimes it happened.* [P1]

The compacted schedule caused others to arrive late to lectures or struggle with back-to-back classes, which they found overwhelming.

> *I'm arriving late to the lectures by around 15 minutes because of the back-to-back classes.* [P2]

Several students reported a decline in their grades, attributing it to increased stress and the tutors' inability to cover all the material during class time, leaving more for self-study.

> *My grades got decreased due to stress.* [P9]
>
> *but for the negative impact, as I mentioned, the tutors are unable to finish materials. So, there are a lot (of) self-study material.* [P10]

Balancing personal responsibilities, family time, and academic workloads became harder for many, as they found it difficult to manage their time effectively within the condensed 4-day academic week. Though the extended weekend was considered a benefit, it often came at the cost of a worsened balance between academic and personal responsibilities.

Discussion

Principal Findings

This study revealed that the shift to a 4-day academic week from a 5-day week had a significant impact on various aspects of our students' academic and personal lives. We found that the 4-day academic week had a positive impact on most of our student participants, enhancing their motivation, engagement, academic work-life balance, and physical and mental well-being. Some students received positive support from faculty and family. However, a few students encountered increased stress and struggled to balance academic and personal commitments with this model.

Most participants reported increased motivation, focus, and engagement in both lectures and laboratory sessions, which positively impacted their academic performance. These findings align with the study by Gaballah et al [13] where most students acknowledged improvements in academic performance with the 4-day academic week, although one-third did not report such benefits.

The shift to a 4-day academic week improved students' academic work-life balance, allowing them to allocate more time to rest, personal commitments (with family and friends), and extracurricular activities. The extra day off provided an opportunity for reflection and relaxation, which contributed to better mental health and well-being. These findings agree with Gaballah et al [13], who reported that around 90% (257 out of 284) of their dental student participants reported stress relief and spending quality time with friends or families, in addition to having extra time for studies and exam preparation because of the long (3-day) weekend [13]. Dormitory students particularly appreciated the reduced need to travel frequently (considering an extra day off), which decreased physical exhaustion. Many participants emphasized that they could now distribute their academic workload more evenly, avoiding burnout, which is supported by the theory of self-regulated learning**,** which focuses on students' ability to manage their own learning processes [22]. The improved balance between academic and personal responsibilities not only boosts productivity but also helps students maintain a healthier lifestyle [23]. However, managing this balance required effective time management and self-care to avoid procrastination during the longer weekends, as some students feel that a 4-day academic week has led to academic overload [24-26]. Academic overload may be regarded as students' feelings of being overwhelmed by their academic requirements or responsibilities while pursuing a degree at university.

While the students enjoyed a 4-day academic week with the exemption of commuting time on an extra day, the long workday and compact schedule took its toll on some students. Entry-level students in higher education have been reported to experience difficulties in managing the academic workload at university [27]. Chambel et al [28] found that students' inability to manage academic workload had a negative impact on academic adjustment to university and academic performance. Some students were happy with the support received from faculty and family, which would have played a pivotal role in helping students adapt to the 4-day academic week. Indeed, some faculty members were flexible in accommodating students' needs, offering extra lab time or adjusting deadlines to help them balance their academic and personal responsibilities. The findings show that faculty members played a key role in providing cognitive flexibility by helping students adjust to the new schedule and commitments while maintaining the professorial standards [29]. However, one-third of the dental students, from the previous survey conducted at our institution, reported limited time to meet their professors and advisors, while the rest expressed satisfaction with the support provided by faculty [13].

Many students appreciated the flexibility offered, such as optional assignments and extended lab hours. These arrangements could be beneficial for students who had other professional obligations to improve their academic achievement [30,31]. Emotional support from family also encouraged students

to stay focused and motivated. With this support system in place, students felt more confident in managing the demands of their studies. The combined support from faculty, family, and friends might have alleviated much of the stress associated with the compressed schedule, helping students navigate challenges more effectively.

Despite some students receiving favorable support, a few participants reported a lack of adequate support from faculty to facilitate their transition to this new framework. Universities using a 4-day academic week model need to improve manual (eg, student advisors, counselors) and digital support systems (through a dedicated service desk), amongst others, to help mitigate or alleviate the stressors and enhance the academic performance of students adjusting to this transition from a 5-day week.

For most students, attendance levels remained unchanged after the shift to a 4-day academic week, indicating that the change did not negatively affect class participation. All faculty members use an online attendance tracking system at our institution, which issues automatic warnings when students reach 10%, 15%, and 20% absence thresholds. Those exceeding the 20% absence limit are prohibited from attending exams, ensuring consistent attendance and academic engagement. However, a minority of students reported improved attendance, attributing this to increased motivation and better preparation. The extra day off might allow students to rest and recharge, which would have made them more consistent in attending classes [32]. This improvement in attendance also reflects a stronger commitment to academic schedules and responsibilities [33]. However, for those whose attendance remained the same, the extra day was more of a personal benefit than an academic one.

While the 4-day academic week offered clear benefits, many students found the condensed schedule challenging, especially in the clinical rotations, which is in line with the findings of the earlier study [34]. The longer days, with back-to-back lectures and lab sessions, left little time for breaks or reflection. This intensified workload increased stress and fatigue, making it harder for students to stay focused during classes. The compressed schedule might pose challenges to faculty in covering all the materials included in their course syllabi, and some students pointed out their reliance on self-study in this regard.

A few participants felt a lack of extra support from faculty to adjust to this transition to the 4-day academic week. Despite these difficulties, many students accepted the trade-off, valuing the extended weekend for relaxation and personal commitments. The challenge lay in balancing the intensified academic schedule with the need for rest and fulfilling other (personal) commitments [35]. The mixed findings of our study provide a nuanced understanding of the 4-day academic week's benefits and drawbacks, particularly in the context of higher education in health care in Sharjah, United Arab Emirates.

Academic overload could lead to lower academic adjustment among university students. This means that those students who feel overwhelmed by their daily academic requirements and responsibilities will have lower academic adjustment. This result is in accordance with previous research by Bitzer et al [27]. Although we did not have access to the grades, this may be a plausible explanation. Students who feel confident about their skills, who are confident about their academic and learning capabilities, and who have a positive attitude or perception toward their abilities will perform better academically and adapt to the academic demands of the university. Moreover, a previous study indicated a positive relationship between self-efficacy and academic adjustment [36].

The more time students allocate toward studying, the higher their level of self-efficacy and the better their academic adjustment. In general, the 3-day extended weekend appeared to be conducive to learning.

Strengths and Limitations of the Study

A variety of students from the different colleges within the medical campus has resulted in rich data, and they provided interesting perspectives on the impact of the 4-day academic week on academic performance across the different colleges. In addition, the study followed a robust protocol to reduce bias and enhance credibility of the findings and adhered to the COREQ guidelines (Consolidated criteria for Reporting Qualitative research) checklist to improve transparency in reporting.

A limitation of the study is that we had included medicine, dentistry, and health science students only from the University of Sharjah, which may not be generalizable to students studying different programs (eg, engineering, law, arts, and science) and other universities and in the other emirates within the United Arab Emirates. However, at present, the Emirate of Sharjah is the only emirate that has implemented the 4-day academic week for universities.

Implications and Future Recommendations

The findings clearly illustrate the nature in which the transition to the 4-day week in the United Arab Emirates has occurred, along with the positives and negative side. The themes and trends identified in this work will inform and guide current and future research studies, thereby broadening the scope of our work. Future studies are needed to test whether the identified themes are common in the emirate of Sharjah and beyond.

Given the mixed results, a need for a study skills program to accompany the transition to the 4-day work week is required to mitigate the issues encountered by some students. This would include prioritizing and scheduling academic activities by creating a weekly schedule that allocates dedicated time to both academic and personal activities, along with prioritizing tasks and setting realistic goals for every identified task. Students need to manage time efficiently by breaking tasks into smaller, manageable segments and using tools like time-blocking and to-do lists. These strategies can help students stay organized and in control of both academic and personal lives. Students would need to clearly define boundaries between academic and personal life, which is an area of life skills for the future. Moreover, students can practice self-care for themselves, like meditation, exercise, and hobbies that rejuvenate them mentally and physically. Students should be able to seek help from the faculty and family whenever required to accommodate and balance their expectations and needs in the best way possible.

Students need to reflect and adjust as necessary, as their academic demands and priorities may change over time.

This study offers recommendations, including that the 4-day academic week, while working in the favor of most university students, needs to have a supportive mechanism. One such support system is an effective advisory system so that students (advisees) can seek help with study skills and time management from their faculty advisors or mentors to mitigate the drawbacks of the extended hours during the 4-day academic week. Our university already has such an advising system (with each student assigned to a faculty advisor), a disability resource center, and counselors in place to support all students, including but not limited to those at risk of academic probation because of their poor performance in studies, to thrive with the current framework. Moreover, our university has introduced peer tutoring activities and is planning peer advising initiatives to further facilitate this transition from a 5-day to a 4-day academic week in the emirate of Sharjah, United Arab Emirates. Further studies involving students from diverse academic disciplines (eg, engineering, arts, literature, science, and law) and from other universities in the United Arab Emirates that have adopted the 4-day academic week are required to generate wider insights and provide implications for policy.

Conclusion

Overall, students studying medicine, dentistry, and health sciences programs perceived that the 4-day academic week had a positive impact by enhancing their motivation, engagement, class attendance, and work-life balance, along with the support of faculty and family. However, a few students encountered increased stress and struggled to balance academic and personal commitments with this model. Universities using a 4-day week model need to improve manual (eg, student advisors and counselors) and digital support systems (through a dedicated service desk), amongst others, to help mitigate or alleviate the stressors and enhance the academic performance of students adjusting to this transition from a 5-day week to a 4-day week.

Acknowledgments

No funding was received for this study.

Data Availability

All data generated or analyzed during this study are included in this published article.

Authors' Contributions

Conceptualization: AA, BMM, BS
Methodology: AA, JMD
Supervision: AA, JMD
Investigation: AA, JMD, RMQ, RJS
Data Curation: AA, JMD, RMQ, RJS, TAMO, SN, SR
Formal Analysis: AA, JMD, RMQ, RJS, TAMO, SN, SR
Validation: AA, JMD, SN, SR
Visualization: AA, JMD, TAMO, SN, SR
Writing – Original Draft: AA, JMD, RMQ, RJS, TAMO, SN, SR
Writing – Review & Editing: AA, JMD, SN, SR, BMM, BS, NRAL
Project Administration: AA, JMD
Resources: AA, JMD

Conflicts of Interest

None declared.

Multimedia Appendix 1
A summary of themes, subthemes, categories, and quotes supporting medical and health sciences students' perception on the impact of a 4-day academic week on their academic performance.
[DOCX File, 26 KB - mededu_v11i1e67775_app1.docx]

References

1. Rodriguez A. Unravelling the 4-day week: a comprehensive study of its implementation and global impacts, with a focus on italy. : POLITesi - Archivio digitale delle tesi di laurea e di dottorato; 2022.
2. Sng M, Khor WJ, Oide T, Suchar SC, Tan BCK. Effectiveness of a four-days/eight hour work week. : Embry-Riddle Aeronautical University Asia; 2021.
3. Iacobucci G. NHS should trial four day week to tackle burnout and improve retention, says report. BMJ 2023:779. [doi: 10.1136/bmj.p779]

4. Hewitt PM, Denny GS. The four-day school week: impact on student academic performance. The Rural Educator 2011(2):23-31. [doi: 10.35608/ruraled.v32i2.431]
5. Hunt R. A fusion high school program. Am Sch Board J 1936;93(5):66.
6. Thompson PN, Ward J. Only a matter of time? The role of time in school on four-day school week achievement impacts. Econ Educ Rev 2022 Feb;86:102198. [doi: 10.1016/j.econedurev.2021.102198]
7. Sagness RL, Salzman SA. Evaluation of the four-day school week in idaho suburban schools. 1993 Presented at: Northern Rocky Mountain Educational Research Association; Oct 1-2, 1993; Jackson, WY URL: http://files.eric.ed.gov/fulltext/ED362995.pdf [accessed 2025-10-01]
8. Anglum JC, Park A. Keeping up with the Joneses: district adoption of the 4-day school week in rural Missouri. AERA Open 2021 Jan;7:23328584211002840. [doi: 10.1177/23328584211002842]
9. Thompson PN. Does a day lost equal dollars saved? The effects of four-day school weeks on school district expenditures. Natl Tax J 2021 Mar 1;74(1):147-183. [doi: 10.1086/712916]
10. Anderson DM, Walker MB. Does shortening the school week impact student performance? Evidence from the four-day school week. Educ Finance Policy 2015 Jul;10(3):314-349. [doi: 10.1162/EDFP_a_00165]
11. Ivancevich JM. Effects of the shorter workweek on selected satisfaction and performance measures. Journal of Applied Psychology 1974;59(6):717-721. [doi: 10.1037/h0037504]
12. Tesorero A. Sharjah: 3-day weekend increases student attendance by 95%; teachers say work-life balance improved. Khaleej Times. URL: https://www.khaleejtimes.com/lifestyle/sharjah-3-day-weekend-increases-student-attendance-by-95-teachers-say-work-life-balance-improved [accessed 2024-10-17]
13. Gaballah K, El Kishawi M, Ibrahim E, Al Kawas S. Impact of a shorter university week on academic performance and study–life balance of dental students. Adv Biomed Health Sci 2024;3(3):111-117. [doi: 10.4103/abhs.abhs_5_24]
14. Nayak B, Bhattacharyya SS, Krishnamoorthy B. Application of digital technologies in health insurance for social good of bottom of pyramid customers in India. IJSSP 2019 Sep 9;39(9/10):752-772. [doi: 10.1108/IJSSP-05-2019-0095]
15. Booth A, Hannes K, Harden A, Noyes J, Harris J, Tong A. COREQ (consolidated criteria for reporting qualitative studies). In: Guidelines for Reporting Health Research: A User's Manual 2014:214-226. [doi: 10.1002/9781118715598]
16. Thompson PN. Is four less than five? Effects of four-day school weeks on student achievement in Oregon. J Public Econ 2021 Jan;193:104308. [doi: 10.1016/j.jpubeco.2020.104308]
17. Jain MJ, Chouliara N, Blake H. From five to four: examining employee perspectives towards the four-day workweek. Adm Sci 2025 Mar;15(3):114. [doi: 10.3390/admsci15030114]
18. Israel W, Mulitauopele C, Ma M, Levinson AH, Cikara L, Brooks-Russell A. Adolescent health behaviors in schools with 4- versus 5-day school weeks. J Sch Health 2020 Oct;90(10):794-801. [doi: 10.1111/josh.12941] [Medline: 32812223]
19. Bowles A, Fisher R, McPhail R, Rosenstreich D, Dobson A. Staying the distance: students' perceptions of enablers of transition to higher education. High Educ Res Dev 2014 Mar 4;33(2):212-225. [doi: 10.1080/07294360.2013.832157]
20. Chapman AL, Hadfield M, Chapman CJ. Qualitative research in healthcare: an introduction to grounded theory using thematic analysis. J R Coll Physicians Edinb 2015;45(3):201-205. [doi: 10.4997/JRCPE.2015.305] [Medline: 26517098]
21. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qualitative Research in Sport, Exercise and Health 2019 Aug 8;11(4):589-597. [doi: 10.1080/2159676X.2019.1628806]
22. Panadero E. A review of self-regulated learning: six models and four directions for research. Front Psychol 2017;8:422. [doi: 10.3389/fpsyg.2017.00422] [Medline: 28503157]
23. Abdelmoteleb AAS, Robby M, Purinton T, Hamdan S, Bourchak M. Supporting schools, educators, students, and families in a transition to a four day week. Presented at: The Asian Conference on Education 2023; Nov 22-25, 2023; Tokyo, Japan. [doi: 10.22492/issn.2186-5892.2024.16]
24. Macan TH, Shahani C, Dipboye RL, Phillips AP. College students' time management: correlations with academic performance and stress. J Educ Psychol 1990;82(4):760-768. [doi: 10.1037//0022-0663.82.4.760]
25. Pedersen M, Muhr SL, Dunne S. The care of the self and the meaningful four-day workweek. Philosophy of Management 2024 Sep;23(3):335-352. [doi: 10.1007/s40926-024-00314-2]
26. Ulriksen L, Nejrup C. Balancing time – university students' study practices and policy perceptions of time. Sociol Res Online 2021 Mar;26(1):166-184. [doi: 10.1177/1360780420957036]
27. Bitzer E, Bruin CTD. The effect of factors related to prior schooling on student persistence in higher education. S Afr J Educ 2004;24(2):119-125 [FREE Full text] [doi: 10.10520/EJC31989]
28. Chambel MJ, Curral L. Stress in academic life: work characteristics as predictors of student well‐being and performance. Applied Psychology 2005 Jan;54(1):135-147. [doi: 10.1111/j.1464-0597.2005.00200.x]
29. Alsaif B, Hassan SUN, Alzain MA, Almishaal AA, Zahra A. Cognitive flexibility's role in reducing academic stress during the COVID-19 pandemic. Psychol Res Behav Manag 2024;17:457-466. [doi: 10.2147/PRBM.S451211] [Medline: 38371712]
30. Toraman Ç, Özdemir HF, Koşan AMA, et al. Relationships between cognitive flexibility, perceived quality of faculty life, learning approaches, and academic achievement. Int J Instruction 2020;13(1):85-100. [doi: 10.29333/iji.2020.1316a]
31. VanWeelden H. School schedules and their impact on teacher job satisfaction. : Digital Collections of Dordt University; 2021.

32. Arsha DZ, Isse B, Moronkeji A. The viability of four-day workweek in the education sector: evaluating managerial views, stakeholder opinions, and the pursuit of work-life balance. : Malmö University; 2024.
33. Bergin J, Ferrara L. How student attendance can improve institutional outcomes. : Macmillan Learning; 2019.
34. Exploring the lived experience of osteopathy students making the transition from working in student clinic (8 hours per week) to working full-time (40 hours per week) during the summer holiday break. : Unitec Institute of Technology; 2019 URL: https://www.researchbank.ac.nz/server/api/core/bitstreams/aab1dcf1-73ec-48f4-86df-9da1f0cbb78a/content
35. Campbell TT. The four-day work week: a chronological, systematic review of the academic literature. Manag Rev Q 2024 Sep;74(3):1791-1807. [doi: 10.1007/s11301-023-00347-3]
36. Yadak SMA. The impact of the perceived self-efficacy on the academic adjustment among Qassim University undergraduates. JSS 2017;05(1):157-174. [doi: 10.4236/jss.2017.51012]

Abbreviations

COREQ : Consolidated Criteria for Reporting Qualitative study

Edited by B Lesselroth; submitted 21.10.24; peer-reviewed by A Davies, LO Alsoud, SH Sung, VF Hanson; revised version received 24.06.25; accepted 23.09.25; published 04.11.25.

Please cite as:
Arumugam A, Dias JM, Narasimhan S, Qadah RM, Shalash RJ, Omran TAM, Mussa BM, Saddik B, Al Mazrouei NR, Ramakrishnan S
Balancing Academics and Life: Qualitative Study of Health Professions Students' Perceptions of a Four-Day Academic Week in the United Arab Emirates
JMIR Med Educ 2025;11:e67775
URL: https://mededu.jmir.org/2025/1/e67775
doi:10.2196/67775

Evaluating Tailored Learning Experiences in Emergency Residency Training Through a Comparative Analysis of Mobile-Based Programs Versus Paper- and Web-Based Approaches: Feasibility Cross-Sectional Questionnaire Study

Hsin-Ling Chen[1*], MD, MS; Chen-Wei Lee[2,3*], MD, MS; Chia-Wen Chang[1,4], BS; Yi-Ching Chiu[1], BS; Tzu-Yao Hung[1,5], MD, PhD

[1]Department of Emergency Medicine, Zhong-Xing branch, Taipei City Hospital, 145 Zhengzhou Road, Taipei, Taiwan

[2]Department of Emergency, Dalin Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, Chiayi County, Taiwan

[3]School of Medicine, Tzu Chi University, Hualien County, Taiwan

[4]Department of Surgery, Zhong-Xing branch, Taipei City Hospital, Taipei, Taiwan

[5]Faculty of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan

[*]these authors contributed equally

Corresponding Author:
Tzu-Yao Hung, MD, PhD
Department of Emergency Medicine, Zhong-Xing branch, Taipei City Hospital, 145 Zhengzhou Road, Taipei, Taiwan

Abstract

Background: In the rapidly changing realm of medical education, Competency-Based Medical Education is emerging as a crucial framework to ensure residents acquire essential competencies efficiently. The advent of mobile-based platforms is seen as a pivotal shift from traditional educational methods, offering more dynamic and accessible learning options. This research aims to evaluate the effectiveness of mobile-based apps in emergency residency programs compared with the traditional paper- and web-based formats. Specifically, it focuses on analyzing their roles in facilitating immediate feedback, tracking educational progress, and personalizing the learning journey to meet the unique needs of each resident.

Objective: This study aimed to compare mobile-based emergency residency training programs with paper- and web-based (programs regarding competency-based medical education core elements.

Methods: A cross-sectional web-based survey (Nov 2022-Jan 2023) across 23 Taiwanese emergency residency sites used stratified random sampling, yielding 74 valid responses (49 educators, 16 residents, and 9 Residency Review Committee hosts). Data were analyzed using Mann-Whitney *U* test, chi-squared tests, and *t* tests.

Results: MB programs (n=14) had fewer missed assessments (P=.02) and greater ease in identifying performance trends (P<.001) and required clinical scenarios (P<.001) compared with paper- and web-based programs (n=60). In addition, mobile-based programs enabled real-time visualization of performance trends and completion rates, facilitating individualized training (P<.001).

Conclusions: In our nationwide pilot study, we observed that the mobile-based interface significantly enhances emergency residency training. It accomplishes this by providing rapid, customized updates, thereby increasing satisfaction and autonomous motivation among participants. This method is markedly different from traditional paper- or web-based approaches, which tend to be slower and less responsive. This difference is particularly evident in settings with limited resources. The mobile-based interface is a crucial tool in modernizing training, as it improves efficiency, boosts engagement, and facilitates collaboration. It plays an essential role in advancing Competency-Based Medical Education, especially concerning tailored learning experiences.

(JMIR Med Educ 2025;11:e57216) doi:10.2196/57216

KEYWORDS

app; mobile; web-based; competency-based medical education; residency training program

Introduction

The competency-based medical education (CBME) model has been a global trend in residency training for over a decade [1-4], with timely, personalized, and meaningful coaching feedback as one of its core elements [4]. At the same time, advances in mobile technology are reshaping both medical education and clinical practice [5-11]. Mobile learning has emerged as a cost-effective, accessible approach that supports context-driven, real-time learning and continuous feedback—despite challenges

like technical limitations and potential distractions [5]. A national survey indicates that tablet use, predominantly iPads, is on the rise, with strong support for further integration to enhance clinical efficiency, particularly among younger clinicians [8]. In addition, guidelines for mobile technologies in workplace-based assessments show that these devices can streamline real-time data capture, reduce administrative burdens, and facilitate competency-based decision-making through intuitive interfaces, robust security, and comprehensive training [12].

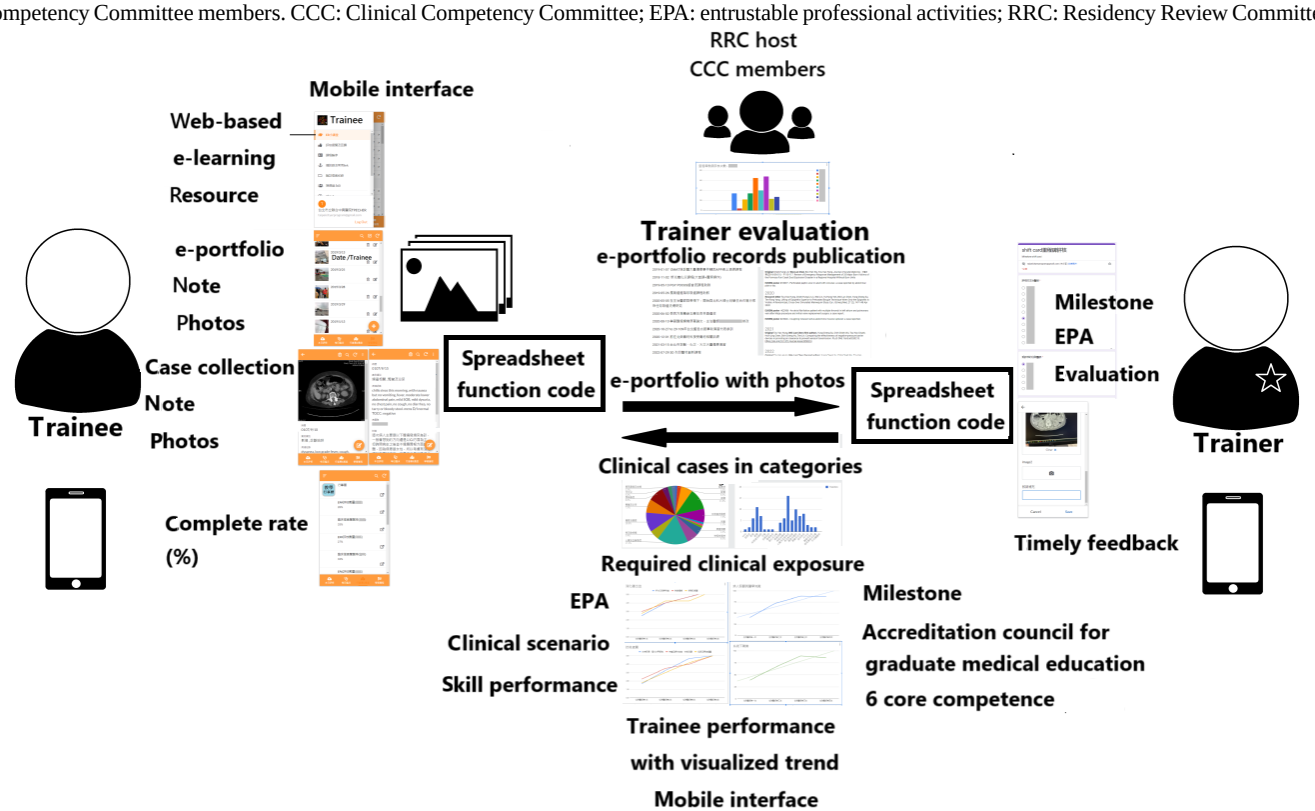
A residency training program is designed to help residents manage patients effectively across diverse scenarios while continuously improving their skills. The goal is to foster intrinsic motivation and enhance performance by creating an environment that supports autonomy, competence, and relatedness [13].

Repeated, multisource evaluations provide a more reliable trainee assessment than one-off reviews [14,15]. Oudkerk Pool et al [14] found that multiple evaluators help mitigate biases by iteratively acquiring, organizing, and integrating evidence. As CBME emphasizes longitudinal performance tracking, effective data management is crucial. Leveraging big data enables objective, evidence-based promotion decisions, reducing reliance on subjective faculty recall [16].

However, performance evaluation demands extensive documentation, data collection, and analysis, which can be time-consuming and resource-intensive, potentially affecting the quality of feedback. In addition, interpreting performance trends is complex and must generate meaningful insights for program committees to tailor and individualize training. For resource-limited programs, the infrastructure required to integrate evaluation interfaces and visualize performance trends across trainers, trainees, and program committees poses a significant challenge [5,12,16].

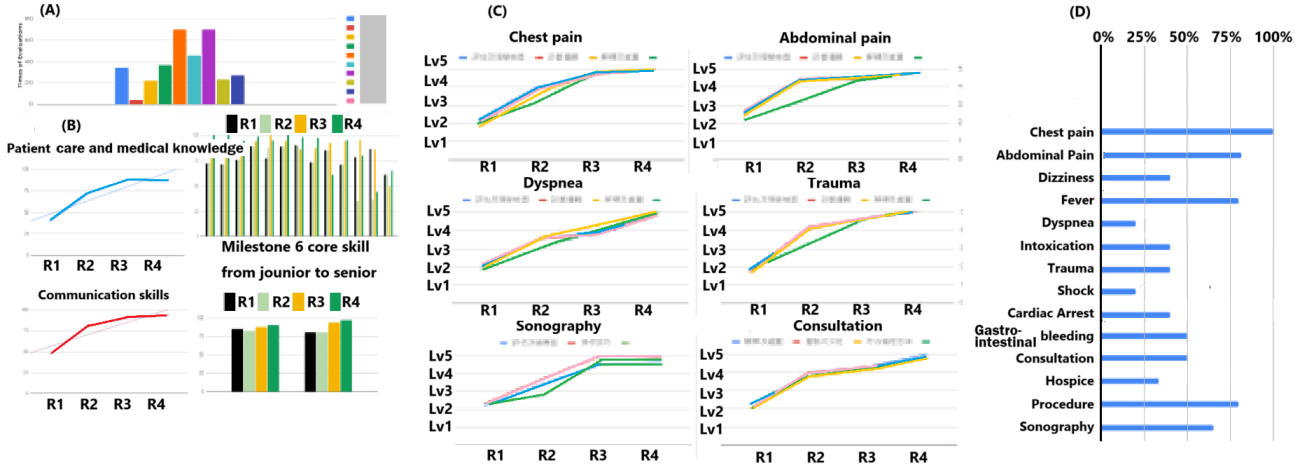
Although mobile apps for medical learning have advanced significantly [5-7,9,17,18], assessment in training programs has largely relied on paper- or web-based systems [19]. However, computers and laptops can be cumbersome and require stable internet access, which is not always available. In contrast, mobile interfaces on smartphones and tablets offer a more accessible and user-friendly alternative without compromising content quality (Figure 1) [5-9,12]. These platforms facilitate frequent performance monitoring and review, promoting continuous improvement in alignment with evolving educational and professional needs [12].

Figure 1. The illustration of the mobile-based interface is shared among trainers, trainees, the Residency Review Committee host, and Clinical Competency Committee members. CCC: Clinical Competency Committee; EPA: entrustable professional activities; RRC: Residency Review Committee.



Ideally, residents and faculty can exchange real-time feedback, allowing both groups to track performance trends and refine learning and coaching strategies. Rather than relying on one-time scores, the Residency Review Committee (RRC) and Clinical Competency Committee (CCC) can assess performance through visual trend analyses (Figure 2), enabling personalized program adjustments and targeted interventions for underperformance [12]. This framework consolidates evaluations from multiple sources, transforming data into motivational insights for trainees and actionable trends for program directors. In addition, visualizing progressive percentages helps trainees understand their clinical diversity requirements and ensures they receive adequate competency evaluations throughout emergency residency training.

Figure 2. (A) The assessment distribution by the trainer is based on one training site with a mobile-based assessment system. The trainee's milestones are assessed weekly, and these data can be accumulated and transformed into visualized trends quarterly and annually to track progress. (B) Entrustable professional activities, with scenarios characterized by varying levels of emergency, can be accumulated and transformed into visualized trends quarterly and annually to gauge progress. (C) The required assessments of entrustable professional activities can be converted into a completion rate to aid trainers in more comprehensive assessments of trainees. EPA: entrustable professional activity.



Workplace-based assessments are a cornerstone of CBME, particularly in emergency residency training, where timely feedback and performance tracking are essential [12]. While paper-based and some web-based platforms (eg, Google Forms and SurveyCake) facilitate structured evaluations, they often present challenges in real-time data integration, accessibility, and administrative burden for trainers and trainees.

Despite advancements in medical education technology, our preliminary data from Taiwan suggest that mobile interface assessments remain significantly underused compared with web-based alternatives, with an approximate ratio of 1:4-1:5. This disparity may reflect institutional technological barriers, faculty adoption challenges, or security and interoperability concerns, all of which hinder the seamless aggregation and visualization of multisource evaluation data. While web-based platforms are widely implemented, the potential of mobile-based platforms to enhance real-time feedback and streamline competency tracking in residency training remains underexplored [16,17].

This pilot study aims to determine whether mobile-based platforms can address these challenges by improving real-time performance monitoring and optimizing feedback mechanisms in emergency residency training. By evaluating user adoption, data integration efficiency, and feedback effectiveness, this study seeks to provide insights into the feasibility and impact of mobile technology in workplace-based medical assessments.

Methods

Overview

A cross-sectional, web-based survey was conducted from November 3, 2022, to January 3, 2023, to explore the perspectives of educators and resident physicians on assessment platforms in emergency residency training. As a preliminary investigation, this study aimed to assess the feasibility of mobile-based assessments, refine survey methodology, and identify key themes for future large-scale research.

We calculated the sample size using a *t* test for 2 independent proportions. Given that mobile-based programs constitute approximately one-fourth of RRC programs, we based our estimation on a 1:4 ratio. Assuming a 1-point difference in evaluations regarding training individualization, we determined that a minimum of 11 mobile-based participants and 44 paper- and web-based participants would be needed to achieve 80% power at α=.05 [20].

The survey targeted program hosts, trainers, and trainees across emergency residency training sites in Taiwan, inviting them to participate anonymously (Multimedia Appendix 1). A stratified random sampling method was applied, proportionally selecting participants from accredited emergency residency training hospitals as recognized by the National Emergency Medicine Association. From a total of 1068 qualified educators and 313 resident physicians, a 5% sampling rate was used, resulting in invitations sent to 55 educators and 20 resident physicians. In addition, to gather insights on the perspectives of RRC hosts, 10 were invited to participate in the survey [20].

This study evaluates assessment designs used across different emergency residency training sites. Participants from diverse programs were categorized into paper- and web-based or mobile-based evaluation methods for detailed analysis. Paper- and web-based programs were grouped together due to shared limitations, such as delayed feedback, lack of real-time performance tracking, and reliance on periodic evaluations rather than continuous competency monitoring. In contrast, mobile-based programs provided immediate feedback and seamless trainee progress tracking, distinguishing them as a separate category.

The questionnaire was developed using the Delphi method, ensuring expert consensus on its content and structure. It included a background survey capturing details, such as the name and level of the training site, the number of trainers and trainees, and the participant's role in the training program. The core survey items consisted of Likert-scale questions (5-point scale, ranging from "Strongly Disagree" to "Strongly Agree") designed to assess perceptions of CBME core elements. These

responses were analyzed and compared to evaluate differences in assessment approaches. To ensure the integrity and reliability of the collected data, only fully completed surveys were included in the final analysis. Incomplete responses were excluded to maintain data accuracy. In addition, responses underwent internal consistency checks to detect and exclude any contradictory or inconsistent answers.

The questionnaire was distributed on the web, with participants invited via email to complete it voluntarily and anonymously. No monetary incentives were provided; however, participants were informed that their input would contribute to improving workplace-based assessments in emergency residency training. While hosts offered additional insights, their responses were categorized under trainers for analysis, as their primary role aligned with faculty responsibilities.

Statistical Analysis

Analyses were conducted using IBM SPSS Statistics (version 23). Dependent variables included assessment satisfaction, assessment and feedback duration, frequency of missed assessments, ability to identify performance trends, ability to review and respond within 24 hours, ability to individualize training programs based on performance results, and ability to track the completion rate of required clinical scenarios and assessments.

Independent variables included the training site level (district, regional, or center hospitals), participants' age, participant role (trainee or trainer), and assessment platform (mobile-based or paper- and web-based).

The Mann-Whitney *U* test was used to compare differences in dependent variables across training sites and participant roles. A linear regression model with repeated measurements at the hospital level was applied to evaluate the ability to individualize training programs based on performance results, adjusting for potential confounders, including the number of trainers and trainees. The chi-squared test was used to analyze differences in assessment platform performance, with statistical significance set at $P<.05$. The chi-square test was used to analyze differences in assessment platform performance, with statistical significance set at $P<.05$.

Ethical Considerations

This pilot study received ethical approval from the Taipei City Hospital institutional review board (TCHIRB-11110007-E). The study was conducted after being approved by the Taipei City Hospital institutional review board on November 24, 2022. The survey was anonymous. The informed consent was waived by the Taipei City Hospital institutional review board.

Results

Demographics

In this study, 62.22% of emergency residency training sites nationwide participated, amounting to 28 out of 45 sites. A total of 74 valid responses were collected from 28 emergency residency training sites across Taiwan, including district, regional, and center hospitals. Among the respondents, 49 trainers participated, reflecting an 89.1% response rate, while 16 resident trainees responded, yielding an 80% response rate. To further explore the perspectives of RRC hosts, 10 were invited to participate, with 9 providing responses. Regarding assessment platform usage, 14 participants reported using mobile-based assessments, whereas the remaining 60 relied on either paper- or web-based platforms. Within this group, 3 participants used paper-based assessments, and 57 used web-based systems.

Survey Validity Results

The questionnaire demonstrated strong reliability for trainee responses, with a Cronbach α of 0.83, indicating high internal consistency. For trainers, the reliability was moderate, with a Cronbach α of 0.61, which is acceptable for exploratory research. In addition, the content validity index of 0.96 confirmed strong expert agreement on item relevance, supporting the overall validity of the questionnaire.

Survey Results

The analysis revealed no significant differences between trainers and trainees concerning the distribution of assessment platforms, satisfaction with the evaluation process, the time taken for evaluation and feedback, the likelihood of forgetting mutual assessments, or responses to core CBME-related questions (Table 1). However, when comparing mobile-based assessments with paper- and web-based platforms, mobile interfaces demonstrated several advantages. Respondents using mobile-based platforms reported higher satisfaction with the assessment process, a lower likelihood of missing assessments, and an improved ability to identify performance trends. In addition, they were more likely to review feedback within 24 hours, found it easier to tailor training programs based on performance results, and experienced greater ease in identifying required clinical scenarios and necessary assessments (Table 2).

Table . Characteristics of internet survey participants from emergency residency training sites. Data are presented as N (%), mean (SD), or median (IQR).

Variables	Total (N=74) (%)	Group		*P* value
		Trainer, (n=58)	Trainee, (n=16)	
The level of the training site, n (%)				__[a]
District hospital	3 (4)	2 (4)	1 (6)	
Regional hospital	45 (61)	33 (61)	12 (75)	
Medical center	26 (35)	23 (35)	3 (19)	
Numbers of trainer in the training site				—
Mean (SD)	16.6 (9.9)	18.1 (10.5)	11.1 (3.3)	
Median (IQR)	13 (3-23)	15 (5-25)	10 (9-11)	
Numbers of trainee in the training site				—
Mean (SD)	7.5 (5.2)	8.2 (5.3)	4.9 (3.7)	
Median (IQR)	5 (1-9)	7 (3-11)	4 (2-6)	
Group, n (%)				.49
Paper and web (Google, Survey Cake etc.)	60 (81)	48 (83)	12 (75)	
Mobile-based	14 (19)	10 (17)	4 (25)	
Assessment interface, n (%)				.46
Mobile-based	14 (19)	10 (17)	4 (25)	
Paper-based	3 (4)	1 (2)	2 (13)	
Web-based	57 (77)	47 (81)	10 (62)	
Degree of satisfaction, n (%)				.06
Very satisfied	12 (16)	8 (14)	4 (25)	
Satisfied	27 (36)	22 (38)	4 (25)	
Neutral	26 (35)	22 (38)	5 (31)	
Dissatisfied	7 (9)	6 (10)	1 (6)	
Very dissatisfied	2 (3)	0 (0)	2 (13)	
The duration of assessment process in each shift (min)				.08
Mean (SD)	16 (10)	16 (11)	13 (9)	
Median (IQR)	15 (5-15)	15 (5-15)	10 (2.5-22.5)	
The duration of feedback in each shift (min) (orally or in text)				.77
Mean (SD)	19 (16)	20 (17)	17 (13)	
Median (IQR)	15 (5-15)	15 (5-15)	15 (5-15)	
Likelihood of forgetting to complete the assessments				.11
Mean (SD)	3.1 (1.3)	3.2 (1.3)	2.6 (1.3)	
Median (IQR)	3 (1-5)	3 (1-5)	3 (1-5)	
Present Method of the performance result, n (%)				
Not seen	4 (5)	2 (3)	2 (13)	.2
With number (0-100)	17 (23)	14 (24)	3 (19)	.75
With level (1-5)	53 (72)	41 (70)	12 (75)	>.99
With number and visualized trend	31 (42)	23 (40)	8 (50)	.57
With number, visualized trend, and completion rate	36 (49)	28 (48)	8 (50)	>.99

Variables	Total (N=74)	Group		*P* value
	(%)	Trainer, (n=58)	Trainee, (n=16)	
Can you identify whether the performance trend is improved or worsened from the assessment result?				.95
Mean (SD)	3.1 (1.3)	3.1 (1.2)	3.1 (1.5)	
Median (IQR)	3 (1-5)	3 (1-5)	3 (1-5)	
Can you review and respond to the feedback within 24 h?				.35
Mean (SD)	2.7 (1.3)	2.6 (1.1)	3.0 (1.6)	
Median (IQR)	3 (1-5)	3 (1-5)	3 (1-5)	
Are you able to individualize the training program based on performance results?				.4
Mean (SD)	2.7 (1.1)	2.7 (1.1)	2.5 (1.4)	
Median (IQR)	2 (1-3)	2 (1-3)	3 (1-5)	
Can you identify the required clinical scenarios and the assessments needed for each trainee?				.24
Mean (SD)	2.6 (1.3)	2.6 (1.2)	2.4 (1.5)	
Median (IQR)	2 (1-3)	2 (1-3)	2 (1-3)	

[a]Not applicable.

Table . The comparison between mobile-based and paper or web-based programs.

Variables	Total	Group		*P* value
	(n=74)	Paper or Web (n=60)	Mobile (n=14)	
The level of the training site, n (%)				—[a]
District hospital	3 (4)	1 (2)	0 (0)	
Regional hospital	45 (61)	32 (53)	14 (100)	
Medical center	26 (35)	26 (43)	0 (0)	
Numbers of trainer in the training site				<.001
Mean (SD)	16.6 (9.9)	18.1 (10.4)	10.0 (0.0)	
Median (IQR)	13 (3-23)	15 (6-24)	10 (10-10)	
Numbers of trainee in the training site				.04
Mean (SD)	7.5 (5.2)	8.1 (5.6)	4.9 (0.4)	
Median (IQR)	5 (1-9)	8 (1-15)	5 (5-5)	
Participant age (year-old)				.65
Mean (SD)	42.1 (8.2)	41.3 (8.4)	42.0 (7.9)	
Median (IQR)	43 (31-55)	42 (31.5-52.5)	45 (35.5-54.5)	
Participant roles, n (%)				.53
Junior resident (R1/R2)	7 (9)	5 (8)	2 (14)	
Senior resident (R3/R4)	9 (12)	7 (12)	2 (14)	
RRC[b] host	9 (12)	8 (13)	1 (7)	
Chief of the department	9 (12)	9 (15)	0 (0)	
Clinical instructor	40 (54)	31 (52)	9 (64)	
Assessment interface, n (%)				—
Mobile-based	14 (19)	0 (0)	14 (100)	
Paper-based	3 (4)	3 (5)	0 (0)	
Web-based	57 (77)	57 (95)	0 (0)	
Degree of satisfaction, n (%)				<.001
Very satisfied	12 (16)	3 (5)	9 (64)	
Satisfied	27 (36)	22 (37)	5 (36)	
Neutral	26 (35)	26 (43)	0 (0)	
Dissatisfied	7 (9)	7 (12)	0 (0)	
Very dissatisfied	2 (3)	2 (3)	0 (0)	
The duration of assessment process in each shift (min)				.54
Mean (SD)	15.5 (10.2)	15.3 (9.0)	16.8 (14.6)	
Median (IQR)	15 (5-15)	15 (5-15)	10 (5-15)	
The duration of feedback in each shift (min) (orally or in text)				.13
Mean (SD)	19.0 (16.0)	17.7 (14.9)	24.6 (19.6)	
Median (IQR)	15 (5-15)	15 (2-28)	18 (8-28)	
Likelihood of forgetting to complete the assessments				.02
Mean (SD)	3.1 (1.3)	3.3 (1.3)	2.4 (1.1)	
Median (IQR)	3 (1-5)	3 (1-5)	3 (1-5)	
Present Method of the performance result, n (%)				—
Not seen	4 (5)	4 (7)	0 (0)	

Variables	Total	Group		*P* value
	(n=74)	Paper or Web (n=60)	Mobile (n=14)	
The level of the training site, n (%)				__[a]
With number (0 - 100)	17 (23)	10 (17)	7 (50)	
With level (1-5)	53 (72)	44 (73)	9 (64)	
With number and visualized trend	31 (42)	19 (32)	12 (86)	
With number, visualized trend, and complete rate	36 (49)	24 (40)	12 (86)	
Can you identify whether the performance trend is improved or worsened from the assessment result?				<.001
Mean (SD)	3.1 (1.3)	2.7 (1.1)	4.7 (0.6)	
Median (IQR)	3 (1-5)	2 (1-3)	5 (5-5)	
Can you review and respond to the feedback within 24 hours?				<.001
Mean (SD)	2.7 (1.3)	2.2 (0.9)	4.5 (0.9)	
Median (IQR)	2 (1-3)	2 (1-3)	5 (4-5)	
Are you able to individualize the training program based on performance results?				<.001
Mean (SD)	2.7 (1.1)	2.3 (0.8)	4.4 (0.8)	
Median (IQR)	3 (2-4)	2 (1-3)	5 (4-5)	
Can you identify the required clinical scenarios and the assessments needed for each trainee?				<.001
Mean (SD)	2.6 (1.3)	2.1 (0.8)	4.8 (0.4)	
Median (IQR)	2 (1-3)	2 (2-2)	5 (5-5)	

[a]Not applicable.

[b]RRC: indicated residency review committee.

Further statistical analysis indicated that the use of MB platforms was significantly associated with a 2.08-point increase (95% CI 1.73 - 2.43, *P*=.002) in the ability to individualize training programs based on performance results (Table 3).

Table . Regression analysis for individualization of training program.

Variables	OR[a] (95% CI)	*P* value
Mobile-based	2.08 (1.73 - 2.43)	.02
Number of trainees	<0.01 (–0.07 to 0.06)	.78
Number of trainers	<0.01 (–0.02 to 0.02)	.89

[a]OR: odds ratio

Discussion

Principal Findings

This pilot study found that mobile-based programs more frequently used visualized performance trends (85.71%) than paper- and web-based programs (73.33%), enhancing trainees' understanding (mean 4.71, SD 0.61 vs mean 2.72, SD 1.06; *P*<.001). Mobile-based platforms also enabled faster review of ad hoc responses within 24 hours (mean 4.5, SD 0.85 vs mean 2.22, SD 0.88; *P*<.001) and better supported individualized training in alignment with CBME principles (mean 4.36, SD 0.84 vs mean 2.27, SD 0.78; *P*<.001). Statistical analysis indicated a significant association between MB assessments and improved training individualization (2.08-point increase, 95% CI 1.73 - 2.43; *P*=.002), suggesting mobile-based platforms facilitate more timely and adaptive program adjustments than paper- and web-based (Table 3).

Comparison With Previous Work

Since 2009, CBME has become a global standard in medical training [1]. Oudkerk Pool et al [14] highlight that competency judgments are formed through an iterative process of acquiring and synthesizing evidence, underscoring the need for structured, multisource assessments. However, effective CBME implementation requires more than data collection; it demands meaningful integration and interpretation among instructors, residents, and program coordinators, including RRC hosts and CCC members [14,16].

Despite advancements in assessment methods, paper- and web-based evaluations remain prevalent in Taiwan's residency programs (Table 2). These static, retrospective tools limit

real-time performance tracking and individualized feedback. Chan et al [16] demonstrated that programmatic workplace-based assessments, such as the McMaster Modular Assessment Program (McMAP), improve competency evaluations by replacing single-assessor recall with multisource, continuous feedback. Likewise, our findings suggest that mobile-based platforms offer a promising alternative by enabling timely communication, visualizing performance trends, and tracking clinical exposures in real time, aligning with CBME's emphasis on progressive competency tracking.

However, while convenience is improved, adopting a mobile-based platform alone does not guarantee the effective integration of summative performance trends. As Oudkerk Pool et al [14] emphasize, competency judgments require active evidence synthesis rather than passive data aggregation. Therefore, mobile-based platforms must be supported by faculty training and institutional commitment to continuous assessment. Beyond technology, mobile-based platform development reflects institutional investment in assessment culture. Successful implementation requires standardized tools, integration, and a culture that embraces real-time feedback. Chan et al [16] highlight that workplace-based assessment systems like McMAP not only streamline logistics but also normalize structured, frequent evaluations, shaping residency training [12]. This may explain why, in our study, the mobile-based platform yielded higher satisfaction and a more tailored training experience compared with the paper- and web-based group.

Compared with mobile-based assessment evaluations, paper- and web-based assessments, though practical, are often less accessible, time-consuming, and associated with lower compliance [19]. Mobile-based platforms provide easy access and real-time feedback, allowing trainers and trainees to review performance trends anytime and anywhere. In addition, their integration with smartphone features—such as cameras, note apps, calendars, GPS, and barcode scanners—enhances efficiency and usability [18,19,21-23]. However, while mobile-based platforms improve accessibility, Walsh [5] noted that mobile learning in medical education may also introduce distractions, potentially affecting engagement and focus during training sessions.

Previous studies have demonstrated the feasibility and effectiveness of mobile-based assessment tools in medical training. Nethala et al [11] found that mobile apps enable real-time assessment and individualized skill training in urology residency programs, facilitating structured competency tracking. Similarly, Sung and Park [22] reported that a mobile-based training program improved nurses' competence through enhanced accessibility and interactive learning. Green highlighted the cost-effectiveness and usability of smartphone platforms for surgical resident evaluations, reinforcing the practical benefits of mobile-based assessments in medical education [23].

In our study, while assessment and feedback durations did not differ significantly between mobile-based and paper- and web-based platforms, mobile-based assessments yielded higher satisfaction and a lower frequency of missed evaluations (P<.001 and P=.02, respectively; Table 1). These findings suggest that the convenience and accessibility of mobile platforms may enhance trainer engagement and encourage more consistent assessment documentation.

Beyond technological advantages, integrating multisource evaluations into visualized trend tracking on mobile-based platforms may enhance participant engagement, foster active participation, and support continuous improvement. According to self-determination theory, providing trainees with tools for real-time performance monitoring fosters autonomy, competence, and relatedness, key psychological needs that enhance motivation and self-directed learning [13]. When trainees and trainers can instantly track progression and performance trends, they gain a clearer perception of their self-determination (Figure 2), which may drive greater accountability and active participation in their own professional growth. This aligns with Association for Medical Education in Europe Guide No. 59, which highlights the role of real-time feedback in enhancing intrinsic motivation and engagement in competency development [12,13].

Our survey found that although some web-based programs included visualized trends and completion rates, mobile-based programs had a significantly greater impact on both trainers and trainees (mean 2.3, SD 0.8 vs mean 4.4, SD 0.8; P<.001; Table 2). This advantage likely stems from the mobile-based platform's accessibility, enabling real-time review compared with paper- and web-based assessments, which are typically evaluated only during biannual or quarterly CCC meetings [24]. Delayed review in paper- and web-based evaluations may hinder timely performance adjustments, whereas immediate access to evaluation results in mobile-based programs enables trainees to modify their performance on the next shift, reinforcing self-directed learning [13].

The observed advantages of mobile-based platforms should not be attributed solely to their technology but rather to the broader institutional commitment to structured assessment and a culture of continuous feedback. While mobile-based platforms facilitate real-time performance tracking and self-directed learning, their effectiveness depends on institutional investment in faculty training, standardized assessment frameworks, and integration within CBME structures [4,16]. The shift from paper- and web-based biannual review model [24] to continuous monitoring reflects not just technological convenience but a deeper transformation in how competency is assessed and supported. As previous studies highlight, frequent, structured feedback is essential for meaningful competency development [12,15], and the adoption of MB platforms likely signifies an institutional prioritization of learner-centered training and assessment rigor rather than merely a technological upgrade. This shift, reinforced by evidence-based coaching and real-time decision-making, aligns with a growing recognition that assessment culture, not just digital tools, drives improved training experience [13,16].

The higher satisfaction with mobile-based platforms compared with paper- and web-based (P<.001, Table 2) may be attributed to their ability to provide real-time accessibility, structured assessments, and visualized performance trends, benefiting trainees, trainers, and RRC hosts. For trainees, mobile-based platforms enabled immediate feedback and continuous

performance tracking, allowing timely adjustments rather than relying on periodic CCC reviews. This aligns with CBME principles by supporting self-directed learning and competency progression. In addition, mobile-based platforms facilitated easier identification of required clinical scenarios and assessments (P<.001), which may contribute to a more structured training experience.

For trainers, mobile-based platforms allowed more flexible and timely assessments, reducing recall bias and documentation burden. The ability to complete evaluations outside clinical hours (P=.02; Table 2) suggests increased convenience and consistency in providing feedback. The integration of smartphone tools, such as cameras and note apps, may have contributed to streamlined documentation and more structured evaluations [21]. For RRC hosts, mobile-based platforms provided a longitudinal perspective on trainee progress, reducing reliance on episodic assessments and allowing competency decisions to be based on continuous performance data rather than retrospective impressions [14,16]. The findings suggest that mobile-based platforms improve accessibility and assessment efficiency, reflecting a shift toward more continuous, data-driven evaluation in residency training.

The preference for mobile-based platforms in tailored training (P<.001; Table 2) stems from their ability to provide real-time feedback, trend visualization, and individualized learning, aligning with CBME's emphasis on continuous assessment [2,4]. Unlike paper- and web-based periodic evaluations, mobile-based platforms empower trainees with self-directed learning and immediate performance adjustments while enabling trainers to deliver data-driven coaching with minimized bias [12,16]. For RRC hosts, mobile-based platforms enhance longitudinal competency tracking, overcoming the limitations of CCC's fixed evaluation intervals and allowing timely interventions [1,3,24]. However, their success depends on structured implementation, faculty engagement, and integration within CBME frameworks to ensure meaningful assessment without cognitive overload [5,6,17].

Overall, while mobile-based platforms show promise in improving feedback timeliness and training individualization in our pilot study, further research is needed to evaluate their long-term impact on competency development, clinical outcomes, and residency training culture.

Highlights

First, mobile-based platforms provide real-time feedback and continuous performance tracking, complementing scheduled CCC reviews and addressing paper- and web-based limitations in integrating multisource evaluation results. Second, structured mobile-based evaluations improve engagement and training individualization, offering more timely and adaptive learning opportunities that better align with CBME principles compared with paper- and web-based platforms. Finally, successful integration depends on faculty training and strategic implementation, enhancing assessment validity, competency tracking, and institutional commitment to resident development, ultimately leading to greater satisfaction.

Structured mobile-based evaluations improve engagement and training individualization, offering more timely and adaptive learning opportunities that better align with CBME principles compared with paper- and web-based platforms. Successful mobile-based integration depends on faculty training and strategic implementation, enhancing assessment validity, competency tracking, and institutional commitment to resident development, ultimately leading to greater satisfaction.

Limitations

First, the study results were based on a convenience sample from an internet survey. However, enrollment only covered 62.22% of the emergency residency training sites in the country. Second, the survey results indicated that only 14 out of 74 responses applied mobile-based assessment. Further investigation to evaluate mobile-based assessment will be needed as more training programs start being delivered through the mobile interface. Finally, in the survey, participants provided responses based on their experience with the interface's ability to facilitate tailored training adjustments, rather than on direct evidence of its actual implementation. Further research is needed to assess the long-term impact of mobile-based platform implementation on training outcomes.

The survey results indicated that only 14 out of 74 responses applied mobile-based assessment. Further investigation to evaluate mobile-based assessment will be needed as more training programs start being delivered through the mobile interface.

In the survey, participants provided responses based on their experience with the interface's ability to facilitate tailored training adjustments, rather than on direct evidence of its actual implementation. Further research is needed to assess the long-term impact of mobile-based platform implementation on training outcomes.

Conclusion

In conclusion, the mobile-based interface emerges as a dynamic and effective platform for emergency residency training programs. It facilitates rapid updates and individualized program modifications, thereby increasing the engagement and satisfaction of trainers, trainees, and other stakeholders. Contrarily, the conventional paper- and web-based methods face limitations due to prolonged review periods in committee meetings, which may lead to delays in crucial program modifications and impede the progression of the program. The adoption of mobile-based technology in this context demonstrates its capacity to greatly enhance the efficiency and efficacy of CBM, particularly in making tailored adjustments. This technology also promotes better-informed and more collaborative interactions among all involved parties.

Acknowledgments

During the preparation of this work, the authors used ChatGPT 4 in order to improve readability and language. After using this tool or service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication. The visualization trend and completion ratio were inspired by the design of a web-based e-portfolio by the Department of Emergency, Dalin Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, Chiayi County, Taiwan. The designers are Cheng-Wei Lee and Yi-Kung Lee.

Data Availability

The data that support the findings of this study are available from the corresponding author (TYH) upon reasonable request.

Authors' Contributions

HLC contributed to conceptualization, investigation, methodology, and writing-original draft. CWL handled conceptualization and software. CWC managed validation, formal analysis, and conceptualization. YCC handled investigation and visualization. TYH contributed to conceptualization, resource, software, supervision, methodology, investigation, visualization, and writing-review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Internet survey form.
[DOCX File, 17 KB - mededu_v11i1e57216_app1.docx]

References

1. Taber S, Frank JR, Harris KA, et al. Identifying the policy implications of competency-based education. Med Teach 2010 Aug;32(8):687-691. [doi: 10.3109/0142159X.2010.500706]
2. Frank JR, Snell LS, Cate OT, et al. Competency-based medical education: theory to practice. Med Teach 2010 Aug;32(8):638-645. [doi: 10.3109/0142159X.2010.501190]
3. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR, for the International CBME Collaborators. The role of assessment in competency-based medical education. Med Teach 2010 Aug;32(8):676-682. [doi: 10.3109/0142159X.2010.500704]
4. Van Melle E, Frank JR, Holmboe ES, et al. A core components framework for evaluating implementation of competency-based medical education programs. Acad Med 2019 Jul;94(7):1002-1009. [doi: 10.1097/ACM.0000000000002743] [Medline: 30973365]
5. Walsh K. Mobile learning in medical education: review. Ethiop J Health Sci 2015 Oct;25(4):363-366. [doi: 10.4314/ejhs.v25i4.10] [Medline: 26949301]
6. Sharma S, Kumari B, Ali A, et al. Mobile technology: a tool for healthcare and a boon in pandemic. J Family Med Prim Care 2022 Jan;11(1):37-43. [doi: 10.4103/jfmpc.jfmpc_1114_21] [Medline: 35309626]
7. Hartman DJ. Mobile technology for the practice of pathology. Adv Anat Pathol 2016 Mar;23(2):118-124. [doi: 10.1097/PAP.0000000000000093] [Medline: 26849818]
8. Sclafani J, Tirrell TF, Franko OI. Mobile tablet use among academic physicians and trainees. J Med Syst 2013 Feb;37(1):9903. [doi: 10.1007/s10916-012-9903-6] [Medline: 23321961]
9. Franko OI, Tirrell TF. Smartphone app use among medical providers in ACGME training programs. J Med Syst 2012 Oct;36(5):3135-3139. [doi: 10.1007/s10916-011-9798-7] [Medline: 22052129]
10. Duggan N, Curran VR, Fairbridge NA, et al. Using mobile technology in assessment of entrustable professional activities in undergraduate medical education. Perspect Med Educ 2021 Dec;10(6):373-377. [doi: 10.1007/s40037-020-00618-9] [Medline: 33095399]
11. Nethala D, Martin C, Griffiths L, et al. Feasibility and utility of mobile applications for the evaluation of urology residents' surgical competence. Urology 2021 Dec;158:11-17. [doi: 10.1016/j.urology.2021.05.112] [Medline: 34437893]
12. Marty AP, Linsenmeyer M, George B, Young JQ, Breckwoldt J, Ten Cate O. Mobile technologies to support workplace-based assessment for entrustment decisions: Guidelines for programs and educators: AMEE Guide No. 154. Med Teach 2023 Nov;45(11):1203-1213. [doi: 10.1080/0142159X.2023.2168527] [Medline: 36706225]
13. Ten Cate TJ, Kusurkar RA, Williams GC. How self-determination theory can assist our understanding of the teaching and learning processes in medical education. AMEE guide No. 59. Med Teach 2011;33(12):961-973. [doi: 10.3109/0142159X.2011.595435] [Medline: 22225433]
14. Oudkerk Pool A, Govaerts MJB, Jaarsma D, Driessen EW. From aggregation to interpretation: how assessors judge complex data in a competency-based portfolio. Adv Health Sci Educ Theory Pract 2018 May;23(2):275-287. [doi: 10.1007/s10459-017-9793-y] [Medline: 29032415]

15. Turnbull J, Gray J, MacFadyen J. Improving in-training evaluation programs. J Gen Intern Med 1998 May;13(5):317-323. [doi: 10.1046/j.1525-1497.1998.00097.x] [Medline: 9613887]
16. Chan T, Sherbino J, Collaborators M. The McMaster Modular Assessment Program (McMAP): a theoretically grounded work-based assessment system for an emergency medicine residency program. Acad Med 2015 Jul;90(7):900-905. [doi: 10.1097/ACM.0000000000000707] [Medline: 25881648]
17. Chandran VP, Balakrishnan A, Rashid M, et al. Mobile applications in medical education: a systematic review and meta-analysis. PLoS ONE 2022;17(3):e0265927. [doi: 10.1371/journal.pone.0265927] [Medline: 35324994]
18. Payne KFB, Wharrad H, Watts K. Smartphone and medical related App use among medical students and junior doctors in the United Kingdom (UK): a regional survey. BMC Med Inform Decis Mak 2012 Oct 30;12:121. [doi: 10.1186/1472-6947-12-121] [Medline: 23110712]
19. Karadeniz Ş. The impacts of paper, web and mobile based assessment on students' achievement and perceptions. Scientific Research and Essay 2009;4:984-991.
20. Hertzog MA. Considerations in determining sample size for pilot studies. Res Nurs Health 2008 Apr;31(2):180-191. [doi: 10.1002/nur.20247] [Medline: 18183564]
21. Gladman T, Tylee G, Gallagher S, Mair J, Rennie SC, Grainger R. A tool for rating the value of health education mobile apps to enhance student learning (MARuL): development and usability study. JMIR Mhealth Uhealth 2020 Jul 31;8(7):e18015. [doi: 10.2196/18015] [Medline: 32735228]
22. Sung S, Park HA. Effect of a mobile app-based cultural competence training program for nurses: a pre- and posttest design. Nurse Educ Today 2021 Apr;99:104795. [doi: 10.1016/j.nedt.2021.104795] [Medline: 33621852]
23. Green JM. An innovative, no-cost, evidence-based smartphone platform for resident evaluation. J Surg Educ 2016;73(6):e14-e18. [doi: 10.1016/j.jsurg.2016.07.016] [Medline: 27651056]
24. Ekpenyong A, Padmore JS, Hauer KE. The purpose, structure, and process of clinical competency committees: guidance for members and program directors. J Grad Med Educ 2021 Apr;13(2 Suppl):45-50. [doi: 10.4300/JGME-D-20-00841.1] [Medline: 33936532]

Abbreviations

CBME: competency-based medical education
CCC: Clinical Competency Committee
McMAP: McMaster Modular Assessment Program
RRC: Residency Review Committee

Edited by B Lesselroth; submitted 08.02.24; peer-reviewed by M Shershneva, YJ Su; revised version received 03.03.25; accepted 01.06.25; published 24.07.25.

Please cite as:
Chen HL, Lee CW, Chang CW, Chiu YC, Hung TY
Evaluating Tailored Learning Experiences in Emergency Residency Training Through a Comparative Analysis of Mobile-Based Programs Versus Paper- and Web-Based Approaches: Feasibility Cross-Sectional Questionnaire Study
JMIR Med Educ 2025;11:e57216
URL: https://mededu.jmir.org/2025/1/e57216
doi:10.2196/57216

Acceptance of AI-Powered Chatbots Among Physiotherapy Students: International Cross-Sectional Study

Salwa B El-Sobkey[1,2*], Prof Dr; Kerolous Ishak Kelini[3*], PhD; Mahmoud ElKholy[2*], PhD; Tayseer Abdeldayem[2*], PhD; Mariam Abdallah[2*], PhD; Dina Al-Amir Mohamed[2*], PhD; Aya Fawzy[4*], PhD; Yomna F Ahmed[4*], PhD; Ayman El Khatib[5*], Prof Dr; Hind Khalid[6*], MSc; Balkhis Banu Shaik[1*], PhD; Ana Anjos[1*], MSc; Mutasim D Alharbi[7*], PhD; Karim Fathy[6*], Prof Dr; Khaled Takey[5,8*], PhD

[1]Department of Physiotherapy, Fatima College of Health Sciences, Khalifa Bin Zayed Street, Al Maqam, PO Box 24162, Al Ain, Abu Dhabi, United Arab Emirates

[2]Faculty of Physical Therapy, Beni-Suef University, Beni-Suef, Egypt

[3]Department of Physical Therapy, Faculty of Applied Medical Sciences, Al-Zaytoonah University of Jordan, Amman, Jordan

[4]Faculty of Physical Therapy, Modern University for Technology and Information, Cairo, Egypt

[5]Physical Therapy Department, Faculty of Health Sciences, Beirut Arab University, Beirut, Lebanon

[6]Faculty of Physical Therapy, October 6 University, October 6, Egypt

[7]Department of Physical Therapy, Faculty of Medical Rehabilitation Sciences, King Abdulaziz University, Jeddah, Saudi Arabia

[8]Faculty of Physical Therapy, Misr University for Sciences and Technology, October 6, Egypt

[*]all authors contributed equally

Corresponding Author:
Salwa B El-Sobkey, Prof Dr
Department of Physiotherapy, Fatima College of Health Sciences, Khalifa Bin Zayed Street, Al Maqam, PO Box 24162, Al Ain, Abu Dhabi, United Arab Emirates

Abstract

Background: Artificial intelligence–powered chatbots (AI-PCs) are increasingly integrated into educational settings, including health care disciplines. Despite their potential to enhance learning, limited research has investigated physiotherapy (PT) students' acceptance of this technology.

Objective: This study aims to assess undergraduate PT students' acceptance of AI-PCs and to identify personal, academic, and technological factors influencing their acceptance.

Methods: Over a 4-month period, a cross-sectional survey was conducted across 7 PT programs in 5 countries. Eligible participants were national undergraduate PT students. The technology acceptance model (TAM)–based questionnaire was used for capturing perceived usefulness, perceived ease of use, attitude, behavioral intention, and actual behavioral use of AI-PCs. The influence of personal, academic, and technological factors was examined. Descriptive and inferential statistics were conducted.

Results: The mean total TAM score was 3.59 (SD 0.82), indicating moderate acceptance. Of the 1066 participants, 375 (35.2%) showed high acceptance, 650 (60.9%) moderate, and 41 (3.9%) low. Prior experience with artificial intelligence (AI) tools emerged as the strongest predictor of acceptance (β=.43; P<.001), followed by university affiliation (ANOVA P<.001). Cumulative grade point average percentage was positively correlated with TAM score (r=0.135; P<.001) but was not a significant predictor in regression (P=.23). Age (P=.54), sex (P=.56), academic level (P=.26), and current use of AI-PCs (P=.10) were not significant predictors.

Conclusions: PT students demonstrated moderate acceptance of AI-PCs. Prior technological experience was the strongest predictor, underscoring the importance of early exposure to AI tools. Educational institutions should consider integrating AI technologies to enhance students' familiarity and foster positive attitudes toward their use.

(JMIR Med Educ 2025;11:e76574) doi:10.2196/76574

KEYWORDS

artificial intelligence; AI-powered chatbots; health care students; physiotherapy education; TAM; technology acceptance model; technology acceptance

Introduction

Background

Artificial intelligence (AI) enables computer systems to imitate human intelligence by processing external data and learning to perform specific tasks [1-4]. Large language models, a type of AI, are trained for human-like communication, with artificial intelligence–powered chatbots (AI-PCs) being one of their key applications [5-7]. AI has gained significant attention in health care and health professional education due to its potential to make learning more accessible, affordable, and effective [8,9]. Universities are increasingly adopting AI technologies, with AI-PC use expected to grow as technology advances [6,10-12].

In the 1970s, chatbots were known as pedagogical agents [13] and are now referred to as conversational agents, tutors, or simply bots [14]. Literature identifies AI-PCs as valuable educational tools that enhance learning experiences and outcomes [6,15,16], though they also face challenges such as accuracy, bias, and user attitudes [14,17]. The acceptance of AI-PCs in education remains underexplored, with a need for further research to clarify their adoption in universities [16,18]. To examine students' acceptance of AI-PCs, it is important to use a well-established technology acceptance model (TAM) [19].

TAM, originally developed by Davis in 1989 and updated in 1993 [20,21], is the most widely used framework for studying technology acceptance in education [22,23]. It is considered a reliable source for understanding the students' acceptance of learning technologies due to its educational focus and robust structure [24,25]. According to TAM, technology acceptance follows a 3-stage process, in which external factors trigger cognitive responses, perceived usefulness (PU) and perceived ease of use (PEU), that subsequently shape affective responses, namely attitude toward using technology and behavioral intention (BI), ultimately resulting in actual behavioral use (ABU) [20,21].

The PU refers to the individual's beliefs that a technology will enhance task performance, while PEU reflects how effortless it is to use [26,27]. While BI and attitude represent the potential consequences of the behavior [28], ABU is the outcome [23]. Investigating external factors such as students' age and prior knowledge could provide valuable insights [6].

AI is among the top 10 priorities for physiotherapy (PT) service improvement [29], and while its role in treatment is documented, research on its application in PT education remains scarce [5]. Further studies are required to explore the role of AI-PCs in educational settings and external factors influencing their adoption [6]. Moreover, understanding how students' characteristics interact with AI-PCs is crucial for educators and policy makers.

Objective

Building on this need, this study aims to assess undergraduate PT students' acceptance of AI-PCs and to identify personal, academic, and technological factors influencing their acceptance.

Methods

Ethical Considerations

Ethics approval was obtained from the respective institutional research boards in each participating institution: Fatima College of Health Sciences (FCHS) Institutional Research Board (approval: FECE-1-24-25ELSOBKEY), Beni-Suef University (BSU) Ethical Committee (approval: 12492024), Al-Zaytoonah University of Jordan (ZUJ) Institutional Research Board (approval: 10/10/2024 - 2025), Modern University for Technology and Information (MTI) Research Ethical Committee (approval: REC/2111/MTI.PT/2410292), Beirut Arab University (BAU) Institutional Research Board (approval: 2024-H-0206-HS-M-0643), October 6 University (O6U) Research Ethical Committee (approval: O6U.PT.REC/024/002009), and King Abdulaziz University (KAU) Research Ethics Committee (approval: 320 - 24). The students' confidentiality measures were explained in a participant information sheet. The questionnaire was anonymous, and the researchers are the only authorized body to have access to the collected data. Additionally, it was emphasized that participation was entirely voluntary, without academic consequences for nonparticipation. Participants signed a consent form, which was provided at the end of the participant's information sheet. For digital surveying, the following statement was included in the first section of the Google Forms: "Responding to this questionnaire will be considered as agreement to participate in the study."

Study Design

This study used an international cross-sectional survey design. The study was conducted across 7 PT programs in 5 countries (Table 1). Data were collected from September 29, 2024, to January 26, 2025. Undergraduate PT students in the participating universities were the target population and were recruited through bulletin boards, college emails, WhatsApp messages, and verbal invitations. A participant information sheet was distributed for paper-based surveying and was the first section on the digital questionnaire. It explained the study's purpose, procedure, and the average time to respond (7, SD 1.5 minutes). Convenience sampling was used. Eligible participants were national students enrolled in a bachelor-level PT program who voluntarily provided informed consent. Exclusion criteria included digital or hybrid learning enrollment, repeating multiple courses, being at the internship level (as not all programs included this level), or not completing the questionnaire. All eligible students were invited to participate.

Table . Countries, Universities, physiotherapy programs, and number of participating students (N=1066).

Country and university		College or department	Participating students, n (%)
1. Egypt			
	Beni-Suef University	Faculty of Physical Therapy	200 (18.8)
	October 6 University	Faculty of Physical Therapy	200 (18.8)
	Modern University for Technology and Information	Faculty of Physical Therapy	200 (18.8)
2. United Arab Emirates			
	Institute of Applied Technology	Physiotherapy Department, Fatima College of Health Sciences	107 (10.0)
3. Saudi Arabia			
	King Abdulaziz University	Department of Physical Therapy, Faculty of Medical Rehabilitation Sciences	67 (6.3)
4. Lebanon			
	Beirut Arab University	Physical Therapy Department, Faculty of Health Sciences	200 (18.8)
5. Jordan			
	Al - Zaytoonah University of Jordan	Department of Physical Therapy-Faculty of Applied Medical Sciences	92 (8.6)

Variables

Variables were aligned with TAM constructs and relevant external factors. Outcome variables included the TAM 5 constructs: PU, defined as students' perception of AI-PCs' benefits to their studies; PEU, referring to how easy students find AI-PCs to use; attitude, representing students' overall feelings toward AI-PCs; BI, assessed as the intention to use AI-PCs among nonusers and the intention to continue use among current users; and ABU, measuring AI-PC use. A total TAM score was calculated as the average of these 5 constructs. External factors included personal: age and sex; academic: university affiliation, academic level, cumulative grade point average percentage (CGPA%), and current use of AI-PCs; and technological: prior experience with technologies or AI tools other than AI-PCs.

Data Collection Process

Students first answered an initial screening question about their prior experience with AI-PCs to determine which version of the questionnaire to complete. Those with experience received questionnaire A, and those without experience received modified questionnaire B. For the digital version, researchers provided clear instructions and links, highlighting the importance of selecting the appropriate questionnaire according to prior AI-PCs' use. The digital questionnaire back button character was enabled to allow respondents to review and change their answers. To avoid multiple entries from the same student, the form setting of a limit to 1 response was activated.

Data Sources or Measurement

Two structured questionnaires were used. Questionnaire A (for students who are currently using or who had previously used AI-PCs) included assessed external factors (7 items) and the basic TAM section (24 items, covering 5 constructs, scored by a 5-point Likert scale) [20,21,27]. Instructions were adapted from Lewis [27]. Mean scores were calculated for each construct and overall acceptance, with scores ranging from 1 to 5. Students were categorized into high (≥4), moderate (2 - 3.9), and low (<2) acceptance. Questionnaire B (for students with no prior AI-PC experience) followed the same structure but excluded the ABU construct. Both questionnaires were anonymous and were available in digital (Google Forms) and paper-based formats. The digital questionnaire was an open survey for students who received the link from the corresponding investigators. In both questionnaire versions, students were instructed that the questionnaire is divided into 2 sections, and they need to respond to each question by choosing the most accurate answer that reflects their personal experience or opinion. The 5-point Likert scale was also explained, and if they have any questions or concerns, they can freely contact the research team at their institution.

Pilot Study

The questionnaire was piloted with at least 5 students per institution. Feedback led to 1 minor revision from King Abdulaziz University. Completion time ranged from 5 to 15 minutes (average 7, SD 1.5 minutes). The final version was used in both formats. Additionally, the digital questionnaire was tested for usability and technical functionality.

Efforts to Address Potential Sources of Bias

Student enrollment varied widely across institutions, ranging from approximately 100 to 1000s of PT students, posing a risk of statistical bias. To reduce overrepresentation, all eligible students from smaller institutions were included, while 200 students were sampled from larger ones. Additionally, to ensure curricular consistency, only bachelor's students were included,

excluding those in Doctor of Physiotherapy programs due to structural and competency differences. Internship-level students were also excluded, as not all institutions offered this academic level. These measures aimed to enhance sample homogeneity and minimize variability that could affect the study's outcomes.

Further, only national students were included to minimize cultural and linguistic variability across diverse settings, as such cross-cultural differences may influence students' acceptance of AI-PCs. Similarly, students who were repeating more than 1 course during the data collection semester were excluded, as repeated exposure to content may lead to an inflated perception of the chatbot's usefulness and a biased response in TAM-related items. These measures were taken to enhance sample homogeneity and reduce variability that could affect the study's outcomes.

Handling of Data

Digital responses were exported to a Microsoft Excel sheet, and paper-based data were entered manually. All entries were reviewed for accuracy before statistical analysis.

Statistical Methods

Descriptive statistics included means and SDs for continuous variables and frequency distributions for categorical variables. Inferential tests included the chi-square test and Pearson or Spearman correlations to assess associations and correlations between total TAM score, acceptance categories, and external factors. An independent 2-tailed *t* test compared TAM scores between users and nonusers. ANOVA with Tukey post hoc tests examined differences in TAM scores, construct scores, CGPA%, and prior experience across universities. The chi-square test also assessed differences in categorical variables like acceptance categories and academic level across universities. Finally, linear regression identified significant predictors of AI-PCs and measured effect sizes. The threshold for statistical significance was set at $P<.05$.

Handling Missing Data

All items of the digital questionnaire were mandatory to be answered to enable the questionnaire submission. Instead, a few number of students, fewer than 10 students, answered by placing a dot for age and cumulative grade point average (CGPA) items. Missing data for these 2 items were replaced with mean values calculated separately for each institution.

Reporting

This study was reported by the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist for cross-sectional studies in addition to the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) [30].

Results

Participants

A total of 1066 PT students participated in the study (Table 1). Four institutions (BSU, O6U, MTI, and BAU) each contributed 200 students, meeting the target sample size. The remaining 3 institutions (FCHS, ZUJ, and KAU) had fewer participants but achieved a 100% response rate among eligible students. These institutions follow a 4-year study plan, unlike the 5-year structure on the other sites. At FCHS and ZUJ, many students were nonnationals and excluded based on eligibility criteria. At KAU, 67 students participated, as PT specialization begins in the second year; this number represented the entire eligible cohort.

Demographic and Academic Characteristics

Table 2 presents the demographic and academic characteristics of participating PT students. The mean age was 20.1 (SD 1.5) years, and 64.1% (n=683) were female. Students were enrolled across all academic years, with the highest proportion in the second year (n=367, 34.4%) and the lowest in the fifth year (n=103, 9.7%). The fifth year applies only to Egyptian universities (BSU, O6U, and MTI), while others follow a 4-year program. Due to differences in CGPA scales (4.0 vs 5.0) across universities, CGPA% was used for standardization, resulting in a mean of 79.7% (SD 14%). CGPA% was not applicable for first-year students in their first semester, as their CGPA was not yet available.

Table . Demographic, academic, and technological characteristics of the participating physiotherapy students (N=1066).

Characteristic		Values
Age (years)		
	Mean (SD)	20.1 (1.5)
	Range	17.0-28.0
Cumulative grade point average percentage (n=924)		
	Mean (SD)	79.7 (14.0)
	Range	26.8-100.0
Sex, n (%)		
	Female	683 (64.1)
	Male	383 (35.9)
Academic level, n (%)		
	First year	176 (16.5)
	Second year	367 (34.4)
	Third year	237 (22.2)
	Fourth year	183 (17.2)
	Fifth year	103 (9.7)
Use of artificial intelligence–powered chatbots, n (%)		
	Yes	586 (55.0)
	No	480 (45.0)

PT Students' Acceptance of AI-PCs

PT students' acceptance of AI-PCs, as measured by the TAM, had a mean total score of 3.59 of 5 (SD 0.81), reflecting overall moderate acceptance across all universities. The mean scores for the 5 TAM constructs were: PU=3.69 (SD 0.92), PEU=3.68 (SD 0.88), attitude=3.58 (SD 0.91), BI=3.57 (SD 0.91), and ABU=3.40 (SD 0.91). Based on TAM categories, 35.2% (n=375) of students demonstrated high acceptance, 60.9% (n=649) moderate, and 3.9% (n=42) low.

External Factors Influencing PT students' Acceptance of AI-PCs

Personal, academic, and technological factors were assessed as potential predictors of acceptance of AI-PCs. Among personal factors, age (Table 3) showed no significant correlation with the TAM mean score (P=.54), and sex showed no significant association with TAM acceptance categories (P=.56).

Table . Correlation between physiotherapy students' total mean score of technology acceptance of artificial intelligence–powered chatbots (AI-PCs) and external factors (N=1066).

External factors	*r*	*P* value
TAM[a] total score of AI-PCs*students' age	−0.019	.54
TAM total score of AI-PCs*students' CGPA%[b]	0.135	<.001
TAM total score of AI-PCs*students' previous experience of using other artificial intelligence–powered learning tools or applications beyond chatbots	0.445	<.001

[a]TAM: technology acceptance model.

[b]CGPA%: cumulative grade point average percentage.

In contrast, academic factors had a notable impact. University affiliation significantly influenced the TAM mean score (P<.001; Table 4), TAM acceptance categories (P<.001; Figure 1), and the 5 constructs mean scores (P<.001; Figure 2). CGPA% (Table 3) was positively correlated with the total TAM score (P<.001), while academic level was not significantly associated with TAM acceptance categories (P=.26).

Table . Mean total score of physiotherapy students' acceptance of artificial intelligence–powered chatbots across universities (N=1066).

University	Values, n	Mean score (SD)	Range	*P* value
Fatima College of Health Sciences	107	3.78 (0.70)	1.0 - 5.0	<.001
King Abdulaziz University	67	3.77 (0.97)	1.0 - 5.0	<.001
October 6 University	200	3.73 (0.79)	1.6 - 5.0	<.001
Beirut Arab University	200	3.66 (0.62)	1.4 - 5.0	<.001
Beni-Suef University	200	3.60 (0.76)	1.1 - 5.0	<.001
Modern University for Technology and Information	200	3.50 (0.85)	1.0 - 5.0	<.001
Al-Zaytoonah University of Jordan	92	3.03 (0.91)	1.0 - 5.0	<.001
Total	1066	3.59 (0.81)	1.0 - 5.0	<.001

Figure 1. Distribution of physiotherapy students' acceptance categories of AI-powered chatbots across universities (N=1066; *P*<.001). AI: artificial intelligence.

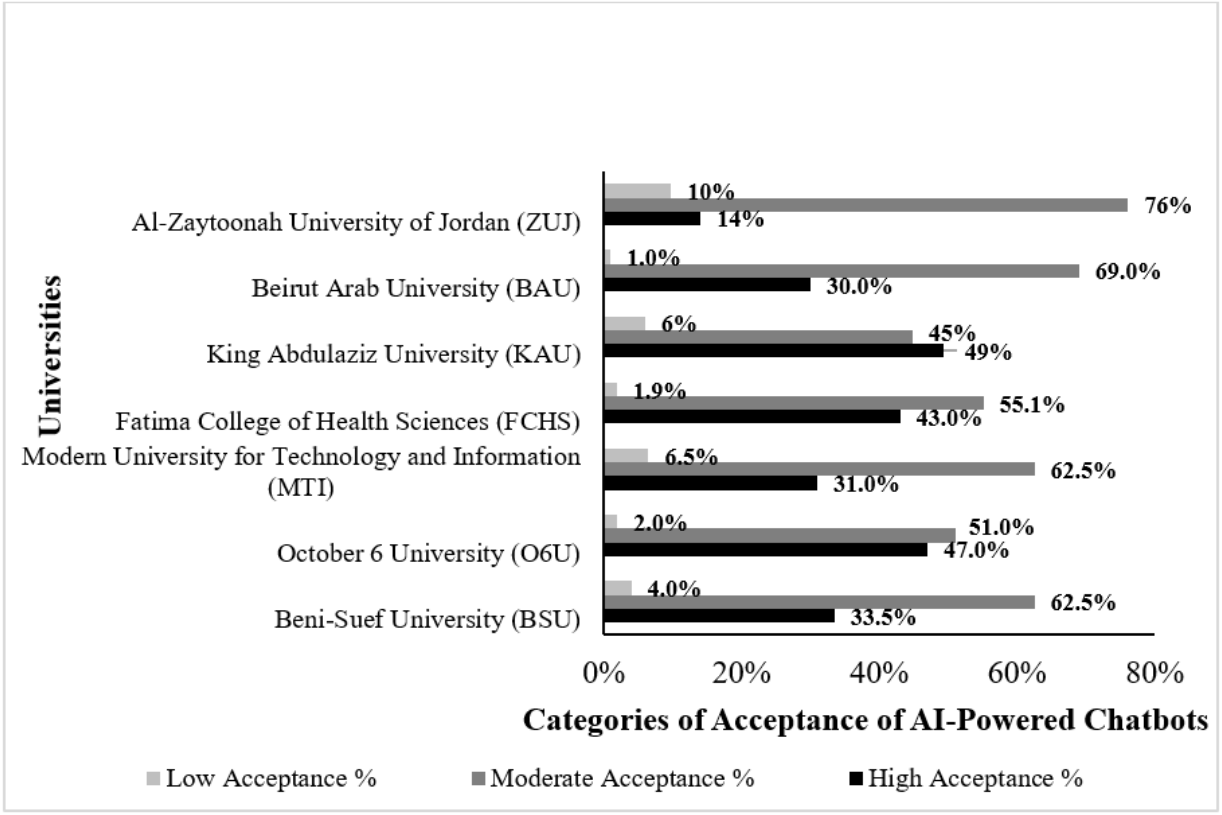
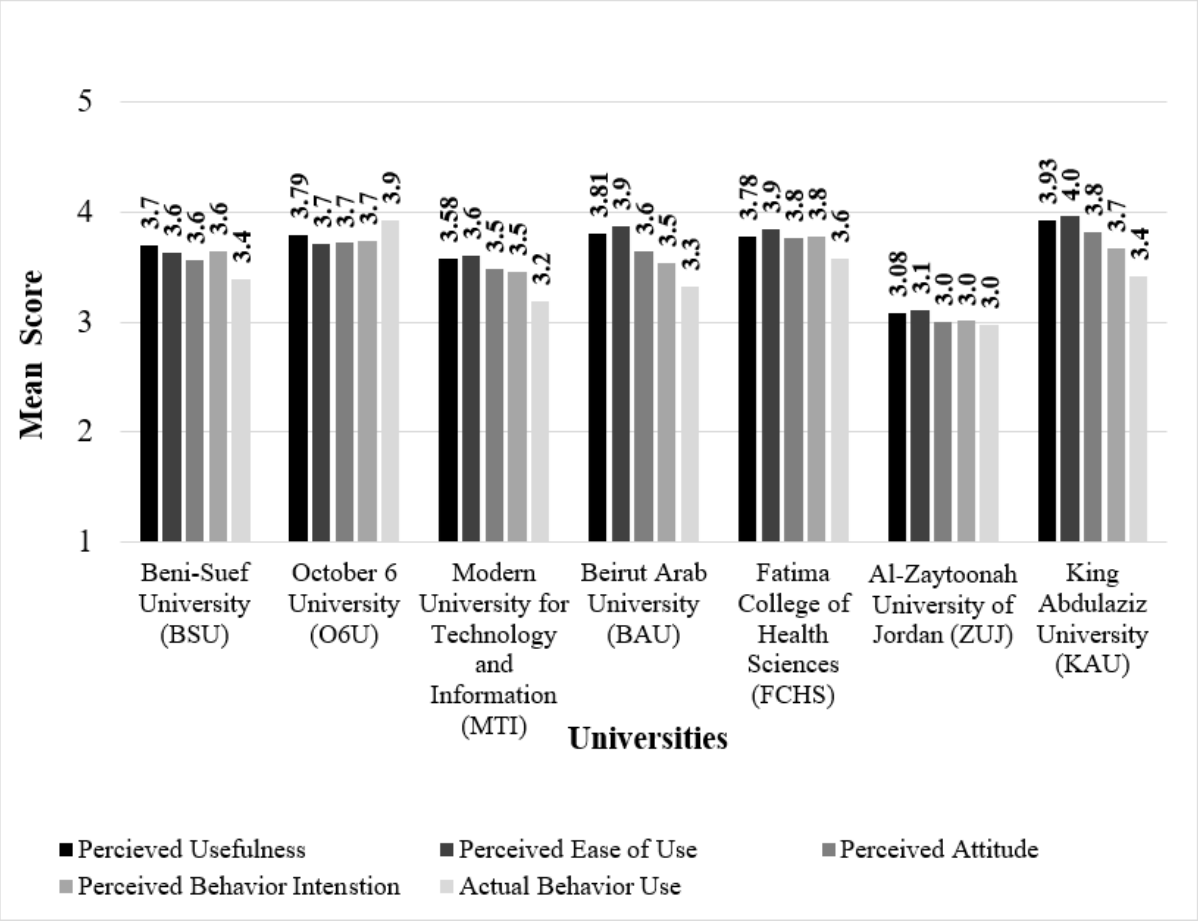


Figure 2. Physiotherapy students' perception of usefulness, ease of use, attitude, behavior, and actual behavior use of AI-powered chatbots mean scores across universities (N=1066 for all and n=586 for actual behavior use; *P*<.001). AI: artificial intelligence.



Technological factors also showed strong effects. A significant positive correlation was found between prior experience with AI-powered tools beyond chatbots and TAM scores (*P*<.001; Table 2). Additionally, current users of AI-PCs had significantly higher TAM scores (mean 3.89, SD 0.81) than nonusers (mean 3.48, SD 0.79; *P*<.001).

In summary, university affiliation, CGPA%, prior technological experience, and current use were the key factors influencing PT students' acceptance of AI-PCs. These variables were compared across institutions to further explore their impact.

University Affiliation and PT Students' Acceptance of AI-PCs

Significant differences in TAM total scores were observed across universities (*P*<.001; Table 4). ZUJ had the lowest score (mean 3.03), and FCHS the highest (mean 3.78). Post hoc Tukey analysis revealed that ZUJ's score was significantly lower than all other universities (*P*<.001), while FCHS did not significantly differ from other universities.

TAM acceptance category also varied significantly by universities (*P*<.001; Figure 1). KAU had the highest proportion of students in the high-acceptance category (n=33, 49.3%), whereas ZUJ had the lowest (n=13, 14.1%) and the highest proportion of students in the low-acceptance category (n=9, 9.8%).

Universities also differed significantly in TAM constructs' scores (*P*<.001; Figure 2 and Table 5). ZUJ had significantly lower PU, PEU, attitude, and BI scores than all other universities (*P*<.001; *P*=.001 for BI vs MTI). It also had the lowest, though nonsignificant, ABU score (2.98). KAU had the highest nonsignificant PU (3.93), PEU (3.96), and attitude (3.82) scores. FCHS had the highest nonsignificant BI score (3.78), while O6U had the highest ABU score (3.93), significantly exceeding all other universities except FCHS (*P*=.42).

Table . Summary of technology acceptance model (TAM) constructs mean scores of artificial intelligence (AI)–powered chatbot across universities (N=1066).

TAM constructs of AI-powered chatbots	Overall mean score out of 5	*P* value	Highest score university	Lowest score university	Significant Tukey post hoc test	*P* value
Perceived usefulness	3.69	<.001	KAU[a]	ZUJ[b]	ZUJ is lower than all other universities	<.001
Perceived ease of use	3.68	<.001	KAU	ZUJ	ZUJ is lower than all other universities	<.001
Perceived ease of use	3.68	<.001	KAU	ZUJ	MTI[c] is lower than BAU[d]	.03
Attitude	3.58	<.001	KAU	ZUJ	ZUJ is lower than all other universities	<.001
Behavior intention	3.57	<.001	FCHS[e]	ZUJ	ZUJ is lower than all other universities	<.001 and .001
Behavior intention	3.57	<.001	FCHS	ZUJ	MTI is lower than O6U[f]	.03
Actual behavioral use	3.40	<.001	O6U	ZUJ	O6U is higher than all other universities except FCHS	<.001

[a]KAU: King Abdulaziz University.

[b]ZUJ: Al - Zaytoonah University of Jordan.

[c]MTI: Modern University for Technology and Information.

[d]BAU: Beirut Arab University.

[e]FCHS: Fatima College of Health Sciences.

[f]O6U: October 6 University.

CGPA% Across Universities

CGPA% differed significantly among students in the academic levels 2-5 (*P*<.001; Table 6), as level 1 students lacked CGPA data. Post hoc test showed that ZUJ had the lowest CGPA% across all universities (*P*<.001), while BSU recorded the highest. BSU's CGPA% was significantly higher than ZUJ, BAU, FCHS (*P*<.001 for each), and MTI (*P*=.002) and higher, though nonsignificant, than KAU (*P*=.10) and O6U (*P*=.10). Additionally, BAU had significantly lower CGPA% than KAU (*P*<.001), MTI (*P*=.03), and O6U (*P*=.001). FCHS also had significantly lower CGPA% than BSU, KAU (*P*<.001 each), MTI (*P*=.04), and O6U (*P*=.003).

Table . Comparison of physiotherapy students' mean cumulative grade point average percentage (CGPA%) and previous technological experience across universities.

Universities	CGPA%			Previous technological experience (1 - 5 scale)		
	Students, n	Mean (SD)	*P* value	Students, n	Mean (SD)	*P* value
Al-Zaytoonah University of Jordan	92	67.2 (17.5)	<.001	92	2.4 (1.6)	<.001
Beirut Arab University	163	76.3 (14.0)	<.001	200	3.3 (1.4)	<.001
Beni-Suef University	171	86.0 (10.6)	<.001	200	3.1 (1.4)	<.001
Fatima College of Health Sciences	80	75.4 (10.4)	<.001	107	3.5 (1.3)	<.001
King Abdulaziz University	65	84.9 (7.5)	<.001	67	3.5 (1.5)	<.001
Modern University for Technology and Information	186	80.6 (12.8)	<.001	200	3.1 (1.5)	<.001
October 6 University	167	82.2 (14.0)	<.001	200	2.8 (1.6)	<.001
Total	924	79.7 (14.0)	<.001	1066	3.1 (1.5)	<.001

Previous Technological Experience Across Universities

PT students' prior experience with AI tools beyond chatbots varied significantly across universities (*P*<.001; Table 6). ZUJ students had the lowest prior experience, significantly lower than all other universities (*P*<.001). Additionally, MTI students had significantly less prior experience than FCHS students (*P*=.04).

Current Use of AI-PCs Across Universities

PT students' use of AI-PCs differed significantly across universities (*P*<.001). KAU students reported the highest use rate (n=56, 83.6%), while FCHS had the lowest (n=34, 31.8%; Table 7).

Table . Physiotherapy students' use of artificial intelligence (AI) chatbots across universities (N=1066).

Use of AI chatbots	Beni-Suef University, n (%)	October 6 University, n (%)	Modern University for Technology and Information, n (%)	Fatima College of Health Sciences, n (%)	King Abdulaziz University, n (%)	Beirut Arab University, n (%)	Al - Zaytoonah University of Jordan, n (%)	Total, n (%)
Users	117 (58.5)	89 (44.5)	129 (64.5)	34 (31.8)	56 (83.6)	120 (60)	41 (44.6)	586 (55)
Nonusers	83 (41.5)	111 (55.5)	71 (35.5)	73 (68.2)	11 (16.4)	80 (40)	51 (55.4)	480 (45)
Total	200 (100)	200 (100)	200 (100)	107 (100)	67 (100)	200 (100)	92 (100)	1066 (100)

Key Predictors of AI-PCs' Acceptance Among PT Students

A multiple linear regression was conducted to identify predictors of AI-PCs' acceptance among PT students. To meet regression analysis requirements, university affiliation was converted to dummy variables (ZUJ as the reference group), and current use was coded numerically. The model was statistically significant ($F_{9,914}$=32.33; *P*<.001), with an R^2 of 0.24, indicating that 24% of the variance in TAM score was explained. Significant predictors included prior experience with AI-PCs (B=0.23; *P*<.001) and university affiliation. Students from all universities scored significantly higher than those from ZUI, with the strongest effects seen at O6U (B=0.59; *P*<.001), FCHS (B=0.44; *P*<.001), and BAU (B=0.42; *P*<.001). CGPA% and current use were not significant predictors (*P*=.23 and *P*=.10, respectively; Table 8).

Table . Regression analysis results: predictor factors of physiotherapy students' acceptance of artificial intelligence–powered chatbots (N=1066).[a]

Predictor	B (unstandardized coefficient)	SE	β (standardized coefficient)	*t* test (*df*)	*P* value
Constant	2.33	0.15	0.00	15.90 (2)	<.001
Beni-Suef University (BSU_dummy)	0.36	0.10	0.17	3.63 (2)	<.001
October 6 University (O6U_dummy)	0.59	0.10	0.27	6.00 (2)	<.001
Modern University for Technology and Information (MTI_dummy)	0.27	0.10	0.13	2.82 (2)	<.001
Beirut Arab University (BAU_dummy)	0.40	0.10	0.19	4.16 (2)	<.001
Fatima College of Health Sciences (FCHS_dummy)	0.44	0.11	0.15	3.86 (2)	<.001
King Abdulaziz University (KAU_dummy)	0.45	0.12	0.14	3.65 (2)	<.001
Prior experience	0.23	0.02	0.43	13.93 (2)	<.001
CGPA%[b]	0.0	0.0	0.04	1.21 (2)	.23
Use	0.00	0.05	0.00	0.01 (2)	.99

[a]Al Zaytoonah University of Jordan is the reference group.

[b]CGPA%: cumulative grade point average percentage.

Regression Equations

Full Model (All Predictors)

TAM score=2.33+0.00 (use)+0.23 (prior experience)+0.36 (BSU)+0.27 (MTI)+0.42 (BAU)+0.44 (FCHS)+0.45 (KAU)+0.59 (O6U)+0.00 (CGPA%)

Simplified Model (Significant Predictors Only)

TAM score=2.33+0.23 (prior experience)+0.36 (BSU)+0.27 (MTI)+0.42 (BAU)+0.44 (FCHS)+0.45 (KAU)+0.59 (O6U)

(Note: To calculate a student's score, assign "1" to their university and "0" to all others.)

Previous experience emerged as the strongest predictor (standardized β=.43), followed by O6U (β=0.27), indicating that students' prior experience with AI tools had a larger impact on their acceptance score than university affiliation.

Discussion

Principal Findings

PT students demonstrated a moderate acceptance of AI-PCs. University affiliation, CGPA%, prior technological experience, and current use were identified as external factors influencing students' acceptance. However, the regression model confirmed that students' prior technological experience is the strongest predictor.

AI-PCs' use is steadily increasing in university education, with students' acceptance playing a key role in their adoption. For PT students, the clinical and practical demands of their curriculum present unique challenges to AI-PCs' acceptance. This study addressed the need to assess undergraduate PT students' acceptance of AI-PCs in their studies and identify personal, academic, and technological factors that influence it.

The multisite design enhances validity and generalizability, offering a broad view of AI-PCs' acceptance among PT students. Including all academic levels captures diverse perceptions, while a mean CGPA% of 79.7% reflects an average profile, minimizing academic performance–related bias.

Despite the rigorous nature of PT programs—with their intensive practical and clinical components—PT students still showed a moderate level of acceptance toward AI-PCs. This suggests a continued openness to AI-PCs, even within a rigorous academic environment. A study was conducted in South Korea among medical students and physicians, and it revealed that although physicians were cautious about the use of AI-PCs, particularly ChatGPT, in guiding patients' treatment, students had a positive perception of using ChatGPT for guiding treatment and medical education [31], which supports the concept that students might be more open to using advanced technology.

TAM constructs' scores provided deeper insight into the PT students' acceptance of AI-PCs. All scores fell within the moderate range, with PU and PEU scoring the highest means, indicating that PT students perceived AI-PCs as both useful and easy to use. However, slightly lower scores for attitude and BI suggest that while students acknowledge these benefits, their enthusiasm and intent to adopt AI-PCs are still developing. The lowest score for ABU highlights limited real-world use, reinforcing the study's overall finding of moderate acceptance.

In line with TAM's 3-stage process [21], external factors play a crucial role in shaping students' acceptance of AI-PCs. The

first stage includes PU and PEU, which reflect students' perceptions of usefulness and ease of use [26,27]. These influence attitude, which may substitute for BI [21,28], forming the second stage. Together, PU, PEU, attitude, and BI predict ABU [23]. Personal, academic, and technological factors influence students' perceptions, attitudes, and adoption behaviors toward AI technologies [6] such as AI-PCs. Among all the external factors, predictors, prior experience with AI-powered tools and applications beyond chatbots, and university affiliation had the strongest predictive influence on students' acceptance of AI-PCs.

Older students might be expected to show greater acceptance of AI-PCs due to increased exposure through academic progression and peer networks. Conversely, younger students, particularly Generation Z, could also be more accepting, having grown up immersed in digital technology. Similarly, male students are often presumed to show higher acceptance due to greater interest in technology and gaming. However, the findings did not support these assumptions, as age, sex, and academic level showed no significant correlation with AI-PCs' acceptance. Several hypotheses may explain this. First, institutional integration of AI tools may provide equal exposure to all students, minimizing demographic differences. Second, although Gen Z is highly technology competent, their adoption of AI-PCs may depend on perceived academic relevance. Third, sex-based assumptions about technology enthusiasm may not translate into actual academic tool use.

Students with greater exposure to AI tools demonstrated higher acceptance of AI-PCs, a logical outcome, as familiarity builds confidence. For these students, chatbots were a natural extension of technologies they already use, making adoption more intuitive. Supporting this, Horowitz et al [32] found that individuals with more familiarity and expertise in AI were more likely to support autonomous technologies than those with limited understanding, suggesting that experience enhances acceptance of new technologies.

University affiliation emerged as a key predictor of AI-PCs' acceptance. ZUJ recorded the significantly lowest total TAM score and the significantly lowest scores in 4 of the 5 constructs (except ABU). It also had the smallest proportion of students in the high-acceptance category and the highest proportion in the low-acceptance group, raising the question: What explains ZUJ's lower acceptance?

One possible factor is CGPA%, as ZUJ had the lowest CGPA%. However, CGPA% alone is unlikely to explain the outcome. For example, BSU, despite having the highest CGPA%, ranked the third lowest in TAM and the fourth lowest in the proportion of high-acceptance students. This suggests that academic performance is not a decisive factor in acceptance.

Prior technological experience, however, appears more influential. ZUJ students had the lowest prior technological experience across all universities, which aligns with their low acceptance rate. While prior experience was notably higher in the 3 universities with better acceptance (FCHS, KAU, and O6U), it was also relatively low in universities with low acceptance, such as BSU and MTI.

Although academic level was not significantly correlated with acceptance, most participating ZUJ students were in levels 1 to 3, as nearly all level 4 students were nonnationals and thus excluded. Additionally, ZUJ's PT department was established only 4 years ago. Together, these factors likely contributed to students' limited prior technological experience, and consequently, their lower acceptance of AI-PCs.

In contrast, examining the current use, ZUJ showed a moderate percentage of users, like O6U, the only university with a significantly higher score in one of the constructs. Meanwhile, FCHS demonstrated the lowest percentage of students currently using AI-PCs. If current use was a key factor in acceptance, both FCHS and O6U would be expected to have a high use rate, which was not the case. These patterns suggest that current use is not a reliable predictor of acceptance.

Hence, linear regression analysis was essential to clarify which external factors significantly predict PT students' acceptance of AI-PCs. Although the model explained 24% of the variance, it identified prior technological experience as the strongest predictor, confirming earlier assumptions about its importance. The analysis also confirmed that CGPA% and current use were not significant predictors and were excluded from the regression equation. In addition to prior experience, university affiliation emerged as an external predictor, though with a smaller effect size. These results confidently attribute ZUJ's low acceptance to its students' limited prior technological experience.

While the findings highlight university affiliation as a key predictor of acceptance, this study did not elaborate on the detailed cultural characteristics of studentship within each country or institution. Cultural norms, instructor-student dynamics, technological infrastructure, and pedagogical styles can certainly influence students' attitudes toward innovation. However, including a comprehensive cultural profile for each country or institution would have lengthened the paper unnecessarily. To reduce cultural bias and isolate academic and technological factors, only national students were recruited. This approach served to eliminate variability related to cross-national cultural differences, providing a more homogeneous sample to assess institutional influence on AI chatbot acceptance.

Interestingly, construct-level results revealed 2 cases where students' ABU did not align with their PU or PEU. These discrepancies suggest gaps between perceptions and actual use of AI-PCs or the influence of external factors on behavior. Supporting this, a study of 399 students in Hong Kong revealed that while attitudes toward AI technologies were generally positive, actual use remained limited, highlighting a gap between favorable perception and limited actual use of AI chatbots in education [33]. Another example is a study conducted in the United Arab Emirates among 265 recently graduated medical students, which reported an overall positive attitude and optimism toward the future of AI in medical education and health care but also revealed limited use of AI-PCs [34]. Similarly, a national survey of 693 medical students across 57 Chinese universities reported that only 28.7% used AI chatbots for studying, despite 91% acknowledging their usefulness for accessing medical information, reinforcing the disconnect

between intention and practice [35]. Specifically, O6U demonstrated the highest ABU across all universities, despite reporting only moderate PU and PEU scores. In other words, O6U students used AI-PCs in their studying, even though they perceived them as only moderately useful or easy to use. Several external factors may explain this pattern. One possibility is a university mandate, such as a course assignment, which may have driven use regardless of personal perception. Another is social influence, where encouragement from peers or faculty members promotes greater use. A third explanation could be habituation exposure; students may have become accustomed to using the AI-PCs through repeated encounters, resulting in higher ABU despite lower PU and PEU.

In contrast, KAU students showed the opposite pattern, high PU and PEU scores but lower ABU. This highlights the common gap between what students perceive and what they do, especially when technology is available but not fully integrated into the learning process. In this case, positive perceptions of AI-PCs were not sufficient to drive consistent use in academic practices. A large-scale study, which analyzed responses from over 34,000 college students, supports this pattern, showing that favorable views of educational technology did not always result in increased use, particularly when the tools were not embedded in instruction [36].

Findings from O6U and KAU highlight the complexity of technology acceptance and the influence of external factors beyond individual perceptions. They illustrate how such external factors can sometimes override personal perceptions in driving actual behavior, warranting further exploration.

Limitation

While this study examined selected personal, academic, and technological factors, many other factors may contribute to PT students' acceptance of AI-PCs. These include students' learning style, personality traits, academic load, technology anxiety, self-regulated learning skills, language proficiency, faculty support, and access to physical resources such as internet connectivity.

Further Work

While this study focused on personal, academic, and technological factors influencing PT students' AI-PCs, several broader contextual variables warrant exploration in future research. These include deeper investigations into the culture of studentship, both within and across countries in the region, such as how students interact with instructors and technology, their expectations, and the learning environments they are embedded in. Additionally, variations in the structure of PT programs, such as 4- versus 5-year tracks, patterns of matriculation, and progression rates, could provide further insight into students' readiness and attitudes toward adopting educational technologies. Future studies may also benefit from considering national regulations and licensing frameworks, which shape the continuum of PT education, including the integration of internship, postgraduate training, and continuing education. These elements, although beyond our study's scope, are crucial for developing a comprehensive understanding of technology acceptance within the broader context of PT education.

Conclusions

PT students demonstrated a moderate acceptance of AI-PCs. University affiliation, CGPA%, prior technological experience, and current use were identified as external factors influencing students' acceptance. However, the regression model confirmed that prior technological experience is the strongest predictor of AI-PCs' acceptance among PT students.

Acknowledgments

The authors acknowledge students participated in this study. ChatGPT was used to improve manuscript clarity, structure, and language. All content, design, data analysis, and interpretation are the authors' original work.

Conflicts of Interest

None declared.

References

1. Mir MM, Mir GM, Raina NT, et al. Application of artificial intelligence in medical education: current scenario and future perspectives. J Adv Med Educ Prof 2023 Jul;11(3):133-140. [doi: 10.30476/JAMP.2023.98655.1803] [Medline: 37469385]
2. Schwalbe N, Wahl B. Artificial intelligence and the future of global health. Lancet 2020 May 16;395(10236):1579-1586. [doi: 10.1016/S0140-6736(20)30226-9] [Medline: 32416782]
3. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022 Jan;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
4. Wang F, Preininger A. AI in health: state of the art, challenges, and future directions. Yearb Med Inform 2019 Aug;28(1):16-26. [doi: 10.1055/s-0039-1677908] [Medline: 31419814]
5. Lowe SW. The role of artificial intelligence in Physical Therapy education. Bull Fac Phys Ther 2024;29(1):13. [doi: 10.1186/s43161-024-00177-8]
6. Labadze L, Grigolia M, Machaidze L. Role of AI chatbots in education: systematic literature review. Int J Educ Technol High Educ 2023;20(1):56. [doi: 10.1186/s41239-023-00426-1]

7. Dharani M, Jyostna J, Sucharitha E, Likitha R, Manne S. Interactive transport enquiry with AI chatbot. Presented at: 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS); May 13-15, 2020; Madurai, India p. 1271-1276. [doi: 10.1109/ICICCS48265.2020.9120905]
8. Veras M, Dyer JO, Rooney M, Barros Silva PG, Rutherford D, Kairy D. Usability and efficacy of artificial intelligence chatbots (ChatGPT) for health sciences students: protocol for a crossover randomized controlled trial. JMIR Res Protoc 2023 Nov 24;12:e51873. [doi: 10.2196/51873] [Medline: 37999958]
9. Crompton H, Burke D. Artificial intelligence in higher education: the state of the field. Int J Educ Technol High Educ 2023;20(1):1-22. [doi: 10.1186/s41239-023-00392-8]
10. Al-Sharafi MA, Al-Emran M, Iranmanesh M, Al-Qaysi N, Iahad NA, Arpaci I. Understanding the impact of knowledge management factors on the sustainable use of AI-based chatbots for educational purposes using a hybrid SEM-ANN approach. Interact Learn Environ 2023 Dec 15;31(10):7491-7510. [doi: 10.1080/10494820.2022.2075014]
11. Lee D, Yeo S. Developing an AI-based chatbot for practicing responsive teaching in mathematics. Comput Educ 2022 Dec;191:104646. [doi: 10.1016/j.compedu.2022.104646]
12. Mohd Rahim NI, A. Iahad N, Yusof AF, A. Al-Sharafi M. AI-based chatbots adoption model for higher-education institutions: a hybrid PLS-SEM-neural network modelling approach. Sustainability 2022;14(19):12726. [doi: 10.3390/su141912726]
13. Laurillard D. Rethinking University Teaching: A Conversational Framework for the Effective Use of Learning Technologies, 2nd edition: Routledge; 2002. [doi: 10.4324/9781315012940]
14. Pérez JQ, Daradoumis T, Puig JMM. Rediscovering the use of chatbots in education: a systematic literature review. Comput Appl Eng Educ 2020 Nov;28(6):1549-1565. [doi: 10.1002/cae.22326]
15. Wollny S, Schneider J, Di Mitri D, Weidlich J, Rittberger M, Drachsler H. Are we there yet?—A systematic literature review on chatbots in education. Front Artif Intell 2021;4:654924. [doi: 10.3389/frai.2021.654924] [Medline: 34337392]
16. Deng X, Yu Z. A meta-analysis and systematic review of the effect of chatbot technology use in sustainable education. Sustainability 2023;15(4):2940. [doi: 10.3390/su15042940]
17. Okonkwo CW, Ade-Ibijola A. Chatbots applications in education: a systematic review. Comput Educ Artif Intell 2021;2:100033. [doi: 10.1016/j.caeai.2021.100033]
18. Bonsu EM, Baffour-Koduah D. From the consumers' side: determining students' perception and intention to use ChatGPT in Ghanaian higher education. J Educ Soc Multicult 2023 Jun 1;4(1):1-29. [doi: 10.2478/jesm-2023-0001]
19. Almahri FAJ, Bell D, Merhi M. Understanding student acceptance and use of chatbots in the united kingdom universities: a structural equation modelling approach. Presented at: 2020 6th International Conference on Information Management (ICIM); Mar 27-29, 2020; London, United Kingdom. [doi: 10.1109/ICIM49319.2020.244712]
20. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep;13(3):319. [doi: 10.2307/249008]
21. Davis FD. User acceptance of information technology: system characteristics, user perceptions and behavioral impacts. Int J Man Mach Stud 1993 Mar;38(3):475-487. [doi: 10.1006/imms.1993.1022]
22. Bizzo E. Acceptance and resistance to e-learning adoption in developing countries: a literature review. Ensaio Aval Polít Públicas Educ 2021;30(115):458-483. [doi: 10.1590/s0104-403620220003003342]
23. Marikyan D, Papagiannidis S. Technology acceptance model. In: Papagiannidis S, editor. TheoryHub Book: Newcastle University; 2022. URL: https://open.ncl.ac.uk/theories/1/technology-acceptance-model/ [accessed 2025-07-28]
24. Granić A, Marangunić N. Technology acceptance model in educational context: a systematic literature review. Br J Educ Technol 2019 Sep;50(5):2572-2593. [doi: 10.1111/bjet.12864]
25. Esiyok E, Gokcearslan S, Kucukergin KG. Acceptance of educational use of AI chatbots in the context of self-directed learning with technology and ICT self-efficacy of undergraduate students. Int J Hum Comput Interact 2025 Jan 2;41(1):641-650. [doi: 10.1080/10447318.2024.2303557]
26. Kulviwat S, Bruner II GC, Neelankavil JP. Self-efficacy as an antecedent of cognition and affect in technology acceptance. J Consum Mark 2014 May 6;31(3):190-199. [doi: 10.1108/JCM-10-2013-0727]
27. Lewis JR. Comparison of four TAM item formats: effect of response option labels and order. J Usability Stud 2019;14(4):224-236. [doi: 10.5555/3542805.3542809]
28. Ajzen I. The theory of planned behaviour: reactions and reflections. Psychol Health 2011 Sep;26(9):1113-1127. [doi: 10.1080/08870446.2011.613995] [Medline: 21929476]
29. Jette AM. Health services research in the 21st century. Phys Ther 2019 Mar 1;99(3):255-257. [doi: 10.1093/ptj/pzz002] [Medline: 30690534]
30. Eysenbach G. Correction: improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;14(1):e8. [doi: 10.2196/jmir.2042]
31. Alkhaaldi SMI, Kassab CH, Dimassi Z, et al. Medical student experiences and perceptions of ChatGPT and artificial intelligence: cross-sectional study. JMIR Med Educ 2023 Dec 22;9:e51302. [doi: 10.2196/51302] [Medline: 38133911]
32. Horowitz MC, Kahn L, Macdonald J, Schneider J. Adopting AI: how familiarity breeds both trust and contempt. AI & Soc 2024 Aug;39(4):1721-1735. [doi: 10.1007/s00146-023-01666-5]
33. Chan CKY, Hu W. Students' voices on generative AI: perceptions, benefits, and challenges in higher education. Int J Educ Technol High Educ 2023;20(1):43. [doi: 10.1186/s41239-023-00411-8]

34. Tangadulrat P, Sono S, Tangtrakulwanich B. Using ChatGPT for clinical practice and medical education: cross-sectional survey of medical students' and physicians' perceptions. JMIR Med Educ 2023 Dec 22;9:e50658. [doi: 10.2196/50658] [Medline: 38133908]
35. Tao W, Yang J, Qu X. Utilization of, perceptions on, and intention to use AI chatbots among medical students in China: national cross-sectional study. JMIR Med Educ 2024 Oct 28;10:e57132. [doi: 10.2196/57132] [Medline: 39466038]
36. Correia AP, Good K. Exploring the relationships between student perceptions and educational technology utilization in higher education. J Educ Technol Dev Exch 2023;16(1):92-107. [doi: 10.18785/jetde.1601.05]

Abbreviations

ABU: actual behavioral use
AI: artificial intelligence
AI-PC: artificial intelligence–powered chatbot
BAU: Beirut Arab University
BI: behavioral intention
BSU: Beni-Suef University
CGPA: cumulative grade point average
CGPA%: cumulative grade point average percentage
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
FCHS: Fatima College of Health Sciences
KAU: King Abdulaziz University
MTI: Modern University for Technology and Information
O6U: October 6 University
PEU: perceived ease of use
PT: physiotherapy
PU: perceived usefulness
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
TAM: technology acceptance model
ZUJ: Al - Zaytoonah University of Jordan

Edited by D Chartash; submitted 26.04.25; peer-reviewed by A Farghaly, M Sayed; revised version received 24.06.25; accepted 24.06.25; published 19.08.25.

Please cite as:
El-Sobkey SB, Kelini KI, ElKholy M, Abdeldayem T, Abdallah M, Mohamed DAA, Fawzy A, Ahmed YF, El Khatib A, Khalid H, Shaik BB, Anjos A, Alharbi MD, Fathy K, Takey K
Acceptance of AI-Powered Chatbots Among Physiotherapy Students: International Cross-Sectional Study
JMIR Med Educ 2025;11:e76574
URL: https://mededu.jmir.org/2025/1/e76574
doi:10.2196/76574

Paradox of AI in Higher Education: Qualitative Inquiry Into AI Dependency Among Educators in Palestine

Anas Ali Alhur[1]; Zuheir N Khlaif[2], PhD; Bilal Hamamra[3], PhD; Elham Hussein[4], PhD

[1]Imam Abdulrahman Bin Faisal University, Damam, Saudi Arabia

[2]Educational Sciences, An-Najah National University, Old Campus Street, Nablus, West Bank, Palestinian Territory

[3]Education Language, An-Najah National University, Nablus, West Bank, Palestinian Territory

[4]Al Ain University, Al Ain, United Arab Emirates

Corresponding Author:
Zuheir N Khlaif, PhD
Educational Sciences, An-Najah National University, Old Campus Street, Nablus, West Bank, Palestinian Territory

Abstract

Background: Artificial intelligence (AI) is increasingly embedded in medical education, providing benefits in instructional design, content creation, and administrative efficiency. Tools like ChatGPT are reshaping training and teaching practices in digital health. However, concerns about faculty overreliance highlight risks to pedagogical autonomy, cognitive engagement, and ethics. Despite global interest, there is limited empirical research on AI dependency among medical educators, particularly in underrepresented regions like the Global South.

Objective: This study focused on Palestine and aimed to (1) identify factors contributing to AI dependency among medical educators, (2) assess its impact on teaching autonomy, decision-making, and professional identity, and (3) propose strategies for sustainable and responsible AI integration in digital medical education.

Methods: A qualitative research design was used, using semistructured interviews (n=22) and focus group discussions (n=24) involving 46 medical educators from nursing, pharmacy, medicine, optometry, and dental sciences. Thematic analysis, supported by NVivo (QSR International), was conducted on 15.5 hours of transcribed data. Participants varied in their frequency of AI use: 45.7% (21/46) used AI daily, 30.4% (14/46) weekly, and 15.2% (7/46) monthly.

Results: In total, 5 major themes were identified as drivers of AI dependency: institutional workload (reported by >80% [37/46] of participants), low academic confidence (noted by 28/46, 60%), and perfectionism-related stress (23/46, 50%). The following 6 broad consequences of AI overreliance were identified: Skills Atrophy (reported by 89% [41/46]): educators reported reduced critical thinking, scientific writing, and decision-making abilities. Pedagogical erosion (35/46, 76%): decreased student interaction and reduced teaching innovation. Motivational decline (31/46, 67%): increased procrastination and reduced intrinsic motivation. Ethical risks (24/46, 52%): concerns about plagiarism and overuse of AI-generated content. Social fragmentation (22/46, 48%): diminished peer collaboration and mentorship. Creativity suppression (20/46, 43%): reliance on AI for content generation diluted instructional originality. Strategies reported by participants to address these issues included establishing boundaries for AI use (n=41), fostering hybrid intelligence (n=37), and integrating AI literacy into teaching practices (n=39).

Conclusions: While AI tools can enhance digital health instruction, unchecked reliance risks eroding essential clinician competencies. This study identifies cognitive, pedagogical, and ethical consequences of AI overuse in medical education and highlights the need for AI literacy, professional development, and ethical frameworks to ensure responsible and balanced integration.

(JMIR Med Educ 2025;11:e74947) doi:10.2196/74947

KEYWORDS

AI dependency; generative AI; procrastination; AI reliance; hybrid intelligence

Introduction

Background

The rapid incorporation of artificial intelligence in education (AIED) has significantly changed the educational landscape, providing unique chances for tailored teaching and learning experiences, instant feedback, and enhanced efficiency [1]. Consequently, institutions across the globe are increasingly using AI-powered tools, including generative models like ChatGPT, to customize the educational process, streamline administrative duties, and reduce faculty workload [2]. Despite important benefits, stakeholders have concerns regarding overdependence on AI, which is simply defined as students' and instructors' uncritical acceptance of AI-generated material

[3]. In addition, overreliance on AIED hinders cognitive involvement, productive critical thinking and decision making, and long-term skill growth [4].

While literature on the topic of AIED is abundant, it primarily focuses on the gains of integrating AI in the educational setting [5]. Ethical, cognitive, and professional challenges related to AIED, on the other hand, are far less explored [6,7]. Furthermore, most of the available research on overreliance on AIED is investigated in relation to students, exploring how excessive dependence on AI can negatively impact their ability to think critically and independently and to perform self-directed learning [4,8]. Other student-related studies have linked AI reliance to psychological factors such as perfectionism, impulsivity, and the need for immediate cognitive relief, drawing comparisons to internet and social media dependence [9].

Empirical research investigating AI dependency among educators is still meager [7], and most of the existing studies explore how AI enhances educators' productivity and decision-making [10]. Risks and challenges pertaining to AIED, on the other hand, are less explored. In addition, most research on AIED has been conducted in Western or Chinese contexts, with limited cross-cultural perspectives [11]. This study has three aims: to examine the causes of educators' overreliance on AI, to assess its impact on teaching autonomy and pedagogical identity, and to identify challenges and strategies for responsible AI use. Focusing on Palestinian higher education, the research offers insights to guide policy, faculty development, and ensure that AI enhances—rather than replaces—educators' professional expertise.

This study conceptualizes the "AI Paradox" in higher education, where AI's benefits in streamlining academic work are offset by concerns over skill atrophy, ethics, and autonomy. It explores how faculty navigate these competing effects, highlighting the coexistence of innovation and emerging professional vulnerabilities.

Research Questions

The research questions are as follows (1) What key factors contribute to AI dependency among educators in higher education? (2) How does AI dependency impact educators' teaching autonomy, decision-making, and professional identity? (3) What strategies help educators maintain a balanced use of AI while maintaining pedagogical creativity and instructional effectiveness?

Theoretical Foundation of the Study

The Interaction of Person-Affect-Cognition-Execution (I-PACE Model; Figure 1), developed by [12], provides a comprehensive framework for understanding problematic technology use, including AI dependency. The I-PACE model explains how individual traits, emotional responses, cognitive processes, and behaviors interact to shape technology reliance. Its 4 components are: person-level predispositions (personality traits and cognitive styles), affect (emotional states like stress or anxiety), cognition (decision-making and impacts on autonomy), and execution (behavioral outcomes such as deskilling). This framework helps analyze how and why AI dependency develops among educators and its effects on autonomy and professional agency.

Figure 1. I-PACE Model. I-PACE: Interaction of Person-Affect-Cognition-Execution.

Previous Studies Using I-PACE

Research using I-PACE has demonstrated how individual predispositions, emotional states, and cognitive processes interact to influence excessive reliance on digital tools. For instance, studies on internet and social media addiction have found that individuals with high neuroticism**,** impulsivity, or perfectionism are more likely to develop problematic usage patterns, often as a coping mechanism for stress [12]. Similarly, research on AI reliance in education suggests that educators and students with low self-efficacy in digital literacy or high performance expectations tend to offload cognitive tasks to AI, reducing engagement in critical thinking and problem-solving [8,13]. Studies show that institutional pressures and workload demands can drive educators to rely on AI for efficiency, often at the expense of pedagogical autonomy. Applying the I-PACE model, these studies highlight how occasional AI use can evolve into habitual dependence, shaped by psychological, cognitive, and behavioral factors.

Reliance on AI in education may lead to unhealthy dependence on AI tools that can reduce student autonomy, hinder skill acquisition, and elevate the risk of academic stress [8,14]. Although tactical application of AI is likely to improve teaching effectiveness and student engagement, students could still delegate crucial cognitive activities to technology rather than participating in reflective or creative thought [9,15,16]. When consistently depending on AI for writing tasks, research projects, or making decisions, students jeopardize their chances of developing critical thinking, creativity, and self-directed learning techniques [4,16]. While AI can serve as a valuable complement to teaching strategies, instructors must ensure that embracing technology does not impede human-centered exploration [17].

The psychological implications of reliance on AI are complex, involving personality characteristics, emotional control, and cognitive distortions [8]. According to [7] and [8], neuroticism, self-critical perfectionism, and impulsivity elevate the chances of excessive dependence on AI-driven tools. Neuroticism, marked by increased responses to stress, frequently leads individuals to pursue immediate technological solutions for anxiety or performance apprehensions [18]. Similarly, self-critical perfectionists might be drawn to the flawless results that AI is thought to produce, seeking to prevent errors and the unease linked to learning through trial and error [19]. Instead of developing study habits, impulsive learners might resort to AI whenever they face challenges, perpetuating a pattern of dependence on external sources. This AI dependency diminishes creativity and independent thought [8].

Scholars use the I-PACE model to explore how personal traits—like perfectionism and impulsivity—interact with emotional states and cognitive functions and sometimes result in maladaptive technology use [8,12]. This model illustrates how some individuals deal with academic pressure by consistently looking for digital shortcuts. Depending on digital shortcuts compromises an individual's need for autonomy, competence, and relatedness, which, according to the Basic Psychological Needs theory, are crucial for one's well-being and motivation [20]. If AI reliably addresses unmet needs—like providing quick responses that protect learners from the unease of ambiguity—an excessive dependence might develop, reducing the intrinsic motivation essential for authentic learning and developmental progress [8,13]. Simultaneously, determining which aspects of AI design—such as user interface, adaptive feedback, or customization features—enhance or reduce dependency could assist developers in creating more ethically responsible platforms [21].

Adverse academic emotions—like anxiety and frustration—can greatly heighten reliance on AI, hindering students' motivation and their capacity to manage their own learning [22]. As unaddressed psychological needs build up, students might seek emotional support or affirmation from AI [8,23]. Performance expectations further amplify this dependence: when students think AI significantly enhances their grades or educational results, they might overrate the technology's necessity [24]. This corresponds with extensive research on performance expectancy, an element demonstrated to be essential in the adoption and ongoing usage of new technologies [25]. Eventually, this persistent trend of pursuing immediate solutions or emotional comforts causes learners to neglect crucial cognitive activities like integrating information or contemplating mistakes [8].

Methods

Overview

This study used a qualitative research approach to explore the factors influencing AI dependency and the consequences of AI dependency in medical education. A mixed approach of semistructured interviews and focus group sessions was used to gain insight into the experiences and perspectives of faculty members and to explore how using AI tools in their practice affects them as educators [26].

To ensure participant anonymity while distinguishing between data sources, a consistent coding system was used. "E" refers to individual educators who participated in semistructured interviews, followed by a number identifying the participant (eg, E5=Educator 5). **"**EFG" refers to participants in focus group discussions (eg, EFG14). This system enabled pattern recognition across data types while maintaining confidentiality.

Semistructured Interviews

Semistructured interviews were conducted with 22 faculty members to explore factors driving AI dependency and its consequences in medical education [27]. Participants were purposively sampled from multiple universities and represented diverse disciplines (eg, nursing, pharmacy, optometry, medicine, and dental sciences). Variation in roles, teaching experience, and frequency of AI use was considered to ensure maximum variation and minimize overlap in perspectives due to shared institutional or disciplinary contexts. The interview protocol (Multimedia Appendix 1) was informed by the study's theoretical framework. Each interview lasted 25 - 35 minutes and was audio-recorded with consent. The interview guide (Multimedia Appendix 1) was informed by the I-PACE framework and developed to explore individual, cognitive, and emotional dimensions of AI use. This guided the formulation of probes around motivation, attitudes, and usage patterns.

Focus Group Sessions

To complement the interviews, 4 focus group sessions (n=24) were held with educators not involved in the interviews. This method enabled reflection and interaction, encouraging participants to build on shared experiences [28-31]. Purposive sampling ensured diversity across institutions, disciplines, and levels of AI engagement. All participants held teaching, research, or administrative roles. Focus group prompts (Multimedia Appendix 2) were generated from preliminary interview analysis to deepen insight into shared institutional experiences. Sessions were audio-recorded and lasted approximately 1 hour. In total, the study involved 46 participants.

Data Collection

In total, 22 semistructured interviews (25-35 min each) were audio-recorded with consent. Questions explored participants' use of generative AI in teaching, research, and administration, including frequency, delegated tasks, and changing reliance. Follow-ups addressed motivations, benefits, and challenges in maintaining professional autonomy alongside AI integration.

Although the average duration of these interviews may appear relatively short for qualitative research, this was a deliberate methodological decision based on both logistical and conceptual considerations. Participants were full-time faculty members in medical and health sciences disciplines with heavy teaching, research, and clinical responsibilities. To respect their limited availability and ensure meaningful engagement, a focused and well-structured interview protocol was designed, aligned with the I-PACE model. This protocol was pilot-tested and refined to elicit conceptually rich responses within a concise time frame.

Despite the shorter duration, the interviews yielded dense, thematically robust narratives. To enrich data depth and validate emerging themes, 4 focus group discussions—each approximately 1 hour—were also conducted. These provided opportunities for reflection, triangulation, and deeper elaboration of core issues. Importantly, thematic saturation was achieved after 18 interviews, with subsequent interviews and focus groups confirming the stability of the thematic framework. Thus, while longer interviews might have allowed for further elaboration, the approach adopted ensured a balance between feasibility and qualitative rigor, without compromising the credibility or conceptual depth of the data.

To supplement these insights, four 1-hour focus groups (2 in person and 2 online) with 24 participants were held, facilitated by 2 researchers using prompts from interview analysis. Discussions explored AI usage patterns, reliance, and impacts on teaching and decision-making. All sessions were audio-recorded, providing broader perspectives that complemented the interview findings.

To ensure conceptual depth and data adequacy, recruitment and data collection continued until thematic saturation was achieved. Saturation was defined as the point at which no new themes or insights emerged during ongoing data analysis. This was observed after conducting 18 interviews, with subsequent interviews and focus group discussions confirming the stability and completeness of the thematic framework.

In this study, AI dependency is defined as overreliance on generative AI tools for academic tasks, leading to reduced cognitive engagement, professional autonomy, or pedagogical creativity. This contrasts with effective AI integration, which involves strategic, ethical use that maintains human oversight. Semistructured interviews probed these distinctions, exploring both the benefits and drawbacks of AI use and assessing impacts on teaching autonomy, motivation, and professional identity.

Data Analysis

An inductive thematic analysis was conducted following Braun and Clarke's 6-phase framework (2006) to analyze data. The dataset comprised 12.25 hours of interview recordings and 3.25 hours of focus group sessions. All recordings were transcribed verbatim and verified through participant member checking to ensure accuracy.

Coding was performed using NVivo, with initial codes iteratively refined into subthemes and themes aligned with study objectives and relevant literature. Data collection and analysis were concurrent, and thematic saturation was reached after the 18th interview, confirmed by later focus groups.

To strengthen credibility, the study used methodological triangulation with interviews, focus groups, and artifacts such as anonymized AI-generated lesson plans and teaching notes. Analyzing these artifacts alongside participant reports provided contextual depth and validated emerging themes. Preserving cultural and linguistic nuances, coding and theme development were conducted in Arabic by bilingual researchers, with themes collaboratively translated into English. Backward translation ensured accuracy and contextual depth, while team discussions resolved discrepancies, maintaining conceptual equivalence and data integrity.

Participant responses were categorized using standardized terms: "most" (35/46, >75%), "many" (50% [23/46,] - 75% [35/46]), and "some" (23/46, <50%). Where participants expressed multiple perspectives, responses were coded under all relevant categories to reflect the complexity of AI integration in practice. This structured and multisource approach ensured the robustness, transparency, and trustworthiness of the findings.

Trustworthiness

To ensure trustworthiness, the study systematically addressed the 4 key criteria of qualitative rigor: credibility, confirmability, dependability, and transferability. Credibility was strengthened through triangulation of data sources—semistructured interviews, focus groups, and AI-generated teaching artifacts—which provided multiple perspectives on the phenomenon. In addition, member checking was conducted by inviting participants to review selected transcripts and thematic summaries to validate interpretations.

The study used a hybrid coding approach, integrating both deductive and inductive strategies. A preliminary codebook based on the I-PACE model guided the initial round of coding, allowing the researchers to identify theoretically grounded patterns. Simultaneously, the analysis remained open to emergent inductive themes not captured by the model, thereby

ensuring that the coding process was both conceptually informed and data-driven.

Confirmability was supported through independent coding by 2 researchers on a subset of transcripts (20/60, 30%), followed by peer debriefing and consensus-building sessions to refine code definitions and resolve discrepancies. While no formal interrater reliability statistic (eg, Cohen Kappa) was calculated, a shared audit trail was maintained to document analytic decisions, code evolution, and theme development. Final coding was conducted collaboratively using NVivo software.

Dependability was reinforced by clearly outlining data collection and analysis procedures, adhering to Braun and Clarke's 6-phase thematic analysis protocol, and using NVivo for systematic coding and data management. Transferability was supported through purposive sampling for maximum variation across discipline, institutional context, and AI use frequency, along with thick descriptions of participants' backgrounds and contextual settings. Linguistic and cultural nuances were preserved through bilingual coding and backward translation, ensuring integrity across languages. Collectively, these procedures establish a rigorous foundation for the trustworthiness of the study's findings.

Ethical Considerations

The study was conducted following approval (approval number Med. Dec. 2024/29) from the institutional review board at An Najah National University. All participants provided informed consent, receiving clear explanations of the study's purpose, voluntary participation, and confidentiality. Consent was indicated by participation, and participants were reminded of their right to withdraw at any time, ensuring ethical compliance and autonomy.

Researchers' Background and Positionality

This study was conducted by an interdisciplinary research team composed of scholars from diverse academic backgrounds and institutions across different countries. The team includes experts in medical education, English language studies, educational technology, and qualitative research methodology, with shared experience in the integration of AI tools in teaching, learning, and research. Members of the team are affiliated with universities in the Middle East and North Africa region, providing a transnational perspective on the evolving role of AI in higher education. The team's varied disciplinary expertise and familiarity with a wide range of learner populations—including undergraduate students, clinical educators, and faculty across Science, Technology, Engineering, and Mathematics (STEM) and humanities disciplines—helped ensure a nuanced and contextually grounded interpretation of the data. This interdisciplinary orientation aligns with the study's broader aim to explore AI dependency as a multidimensional and cross-sectoral phenomenon.

Results

Overview

To contextualize the qualitative findings, Table 1 presents the demographic characteristics of the 46 participants who took part in the study, including faculty from diverse medical and health sciences disciplines such as nursing, pharmacy, optometry, medicine, and dental sciences. Participants varied in gender, age, teaching experience, frequency of AI use, and level of AI familiarity, ensuring a wide range of perspectives on the phenomenon of AI dependency in academic practice.

Table . Demographic information about the participants in the semistructured interviews and focus group sessions.

Variable	Frequency (%)
Sex	
Male	20 (43.5)
Female	26 (56.5)
Age (years)	
25 - 35	5 (10.9)
36 - 45	15 (32.6)
46 - 55	18 (39.1)
56+	8 (17.4)
Frequency of gen AI[a] use	
Daily	21 (45.7)
Weekly	14 (30.4)
Monthly	7 (15.2)
Occasionally	4 (8.7)
Teaching experience	
1 - 5 years	16 (34.8)
6 - 10 years	6 (13.0)
11 - 15 years	15 (32.6)
16+ years	9 (19.6)
AI familiarity level	
Low	19 (41.3)
Moderate	16 (34.8)
High	11 (23.9)
Medical science discipline	
Nursing	8 (17.4)
Pharmacy	8 (17.4)
Optometry	9 (19.6)
Doctor of Medicine	7 (15.2)
Medical Laboratory Sciences	6 (13.0)
Doctor of Dental	8 (17.4)

[a]AI: artificial intelligence.

Building on this diverse dataset, the findings are structured and interpreted through the lens of the I-PACE model. This framework enabled a systematic mapping of themes across 4 interrelated components: institutional and individual pressures were aligned with person variables; anxiety and motivation reflected the affect dimension; cognitive offloading and creativity suppression were linked to cognition; and habitual overuse of AI tools corresponded to execution functions. Together, these findings illustrate how psychological, cognitive, and behavioral mechanisms interact to shape educators' dependency on AI technologies in higher education.

Research question #1*:* What key factors contribute to AI dependency among educators in higher education?

The first research question aimed to explore the factors influencing AI dependency among educators. The analysis revealed that AI dependency is shaped by institutional, psychological, cognitive, technological, and individual factors. Table 2 presents the coding book for the first research question.

Table . Coding book for the reasons for AI dependency research question.

Theme and subtheme	Quotation
Institutional factors	
Heavy workload and time constraints	*Without AI, I would spend hours grading assignments and drafting course materials. Now, I can do it in minutes*. [E5]
Institutional expectation of using AI[a]	*Our university expects us to use AI, but there's little guidance on how to do it effectively*. [E21]
Lack of institutional guidance	*There are no clear policies on how much AI use is appropriate, so I just use it however I think best*. [E8]
Psychological factors	
Anxiety and performance pressure	*Everyone around me is using AI, and I don't want to be left behind.* [E2]
Perfectionism and fear of errors	*I feel like I've lost my ability to draft a paper from scratch. I always turn to AI first*. [E6]
Cognitive factors	
Cognitive offloading	*I used to brainstorm lesson plans myself. Now, I ask AI, and it gives me something within seconds*. [E15]
Loss of pedagogical creativity	*I used to create my own case studies and interactive activities. Now, I just modify what AI generates*. [E17]
Technological factors	
Ease of access and automation	*Why spend hours writing something when AI can give me a great draft in seconds?* [E21]
Lack of AI literacy	*I know AI isn't perfect, but I don't always know how to verify the accuracy of what it generates*. [E18]
Feeling powerful and capable	*With AI, I can complete tasks that would have taken me hours in just a few minutes*. [E3]
Individual factors	
Academic self-efficacy	*AI helps me refine my ideas, making my work more professional and well-structured*. [E6]
Academic reputation	*I rely on AI in writing research papers due to my reputation and publication pressure*. [E9]
High performance expectations	*There's a lot of pressure to produce high-quality work, so I use AI to meet expectations*. [E11]
Low academic confidence	*I don't always trust my writing skills, so AI gives me the reassurance I need to finalize my work*. [E4]
Lack of scientific research engagement	*I rarely engage in deep scientific research anymore because AI provides quick summaries*. [E13]

[a]AI: artificial intelligence.

Institutional Factors

Heavy Workload and Time Constraints

Many participants viewed AI as essential for managing academic workloads, highlighting its role in reducing time spent on repetitive tasks like grading, report writing, and content creation.

> *Without AI, I would spend hours grading assignments and drafting course materials. Now, I can do it in minutes, which helps me keep up with my workload.* [E5]

Participants noted that institutional pressure to publish and show teaching effectiveness drives greater AI adoption. However, some—especially in the social sciences—found that unfamiliarity with AI increased their workload and added stress.

> *It's a double-edged sword—AI helps me finish tasks faster, but I sometimes feel I need more time to understand it, which increases the time I need to finish writing the tasks.* [E19]

Lack of Institutional Guidance

A recurring concern was the absence of institutional policies, support, or structured training for integrating AI responsibly. This lack of formal guidance often left participants unsure of how to use AI tools effectively and ethically.

> *There are no clear policies on how much AI use is appropriate, so I just use it however I think best.* [E8]

A few also expressed concern that future policy changes might disrupt their current dependency on AI.

> *I depend on AI now, but if my university decides to impose restrictions, I might struggle to adjust back to traditional methods.* [E5]

Psychological Factors

Anxiety and Performance Pressure

Several participants pointed out that AI helps them manage anxiety related to teaching and research expectations. Some also reported feeling pressured to stay updated with AI advancements and integrating AI tools in order to be adequately tech-savvy.

> *Everyone around me is using AI, and I don't want to be left behind. I feel like I need to use it to stay relevant.* [E2]

Nevertheless, a few participants argued that AI dependency sometimes increases stress, particularly when the accuracy of AI is uncertain.

> *I depend on AI, but at times, I worry that I might be using incorrect or biased information without realizing it.* [E9]

Perfectionism and Fear of Errors

Some participants described themselves as perfectionists, adding that AI helps them ensure accuracy in research, lesson planning, and assessment design. In addition, AI increases their confidence as it helps them refine content and eliminate errors.

> *I use AI to check everything. I don't want to submit anything that isn't polished and error-free.*

Conversely, a few expressed concern over the way AI dependency negatively impacts their academic and scholarly confidence.

> *I feel like I've lost my ability to draft a paper from scratch. I always turn to AI first.* [E 6]

Cognitive Factors

Cognitive Offloading

Most participants acknowledged that AI has become their default tool for information retrieval and content creation, reducing the effort required for critical thinking and problem-solving.

> *I used to brainstorm lesson plans myself. Now, I ask AI, and it gives me something within seconds.* [E15]

However, some educators worried that relying on AI too frequently might weaken their cognitive engagement with their work.

> *I wonder if I'm losing my ability to think critically because AI does the thinking for me.* [E 10]

Loss of Pedagogical Creativity

A few pointed out that AI-generated content makes teaching more convenient but limits their creativity in designing course materials.

> *I used to create my own case studies and interactive activities. Now, I just modify what AI generates.* [E17]

On the other hand, some educators mentioned that AI enhances their creativity by providing new perspectives and ideas they would not have considered otherwise.

> *AI gives me different angles to approach a topic, which actually improves my lesson planning.* [E22]

Technological Factors

Ease of Access and Automation

Many participants indicated that AI tools are readily available and easy to use, making them convenient for performing academic tasks. According to these participants, AI helps them save time, generate ideas, and organize content efficiently.

> *Why spend hours writing something when AI can give me a great draft in seconds?* [E21]

However, a few pointed out that the ease of access makes it tempting to rely on AI for everything, leading to unintentional dependency.

> *The more I use AI, the harder it becomes to work without it. It's like having a calculator for everything.* [E20]

Lack of AI Literacy

Some participants admitted that they lack a deep understanding of AI's inner workings**,** leading them to unquestioningly trust its outputs.

> *I know AI isn't perfect, but I don't always know how to verify the accuracy of what it generates.* [E18]

A few educators expressed concern that inadequate AI literacy might make them overly dependent on AI-generated insights.

> *I sometimes use AI without thinking critically because I assume it knows better than I do.* [E1]

Feeling Powerful and Capable

Some educators pointed out that AI empowers them and allows them to produce high-quality content more efficiently.

> *With AI, I can complete tasks that would have taken me hours in just a few minutes. It makes me feel more capable.* [E3]

Moreover, a few were concerned that this sense of control is misleading, as they might be overestimating AI's reliability.

> *AI makes me feel smarter, but I sometimes wonder if I'm just relying on it too much without questioning its accuracy.* [E 3]

Individual Factors

Based on the coding book (Table 2), individual factors include academic reputation, high performance expectation, low academic self-efficacy, and lack of scientific research engagement.

Academic Reputation

Most participants reported that academic reputation is a key driver of their reliance on AI, particularly for scientific research. One prominent researcher shared that, because of his demanding

teaching and research schedule, he relies on AI to assist with writing.

Academic Self-Efficacy

Some participants mentioned that using AI enhances their confidence in their academic abilities.

> *AI helps me refine my ideas, making my work more professional and well-structured.* [E6]

Other participants expressed concerns that relying on AI might affect their academic reputation, as AI-generated content could be perceived as lacking originality.

> *I worry that using AI too much might make others question the authenticity of my work.* [E7]

Low Academic Confidence

A few participants admitted that AI serves as a tool for procrastination, allowing them to delay tasks while still producing quick results when needed.

> *"I put off writing papers because I know AI can help me generate content at the last minute."*

On the other hand, some educators maintained that AI use compensates for their low academic confidence, making them feel more capable in completing their work.

> *"I don't always trust my writing skills, so AI gives me the reassurance I need to finalize my work."*

Research question #2: How does AI dependency impact educators' teaching autonomy, decision-making, and professional identity?

Findings from the second research question revealed that AI dependency has complex and varied consequences for educators. While participants acknowledged AI's efficiency and convenience, they also raised concerns about its negative effects. More than 30 subthemes emerged, categorized into 6 main themes: skills atrophy, pedagogical erosion, motivation decline, ethical and integrity risk, social fragmentation, and creativity suppression. Table 3 details these subthemes, which are outlined in the following sections.

Table . Consequences of AI dependency.

Main theme	Subtheme
Skills atrophy	• Loss of critical thinking • Weakened problem-solving abilities • Diminished writing and research skills • Reduced analytical skills • Loss of decision-making abilities • Reduced ability to synthesize independently • Weakened judgment
Motivational decline	• Over-reliance on AI[a] for simple tasks • Procrastination • Reduced motivation • Increased laziness • Reduced individual initiatives
Pedagogical erosion	• Erosion of pedagogical autonomy • Reduced active engagement with students • Weakened academic mentorship • Reduced self-confidence • Perceived professional inadequacy
Ethical and integrity risks	• Increased plagiarism rate • Increased copyright infringement • Increased academic misconduct • Diminished human responsibility • Increased misleading information
Social fragmentation	• Affects students' emotional state negatively • Diminished social development • Reduced human interaction • Decreased collaboration among humans
Creativity suppression	• Homogenization of knowledge production • Restrict creativity • Restrict information seeking

[a]AI: artificial intelligence.

Skills Atrophy

In this study, skills atrophy refers to the gradual decline of previously acquired skills and competencies due to lack of use or engagement. Related subthemes include weakened thinking skills, reduced problem-solving abilities, and diminished analytical synthesis.

All participants in the interviews and a few of the participants in the focus group discussions reported a decline in their skills in writing scientific research due to their frequent use of AI in scientific research. In addition, most participants reported that frequent reliance on AI has reduced their engagement in deep critical thinking.

> *I used to spend time reading multiple sources and forming my own analysis. Now, I just ask AI to summarize everything for me, and I take that as my starting point.* [EFG18]

Several participants in the interviews admitted that instead of independently evaluating and synthesizing information, they now trust AI-generated summaries and responses.

> *I don't challenge myself as much anymore. If AI gives me a structured answer, I don't always question it, I just go with it.* [EFG26]

In addition, several participants stated that using AI applications negatively affects their decision-making abilities, making them more reliant on automated solutions.

> *When I come across a difficult question in my field, my first instinct is to ask AI instead of thinking through possible solutions myself.* [EFG36]

A few participants pointed out that while AI speeds up the problem-solving process, it sometimes limits their ability to engage with complex academic challenges.

> *AI gives great answers quickly, but I worry that I'm losing my ability to navigate difficult academic discussions without its help.* [EFG13]

Most participants reported that AI dependence has weakened their academic writing and research skills. Some found it difficult to start writing without AI-generated drafts after extended use, while others rarely conduct extensive literature reviews themselves.

> *I used to write everything from scratch, but now I start with an AI-generated outline. I wonder if I'm losing my ability to structure my thoughts clearly on my own.* [EFG25]

A few participants mentioned that AI-generated summaries are convenient but may lead to a superficial understanding of academic topics.

> *Instead of reading full papers, I rely on AI to summarize them. I get the key points, but I feel like I'm missing the depth of the discussion.* [E17]

Motivational Decline

In the context of this study, motivation decline refers to the gradual reduction in an individual's drive, enthusiasm, or willingness to engage in various academic activities. Subthemes include procrastination, weaker intrinsic motivation, reduced individual initiatives, and overreliance on AI.

Several participants mentioned that AI has encouraged procrastination, as they know they can complete tasks quickly with AI assistance.

> *Before AI, I would start preparing my lectures well in advance. Now, I just rely on AI to generate material the night before.* [EFG28]

A few educators pointed out that while AI helps them meet deadlines, it sometimes leads to rushed and poor-quality work.

> *I meet my deadlines, but sometimes I feel like my work isn't as thoughtful as it used to be because I rely on AI to speed things up.* [E15]

A few educators mentioned that AI has made them more efficient but acknowledged that it can be tempting to take shortcuts.

> *I'm more productive with AI, but I also recognize that I use it to avoid thinking through certain problems myself.* [EFG13]

Pedagogical Erosion

In this study, pedagogical erosion refers to the gradual decline in the effectiveness, relevance, and innovation of teaching practices over time. Subthemes categorized under this theme include perceived professional inadequacy, erosion of pedagogy autonomy, reduced engagement with students, weakened academic mentorship, and reduced self-confidence.

Some participants pointed out that AI efficiently helps them improve the quality of their teaching material.

> *"AI gives me a strong starting point for my lessons. I still personalize them, but it definitely saves me time."*

According to a few participants, interactions with students became less personal since integrating AI into their workflow.

> *I used to spend more time giving individualized feedback. Now, I use AI-generated comments, and I feel like I'm not engaging with my students as much.* [E11]

On the other hand, some educators argued that AI allows them to focus on more meaningful discussions instead of wasting time on repetitive tasks.

> *AI handles the routine work, so I can dedicate more time to having deeper discussions with my students.* [E20]

Many participants admitted that AI has made them doubt their own expertise, as they often feel that AI-generated content is superior to their own in structure and insight.

> *Sometimes I feel like AI writes better than I do. It's unsettling to think that I might not be as good as I used to be.* [EFG35]

Similarly, some mentioned that they feel they must use AI particularly when completing research or complex writing tasks.

> *"If AI was suddenly unavailable, I don't know how I'd manage my workload. I've started to rely on it too much."*

However, a few pointed out that AI boosts their confidence by helping them refine their work.

> *AI isn't replacing my skills—it's just giving me an extra layer of support to make my work better.* [EFG14]

A few educators stressed that AI can and should be used wisely without hindering professional competence.

> *AI is a tool, not a replacement. As long as we use it strategically, we can avoid becoming too dependent.* [E16]

Ethical and Integrity Risks

Ethical and integrity risks in this study refer to the challenges that arise when AI influences academic practices. Subthemes that emerged under this category include increased plagiarism rate and copyright infringement, increased academic misconduct, decreased human responsibility, and increased misleading information.

Some participants expressed concerns that AI facilitates plagiarism and reduces students' accountability for their work.

> *It's so easy for students to copy-paste AI-generated responses without truly engaging with the material.* [E7]

Others pointed out that AI tools blur the lines of originality and authorship, making it difficult to differentiate between human and machine-generated work.

> *I worry that students are submitting AI-generated essays without understanding the concepts themselves.* [E12]

A few highlighted the fact that AI-generated content sometimes lacks accountability, as it can produce misleading or biased information.

> *AI sometimes generates convincing but factually incorrect statements, and students don't always verify them.* [EFG21]

On the other hand, some faculty members believe AI can be ethically leveraged if students are taught responsible AI use.

> *We should integrate AI literacy into our curriculum so students learn how to use these tools without compromising integrity.* [E26]

Social Fragmentation

Social fragmentation in this research refers to the weakening of human connections and interactions due to AIED. Subthemes include negatively impacting students' emotional state, diminishing social development, reducing human interaction, and decreasing collaboration among humans.

Many participants noted that students are becoming more isolated as AI tools replace traditional peer interactions in collaborative assignments.

> *Group discussions are not as engaging anymore because students rely on AI instead of exchanging ideas with their peers.* [E9]

However, some suggested that AI could be used to enhance social learning if implemented thoughtfully.

> *If we integrate AI strategically—such as using it to facilitate discussions rather than replace them—it can actually strengthen collaboration.* [E30]

Creativity Suppression

Creativity suppression refers to the risk that AI may homogenize knowledge production, restrict creativity, and limit students' ability to seek information. Subthemes include homogenizing knowledge production, restricting creativity, and limiting search for information from diverse sources.

Many participants reported that AI-generated content tends to produce generic, standardized responses, leading to a decline in original thought.

> *Students' assignments are starting to look the same because they rely on AI-generated structures.* [EFG15]

Some educators emphasized that AI discourages thorough research, as students may settle for AI-generated summaries instead of exploring diverse sources.

> *Before AI, students would read multiple papers. Now, they just use AI to summarize, and I feel like they're missing out on critical engagement.* [E14]

Research question # 3: What strategies can help educators balance AI use while maintaining pedagogical creativity and instructional effectiveness?

The third research question explored strategies educators use to balance AI integration with traditional teaching while maintaining effectiveness and autonomy. Findings show that most view AI as a valuable tool but use strategies to prevent overreliance and preserve human-centered teaching, with some embracing its efficiency and others expressing concerns about its effects on critical thinking, creativity, and student-teacher interactions. Key themes are summarized below.

Establishing Clear Boundaries for AI Use: Selective Use of AI for Instructional Tasks

Most participants said they use AI selectively, viewing it as a support tool rather than a replacement for their instructional role. They emphasized that core teaching activities—mentoring, leading discussions, and assessing student progress—should remain human-led.

> *AI helps with structuring lesson plans and summarizing content, but I make sure that my role as an educator remains central in guiding discussions and engaging with students.* [E4]

Some participants pointed out that AI is particularly useful for administrative tasks and content organization but should not dictate teaching methods.

> *I let AI handle routine tasks like scheduling and summarizing, but when it comes to interactive*

> *teaching, I rely on my own expertise and instincts.* [EFG33]

A few participants expressed concerns that without defined limits, educators might unconsciously begin to depend too much on AI-generated content.

> *I initially used AI just for support, but over time, I found myself relying on it more than I intended. Now, I consciously limit my AI use to avoid dependency.* [EFG28]

Encouraging Critical Engagement With AI: Verifying AI-Generated Content

Many participants stated that they actively verify and refine AI-generated outputs before using them in teaching, noting that AI can produce misleading, biased, or overly simplified content. This requires educators to serve as content curators rather than passive users.

> *I never use AI-generated content without reviewing it first. It can be a great starting point, but I always fact-check and refine the materials to align with my course objectives.* [EFG31]

Some educators pointed out that they encourage students to critically analyze AI-generated responses rather than accepting them at face value.

> *I tell my students that AI is a tool, not an authority. They need to critique what it produces, question its assumptions, and think beyond the outputs.* [EFG32]

On the other hand, a few participants expressed concern that not all educators take the time to evaluate AI content carefully, which could lead to inaccuracies in teaching.

> *One of my worries is that some educators might blindly trust AI-generated materials, which could introduce errors or outdated information into their lessons.* [EFG17]

Balancing AI With Human-Centered Teaching

Hybrid Intelligence

Most participants stressed that hybrid intelligence—integrating human and machine intelligence—is vital to prevent overreliance on AI. They defined it as critically assessing AI outputs, promoting deeper reflection and independent analysis rather than passive acceptance.

Prioritizing Interactive and Discussion-Based Learning

Most participants reported intentionally designing courses to prioritize live discussions, debates, and student-driven learning to counterbalance AI-generated content, emphasizing that human interaction is essential for deep understanding and critical thinking.

> *AI is great for providing structured information, but meaningful learning happens when students engage in discussions, challenge ideas, and interact with their peers and instructors.* [E15]

Some participants mentioned that they use AI to supplement discussions by generating diverse perspectives, but they ensure that students analyze and debate those perspectives rather than passively accepting them.

> *I sometimes use AI to provide multiple viewpoints on a topic, but I make sure my students critically evaluate and compare them rather than just absorbing the information.* [EFG25]

However, a few participants pointed out that some educators struggle to balance AI integration with traditional teaching, leading to a more passive learning environment.

> *I've seen cases where AI-generated lectures replace interactive teaching, and I worry that students might become passive learners rather than active thinkers.* [EFG32]

Maintaining Pedagogical Creativity: Using AI as a Spark for Innovation, Not a Substitute for Creativity

Many participants reported that they leverage AI to generate new ideas for lesson planning but ensure that their personal creativity remains the driving force. They view AI as a brainstorming partner rather than a content creator.

> *I use AI to generate multiple lesson ideas, but I always customize them to fit my teaching style and my students' needs.* [E3]

Some educators pointed out that AI helps them explore innovative approaches to teaching but emphasized that human creativity is irreplaceable.

> *AI can suggest engaging activities, but it's my job to refine them and add the human element that makes learning meaningful.* [E22]

On the other hand, a few participants expressed concerns that excessive AI use might lead to a decline in originality if educators become overly reliant on AI-generated content.

> *I fear that if we depend too much on AI, we might lose the uniqueness of our teaching styles. That's why I try to balance AI use with my own creative input.* [EFG26]

Fostering Student AI Literacy and Ethical Use: Teaching Students to Use AI Responsibly

Some participants expressed that they incorporate AI literacy into their courses to help students use AI effectively while avoiding over-reliance. They believe that educators should guide students in understanding AI's strengths and limitations.

> *I teach my students how to use AI as a research tool but also emphasize that they must engage critically with the results rather than just accepting AI-generated answers.* [EFG32]

Most participants reported that they actively discourage students from using AI for academic shortcuts, instead encouraging them to use AI as a means of enhancing learning.

> *AI can help students brainstorm ideas, but they need to put in the effort to analyze and expand on those ideas instead of just submitting AI-generated content.* *[EFG34]*

However, a few participants pointed out that some students misuse AI for assignments, and educators must establish clear guidelines on ethical AI use.

> *We need to teach students that AI is a tool to assist learning, not a way to bypass intellectual effort.* *[EFG11]*

Continuous Professional Development and Peer Collaboration: Engaging in AI Training and Professional Learning Communities

Several participants mentioned that they actively seek professional development opportunities to improve their AI literacy and refine their AI integration strategies. They believe that educators must continuously learn to ensure that AI is used responsibly and effectively.

> *I regularly attend AI workshops to stay updated on best practices. The more I understand AI, the better I can integrate it into my teaching without becoming dependent on it.* [E5]

Some educators pointed out that they collaborate with colleagues to share AI integration strategies, discuss ethical concerns, and develop guidelines for responsible AI use.

> *We have faculty discussions on AI use where we exchange ideas on how to integrate AI without compromising teaching quality. These discussions help us find the right balance.* [EFG36]

On the other hand, a few participants expressed that some institutions lack adequate AI training programs, leaving educators to navigate AI integration on their own.

> *I wish there were more structured AI training for educators. Right now, we mostly figure it out through trial and error.* [EFG35]

Discussion

Overview

The recent study examines the drivers and consequences of AI dependency among medical and health sciences faculty in Palestinian higher education, offering critical insights into over-reliance on generative AI and strategies for sustainable integration. This study contributes to the growing body of literature on digital transformation in medical education by identifying multidimensional factors that influence educators' reliance on AI tools and their implications for pedagogical integrity and professional agency.

Principal Findings

Institutional pressures—particularly workload intensity and a lack of clear AI governance—were identified as primary catalysts for increased AI reliance. These findings mirror previous studies [4,10] highlighting how time-saving AI applications appeal to overburdened faculty. Participants described using AI to manage grading, feedback, and administrative documentation, often without structured support or guidance. The absence of institutional frameworks creates ambiguity, allowing overuse and reinforcing dependency—a concern consistent with the findings of [7].

Psychological and individual factors also emerged as significant influences. Echoing previous research [8,9,32], educators reported that anxiety, performance pressure, and perfectionism motivated AI use as a coping mechanism. High expectations for output quality and academic reputation intensified AI reliance, particularly in competitive or high-stakes disciplines. While some participants viewed AI as a confidence booster, others reported reduced academic self-efficacy and increased procrastination, reflecting problematic patterns of technology use.

At the cognitive level, many educators admitted to cognitive offloading, allowing AI to complete tasks once handled through personal reflection or pedagogical creativity. These practices align with concerns from [33,34] about diminished analytical and innovative capacity when AI is used uncritically. Participants noted that AI-generated lesson plans increased efficiency but limited deep thinking and originality, a paradox especially concerning in health education where critical reasoning and ethical judgment are essential.

Technological factors, especially ease of access and automation, facilitated habitual AI use, often unconsciously. Similar to findings on digital tool overuse [25], educators recognized both the benefits and risks of AI ubiquity. Lack of AI literacy compounded the issue: while participants used GenAI tools regularly, many admitted limited understanding of how algorithms operate or how to evaluate their outputs critically [17]. This blind trust not only undermines informed decision-making but also amplifies the risk of propagating misinformation or algorithmic bias in academic work.

Consequences of AI Dependency

This study identified 6 major consequences of AI overreliance: skills atrophy, motivational decline, pedagogical erosion, ethical risks, social fragmentation, and creativity suppression. Faculty reported declining engagement in original writing, idea generation, and academic rigor—findings echoed by [8,34,35]. In some cases, AI was used as a substitute for personal initiative, weakening academic identity and self-reliance.

Social fragmentation was also discussed as educators expressed concern about diminished student-teacher interaction and reduced peer collaboration. This is consistent with [36] noting how digital tools can isolate learners. Our findings suggest AI can support collaborative learning by matching students for group work and distributing roles equitably, potentially countering its fragmenting effects.

Participants frequently cited ethical risks such as plagiarism, misinformation, and authorship concerns, expressing uncertainty about acceptable boundaries for AI use in assessment and publication. These concerns align with calls for regular AI audits [34], ensuring educational technologies are ethically vetted and contextually appropriate.

Comparison to Prior Work

The findings reinforce and expand on theoretical models such as the I-PACE framework [12], which links personality traits,

affective responses, and cognitive control to problematic technology use. This study shows that in Palestinian higher education, AI dependency is a systemic issue shaped by resource constraints and institutional pressures, not just individual behavior. It highlights the interplay between personal motivations and structural limitations in AI use.

Limitations and Future Research

This study's findings are limited by its focus on Palestinian higher education, which may affect transferability to other contexts. Variations in resources, policies, and academic cultures could lead to different experiences with AI dependency. Future research should explore these issues in diverse settings to broaden understanding.

Second, the use of self-reported data from 46 faculty members may introduce social desirability bias, with participants potentially misrepresenting their AI use. Conducting and translating interviews from Arabic to English may also have affected the preservation of linguistic and cultural nuances. Future studies should consider bilingual coding teams and cultural consultants for greater interpretive accuracy.

Third, this study focused solely on educators' perspectives, excluding student voices. Future research should examine AI dependency among students to understand its effects on motivation, learning outcomes, and academic integrity.

Furthermore, while this study examined AI dependency broadly, it did not address disciplinary differences. Faculty in humanities, social sciences, and STEM may experience unique challenges and opportunities with AI. Future research should compare disciplines to uncover domain-specific dynamics of AI dependency.

Finally, this study provides a snapshot of current practices but does not assess the long-term effects of AI dependency on faculty roles, creativity, or knowledge production. Longitudinal and mixed methods research is needed to track changes over time and evaluate the impact of institutional interventions on pedagogical autonomy and professional integrity.

Theoretical and Practical Implications

This study contributes to theoretical discourse on technology adoption in higher education by framing AI dependency as a multidimensional construct shaped by institutional, psychological, cognitive, and technological factors. Building on the I-PACE model, the findings demonstrate how overreliance on generative AI aligns with patterns of problematic technology use, particularly among faculty navigating high academic demands and limited institutional support. The study reinforces and extends the I-PACE framework by showing how its four components—Person, Affect, Cognition, and Execution—interact dynamically within educational environments to shape AI dependency.

Under the Person component, both individual (eg, academic self-efficacy, perfectionism, low confidence) and institutional drivers (eg, workload pressures, unclear policies) were identified as predisposing factors for AI overuse. These findings extend I-PACE by highlighting that external academic structures—not just personal traits—can significantly shape technology dependency. The Affect domain was reflected in emotional responses such as anxiety, fear of inadequacy, and pressure to meet performance expectations. Rather than being purely internal, these emotions were often shaped by systemic academic stressors, suggesting that affective triggers in the I-PACE model must be understood within broader socio-institutional contexts.

In the Cognition domain, participants frequently described cognitive offloading and declining engagement in creative and reflective tasks. Notably, this study introduces "creativity suppression" as a cognitive outcome not emphasized in the original model, pointing to the need for an update to account for AI's influence on ideation and pedagogical originality. The Execution component was evident in behavioral outcomes such as diminished teaching autonomy, motivational decline, and habitual reliance on AI. These confirm I-PACE's focus on maladaptive behavior patterns but also reveal participants' active strategies—like hybrid intelligence and critical engagement—to retain pedagogical agency.

Overall, the study affirms the I-PACE model's value while proposing a broader interpretation that incorporates institutional, technological, and cultural factors specific to under-resourced higher education systems. By doing so, it deepens our understanding of how AI transforms academic identity, decision-making, and autonomy in the evolving landscape of digital education.

On a practical level, the study offers several actionable recommendations:

Faculty Development

Higher education institutions should invest in personalized professional development to enhance AI literacy, ethical awareness, and pedagogical resilience. Training should include modules on AI ethics, data privacy, and algorithmic bias, tailored to each discipline's needs. Institutions can also use AI to personalize learning pathways for educators based on their backgrounds and evolving instructional needs.

Institutional Policy Design

Institutions must establish clear and adaptive AI governance frameworks to guide responsible AI use among faculty and students. These frameworks should define acceptable use policies, clarify the boundaries of AI-generated content, and promote transparency in AI-supported academic work. To uphold ethical standards—especially in high-stakes fields like medical education—regular audits of AI tools should be implemented to assess potential biases, misinformation, and integrity risks. Moreover, developing such policies should be a collaborative effort involving faculty members, administrators, IT professionals, and ethicists to ensure both institutional trust and community-wide compliance.

Student–Teacher Interaction in AI Contexts

To reduce risks of pedagogical erosion and social fragmentation, institutions should set guidelines that preserve student–teacher engagement in AI-rich environments. Human–AI collaboration, reflective use of AI, and AI-facilitated group work can enhance accountability and peer interaction. Prioritizing feedback, mentorship, and dialogic teaching ensures AI remains a

supplement, supporting balanced integration and safeguarding academic integrity.

Conclusion

This study provides a comprehensive exploration of AI dependency among educators, highlighting the institutional, psychological, and cognitive factors that contribute to reliance on AI in academic practices. While AI offers significant benefits in terms of efficiency, knowledge management, and academic productivity, unchecked dependence can pose risks to pedagogical autonomy, critical thinking, and professional identity. The findings emphasize the importance of hybrid intelligence as a strategy to balance AI use, ensuring that technology enhances rather than replaces human intellectual engagement. The study's contributions to the theoretical discourse on AI adoption extend existing models by integrating the concept of AI dependency within a broader framework of professional agency and academic integrity. From a practical standpoint, these findings underscore the need for institutional policies, faculty training, and AI literacy initiatives that foster responsible AI use. Ultimately, this study advocates for a balanced approach where AI serves as a tool to augment human intelligence rather than replace it. Future research should continue to explore AI's evolving role in higher education, examining its long-term impact on faculty development, student learning, and knowledge creation in an increasingly AI-driven academic landscape.

Disclaimer

The authors use ChatGPT for proofreading of the introduction and organizing ideas in the literature review, all the authors take the responsibility of the accuracy of the content after proofreading and organizing the ideas of the introduction and literature review.

Data Availability

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

Authors' Contributions

The authors are responsible for all aspects of the research, including conceptualization, design, data collection, data analysis, and manuscript preparation. The authors have read and approved the final version of the manuscript.

Conflicts of Interest

The authors declared there is no conflict of interest

Multimedia Appendix 1
Semistructured protocol.
[DOCX File, 16 KB - mededu_v11i1e74947_app1.docx]

Multimedia Appendix 2
Focus group discussion guide.
[DOCX File, 16 KB - mededu_v11i1e74947_app2.docx]

References

1. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature New Biol 2023 Aug;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
2. Ansari AN, Ahmad S, Bhutta SM. Mapping the global evidence around the use of ChatGPT in higher education: a systematic scoping review. Educ Inf Technol 2024 Jun;29(9):11281-11321. [doi: 10.1007/s10639-023-12223-4]
3. Zhai C, Wibowo S, Li LD. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. Smart Learn Environ 2024;11(1):28. [doi: 10.1186/s40561-024-00316-7]
4. Hess BJ, Cupido N, Ross S, Kvern B. Becoming adaptive experts in an era of rapid advances in generative artificial intelligence. Med Teach 2024 Mar;46(3):300-303. [doi: 10.1080/0142159X.2023.2289844] [Medline: 38092006]
5. Creely E, Blannin J. Creative partnerships with generative AI. Possibilities for education and beyond. Think Skills Creat 2025 Jun;56:101727. [doi: 10.1016/j.tsc.2024.101727]
6. AlAli R, Wardat Y. Enhancing classroom learning: ChatGPT's integration and educational challenges. Intl J Rel 2024;5(6):971-985. [doi: 10.61707/znwnxd43]
7. Al-Zahrani AM, Alasmari TM. Exploring the impact of artificial intelligence on higher education: the dynamics of ethical, social, and educational implications. Humanit Soc Sci Commun 2024;11(1):1-12. [doi: 10.1057/s41599-024-03432-4]

8. Zhong W, Luo J, Lyu Y. How do personal attributes shape ai dependency in Chinese higher education context? Insights from needs frustration perspective. PLOS ONE 2024;19(11):e0313314. [doi: 10.1371/journal.pone.0313314] [Medline: 39485818]
9. Marciano L, Camerini AL, Schulz PJ. Neuroticism and internet addiction: what is next? A systematic conceptual review. Pers Individ Dif 2022 Feb;185:111260. [doi: 10.1016/j.paid.2021.111260]
10. Zawacki-Richter O, Marín VI, Bond M, Gouverneur F. Systematic review of research on artificial intelligence applications in higher education – where are the educators? Int J Educ Technol High Educ 2019 Dec;16(1):1-27. [doi: 10.1186/s41239-019-0171-0]
11. Huang Lecturer W, Zhang Z, Zeng Y. Research on university students' intention to use AI assistance in cross-cultural learning. Presented at: Proceedings of The International Conference on Electronic Business, ICEB'24; Oct 24-28, 2024; Zhuhai, Guangdong.
12. Brand M, Young KS, Laier C, Wölfling K, Potenza MN. Integrating psychological and neurobiological considerations regarding the development and maintenance of specific Internet-use disorders: an interaction of person-affect-cognition-execution (I-PACE) model. Neuroscience & Biobehavioral Reviews 2016 Dec;71:252-266. [doi: 10.1016/j.neubiorev.2016.08.033]
13. Yu Q, Xu Y. Awareness of responsibility and internet gaming addiction: the mediating role of cognitive emotion regulation strategies. Presented at: In 2024 9th International Conference on Modern Management, Education and Social Sciences (MMET 2024); Sep 20-22, 2024; Xiamen, China p. 61-70. [doi: 10.2991/978-2-38476-309-2_8]
14. Zhang S, Zhao X, Zhou T, Kim JH. Do you have AI dependency? The roles of academic self-efficacy, academic stress, and performance expectations on problematic AI usage behavior. Int J Educ Technol High Educ 2024;21(1):34. [doi: 10.1186/s41239-024-00467-0]
15. Hamamra B, Mayaleh A, Khlaif ZN. Between tech and text: the use of generative AI in Palestinian universities - a ChatGPT case study. Cogent Education 2024 Dec 31;11(1):2380622. [doi: 10.1080/2331186X.2024.2380622]
16. Omar A, Shaqour AZ, Khlaif ZN. Attitudes of faculty members in Palestinian universities toward employing artificial intelligence applications in higher education: opportunities and challenges. Front Educ 2024;9. [doi: 10.3389/feduc.2024.1414606]
17. Krou MR, Fong CJ, Hoff MA. Achievement motivation and academic dishonesty: a meta-analytic investigation. Educ Psychol Rev 2021 Jun;33(2):427-458. [doi: 10.1007/s10648-020-09557-7]
18. Stein DJ, Craske MG, Rothbaum BO, et al. The clinical characterization of the adult patient with an anxiety or related disorder aimed at personalization of management. World Psychiatry 2021 Oct;20(3):336-356. [doi: 10.1002/wps.20919] [Medline: 34505377]
19. Feher A, Smith MM, Saklofske DH, Plouffe RA, Wilson CA, Sherry SB. The big three perfectionism scale–short form (BTPS-SF): development of a brief self-report measure of multidimensional perfectionism. J Psychoeduc Assess 2020 Feb;38(1):37-52. [doi: 10.1177/0734282919878553]
20. Ryan RM, Deci EL, Vansteenkiste M, Soenens B. Building a science of motivated persons: Self-determination theory's empirical approach to human experience and the regulation of behavior. Motiv Sci 2021;7(2):97-110. [doi: 10.1037/mot0000194]
21. Jia J, Chen L, Zhang L, Xiao M, Wu C. A study on the factors that influence consumers' continuance intention to use artificial intelligence chatbots in a pharmaceutical e-commerce context. The Electronic Library 2025 May 22;43(3):303-321. [doi: 10.1108/EL-09-2024-0275]
22. Caporusso N. Generative artificial intelligence and the emergence of creative displacement anxiety. RDPB 2023;3(1). [doi: 10.53520/rdpb2023.10795]
23. Meng J, Rheu M(, Zhang Y, Dai Y, Peng W. Mediated social support for distress reduction: AI chatbots vs. human. Proc ACM Hum-Comput Interact 2023 Apr 14;7(CSCW1):1-25. [doi: 10.1145/3579505]
24. Lepp M, Kaimre J. Does generative AI help in learning programming: Students' perceptions, reported use and relation to performance. Computers in Human Behavior Reports 2025 May;18:100642. [doi: 10.1016/j.chbr.2025.100642]
25. Casale S, Caplan SE, Fioravanti G. Positive metacognitions about internet use: A new, specific form of metacognitions related to problematic internet use. Comput Human Behav 2016;59:1-7. [doi: 10.1016/j.addbeh.2016.03.014]
26. Tümen Akyildiz S, Ahmed KH. An overview of qualitative research and focus group discussion. International Journal of Academic Research in Education 2021;7(1):1-15. [doi: 10.17985/ijare.866762]
27. Yin RK. Validity and generalization in future case study evaluations. Evaluation (Lond) 2013 Jul;19(3):321-332. [doi: 10.1177/1356389013497081]
28. Lambert SD, Loiselle CG. Combining individual interviews and focus groups to enhance data richness. J Adv Nurs 2008 Apr;62(2):228-237. [doi: 10.1111/j.1365-2648.2007.04559.x] [Medline: 18394035]
29. Geampana A, Perrotta M. Using interview excerpts to facilitate focus group discussion. Qual Res 2025 Feb;25(1):130-146. [doi: 10.1177/14687941241234283] [Medline: 40028392]
30. Khlaif ZN, Alkouk WA, Salama N, Abu Eideh B. Redesigning assessments for AI-enhanced learning: a framework for educators in the generative AI era. Education Sciences 2025;15(2):174. [doi: 10.3390/educsci15020174]

31. Poliandri D, Perazzolo M, Pillera GC, Giampietro L. Dematerialized participation challenges: Methods and practices for online focus groups. Front Sociol 2023;8:1145264. [doi: 10.3389/fsoc.2023.1145264] [Medline: 37091722]
32. Fontana A, Benzi IMA, Cipresso P. Problematic internet use as a moderator between personality dimensions and internalizing and externalizing symptoms in adolescence. Curr Psychol 2023 Aug;42(22):19419-19428. [doi: 10.1007/s12144-021-02409-9]
33. Brand M, Wegmann E, Stark R, et al. The interaction of person-affect-cognition-execution (I-PACE) model for addictive behaviors: Update, generalization to addictive behaviors beyond internet-use disorders, and specification of the process character of addictive behaviors. Neuroscience & Biobehavioral Reviews 2019 Sep;104:1-10. [doi: 10.1016/j.neubiorev.2019.06.032]
34. Chaudhry IS, Sarwary SAM, El Refae GA, Chabchoub H. Time to Revisit Existing Student's Performance Evaluation Approach in Higher Education Sector in a New Era of ChatGPT — a case study. Cogent Education 2023 Dec 31;10(1):2210461. [doi: 10.1080/2331186X.2023.2210461]
35. Baumeister RF, Vohs KD. Self‐regulation, ego depletion, and motivation. Social & Personality Psych 2007 Nov;1(1):115-128. [doi: 10.1111/j.1751-9004.2007.00001.x]
36. Simonÿ C, Damsgaard JB, Johansen K, et al. The power of a ricoeur-inspired phenomenological-hermeneutic approach to focus group interviews. J Adv Nurs 2025 Aug 13. [doi: 10.1111/jan.70133] [Medline: 40801446]

Abbreviations

AI: Artificial Intelligence
AIED: Artificial Intelligence in Education
Gen AI: Generative Artificial Intelligence
IPACE: Interaction of Person-Affect-Cognition-Execution
STEM: Science, Technology, Engineering, and Mathematics

Edited by B Lesselroth; submitted 25.03.25; peer-reviewed by S Sivarajkumar, S Kath, TA Elwi, Z Yu; revised version received 17.07.25; accepted 15.08.25; published 15.09.25.

Please cite as:
Alhur AA, Khlaif ZN, Hamamra B, Hussein E
Paradox of AI in Higher Education: Qualitative Inquiry Into AI Dependency Among Educators in Palestine
JMIR Med Educ 2025;11:e74947
URL: https://mededu.jmir.org/2025/1/e74947
doi:10.2196/74947

Preclinical Medical Students' Perspectives and Experiences With Structured Web-Based English for Medical Purposes Courses: Cross-Sectional Study

Radhakrishnan Muthukumar[1], MBBS, MClinEmbryol, PhD; Isaraporn Thepwongsa[2], MD, MFM, PhD; Poompong Sripa[3], MD; Bangonsri Jindawong[2], MSc, PhD; Kamonwan Jenwitheesuk[1], MD; Surapol Virasiri[1], MD

[1]Academic Affairs, Faculty of Medicine, Khon Kaen University, Khon Kaen, Thailand

[2]Department of Community, Family and Occupational Medicine, Faculty of Medicine, Khon Kaen University, 123 Mittraparb Road, Khon Kaen, Thailand

[3]Inverkeithing Medical Group, Inverkeithing, United Kingdom

Corresponding Author:
Isaraporn Thepwongsa, MD, MFM, PhD
Department of Community, Family and Occupational Medicine, Faculty of Medicine, Khon Kaen University, 123 Mittraparb Road, Khon Kaen, Thailand

Abstract

Background: English for medical purposes (EMP) is essential for medical students as it serves as a foundational language for medical communication and education. However, students often undervalue its importance within the medical curriculum. Given their demanding schedules and workload, educational methods for EMP must align with their needs. Structured web-based learning offers flexibility and convenience, yet limited research has explored its exclusive application for EMP in undergraduate medical education.

Objective: This study aimed to evaluate medical students' perspectives on structured web-based EMP courses and assess their impact on medical English proficiency using objective and subjective measures.

Methods: Structured web-based EMP courses were developed based on evidence-based guidelines, addressing barriers to web-based learning during development and implementation. A cross-sectional study was conducted with 535 medical students who completed these courses. Data were collected via questionnaires, the learning management system, and the Khon Kaen University Medical English Test (KKUMET), which assessed proficiency in listening, reading, writing, and speaking. Data were analyzed using descriptive statistics.

Results: Of the 535 students, 452 (84.5%) completed the survey. Participants reported confidence in reading (mean 4.11, SD 0.87), vocabulary (mean 4.04, SD 0.84), and listening skills (mean 4, SD 0.89), but lower confidence in writing skills (mean 3.46, SD 1.07). The KKUMET results showed statistically significant improvements in all 4 language skills after course completion ($P<.001$). The top-rated benefits of the courses were convenience (mean 4.77, SD 0.59), sufficient instruction (mean 4.5, SD 0.85), and clear content (mean 4.41, SD 0.80).

Conclusions: Structured web-based EMP courses are relevant and well received by medical students. These courses significantly improve students' medical English proficiency, as evidenced by both subjective feedback and objective measures. Medical educators should consider integrating structured web-based EMP programs to better support students' language proficiency in medical contexts.

(JMIR Med Educ 2025;11:e65779) doi:10.2196/65779

KEYWORDS

English for medical purposes; online course; online learning; online education; medical students; medical school; online; online learners; perspectives; English; English language; medical research; educational method; lesson; course; instructional designs; English for medical professional; EMP; barriers; web-based

Introduction

Background

English is the international language of medicine [1,2]. It is the dominant language used in medical research and publications [1,3]. A growing number of medical journals worldwide have been published in English [4]. Almost 90% of the world's publications indexed in the MEDLINE database are published in the English language [3,5]. In addition, considering their high impact factors and citation scores, almost all quality research articles have been published in English-medium journals [1].

The language of medicine has shifted from Latin and Greek to English, influencing the development of modern medical terminology [3]. The use of English in a medical community involves communication with colleagues and staff, reading medical journals, giving conference presentations, or pursuing postgraduate education in English-speaking countries [1,3]. Therefore, the English language and literacy are essential in medical curricula [2,4,6].

General English focuses on overall language competency, emphasizing foundational grammar, vocabulary, and communication skills applicable to a wide range of everyday and academic contexts. English for medical purposes (EMP), on the other hand, is a specialized subset of English tailored to the medical field. It emphasizes medical terminology, reading and comprehension of medical texts and research articles, professional communication skills that are required for clinical and academic purposes, such as writing case reports, interacting with colleagues, and participating in medical discussions [7].

Approximately half of medical schools in the world use English as a medium of instruction [8]. The current status of English use with other languages in medicine is that "English is one lingua franca in medicine but speaks many tongues" [5]. This means that although medical teaching and patient interaction are in the local language, medical professionals still use medical languages with specialized vocabulary in their professional communication [3]. For students or professionals whose English is not their first language, using medical English for their professional-related activities is much more challenging and demanding than it is for those whose English is their first language [3].

The assessment of learning needs is considered fundamental to EMP curriculum design [2,4,7,9] and should be identified early [10]. The learner-centered approach is used to obtain needs from the learner's needs assessment. The learner-centered approach empowers learners to decide what, how, and where to learn, while teachers act as facilitators [11,12]. Learning needs assessments help define program goals and specific teaching objectives [7,13], which leads to the development of lesson plans, materials, tests, assignments, and activities. Many studies have examined the need for EMP [2,9,13-15]. Medical students need to read medical literature, listen to lectures, and write clinical reports and short essays, whereas practitioners need patient interaction and conference participation [1,7]. The development of EMP courses addresses specific language challenges that hinder medical students' learning, particularly in contexts where English is partially used as the medium of instruction, while local languages dominate communication between instructors, students, and patients [16].

Teaching an EMP course differs from teaching general English [7,14]. Instead of learning grammar and fundamental structures, the goal of learning English at this level is to apply the language to medical studies [7]. Previous studies have described the development of EMP courses, including detailed information about curriculum development processes based on students' learning needs [2,16-18], as well as EMP courses developed for international medical graduates [18] and undergraduate medical students [2,16,17]. However, these EMP courses were taught in medical curricula in countries where English was the official language or medium of instruction. In other countries, the EMP course design has not yet been clearly explored. As a result, the requirements of EMP for medical students must be investigated [19], and EMP instruction must be tailored to the intended audience needs. In addition, many studies have focused on preparing students with linguistic tasks for their future careers [9,15]. Less is focused on the design of EMP courses aiming to improve academic performance during undergraduate studies, preclinical years in particular, where basic science knowledge is most of the content rather than clinical application.

Educational methods for EMP should also be matched with learners' needs and preferences. Online and web-based educational methods have been adopted in medical education for the past decade [20]. Online teaching was the primary modality of instruction offered to undergraduate medical students during the COVID-19 pandemic [21]. The web-based education methods have been tested for their effectiveness [22-24]. The advantages of web-based educational methods include improving medical students' knowledge and skills [24]. However, prior meta-analyses have demonstrated that web-based learning is just as effective as the traditional learning methods [23,24]. There are few studies on the exclusive use of web-based learning techniques in EMP. One study explored students' readiness for internet-based learning in EMP [25] or internet- and computer-based as part of blended learning for general English [26,27]; however, none has solely used web-based educational methods in EMP courses for medical students.

Problem Statement

Despite the growing adoption of web-based learning methods in medical education [28], limited research has explored their exclusive use in teaching EMP, especially for preclinical medical students [24]. Web-based education, proven to be as effective as traditional methods [22,23,29,30], offers unique advantages such as flexibility and accessibility [24]. However, few studies have investigated the development and evaluation of structured web-based EMP courses tailored to students in non–English-speaking environments.

Thus, this study developed an exclusive structured web-based learning program for EMP and evaluated its effectiveness. Grounded in constructivist principles, the study emphasizes active, self-directed learning. The structured web-based EMP courses were designed to enable students to construct knowledge through interactive activities, scenario-based learning, and multimedia resources. These approaches foster engagement and empower learners to take ownership of their educational journey, ensuring the content is relevant and applicable to their academic and professional needs. The guiding research question for this study was whether a structured web-based learning program exclusively for EMP would be accepted, considered relevant, and meet the satisfaction of preclinical medical students. We hypothesized that learning EMP through structured web-based courses would be both relevant and well-received by preclinical medical students. This investigation formed the basis for evaluating the course. To comprehensively evaluate the course, the research is divided into 2 parts. The first part, the present study, focuses on assessing the relevance and acceptance of

structured web-based EMP courses based on students' perceptions of the web-based learning mode. The second part examines the effectiveness of the structured web-based EMP courses in improving students' medical English proficiency. To evaluate the first part, a cross-sectional descriptive study was conducted to assess the relevance and acceptance of the structured web-based EMP courses among participating medical students. Additionally, feedback from course instructors, gathered during the course design and development phases, was incorporated to capture their perceptions of the web-based EMP program. The findings from the second part are discussed in a separate article; however, this study included significant findings on the program's effects on students' medical English proficiency. The findings are intended for medical educators, curriculum designers, and policy makers in medical education, particularly those serving non–English-speaking regions.

Our Medical School Curriculum

Our medical school recruits students directly from high school through the Thailand University Central Admission System (TCAS). TCAS, implemented in 2018, is an admission framework consisting of 5 rounds held annually. Medical schools in Thailand often adopt unique criteria and processes within TCAS, reflecting their specific admission requirements and competitive nature [31]. With 288 students enrolled each year, our medical school is one of the regional institutions in northeastern Thailand. The undergraduate medical curriculum spans 6 years, including the first 3 years focused on medical sciences and the last 3 years on clinical practice [32]. Our 2019 updated medical curriculum states that medical students must earn 259 credits over 6 years of medical school. In the first year, premedical education courses (the majority are general education and general principles for medical sciences) account for 38.5 credits. In the second and third years, preclinical education courses comprise 76.5 credits. Clinical rotations, accounting for 48 credits, are conducted in the sixth year, while clinical education courses make up 96 credits across the fourth and fifth years [33]. The majority of the lecture slides, suggested texts, teaching and learning materials, and exam questions are in English, even though the instruction at our school is in Thai.

It is mandatory for Thais to learn English from primary school till higher education [34,35]. In Thailand, the English component taught in high school is of a very basic level, the general English proficiency is low and unsatisfactory [36,37], and by itself is not sufficient for medical professional courses. English proficiency is considered as a part of the entry criteria to the medical schools. Despite English proficiency being a part of the weighted formula for entering medical schools, at present, there is no specific cutoff score or a minimal standardized English language proficiency requirement for entering a medical school in Thailand [38]. English as a foreign language, therefore, is a requirement for students at our medical school [33]. Once students are admitted, they begin enrolling in English courses starting in their first year of study. During medical school, students are required to complete 6 English courses. Four general English courses are taken during the first 3 years of medical school and are taught by nonmedical staff from the university's language institute. They focus on foundational grammar, vocabulary, and communication skills, providing a general linguistic foundation for academic and social contexts. Two EMP courses are introduced in the second and third years of medical school. The content focuses on medical terminology and language skills necessary for academic and clinical tasks, such as reading medical literature, writing reports, and engaging in clinical communication. The curriculum shifts from general language acquisition in the initial years to specialized language skills tailored to medical contexts as students advance in their studies.

Annual feedback from our students consistently indicates that they perceive limited relevance of general English to their medical studies. This perception may stem from several factors. Despite years of English education, many students struggle with basic speaking skills due to insufficient vocabulary and grammar knowledge [39]. Additionally, a lack of intrinsic motivation often leads students to view general English as a mandatory requirement rather than a valuable skill for personal or professional development [40]. The culture within medical programs often emphasizes the need for English proficiency tailored to medical contexts, which shapes students' perceptions and priorities [41]. Therefore, they could not match learning general English in the medical curriculum.

The findings from our students' needs assessments revealed that our medical students needed to improve their English proficiency and wanted the school to organize a test for their English proficiency [42]. They preferred a self-directed web-based learning method and teachers who were both English language experts and medical professionals [42]. Based on the needs assessments, students value medical English because it directly aligns with their academic and professional goals. Unlike general English, medical English equips them with the specific skills needed to read medical literature, participate in clinical discussions, and engage effectively within global medical communities [41,43]. Needs assessments and feedback consistently emphasize students' preference for EMP courses over general English, highlighting the practical benefits they associate with EMP in their medical studies and future careers. Based on learning needs assessments, 2 additional EMP courses were developed while 2 general English courses were removed from the medical curriculum. The newly developed EMP courses were English for Medical Purposes I for the second-year medical students and English for Medical Purposes II for the third-year medical students, which were launched at the same time in the academic year 2021.

Proficiency in medical English is considered to enhance students' academic performance, evidenced by a positive correlation between English proficiency and academic success among medical students [44,45], and to support their development as independent learners [15]. However, evaluating the extent to which students achieve independent learning is beyond the scope of this study.

Methods

Study Objectives

This study aims to evaluate the relevance, acceptance, and satisfaction of structured web-based EMP courses among preclinical medical students.

Study Design

A cross-sectional study was conducted to explore medical students' opinions on learning EMP in a structured web-based course, and qualitative insights from instructors' feedback were obtained.

Participants

The participants were second-year medical students enrolled in the English for Medical Purposes I course and third-year medical students enrolled in the English for Medical Purposes II course at the time of the academic year 2021.

The WinPepi, a statistical software package, was used to calculate sample size based on the proportion of medical students who acknowledged web-based learning [46]. Assuming a proportion of 0.73, a population size of 535, a design effect of 1, and a significance level of 0.05, a total number of 121 was sufficient. However, all 535 students were included in this study to avoid a potential source of selection bias. It is important to clarify that this power analysis was performed for the broader research project, which consists of 2 parts. The first part, detailed in this manuscript, explores medical students' perceptions and experiences with web-based EMP learning. The second part examines the effects of the structured web-based EMP courses on changes in students' English language proficiency after course completion, which is reported in a separate article.

Development of Structured Web-Based EMP Courses

Based on our medical students' needs, EMP course objectives were set accordingly. The main objective of the EMP course was to enhance students' medical English proficiency in all 4 core English language skills (reading, writing, listening, and speaking) through structured, targeted, and interactive web-based practice.

An overview of the four main English language skills is as follows. (1) Reading: Defined as the comprehension of medical literature and academic texts relevant to the students' year of medical study. Instructional strategies include guided analysis of medical articles, scenario-based reading exercises, and formative assessments such as quizzes. (2) Writing: Focuses on effective medical note-taking and medical essays, emphasizing medical content, structure, vocabulary, readability, precision, and clarity. Instructional strategies involve structured writing guides, assignments, peer reviews, and scenario-based tasks such as summarizing patient cases. (3) Listening: Aimed at developing comprehension of medical conversations from case studies and lectures. Instructional methods include exposure to scenario-based audio materials, repeated listening tasks, and interactive multimedia resources. (4) Speaking: Enhances verbal communication in medical contexts, including explaining medical terms, conducting history taking, discussing treatment plans, and educating patients. Instructional strategies feature medical speaking guides, public speaking guides, role-play, video-recorded practice sessions, and feedback loops. Performance improvement for each skill is measured through targeted assessments conducted before and after the course. Standardized testing for reading and listening comprehension, comparison of baseline and final writing samples to assess coherence and technical accuracy, and evaluations of verbal communication in simulated scenarios are used to document and demonstrate progress effectively. These details ensure that the EMP course objectives are transparent and tied directly to improving students' proficiency in medical English through targeted and practical methods.

To develop an effective structured web-based course, we followed the steps of Hays and Veitch's recommendations [47] and applied Cook and Dupras's guide [48]. Key principles of Hays and Veitch emphasize the importance of conducting a thorough needs assessment to identify gaps in knowledge, skills, and practices while ensuring relevance to participants' professional roles and daily practices. They recommend interactive methods such as case studies, group discussions, and workshops to foster engagement and promote practical learning. Flexible delivery formats accommodate diverse learning preferences and schedules, while regular program evaluations ensure continuous improvement, sustainability, and adaptability to emerging needs and advancements in medical education. Cook and Dupras provide a guide specifically for developing web-based learning programs. Their framework focuses on conducting needs assessments, defining clear and measurable learning objectives, and designing content tailored to learners' needs. They highlight the importance of interactive and multimedia-rich content to enhance engagement, as well as ensuring usability and accessibility for diverse learners. Evaluation methods, including pre- and postassessments, combined with continuous feedback and iterative improvements, are essential to measure learning outcomes and maintain program quality. Together, these frameworks provide a robust foundation for creating effective, learner-centered programs that meet educational goals and align with the needs of preclinical medical students. Barriers to web-based learning were considered [49]. More details regarding the development of EMP courses are provided in Multimedia Appendix 1.

The EMP courses were uploaded to a customized web-based learning management system (LMS) developed by our medical school [29]. Each EMP course ran throughout the semester for approximately 48 weeks, and the students' learning schedule was approximately 3 hours per week. However, students can manage their time to study as scheduled or at any time that suits them (see more details on developing the web-based EMP course in Multimedia Appendix 1). Data from the LMS monitoring system were used to report student enrollment and completion of the EMP courses.

Data Sources, Questionnaire, and Assessment

To evaluate the course, questionnaires were used as tools to gather student perspectives on various aspects of the courses, including their confidence in medical English skills, satisfaction with the content, and perceived benefits of web-based learning. These tools provided quantitative data on student experiences

and attitudes. Additionally, feedback from instructors on course design and delivery was integrated into the evaluation process to ensure alignment with educational objectives and identify areas for improvement. Together, these methods offered both quantitative and qualitative insights into the evaluation of the web-based EMP courses.

An online administered questionnaire was developed based on the literature on the use and the evaluation of web-based learning [46,50,51]. The following constructs of web-based learning effectiveness and attitudes toward the structured web-based EMP course were included in the questionnaire: individual learners, confidence in medical English skills, perceived factors influencing web-based participation, perceived satisfaction with content and instructional designs, perceived ease of use, perceived advantages and barriers to web-based learning, and perceived outcomes and benefits of web-based EMP courses. The questionnaire included a 5-point Likert scale ranging from disagree to agree, which was used to assess learners' agreement with each item. For learners' confidence in their medical English skills, the constructs included a 5-point Likert scale ranging from not confident to strongly confident. For the factors influencing web-based learning participation, the constructs included a 5-point Likert scale, ranging from no influence to influence. For learners' perceived benefits of web-based EMP courses, the constructs included a 5-point Likert scale, ranging from not beneficial to strongly beneficial.

To ensure the face and content validity of the questionnaire, each item was evaluated thoroughly by 3 separate experts in the field of medical education at our school. Thirty students participated in a pilot test of the questionnaire, which led to revisions. The Cronbach α coefficient of the attitude and experience part of the questionnaire was 0.93.

The details of the course instructions, content, and materials are provided in Multimedia Appendix 1. The course media included PDF files, audio, videos, and external links for the downloads. The second-year medical students completed a 45-hour web-based asynchronous English for Medical Purposes I course. The third-year medical students completed a 45-hour web-based asynchronous English for Medical Purposes II course. The students would be considered to have completed the module if that module was accessed at least 50% of the total learning time for that module because the LMS has a double speed-up function for playing video or audio clips.

After participants completed the web-based EMP courses, they were invited to complete the questionnaire. A link to the online questionnaire on Google Forms was added at the end of each course. The students completed the questionnaire between December 2021 and April 2022. The questionnaire data were transferred from Google Sheets, and data from the LMS monitoring system were exported to Microsoft Excel. The data were then gathered and checked for completion before being transferred to IBM SPSS for Windows.

Of note, the summative assessment for medical English proficiency was conducted using the Khon Kaen University Medical English Test (KKUMET), which examined proficiency in listening, reading, writing, and speaking. Each component was assessed before and after completing the EMP courses. The baseline EMP proficiency of the participating students, as assessed before enrollment in the 2 EMP courses, revealed that more than two-thirds of the students were at beginner or intermediate levels. The instructional goal of these EMP courses was to enhance their proficiency to intermediate and advanced levels. This study focuses on highlighting final outcomes; detailed results of the EMP proficiency tests conducted before and after these web-based EMP courses are reported in a separate article.

Statistical Analysis

Data from questionnaires were analyzed using IBM SPSS for Windows version 26.0. A pairwise deletion strategy was applied to handle the missing data. The demographic data were described using descriptive statistics. The participants' responses on a 5-point Likert scale to a 42-item questionnaire on their web-based learning experiences were dichotomized by calculating the mean scores and SD. Mean scores of 3.5 and above were considered agreed upon.

Ethical Considerations

This study was approved by the Human Research Ethics Committee of Khon Kaen University (project number: HE631465). Students were recruited through the researcher's assistant, who invited volunteers to participate. Before completing the questionnaires, participants were informed that participation was voluntary and that they could drop out of the study at any time. They were informed that their opinions were important for enhancing the medical English courses and were therefore encouraged to express them. All students voluntarily agreed to participate without receiving compensation. The participants' privacy and identity were protected, and confidentiality was assured in that no identifying information was asked. The study objectives were explained to the participants, and the study was conducted according to the academic ethical code.

Results

Participating Medical Student Demographics

A total of 535 medical students completed the web-based EMP courses, 452 of whom started and returned the completed questionnaires (response rate: 84.5%). Table 1 presents the participants' demographic data. The numbers of male and female participants were relatively similar, as were the numbers of medical students each year.

Table . Demographic data of the participants.

Demographics	Values
Age (n=451), years	
Range	18 - 24
Mean (SD)	20.37 (0.74)
Sex (n=452), n (%)	
Male	219 (48.5)
Female	222 (49.1)
Prefer not to say	11 (2.4)
Year of study (n=450)	
2	231 (51.3)
3	219 (48.7)

Influence of Web-Based Learning Characteristics

Respondents rated the degree of influence of the characteristics of web-based learning on their participation in web-based EMP courses (Table 2). Convenience, flexibility, and accessibility were rated the highest, while facilitator interactions received lower ratings.

Table . Rating of the influence of web-based learning characteristics on participation in web-based EMP[a] courses (n=452).

Characteristics of web-based learning	The degree of influence on adoption of web-based EMP courses, mean[b] (SD)
The convenience of completion the courses at any time or place	4.82 (0.49)
Flexibility to complete and save small sections at a time	4.80 (0.56)
Easy to access the course content	4.60 (0.75)
Easy to use/complete the course	4.58 (0.71)
Access to other useful links and resources	4.53 (0.84)
Instant access to feedback and the right answers when completing quizzes	4.50 (0.85)
The quality of content	4.46 (0.75)
Accessibility to technical support if difficulties are encountered	4.35 (0.94)
The use of case-based information and discussion	3.99 (1.08)
Facilitator's regular input/participation	3.77 (1.17)
The opportunity to communicate/interact with the facilitator	3.75 (1.14)

[a]EMP: English for medical purposes.

[b]Mean was calculated using a 5-point Likert scale ranging from 1 (strongly no influence), 2 (no influence), 3 (neutral), 4 (influence), and 5 (strongly influence).

Confidence in Medical English Skills

Respondents' confidence in their medical English skills is shown in Table 3. The participating medical students reported feeling confident about medical English reading, vocabulary, and listening skills but were not sure about their writing skills.

Table . The participant's confidence in their medical English skills after completing the medical English courses (n=452).

Confidence in medical English skills after completing the modules	Values, mean[a] (SD)
Medical English reading skill	4.11 (0.87)
The use of medical English vocabulary	4.04 (0.84)
Medical English listening skill	4.00 (0.89)
Applying professional-specific knowledge and skills in English	3.92 (0.91)
Medical English speaking skill	3.50 (1.05)
Medical English writing skill	3.46 (1.07)

[a]Mean was calculated using a 5-point Likert scale ranging from 1 (strongly not confident), 2 (not confident), 3 (not sure), 4 (confident), and 5 (strongly confident).

Perceived Advantages, Barriers To, and Attitudes Toward Instructional Designs of Web-Based EMP Courses

The participants identified useful and beneficial aspects of web-based EMP courses. The top 3 highly rated participants agreed on the advantages of convenience, sufficient instructions, and clear and easy-to-understand content (Table 4). For the instructional designs of the web-based EMP courses, the participants agreed on the clarity and appropriateness of the overall design of the courses, including clear objectives and content, appropriate content, arrangement, instruction, media, delivery method, course assessment, and grading (Table 5).

Table . The participants' web-based learning experience of the medical English courses (n=452).

	Values, mean[a] (SD)
Advantages of the web-based medical English modules	
I was able to learn at any place	4.77 (0.59)
Overall, this web-based course provided me with adequate instruction	4.50 (0.85)
The content was easy to understand and clear	4.41 (0.80)
Overall, the course contents covered its objectives	4.30 (0.87)
I felt more comfortable learning in this web-based course than in the face-to-face session	4.19 (1.13)
Overall, the instruction I obtained from this web-based program was motivating	4.16 (0.99)
I knew how to contact the facilitator and the facilitator responses promptly to my questions	4.04 (1.02)
The web-based program fulfilled my learning needs to improve my medical English skills	3.93 (1.04)
If I had technical problems during participating in this program, I received adequate help with technical problems (n=66)[b]	3.30 (1.05)
Difficulties in accessing and completing the course	
There was too much basic, well-known information in the course	3.06 (1.14)
The module took too long to complete	3.03 (1.28)
I spent more time in access to a computer to access this web-based program	2.31 (1.46)
The internet connection was very slow	2.25 (1.42)
I spent more time in downloading external links	2.08 (1.35)
The course was not useful for me because I do not have adequate computer skills to complete the course	1.95 (1.41)

[a]Mean was calculated using a 5-point Likert scale ranging from 1 (Strongly disagree), 2 (disagree), 3 (neutral), 4 (agree), and 5 (strongly agree).

[b]When the participants were asked if they had technical problems during participating in this program, of the 452 participants, 66 (14.6%) had, while 386 (85.4%) had not.

Table . The participants' rating on the web-based medical English course contents and instructional designs (n=452).

The web-based course contents and instructional designs	Values, mean[a] (SD)
The course had clear learning objectives	4.48 (0.79)
The contents were complete, appropriate, and relevant to the objectives	4.41 (0.82)
The order of contents was arranged properly	4.41 (0.82)
The learning media (eg, audios, videos, and PDF files) were appropriate	4.39 (0.90)
Overall, the instructional design of the program was appropriate	4.38 (0.85)
The web-based teaching was appropriate	4.37 (0.86)
The course's assessments (formative assessments on listening, reading, writing, and speaking) were appropriate	3.97 (1.05)
Grading criteria were appropriate	3.74 (1.07)

[a]Mean was calculated using a 5-point Likert scale ranging from 1 (Strongly disagree), 2 (disagree), 3 (neutral), 4 (agree), and 5 (strongly agree).

Perceived Learning Outcomes

The participants agreed that their 4 main skills improved. They also believed that all types of teaching media and lectures benefited them (Table 6).

Table . The participants' perceived outcomes to and benefits of the web-based medical English courses after the course completion (n=452).

	Values, mean (SD)
The participants' perceived outcomes after the course completion[a]	
I improved my listening skills	4.37 (0.85)
I improved my reading skills	4.23 (0.90)
I gained knowledge on how to practice my medical English skills	4.21 (0.89)
I improved my writing skills	3.74 (1.15)
I improved my speaking skills	3.68 (1.21)
Benefits of the web-based medical English courses after the course completion[b]	
The audios to practice listening skills	4.45 (0.83)
The media for medical terms to learn medical terms and practice reading, speaking, and writing skills	4.36 (0.83)
The videos to practice reading, listening, and speaking skills	4.35 (0.91)
The lectures on listening and reading	4.34 (0.84)
The course was organized in the modules	4.28 (0.92)
The recommended reading articles to practice reading and writing skills	4.07 (0.99)
The lectures on scientific writing	4.07 (1.00)
The lectures on speaking	4.02 (1.04)

[a]Mean was calculated using a 5-point Likert scale ranging from 1 (strongly disagree), 2 (disagree), 3 (neutral), 4 (agree), and 5 (strongly agree).

[b]Mean was calculated using a 5-point Likert scale ranging from 1 (strongly not beneficial), 2 (not beneficial), 3 (neutral), 4 (beneficial), and 5 (strongly beneficial).

Learning Behaviors and Engagement Patterns (Data From the LMS Monitoring System)

The second-year students (n=263) completed the learning module ranging from 0 to 7 modules (mean 6.54, SD 1.41). The total learning time ranged from 172 to 24,439 minutes, with a mean of 925.02 (SD 1759.80) minutes. Regarding students' learning behaviors, 153 (58.2%) out of 263 students spent less than 10 hours on the course and often logged in to the course during the afternoon (1069/2993 logins, 35.72%) and evening hours (1005/2993 logins, 33.58%).

The third-year students (n=272) completed the learning module ranging from 0 to 7 modules (mean 6.61, SD 1.41). The total learning time ranged from 0 to 14,100 minutes, with a mean of 768.59 (SD 930.56) minutes. Regarding students' learning behaviors, 136 (50%) students spent less than 10 hours on the course and often logged in to the course during the evening

(1000/2962 logins, 33.76%) and morning hours (655/2962 logins, 22.11%).

Improvements in Medical English Proficiency

The summative assessment of medical English proficiency was conducted using KKUMET, which evaluates listening, reading, writing, and speaking skills. Each component was assessed before and after the EMP courses. The results showed statistically significant improvements in students' scores across all 4 skills, as well as in the total test scores (P<.001), underscoring the effectiveness of the courses. Additional details about the summative testing results are presented in a separate article.

Instructor Feedback on Course Design and Implementation

During the development and implementation of the structured web-based EMP courses, all course instructors actively participated in providing feedback. This feedback was gathered through regular meetings, written evaluations, and iterative reviews of course content and delivery methods. Instructors emphasized the importance of aligning the courses with medical students' academic and professional needs. Their insights directly influenced several key aspects of the course design.

First, scenario-based learning was integrated to make the content more applicable to clinical practice. Case-based scenarios were incorporated into reading and listening exercises to enhance the relevance and practicality of the material. Second, interactive assessments were adapted based on instructor feedback to emphasize practical application. These included the use of peer reviews for writing assignments and role-play activities for speaking exercises, allowing students to practice and refine their skills in realistic scenarios. Third, instructors highlighted the need for multimedia resources to create a more engaging learning experience. This led to the inclusion of audio clips, videos, and interactive tasks designed to cater to various learning preferences. Finally, adaptive scheduling was implemented to accommodate the heavy workloads of medical students. The courses were designed to be asynchronous, enabling students to learn at their own pace and manage their schedules effectively.

Instructors reported that the structured design of the courses, combined with the iterative feedback process, resulted in a cohesive program that effectively addressed students' learning needs. Their ongoing involvement ensured that the courses remained both relevant and practical for medical students.

Discussion

Principal Findings

This study explored the perspectives of preclinical medical students on structured web-based EMP courses and evaluated their proficiency improvements. A total of 535 students enrolled in and completed the courses by the published due dates, of which 452 (84.5%) students completed the questionnaire. The high response rate can be attributed to several factors. The survey's integration at the end of the web-based EMP courses ensured participants encountered it immediately after course completion, when their experiences were fresh and engagement levels were high [52]. The web-based format allowed flexibility, enabling participants to complete the survey at their convenience [53]. Clear instructions, the assurance of anonymity, and the perceived relevance of the survey further encouraged participation [54]. Additionally, students' satisfaction with the course content and instructional design likely motivated them to provide feedback [52].

The summative assessment, conducted using KKUMET, revealed statistically significant improvements across all 4 English language skills (listening, reading, writing, and speaking). Participants reported high confidence in reading, vocabulary, and listening skills but expressed lower confidence in writing and speaking skills. Students rated convenience, clarity of content, and sufficient instruction as the top benefits of the courses. These results affirm the relevance, acceptance, and satisfaction of structured web-based EMP courses for medical students, aligning with the primary objectives of improving proficiency and fostering recognition of EMP's importance in supporting their academic learning.

Evaluation of Course Effectiveness

The primary goal of the EMP courses was to improve students' proficiency in all 4 English language skills and introduce relevant medical terminology tailored to their academic content. This goal was achieved, as evidenced by measurable improvements in KKUMET scores and student-reported confidence levels.

We evaluated the effectiveness of web-based EMP courses, finding significant improvements in students' medical English proficiency across all skills, as measured by KKUMET scores, and increased confidence in reading, vocabulary, and listening. By comparing subjective learner feedback with objective results, we confirmed that reported confidence aligned with measurable gains, minimizing potential overestimation from cognitive biases like the Dunning-Kruger effect [55,56]. These findings underscore the importance of validating self-reported outcomes with objective data to ensure reliable assessments of learning impact [48,57], with future research recommended to explore the influence of cognitive biases and interventions to enhance learning outcomes [55,58].

Furthermore, the courses aimed to highlight the importance of EMP in supporting learning across other subjects, particularly through contextual and content-based design [7,15]. Students expressed initial difficulty in understanding the importance of general English and its relevance within the medical curriculum. However, they highly valued the structured web-based EMP courses and expressed satisfaction with the content, particularly due to its alignment with their academic and professional needs. This alignment was achieved by designing and developing courses that were self-directed and created by medical professionals, which students recognized as enhancing their engagement and acceptance of the educational approach. While students did not explicitly indicate a need for increased interaction with faculty facilitators, they expressed a preference for instructors who were both English language experts and knowledgeable about medical content. This preference suggests an implicit desire for contextualized guidance, highlighting the

value students place on instructors with dual expertise. The structured web-based EMP courses effectively met these needs, motivating them to integrate EMP learning into their study schedules, underscoring the importance of tailoring course design to align with student preferences.

Comparison With Previous Studies

The findings of this study align with existing research on the effectiveness of web-based learning in medical education while addressing gaps specific to EMP. Prior studies have emphasized the effectiveness of web-based learning in enhancing knowledge and skills in medical education [23,24]. Similarly, this study demonstrated statistically significant improvements in all 4 core English language skills (listening, reading, writing, and speaking) among medical students enrolled in structured web-based EMP courses. These findings confirm that well-designed web-based educational modalities can be as effective as traditional methods.

Students greatly appreciated the flexibility and accessibility offered by web-based courses, a perspective consistently supported in previous research [25,26,29,30]. Students appreciated the ability to learn at their own pace, access course materials conveniently, and engage with multimedia content tailored to their needs. The use of diverse teaching media, such as audio, video, and interactive exercises, aligns with widely recognized recommendations for enhancing learning engagement and retention [59,60]. This highlights the role of self-directed learning in increasing engagement and satisfaction.

Unlike previous studies focusing on blended learning or face-to-face instruction [25-27], this study uniquely explored the exclusive use of structured web-based courses for EMP. The findings underscore the potential of this approach to address the challenges faced by medical students in non–English-speaking regions, thereby filling a critical gap in the literature.

This study also found that students reported greater confidence in receptive skills (reading and listening) than in productive skills (writing and speaking), a pattern consistent with findings from previous research [7,15,17]. This underscores the need for targeted instructional strategies to support the development of productive language skills in EMP courses [2,17].

By situating the findings within the broader body of research, this study contributes to the evolving understanding of effective teaching strategies for EMP and web-based learning in medical education.

Implications of Findings for Practice

The findings of this study have important implications for curriculum development and teaching strategies in medical schools, particularly in non–English-speaking contexts. Structured web-based EMP courses significantly improved students' medical English proficiency, demonstrating their potential to meet academic and professional needs effectively. These results suggest that similar approaches could be adopted by other medical schools to enhance student engagement and learning outcomes.

Medical schools can leverage the flexibility and accessibility of web-based learning to design EMP courses that accommodate students' demanding academic schedules. By focusing on targeted skill development and incorporating medical terminology into course content, institutions can create tailored programs that address specific language needs [14,19]. The self-paced nature of web-based learning further enables students to manage their time effectively, aligning with their individual schedules and learning preferences. This adaptability can increase motivation and reduce barriers to learning, particularly in resource-constrained settings.

Integrating EMP with other areas of medical education, such as reading medical literature and writing clinical reports, can help students perceive medical English as an essential part of their academic and professional journey rather than as a standalone requirement. Additionally, structured web-based courses offer scalable and accessible solutions for institutions with large student cohorts, ensuring consistent, high-quality content delivery while reducing the resource burden on faculty and support staff.

Despite the benefits of self-directed learning, low engagement with "consultant hours" highlights the need for integrating opportunities for active faculty-student interaction within course designs. Addressing this issue could involve developing mechanisms that encourage and normalize faculty interaction, which may be especially beneficial in contexts where cultural preferences for independence or heavy academic workloads limit voluntary engagement with support services [23,24,26]. Such measures could enhance overall student support and satisfaction.

These insights underscore the value of structured web-based EMP courses as a model for improving language proficiency and supporting broader academic goals in medical education.

Strengths and Limitations

This study offers a comprehensive evaluation of structured web-based EMP courses, combining subjective learner feedback with objective proficiency measures. A key strength lies in the design of the courses, informed by needs assessments and evidence-based guidelines [47,48]. Moreover, the inclusion of summative assessments provides robust evidence of the courses' effectiveness. This study had a large sample size and response rate, which ensures robust findings. However, the study's scope was limited to a single medical school, potentially affecting generalizability [11,57].

Future research should focus on strategies to enhance productive skills (writing and speaking) while continuing to explore the impact of cognitive biases, such as the Dunning-Kruger effect, on the gap between perceived and actual proficiency [56]. Large-scale, multi-institutional studies are warranted to validate these findings and provide broader recommendations for integrating EMP into medical curricula. Additionally, research should investigate the long-term impact and scalability of such courses across diverse educational settings, as well as the development of adaptive learning technologies to customize course content based on students' baseline proficiency levels, effectively addressing specific skill gaps.

Conclusions

Structured web-based EMP courses are highly relevant, widely accepted, and well-received by medical students, demonstrating significant improvements in their medical English proficiency, particularly in reading, vocabulary, and listening skills, as evidenced by both subjective feedback and objective measures. The flexibility, accessibility, and practicality of structured web-based learning make it an effective approach to address the unique challenges faced by medical students with demanding schedules. By tailoring course content to meet students' academic and professional needs and incorporating engaging instructional designs, these courses provide a scalable and sustainable solution for medical education in non–English-speaking regions. Future developments in EMP course design should focus on enhancing productive language skills, such as writing and speaking, while maintaining the balance between self-directed learning and faculty support, integrating these courses into medical curricula as an essential component to equip students with the language skills necessary for academic success and global medical practice.

Data Availability

The data from this research project are available upon reasonable request.

Conflicts of Interest

The English for Medical Purposes I and II online courses are copyrighted to Khon Kaen University. RM, IT, and KJ have patents for the English for Medical Purposes I and II online courses.

Multimedia Appendix 1
Development of structured web-based English for medical purpose courses.
[DOCX File, 41 KB - mededu_v11i1e65779_app1.docx]

References

1. Outemzabet B, Sarnou H. Exploring the significance of English-based communication for a community of medical academics in a public university teaching hospital in Algeria. Engl Specif Purp 2023 Apr;70:116-130. [doi: 10.1016/j.esp.2022.12.001]
2. Antic Z, Milosavljevic N. Some suggestions for modelling a contemporary medical English course design based on need analysis. Lingua 2016 Dec;184:69-78. [doi: 10.1016/j.lingua.2016.06.002]
3. Džuganová B. Medical language – a unique linguistic phenomenon. JAHR Eur J Bioeth 2019 Jul 29;10(1):129-145. [doi: 10.21860/j.10.1.7]
4. Gotti M, Salager-Meyer F. Teaching medical discourse in higher education: an introduction. Lang Learn High Educ ;6(1):1-13. [doi: 10.1515/cercles-2016-5001]
5. Baethge C. The languages of medicine. Dtsch Arztebl Int 2008 Jan;105(3):37-40. [doi: 10.3238/arztebl.2008.0037] [Medline: 19633751]
6. Tsai PH. A pedagogical dialogue between English for general purposes and English for medical purposes: marching from short story reading and art practice to the writing of a history of present illness. Taiwan J TESOL 2022;19(2):65-106. [doi: 10.30397/TJTESOL.202210_19(2).0003]
7. Lodhi MA, Shamim M, Robab M, Shahzad S, Ashraf A. English for doctors: an ESP approach to needs analysis and course design for medical students. Int J Engl Linguist 2018;8(5):205. [doi: 10.5539/ijel.v8n5p205]
8. Skelton J, Richards C. Communication for medicine: state-of-the-art. ESP Today 2021;9(1):9-29. [doi: 10.18485/esptoday.2021.9.1.1]
9. Alqurashi F. English for medical purposes for Saudi medical and health professionals. Adv Lang Lit Stud 2016;7(6):243-252. [doi: 10.7575/aiac.alls.v.7n.6p.243]
10. Maher J. English for medical purposes. Lang Teach 1986 Apr;19(2):112-145. [doi: 10.1017/S0261444800012003]
11. Liang H. A critical discourse analysis of medical English course syllabuses. J Lang Tech Res 2023;14(4):865-870. [doi: 10.17507/jltr.1404.02]
12. Antić Z. Teacher education in English for special purposes. Acta Fac Med Naiss 2016;33(3):211-215. [doi: 10.1515/afmnai-2016-0022]
13. Antic Z. Forward in teaching English for medical purposes. Med Biol 2007;14(3):141-147 [FREE Full text]
14. Khalid A. Needs assessment in ESP: a review. Stud Lit Lang 2016;12(6):38-46. [doi: 10.3968/8161]
15. Antić Z. Towards uniformity in English for medical purposes: evaluation and design. Srp Arh Celok Lek 2009;137(7-8):454-457. [Medline: 19764605]
16. Shi L, Corcos R, Storey A. Using student performance data to develop an English course for clinical training. Engl Specif Purp 2001 Jan;20(3):267-291. [doi: 10.1016/S0889-4906(00)00002-8]
17. Wette R, Hawken SJ. Measuring gains in an EMP course and the perspectives of language and medical educators as assessors. Engl Specif Purp 2016 Apr;42:38-49. [doi: 10.1016/j.esp.2015.11.002]

18. Hoekje BJ. Medical discourse and ESP courses for international medical graduates (IMGs). Engl Specif Purp 2007 Jan;26(3):327-343. [doi: 10.1016/j.esp.2006.09.002]
19. Willey I, Tanimoto K, McCrohan G, Nishiya K. An English needs analysis of medical doctors in western Japan. JALT J 2020 Nov 1;42(2):143. [doi: 10.37546/JALTJJ42.2-3]
20. Cook DA, Garside S, Levinson AJ, Dupras DM, Montori VM. What do we mean by web-based learning? A systematic review of the variability of interventions. Med Educ 2010;44(8):765-774. [doi: 10.1111/j.1365-2923.2010.03723.x]
21. Rose S. Medical student education in the time of COVID-19. JAMA 2020 Jun 2;323(21):2131-2132. [doi: 10.1001/jama.2020.5227]
22. Thepwongsa I, Kirby CN, Schattner P, Piterman L. Online continuing medical education (CME) for GPs: does it work? A systematic review. Aust Fam Physician 2014;43(10):717-721. [Medline: 25286431]
23. Cook DA, Levinson AJ, Garside S, Dupras DM, Erwin PJ, Montori VM. Internet-based learning in the health professions: a meta-analysis. JAMA 2008 Sep 10;300(10):1181-1196. [doi: 10.1001/jama.300.10.1181] [Medline: 18780847]
24. Pei L, Wu H. Does online learning work better than offline learning in undergraduate medical education? A systematic review and meta-analysis. Med Educ Online 2019 Dec;24(1):1666538. [doi: 10.1080/10872981.2019.1666538] [Medline: 31526248]
25. Mustafa N, Raside D. Online learning readiness: a case study in the field of English for medical purposes. Particip Educ Res 2016:212-220 [FREE Full text]
26. Navidinia H, Zare Bidaki M, Hekmati N. Incorporating e-learning in teaching English language to medical students: exploring its potential contributions. Med J Islam Repub Iran 2016;30(1):462. [Medline: 28491837]
27. Sriwichai C. Students' readiness and problems in learning English through blended learning environment. Asian J Educ Train 2020;6(1):23-34. [doi: 10.20448/journal.522.2020.61.23.34]
28. Ogundiya O, Rahman TJ, Valnarov-Boulter I, Young TM. Looking back on digital medical education over the last 25 years and looking to the future: narrative review. J Med Internet Res 2024 Dec 19;26(1):e60312. [doi: 10.2196/60312] [Medline: 39700490]
29. Thepwongsa I, Muthukumar R, Sripa P, et al. The perspectives of learners at a public medical school on the evaluation of an online learning management system for degree and non-degree courses. Med Educ Online 2024 Dec 31;29(1):2299535. [doi: 10.1080/10872981.2023.2299535]
30. Thepwongsa I, Sripa P, Muthukumar R, Jenwitheesuk K, Virasiri S, Nonjui P. The effects of a newly established online learning management system: the perspectives of Thai medical students in a public medical school. Heliyon 2021 Oct;7(10):e08182. [doi: 10.1016/j.heliyon.2021.e08182]
31. Thammaboosadee S, Yanta S. An analytical data monetization value chain for educational process improvement under Thai University Central Admission System. Presented at: 2019 4th Technology Innovation Management and Engineering Science International Conference (TIMES-iCON); Dec 11-13, 2019; Bangkok, Thailand. [doi: 10.1109/TIMES-iCON47539.2019.9024499]
32. Sripa P, Thepwongsa I, Muthukumar R. Factors associated with the entry to general practice training: a multicentre study. Med Teach 2020 Dec 1;42(12):1394-1400. [doi: 10.1080/0142159X.2020.1811846]
33. Mei A, Gao D, Jiang J, Qiao T, Wang F, Li D. The medical education systems in China and Thailand: a comparative study. Health Sci Rep 2022;5(6):e826. [doi: 10.1002/hsr2.826]
34. Jarunthawatchai W, Baker W. English language education and educational policy in Thailand. In: Moody AJ, editor. The Oxford Handbook of Southeast Asian Englishes: Oxford University Press; 2024:557-574. [doi: 10.1093/oxfordhb/9780192855282.013.30]
35. Baker W, Jarunthawatchai W. English language policy in Thailand. Eur J Lang Policy 2017 Apr;9(1):27-44. [doi: 10.3828/ejlp.2017.3]
36. Teng B, Sinwongsuwat K. Teaching and learning English in Thailand and the integration of conversation analysis (CA) into the classroom. Engl Lang Teach 2015;8(3). [doi: 10.5539/elt.v8n3p13]
37. Noom-ura S. English-teaching problems in Thailand and Thai teachers' professional development needs. Engl Lang Teach 2013;6(11):139. [doi: 10.5539/elt.v6n11p139]
38. Student admissions. Khon Kaen University. URL: https://apps.admissions.kku.ac.th/web/Port/PortProject/7 [accessed 2023-05-14]
39. Chulee W, Khanom K, Chesa N, Sha'ar M, Buddharat C. "Why still we cannot speak English?" Examining internal demotivating factors among Thai tertiary learners. MEXTESOL J 2023 Nov 1;47(4):1-17. [doi: 10.61871/mj.v47n4-5]
40. Na Nongkhai A. An investigation into English language motivation of Thai university students: understanding students' motivation over time, and their visions of future L2 selves, through narrative inquiry [PhD thesis]. : University of Sheffield; 2017.
41. Chan SMH, Mamat NH, Nadarajah VD. Mind your language: the importance of English language skills in an international medical programme (IMP). BMC Med Educ 2022 May 26;22(1):405. [doi: 10.1186/s12909-022-03481-w] [Medline: 35619080]

42. Suwanrot K, Sausukpaiboon K, Ketdao N, et al. Learning needs assessment for English language of medical students and residents in Srinagarind Hospital, Faculty of Medicine, Khon Kaen University [Article in Thai]. Srinagarind Med J 2017;32(5):454-460 [FREE Full text]
43. Wahyuni S. English language needs for medical students: a link and match of academic and professional career. ENGLISH FRANCA Acad J Engl Lang Educ 2021;5(1):169-184. [doi: 10.29240/ef.v5i1.2146]
44. Kaliyadan F, Thalamkandathil N, Parupalli SR, Amin TT, Balaha MH, Al Bu Ali WH. English language proficiency and academic performance: a study of a medical preparatory year program in Saudi Arabia. Avicenna J Med 2015;5(4):140-144. [doi: 10.4103/2231-0770.165126] [Medline: 26629471]
45. Sadeghi B, Kashanian NM, Maleki A, Haghdoost A. English language proficiency as a predictor of academic achievement among medical students in Iran. Theory Pract Lang Stud 2013;3(12):2315-2321. [doi: 10.4304/tpls.3.12.2315-2321]
46. Bączek M, Zagańczyk-Bączek M, Szpringer M, Jaroszyński A, Wożakowska-Kapłon B. Students' perception of online learning during the COVID-19 pandemic: a survey study of Polish medical students. Research Square. Preprint posted online on Jul 14, 2020. [doi: 10.21203/rs.3.rs-41178/v1]
47. Hays R, Veitch C. Continuing medical education and divisions. In: Hays R, Veitch C, editors. Continuing Medical Education for General Practitioners: James Cook University; 1999:5-17.
48. Cook DA, Dupras DM. A practical guide to developing effective web-based learning. J Gen Intern Med 2004 Jun;19(6):698-707. [doi: 10.1111/j.1525-1497.2004.30029.x] [Medline: 15209610]
49. Thepwongsa I. Education of rural and remote general practitioners (GPs) in Australia on type 2 diabetes: impact of online continuing medical education on GPs' knowledge, attitudes and practices and barriers to online learning [Thesis]. : Monash University; 2017 Feb 23. [doi: 10.4225/03/58ae4580ba641]
50. Abbasi S, Ayoob T, Malik A, Memon SI. Perceptions of students regarding e-learning during Covid-19 at a private medical college. Pak J Med Sci 2020 May;36(COVID19-S4):S57-S61. [doi: 10.12669/pjms.36.COVID19-S4.2766] [Medline: 32582315]
51. Lasanthika W, Tennakoon W. Assessing the adoption of learning management systems in higher education. Global J Bus Soc Sci Review 2019 Sep 27;7(3):204-209. [doi: 10.35609/gjbssr.2019.7.3(5)]
52. Dillman DA, Smyth JD, Christian LM. Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method, 4th edition: Wiley; 2014. [doi: 10.1002/9781394260645]
53. Fan W, Yan Z. Factors affecting response rates of the web survey: a systematic review. Comput Human Behav 2010 Mar;26(2):132-139. [doi: 10.1016/j.chb.2009.10.015]
54. Porter SR, Whitcomb ME. The impact of contact type on web survey response rates. Public Opin Q 2003;67(4):579-588. [doi: 10.1086/378964]
55. Dunning D. The Dunning–Kruger effect: on being ignorant of one's own ignorance. In: Olson JM, Zanna MP, editors. Advances in Experimental Social Psychology: Academic Press; 2011, Vol. 44:247-296. [doi: 10.1016/B978-0-12-385522-0.00005-6]
56. Kruger J, Dunning D. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. J Pers Soc Psychol 1999;77(6):1121-1134. [doi: 10.1037/0022-3514.77.6.1121]
57. Eva KW, Regehr G. Self-assessment in the health professions: a reformulation and research agenda. Acad Med 2005 Oct;80(10 Suppl):S46-S54. [doi: 10.1097/00001888-200510001-00015] [Medline: 16199457]
58. Norman G. Research in clinical reasoning: past history and current trends. Med Educ 2005 Apr;39(4):418-427. [doi: 10.1111/j.1365-2929.2005.02127.x] [Medline: 15813765]
59. Arani JA. Mobile educational SMSs as supplementary means to teach sentence paraphrasing in EMP course. Int J Interact Mob Technol 2016;10(1):45-51. [doi: 10.3991/ijim.v10i1.5188]
60. Mayer RE. Evidence-based principles for how to design effective instructional videos. J Appl Res Mem Cogn 2021 Jun 1;10(2):229-240. [doi: 10.1016/j.jarmac.2021.03.007]

Abbreviations

EMP: English for medical purposes
KKUMET: Khon Kaen University Medical English Test
LMS: learning management system
TCAS: Thailand University Central Admission System

Edited by B Lesselroth; submitted 26.08.24; peer-reviewed by N Riapina, Q Xie; revised version received 20.01.25; accepted 25.02.25; published 27.03.25.

Please cite as:
Muthukumar R, Thepwongsa I, Sripa P, Jindawong B, Jenwitheesuk K, Virasiri S
Preclinical Medical Students' Perspectives and Experiences With Structured Web-Based English for Medical Purposes Courses: Cross-Sectional Study
JMIR Med Educ 2025;11:e65779
URL: https://mededu.jmir.org/2025/1/e65779
doi:10.2196/65779

Original Paper

Health Workers' Perspectives on Mobile Health Care Learning Stickiness: Mixed Methods Study

Sabila Nurwardani[1*], MTI; Putu Wuri Handayani[1*], Prof Dr

Faculty of Computer Science University of Indonesia, Depok, Indonesia

[*]all authors contributed equally

Corresponding Author:
Putu Wuri Handayani, Prof Dr
Faculty of Computer Science University of Indonesia
Jl. Kampus UI
Depok, 16424
Indonesia
Phone: 62 7863419 ext 3200
Email: Putu.wuri@cs.ui.ac.id

Abstract

Background: Doctor-to-Doctor (D2D) is a mobile learning app that aims to support continuous learning in health care, commonly known as continuing medical education. One of the metrics of success in mobile learning is the average amount of time spent each month on the app, which is a component of stickiness, the tendency of users to use apps repeatedly. Stickiness metrics are important because stickiness has a direct effect on user retention.

Objective: This study aimed to determine the factors influencing user stickiness of the D2D mobile learning app. The research framework was based on the stimulus-organism-response theory.

Methods: This study used a mixed methods approach, including a web-based questionnaire (quantitative data) and interviews (qualitative data). We recruited 520 health worker respondents, including general practitioners, dentists, specialists, and medical students, as users of the D2D app. Quantitative data processing was conducted using covariance-based structural equation modeling, whereas qualitative analysis was conducted on the data from 15 respondents using the content analysis method.

Results: On the basis of the web-based questionnaire (quantitative) results, we found that cognitive (P=.01) and emotional (P=.004) app relationship quality affected health workers' stickiness in mobile learning. On the other hand, factors related to the functionality of the app and health workers' experience were proven to affect cognitive and emotional app relationship quality (P<.005). In addition, according to interview (qualitative) data, the performance of apps for mobile learning is influenced by information quality and information processing speed, which are needed to deliver a more efficient learning process and reduce the possibility of misunderstanding in the interpretation of learning materials. The user experience is influenced by gamification factors to make the learning process more fun, especially for medical students who do not have to obtain professional credit units (referred to as *satuan kredit profesional* in Indonesia), unlike physicians or specialists.

Conclusions: The results of this study will help mobile learning service providers increase user stickiness in mobile learning, for example, through processing speed, the quality of the information presented, security features, personalized content recommendations, and gamification.

(JMIR Med Educ 2025;11:e63827) doi:10.2196/63827

KEYWORDS

mobile health care learning; mobile learning; health worker; stickiness

Introduction

Background

To date, the health worker and medical student communities have used e-learning and tele-education to enable continuous education and training in the health sector, which is commonly known as continuing medical education (CME) [1]. One form of tele-education is mobile learning [2]. A 2020 report showed that the need to provide CME programs has increased; moreover, the number of medical personnel in Indonesia is always increasing [3]. In 2022, there was an increase in medical personnel in Indonesia by 3.3%, encompassing general practitioners, specialists, dentists, and specialist dentists [4].

Although the number of medical personnel has increased, Indonesia has a ratio of only 0.47 physicians per 1000 population [4]. One of the challenges faced by the health care sector in Indonesia is the uneven distribution of physicians—as many as 71,286 physicians, or approximately 57.63%, are located on the island of Java [5]. This challenge is exacerbated by the geography of Indonesia as an archipelagic country, which makes many physicians reluctant to work in remote areas due to inadequate regional infrastructure [5]. Thus, physicians in regions outside Java face challenges in providing health services and participating in CME programs, which are generally held only in big cities.

In Indonesia, the implementation of a CME program is regulated through Article 28 of Law 29 of 2004, which concerns medical practice. On the basis of that law, all physicians, dentists, and specialists are required to engage in continuous learning via programs conducted by professional organizations on current advancements in science and technology in their respective fields. The implementation of CME includes skill training, webinars, and various events through which attendees earn professional credit units (referred to as *satuan kredit profesional* [SKP] in Indonesia). On the basis of the Decree of the Minister of Health of the Republic of Indonesia HK.01.07/Menkes/1561/2024 concerning the *Guidelines for the Management of the Fulfilment of the Adequacy of Professional Credit Units for Medical Personnel and Health Workers*, SKP points are required to obtain or renew the certificate of expertise, namely, the Registration Certificate issued by the Indonesian Health Workers Council. In Indonesia, physicians and specialist physicians must obtain a score of 250 SKP points for 5 years and dentists and specialist dentists must obtain a score of 100 SKP points for 5 years for clinical (related to direct and indirect medical services) and nonclinical (eg, teaching, researching, conducting health managerial activities, and conducting professional or community service) activities. SKP points are reviewed by the chairman of the Indonesian physician association per region for general practitioners and by the chairman of the association of physicians for specialist physicians.

User stickiness to an app can be considered a key metric for evaluating the relationship between a user and an app [6]. It may influence their decision to adopt or discontinue the use of the app. According to Hsu and Tang [7], the stickiness of a mobile app is its ability to encourage continuous user interaction, which requires that the app maintain users' interest. Chen et al [8] argue that stickiness is the tendency of users to continue to visit apps or websites. One of the strategies that have been implemented to increase the retention rate (ie, stickiness rate) of apps is reinforcing the desire of the user to use the app. These efforts can take the form of direct promotion through webinars, recommendations and reviews by physicians and medical students, and reminder activity through the notification feature on the app [9].

The Doctor-to-Doctor (D2D) app (PT Global Urban Esensial), which competes with various other apps in Indonesia, including Docquity, Halodoc, and Alomedika, is a pioneer in obtaining SKP points for physicians. These 4 apps are the most popular mobile learning apps in Indonesia, especially regarding support for CME programs and webinars [10]. However, D2D is a pioneer in mobile learning and has high ratings and a large number of users in Indonesia compared to other apps. When developing health-related content for physicians, the materials are curated and validated by the company's internal medical team, where each piece of content is produced based on current and relevant medical topics [11]. Approximately 88% of D2D users are physicians practicing in Indonesia [11]. Physicians are also encouraged to engage in CME as part of their ongoing professional development. The D2D app offers various webinars tailored to specific target groups, such as webinars specifically designed for pediatricians [11].

The D2D app was first released in 2018, and to date, it already has >80,000 users on the Google Play Store and Apple App Store [11]. In addition, the D2D app is integrated with medical organizations in Indonesia, namely, the Indonesian Medical Association, and has been certified by the Indonesian Ministry of Health as one of the platforms that can be used to obtain SKP points [11]. Additional benefits for D2D users include access to the latest health content and information, up-to-date information on medical events, and the opportunity to attend free and certified webinars. Features of the D2D app include access to free webinars, health journals, medical discussion forums, and loyalty programs. However, the results of the study by Halim et al [12] show that D2D is only the fourth most frequently accessed app by physicians and medical students. Even though D2D is a pioneer in mobile learning and has amassed high ratings and a large number of users in Indonesia, this does not guarantee that the app will have a high level of user attachment or stickiness. Figures 1 and 2 show example webinar and CME features, respectively. The D2D app's health content is not limited to the D2D health worker community but can also be accessed by all health workers, which is not the case for its competitors such as Alomedika, as shown in Figure 3.

Figure 1. Webinar features of Doctor-to-Doctor.

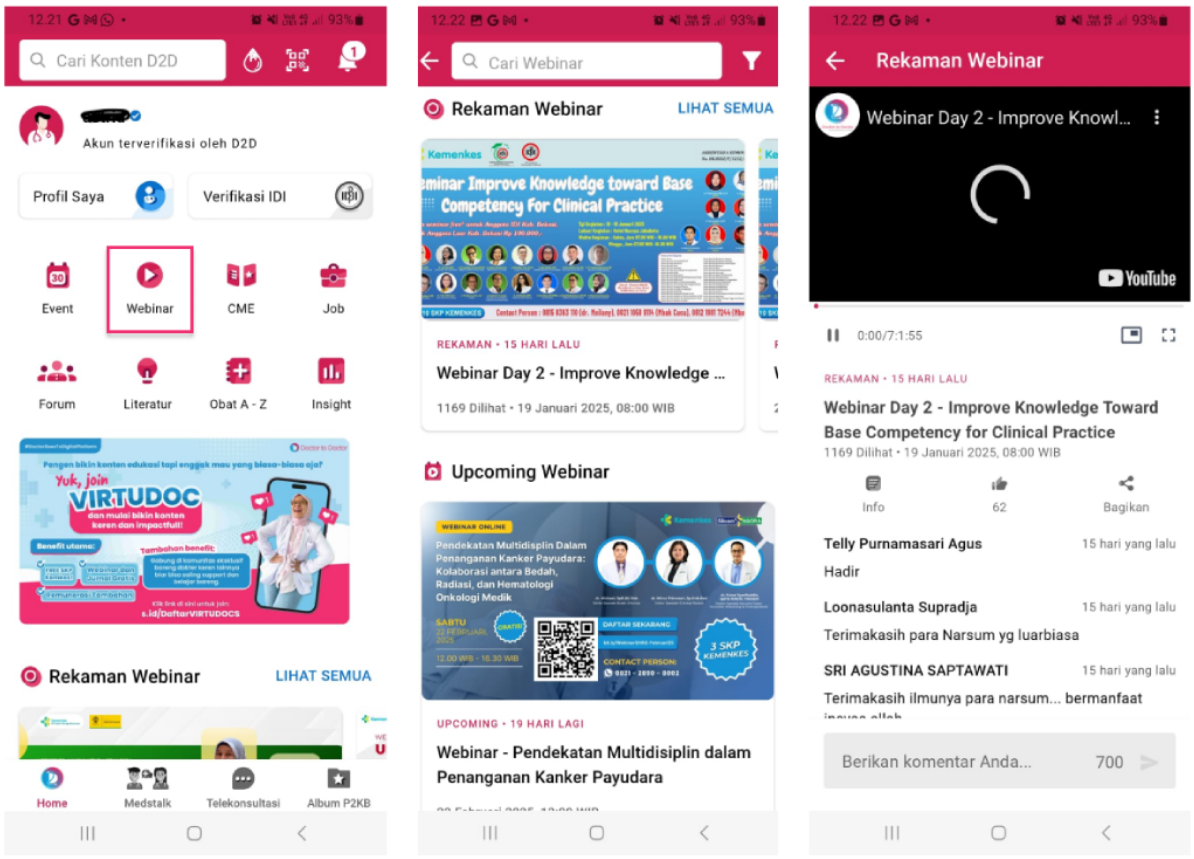


Figure 2. Continuing medical education features of Doctor-to-Doctor.

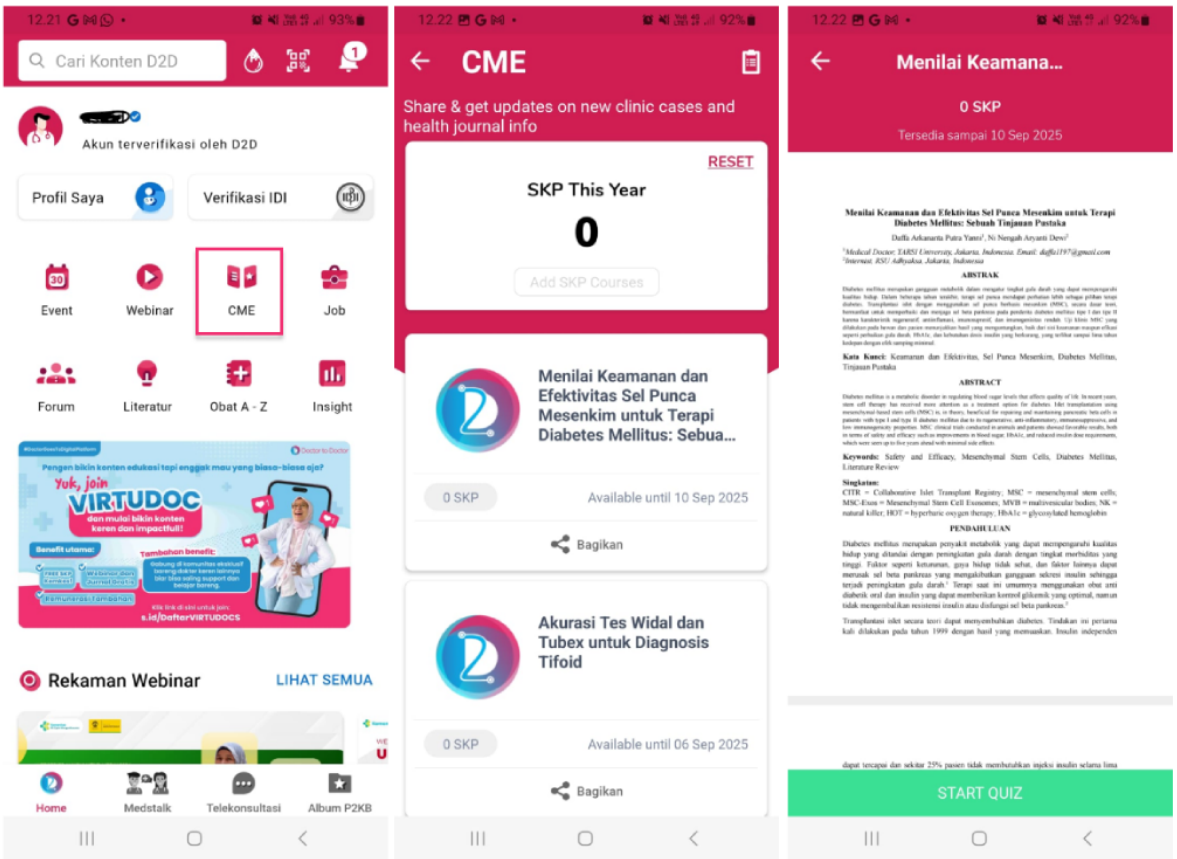
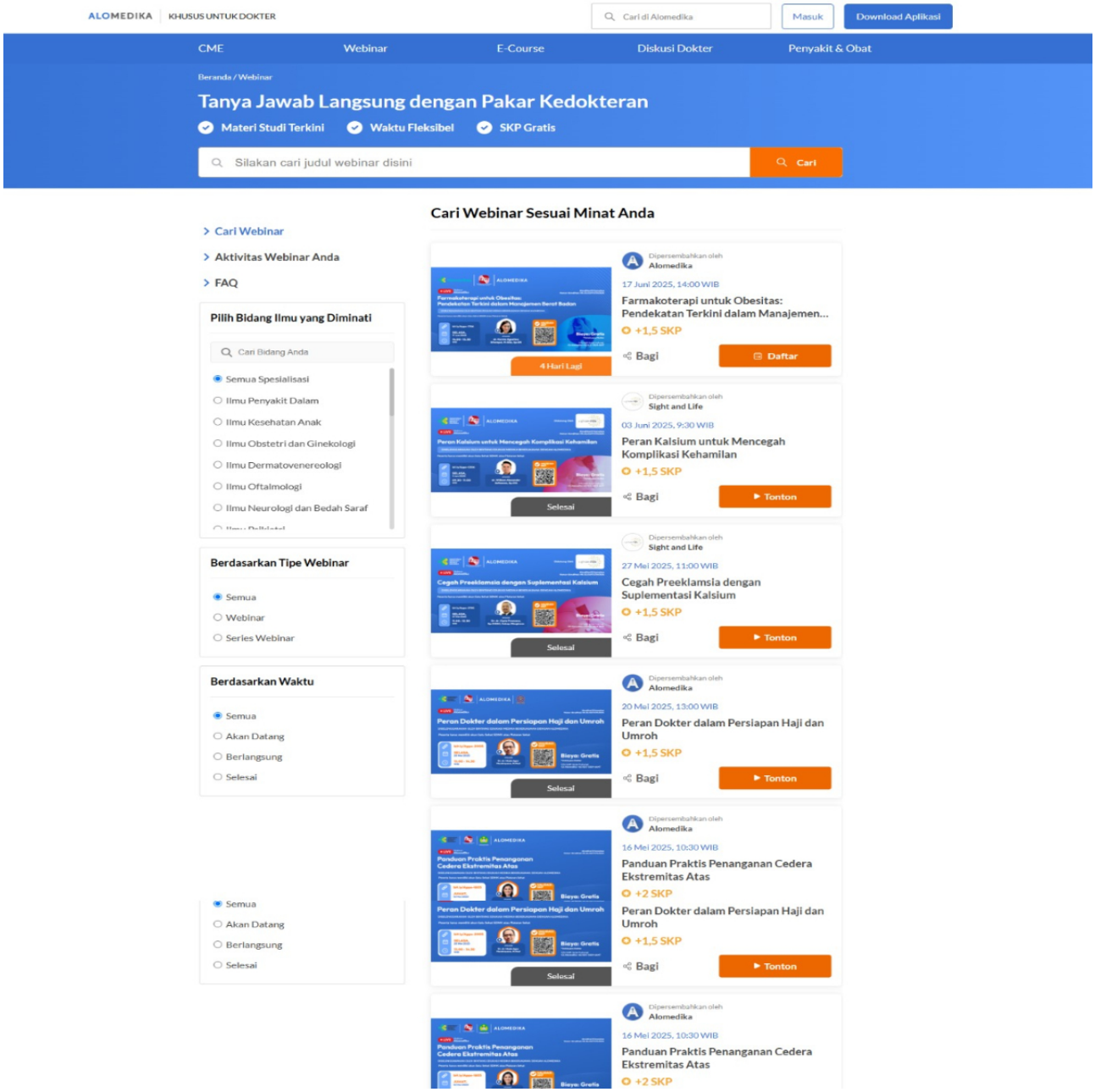


Figure 3. Webinar features of the Doctor-to-Doctor competitor.



Objectives

In the context of e-learning, stickiness can also be affected by the emotions of the users when using an app, such as those that drive a user's decision to continue using or delete an app [13,14]. Moreover, the system performance of the mobile learning app plays an important role in user attachment, specifically in meeting user needs [9]. Thus, stickiness, which is related to both the user's emotions and the performance of the app, is an important benchmark for measuring the intensity of mobile learning use. However, research related to user stickiness is still focused on retail apps [6,15], massive open online courses [16-18], and fitness apps [9]. Therefore, this research sought to answer the following question: What factors affect health workers' stickiness regarding mobile learning apps? We used a web-based questionnaire and interviews to reach a large number of respondents and triangulated quantitative results with interview data to better understand the findings of this study. We formed the hypotheses underlying this study by analyzing the relationship between functional and experiential factors that could influence cognitive app relationship quality (CARQ) and emotional app relationship quality (EARQ), which influence app stickiness among users of mobile learning apps. The results of this study can provide guidance for mobile learning service providers to formulate strategies for improving stickiness rates.

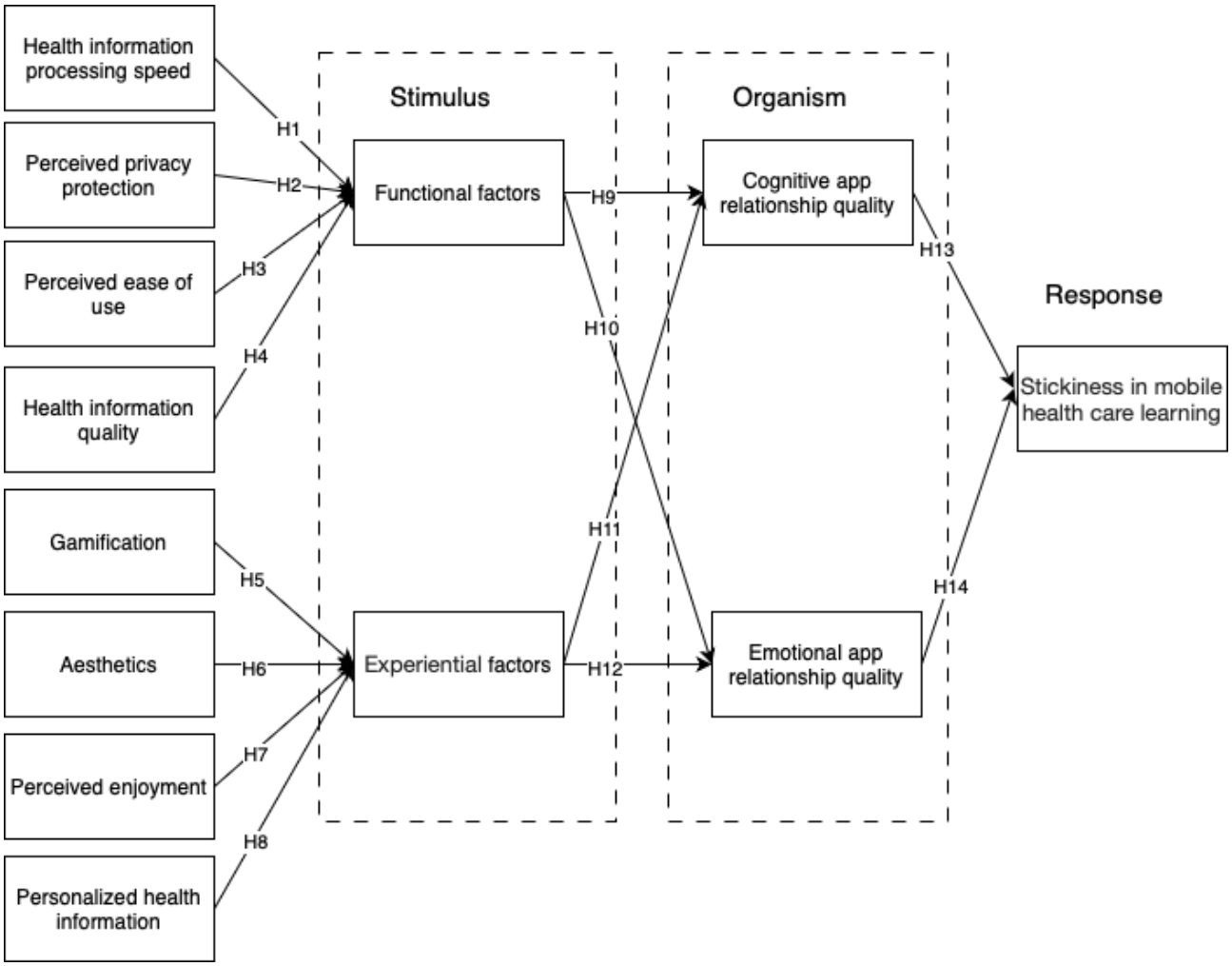
Methods

Research Model

This study's research model was based on the stimulus-organism-response theory [19] that explains the influence of a stimulus on the user assessment process and the relationship between the results of the individual assessment process and user habits [19]. The design of the proposed model

is referred to in multiple studies [6,13,20,21]. The selection of factors was adjusted to the scope of this study, namely, factors that affect user behavior regarding continuous use of an app (ie, user stickiness). This research focused on the theory of app relationship quality, namely, CARQ and EARQ, where the quality of the app relationship is driven by functional and emotional factors, respectively. Figure 4 shows the research model used in this study.

Figure 4. The proposed research model. H1 to H14: hypothesis 1 to hypothesis 14.



In this study, health information processing speed (HIPS) refers to the level of information processing speed in a given system [6]. This constitutes the ability of the system to process information, load pages, display content, or perform other functions [15]. The system speed is the speed at which a system can access information via an access link [22]. When an app can process user input or requests quickly and is able to present results without a delay, it becomes more convenient and efficient for users to carry out activities on the app [23]. Therefore, it can be concluded that the app response speed is one of the most important functional criteria for users of an information system. Indeed, Alnawas et al [6] mention that speed is one of the factors that affect product functionality. Thus, we propose the following hypothesis: HIPS influences functional factors in mobile learning performance assessment (hypothesis 1).

Perceived privacy protection (PPP) is defined as an app's ability to protect users' personal data [24]. This includes protection of data related to the users' browsing history and transaction history and other personal information [24]. On the other hand, PPP is also part of the security dimension of an app [6], and it can include the extent to which an app is able to protect users' personal data from unauthorized access by other parties [25]. Implementation of PPP can include notifying users that their personal data will not be shared with other parties [26]. Such implementation is also critical to preventing data problems caused by the unauthorized use of personal data [27]. Therefore, we propose the following hypothesis: PPP influences functional factors in mobile learning performance assessment (hypothesis 2).

Perceived ease of use (PEU) is defined as a user's assessment of an app based on its ease of use or performance [6]. PEU can also refer to the level of convenience that users feel when operating the app, such as not being confused when using or searching for information on the app for the first time [7]. The convenience that users feel will have a direct effect on their perception of the benefits obtained from the app [28]. The easier an app is to use, the greater the incentive for users to continue carrying out activities without feeling confused about the appearance or features [23]. Ease of use also reduces the time it takes for users to understand the app and its functionality and find solutions when facing problems with the app [25]. Previous research has shown that PEU has an influence on user

satisfaction and stickiness [29]. This is also supported by the work by Hsu and Tang [7], who included PEU in the top 5 factors considered to have the most influence on the user experience ratings of retail apps. PEU is also related to the functionality of an app as users prefer that an app be easy to operate—the app should be pleasing to use and should provide the requested information quickly and accurately [6]. Therefore, we propose the following hypothesis: PEU influences functional factors in mobile learning performance assessment (hypothesis 3).

Health information quality (HIQ) is defined as the level of information quality on an app, including the accuracy, reliability, clarity, and relevance of the content [29]. Such quality can be measured based on the speed at which the user reads the information and on the richness and reliability of the content (trusted sources) [29]. The content presented on an app must be informative and provide new insights to users [12]. Cognitive judgment is influenced by product functionality, an aspect of which is the ease with which users can find information [21]. In the context of mobile learning research, information is obviously a very important element in user learning [8]. The perceived usefulness (utilitarian benefit) of the information obtained during the learning process is the main focus of users of a given app [8]. Therefore, we formulate the following hypothesis: HIQ influences functional factors in mobile learning performance assessment (hypothesis 4).

Gamification is the implementation of game features in an app to increase motivation and interaction between users [13]. It can also refer to the adoption of a certain game design to influence self-awareness and individual behavior during the use of an app [13]. Gamification includes elements such as rewards or challenges that motivate users to continue interacting with an app [13]. Yang and Li [30] described the influence that gamification has on the user's experience and their decision to continue using a given mobile learning app. Aparicio et al [16] have similarly shown that gamification and personalization have a positive influence on the user app experience in massive open online courses. In the field of education, gamification features can encourage learners to become more involved in the learning process [16]. The adoption of gamification has been proven to encourage users to engage in more web-based learning activities, provide an enjoyable learning experience, and stimulate users to learn more [16,31]. Djohan et al [32] found that gamification also influences the user experience of an app. Hence, we propose the following hypothesis: gamification influences the experiential factors of the user's perceived experience of mobile learning (hypothesis 5).

Aesthetics, for the purposes of this study, are a perception produced by the visual appeal of an app through the selection of colors and illustrations, among other things [28]. Through the experience of visual beauty, aesthetics are indirectly able to attract the attention of users [33]. Aesthetics also provide immediate pleasure that is unrelated to the app's functionality or performance assessment [6]. According to Huang et al [33], users feel pleasure when they observe a beautiful app design display. What constitutes a beautiful display is influenced by a number of factors, including strategically arranged feature layouts, easy-to-reach search bar placement, feature placement design based on user need priorities, the use of unobtrusive colors and attractive icons, and the effective use of typography [24]. Huang et al [33] showed that the aesthetic value of an app has a significant influence on user experience. These findings are also supported by the work by Rose et al [34], who found that the aesthetic value of an app's display can affect the user's emotions while using the app. Therefore, we propose the following hypothesis: aesthetics influence the experiential factors of the user's perceived experience of mobile learning (hypothesis 6).

Perceived enjoyment (PE) is the user's perception of feeling happy or entertained or of enjoying the content or features provided by an app [7]. PE tends to be related to the user's emotions (eg, pleasure, happiness, and being entertained) while they are using the system. Pleasure or satisfaction in using an app is not determined by system performance alone but is more often influenced by the overall user experience [35]. On the basis of the work by Hsu and Tang [7], PE is one of the top 5 most important factors in creating a sticky app. The enjoyment that users feel can also increase interactivity when using an app [7]. Thus, we propose the following hypothesis: PE influences the experiential factors of the user's perceived experience of mobile learning (hypothesis 7).

Personalized health information (PHI) is a part of information personalization, the automatic provision of content based on user preferences and certain classifications [6,7]. Personalization refers to the app's ability to offer customized information, services, design, and content to meet user needs and preferences [36]. This helps users focus on what they want, determine the fitness of a product or service, and complete tasks efficiently [37]. Personalization provides customized content to improve the quality of the provided information and encourage users to complete certain tasks; personalization also encompasses user comfort [38]. The benefits offered by the adoption of personalization include the enjoyment that users feel [39], positive emotions [38], and feelings of joy [40]. Alnawas et al [6] found that personalization has a significant influence on the user experience. This is supported by the work by Cheng [13], who indicated that information personalization can affect the user experience and significantly affect the user's habit of using the app. Thus, we propose the following hypothesis: PHI influences the experiential factors of the user's perceived experience of mobile learning (hypothesis 8).

Functional factors are factors that directly affect the success of an app's performance [6]. Functionality is the main factor in evaluating the usefulness (ie, utilitarian benefit) of an app [41]. This is because app functionality influences the app's ability to help users achieve their goals safely and efficiently [6,41]. Cognitive assessment, the result of a comparison of expectations with app capabilities, can be conducted based on app functionality [7]. Wu [42] found that the main factor that significantly influences user engagement with an app is users' performance expectations. Therefore, we propose the following hypothesis: functional factors influence CARQ in the context of user assessment of the perceived usefulness of mobile learning (hypothesis 9).

Al-Nabhani et al [43] noted that functional factors also influence user satisfaction, which significantly affects users' emotions. The functionality of the app is another factor that users feel is related to their sense of pleasure or satisfaction with the app [44]. Responses related to the perception of emotions are called "judgment affective" responses and are related to EARQ [6]. Alnawas et al [6] state that a user's assessment of an app's functionality affects the emotional impact of a retail app, although not as strongly as its cognitive impact. Similarly, Molinillo et al [21] found that system functionality influences affective experience when using an app. Therefore, we propose the following hypothesis: functional factors influence EARQ in the context of user assessment of mobile learning based on emotional perception (hypothesis 10).

Experiential factors are defined as factors related to the user's subjective experience, including aspects such as perception, emotion, preferences, and user interaction with a product [6]. User experience can affect the sensation of using the app [23]. This refers to the psychological state experienced by users when evaluating the app, that is, the emotions felt while using the app [6]. Customer experience is subjective and contextual because it can differ among customers, cultures, and situations [32]. Each individual may have different feelings because their experience affects their assessment of the app [32]. Any emotions that users experience using an app generally affect the perception of pleasure and can encourage users to continue using the app [45]. Although user experience is not directly related to app performance, it affects the user's assessment of the usefulness of the system [6]. In this case, emotions become an important part of cognitive assessment, specifically related to processing the value of an experience, and can form the basis for a deep and intense relationship with the app [6]. User experience also significantly influences satisfaction and trust, to the point that experiences can form the basis for an emotional connection and a deep interest in a given product [6]. Thus, we suggest the following hypothesis: experiential factors influence CARQ in the context of user assessment of the perceived usefulness of mobile learning (hypothesis 11).

Experiential factors include users' emotions in relation to their assessment of the usefulness of an app [6]. Indeed, the emotions felt by users occur and are processed through human senses, especially those related to the appearance of the app [38]. User experience is the experience of users when they do something that involves them psychologically in activities requiring concentration, motivation, and enthusiasm [6,32]. This user experience includes aesthetic value, perception of pleasure (ie, enjoyment), and gamification [6,46]. Moreover, van Noort and van Reijmersdal [45] found that, if an app offers entertainment, it can produce a fairly good level of app performance satisfaction based on the user's emotional perception. User experience, derived from users' emotional assessments of a technology, plays a critical role in determining the use and benefits of the technology, including in the context of a mobile app [7]. Thus, we propose the following hypothesis: experiential factors influence EARQ in the context of user assessment of mobile learning based on emotional perception (hypothesis 12).

Alnawas et al [6] described how EARQ and CARQ affect the level of attachment of app users (ie, stickiness). Zolkepli et al [44] explained that this effect is related to differentiating user behavior based on feelings (affective) and thoughts (cognitive). Cognitive assessment is based on rational thinking, so it is more changeable and vulnerable to external influences such as other people's opinions [47]. The cognitive assessment of apps is usually related to the ease with which users can find factual information in accordance with their needs [48]. A relationship based on CARQ emphasizes analysis of the benefits obtained by the user, so it is necessary to conduct periodic evaluations of the usefulness of the available features [47]. In addition, it has been proven that user satisfaction with app performance—in the context of utilitarian benefits—positively affects the intensity of use of mobile apps [6]. Thus, both the quality of the app's performance and its emotional approach to the user experience can influence the user's desire to invest time and effort in using the app [49]. Therefore, we propose the following hypothesis: CARQ influences stickiness in mobile learning apps and determines the intensity of daily use (hypothesis 13).

Behaviors based on feelings tend to be more subjective because they relate to aspects of positive or negative judgment, whereas behaviors based on thoughts tend to be more objective and relate to notions of right or wrong [44]. As responses that arise due to feelings or emotions tend to be more subjective, it is quite difficult to change affective perceptions [21]. Emotional perceptions experienced while using mobile learning apps greatly affects app access intensity during the web-based learning process [13]. Perceptions that increase the duration of app access generally tend to be positive, such as feeling happy and entertained and enjoying using the app [50]. Nyffenegger et al [51] stated that user attachment to an app can be reflected through satisfaction and trust built based on the intensity of app use. User engagement with the app is the result of user assessments based on experience (ie, affective experience) [21]. Affective judgment is judgment based on the feelings of an individual, which makes influence by external forces more difficult [48]. Decisions based on experience are difficult to change as they shape individual beliefs and perceptions, including the use of apps [48]. Fernandes and Proença [47] explained that, when attitudes are dominated by feelings, individuals tend to behave habitually, making those feelings resistant to change and resulting in a much stronger response compared to attitudes based purely on rational thinking. Thus, we propose the following hypothesis: EARQ influences stickiness in mobile learning apps and determines the intensity of daily use (hypothesis 14).

Research Process

We used a quantitative approach to test the hypotheses and a qualitative approach to more deeply analyze the results of the hypothesis testing. An online questionnaire was used as a quantitative research instrument, and interviews were used to collect qualitative data. The focus of this study was on mobile learning, namely, the D2D app. We used purposive sampling, where all respondents involved in this study were D2D users—physicians, dentists, specialists, and medical students (eg, coassistants, residents, dental students, and medical undergraduate students). Furthermore, we asked the D2D company to help distribute the questionnaire links through push notifications on the D2D app. We also used social media

platforms such as WhatsApp, X (formerly known as Twitter), and Instagram, which are popular in Indonesia, to reach more health care workers.

We conducted a readability test on the questionnaire to ensure that the research instrument was easily understandable, in accordance with writing standards, and relevant to the context of the research. The readability test was carried out from February 1, 2024, to February 7, 2024. In total, 6 respondents participated in the readability test process. Most of the changes after the readability test were made due to ambiguous word choice or unclear syntax. After the readability test, a pilot study was conducted to test the reliability of the research instrument.

The pilot study was conducted to find whether participants experienced problems filling out the questionnaire before it was disseminated to more participants. The pilot study was conducted from February 19, 2024, to February 23, 2024, with a total of 50 respondents. The total Cronbach α value obtained in the pilot study was 0.981; thus, we distributed the questionnaire link to a larger number of respondents. Finally, based on the results of the quantitative analysis, hypotheses 2, 3, 6, 7, 8, and 12 were rejected. We used a qualitative instrument (semistructured interviews) to determine the reason for these rejections and obtain a deeper understanding of the results. Online and offline interviews were conducted with 15 respondents.

Research Instruments

The web-based questionnaire included demographic questions and measurement items. The preparation of the measurement items was carried out with reference to previous research, such as the studies by Alnawas et al [6], Chen et al [8], Cheng [13], Hsu and Chen [52], Huang et al [33], Elsotouhy et al [9], and Yang et al [53], and contextualized for this study, namely, mobile learning in health care. To make it easier for respondents to fill out the questionnaire, we translated the measurement items in Indonesian and tested them by conducting readability tests and pilot studies to ensure that respondents could understand the questionnaire. Each measurement item was given a code according to the variables it represented to make it easier to process the data. The assessment of the measurement items was then carried out using a 5-point Likert scale that included options ranging from “strongly disagree” to “strongly agree.” Multimedia Appendix 1 contains the questionnaire instrument and the 13 interview questions. We defined the interview questions according to the variables used in this study. The results of the questionnaire were analyzed using covariance-based structural equation modeling in SPSS Amos (version 24; IBM Corp).

In addition, the qualitative data were analyzed using content analysis. The results of the semistructured interviews were interpreted and linked to the hypotheses identified during the quantitative data analysis stage. The qualitative codification process was carried out using interview quotes that were grouped according to the specific hypotheses. Multimedia Appendix 1 provides the details of the qualitative results. An example of the codification results can be found in the Qualitative Results and Validity of Hypothesis Testing section.

Ethical Considerations

This study received approval from the research unit of the Faculty of Computer Science, University of Indonesia (reference S-17/UN2.F11.D1.5/PPM.00.00/2024). In line with university policy, the Research and Community Service Department, Faculty of Computer Science, University of Indonesia, adhered to the guidelines and procedures established by the faculty and provided ethics approval for this study. This study was also approved by the D2D company (reference 761/GUE/IX/2024/E). All respondent data were anonymized, all questionnaire respondents provided written informed consent, and all the interview participants provided verbal informed consent to take part in this study. Participants did not receive compensation.

Results

Quantitative Results

Respondents’ Demographics

Data collection for the primary study was carried out for approximately 25 days, from February 29, 2024, to March 21, 2024. The number of respondents who filled out the complete questionnaire was 520, which fulfills the requirements by Hair et al [54], where the minimum number of respondents required is the number of measurement items multiplied by 10. Detailed respondent demographic characteristics are shown in Table 1, and Table 2 shows the demographic characteristics of the interviewees. The interviews lasted between 30 and 60 minutes and took place between April 19, 2024, and April 27, 2024.

Table 1. Questionnaire respondents' demographic characteristics (N=520).

Variable	Respondents, n (%)
Gender	
Woman	311 (59.8)
Man	209 (40.2)
Age (y)	
<20	3 (0.6)
20-30	206 (39.6)
31-40	132 (25.4)
41-50	96 (18.5)
>50	83 (16)
Domicile	
Greater Jakarta	280 (53.8)
Java island outside Greater Jakarta	103 (19.8)
Outside Java island	137 (26.3)
Occupation	
General practitioner	358 (68.8)
Specialist	59 (11.3)
Dentist	4 (0.8)
Coassistant	46 (8.8)
Resident	7 (1.3)
Dental student	6 (1.2)
Medical undergraduate student	40 (7.7)
D2D[a] app use period (mo)	
<6	97 (18.7)
6-12	94 (18.1)
13-24	131 (25.2)
>24	198 (38.1)

[a]D2D: Doctor-to-Doctor.

Table 2. Summary of the interviewees' demographic characteristics.

Respondent	Gender	Age (y)	Occupation
1	Woman	23	General practitioner
2	Woman	25	General practitioner
3	Woman	23	Dentist coassistant
4	Man	23	Coassistant
5	Woman	19	Medical student
6	Man	24	General practitioner
7	Woman	23	General practitioner
8	Woman	23	Dentist coassistant
9	Man	25	General practitioner
10	Woman	20	Medical student
11	Woman	24	General practitioner
12	Woman	36	General practitioner
13	Woman	24	General practitioner
14	Man	19	Medical student
15	Man	19	Medical student

Measurement and Structural Model Testing

The average variance extracted (AVE) value is obtained from the sum of the squares of the values of the factor loadings of each indicator divided by the sum of the squares of the values of the factor loadings of each indicator, and this is added to the total measurement error of all indicators on one variable [55]. The AVE value of a variable is declared valid if it is >0.5 [54,55]. On the basis of the results, all variables in the research model passed the AVE value test (Table 3). The reliability test stage was carried out by checking the value of the composite reliability on each latent variable. The composite reliability of each latent variable is said to be valid if it has a value of >0.7 [54]. On the basis of the results shown in Table 3, all latent variables in this study passed the reliability test.

Table 3. Average variance extracted (AVE) and composite reliability (CR) test results.

Variables	CR	AVE
Stickiness in mobile learning	0.736	0.583
Cognitive app relationship quality	0.79	0.654
Emotional app relationship quality	0.991	0.983
Functional factors	0.99	0.979
Experiential factors	0.752	0.603
Health information processing speed	0.718	0.561
Perceived privacy protection	0.838	0.72
Perceived ease of use	0.989	0.979
Health information quality	0.739	0.588
Gamification	0.802	0.67
Aesthetics	0.849	0.585
Perceived enjoyment	0.814	0.687
Personalized health information	0.857	0.601

Hypothesis Testing

After the research model passed the validity and reliability tests, the next step was to check the goodness of fit (GOF) of the model. The GOF test consists of absolute fit indexes, incremental fit indexes, and parsimony fit indexes. Absolute fit indexes determine model compatibility by checking the values of the root mean square residual, GOF index, adjusted GOF index, and root mean square error of approximation. Incremental fit indexes are tests of normed fit index, comparative fit index, and Tucker-Lewis index values. Meanwhile, parsimony fit

indexes test the parsimonious normed fit index. The results of the GOF test are shown in Table 4.

The determination coefficient test was carried out to determine how well the endogenous variables simultaneously explained the exogenous variables [54]. The determination coefficient, which ranges from 0 to 1, is calculated from the square of the correlation (R^2) between its dependent and independent variables [54]. The R^2 value is categorized as strong if it is >0.67, moderate if it is between 0.33 and 0.67, and weak if it is >0.19 but <0.33 [54]. The results of the determination coefficient test are shown in Table 5.

In this study, we used a 2-tailed hypothesis test. If the *P* value was <.05, then the hypothesis was accepted; if the *P* value was >.05, the hypothesis was rejected [54]. The results of the hypothesis test are shown in Table 6.

Table 4. Goodness-of-fit test results.

Test	Requirement	Result	Description
Chi-square	>0.05	315.5	Good fit
Chi-square divided by *df*	<2.0	1.024	Good fit
Goodness-of-fit index	>0.9	0.949	Good fit
Root mean square error of approximation	≤0.08	0.008	Good fit
Root mean square residual	≤0.05	0.02	Good fit
Normed fit index	≥0.9	0.943	Good fit
Relative fit index	≥0.9	0.92	Good fit
Tucker-Lewis index	≥0.9	0.998	Good fit
Comparative fit index	≥0.9	0.999	Good fit
Parsimonious normed fit index	0-1	0.668	Good fit
Adjusted goodness-of-fit index	≥0.9	0.923	Good fit

Table 5. Results of the determination coefficient test (R2).

Variables	R^2 value	Description
Experiential factors	0.342	Moderate
Functional factors	0.201	Weak
Emotional app relationship quality	0.170	Weak
Cognitive app relationship quality	0.287	Weak
Stickiness in mobile learning	0.421	Moderate

Table 6. Hypothesis test results. The arrows represent the direction of influence.

Hypothesis	Estimate	*P* value	Result
Hypothesis 1: HIPS[a] →FFs[b]	0.282	.01	Accepted
Hypothesis 2: PPP[c]→FFs	0.055	.58	Rejected
Hypothesis 3: PEU[d]→FFs	−0.036	.49	Rejected
Hypothesis 4: HIQ[e]→FFs	0.177	.04	Accepted
Hypothesis 5: GM[f]→EFs[g]	0.337	.005	Accepted
Hypothesis 6: AE[h]→EFs	0.214	.07	Rejected
Hypothesis 7: PE[i]→EFs	0.081	.43	Rejected
Hypothesis 8: PHI[j]→EFs	0.071	.55	Rejected
Hypothesis 9: FFs→CARQ[k]	0.343	.006	Accepted
Hypothesis 10: FFs→EARQ[l]	0.505	.003	Accepted
Hypothesis 11: EFs→CARQ	0.143	.003	Accepted
Hypothesis 12: EFs→EARQ	0.102	.11	Rejected
Hypothesis 13: CARQ→SHL[m]	0.412	.01	Accepted
Hypothesis 14: EARQ→SHL	0.416	.004	Accepted

[a]HIPS: health information processing speed.
[b]FF: functional factor.
[c]PPP: perceived privacy protection.
[d]PEU: perceived ease of use.
[e]HIQ: health information quality.
[f]GM: gamification.
[g]EF: experiential factor.
[h]AE: aesthetics.
[i]PE: perceived enjoyment.
[j]PHI: personalized health information.
[k]EARQ: emotional app relationship quality.
[l]CARQ: cognitive app relationship quality.
[m]SHL: stickiness in mobile learning.

Qualitative Results and Validity of Hypothesis Testing

Hypothesis 1: HIPS and Functional Factors

This study showed that health information processing has an influence on functional factors (hypothesis 1). This is in accordance with the results of previous research related to the effect of stickiness on retail mobile apps, where information processing speed significantly affects app performance (functional factor) [6]. This is also supported by the work by Hsu and Tang [7], who found that response speed in an app is an important indicator of app functionality in terms of responsiveness. In health app studies, speed has an influence on the user's attachment and decision to continue using those apps [9]. In the context of mobile learning by medical students, the speed of information processing plays an important role in facilitating a more efficient learning process, allowing students to receive information faster, thereby reducing the possibility of misunderstandings in the interpretation of learning materials [1].

The results of the interviews showed that interviewees agreed that the speed of information processing was an important part of the functionality of the app—the app should be fast and meet user expectations:

> *The response speed of the app is very good and fast. There are no obstacles in accessing all the features in the D2D app.* [Respondent 1]

On the other hand, users were also happy if the presentation of the information was fast and occurred without advertisements:

> *Yes, it has to be fast and not load without ads.* [Respondent 4]

However, from the users' perspective, there were network constraints that sometimes made it difficult to access information or content on the mobile learning app:

> *...Never have experienced a loading failure on the app page or GT error. Most of all, what I have told you is that the network in my area is bad.* [Respondent 2]

Hypothesis 2: PPP and Functional Factors

PPP did not influence functional factors in this study (hypothesis 2). According to the *Harvard Business Review* [56], PPP is not a priority for users—because users are in a hurry to use the app, they do not pay much attention to the security aspect of personal data [56]. On the basis of the results of the interviews, there is a growing sense that users' trust in the app's ability to keep personal data secure does not affect their assessment of the app's functionality:

> *...Honestly, the functionality is more of a process to create a password and verify it is not complicated, which makes it easier for users. For data security, it's more about creating a sense of trust in users.* [Respondent 13]

Regarding the protection of personal data, users trusted the app more when it partnered with a trusted organization than when the capabilities of the security system were invisible in the user interface:

> *As long as the app is already partnering with a trusted organization.* [Respondent 8]

However, users also wanted the appearance and flow of the app to remain easy to understand even when there was a process of verification and password reset, which is generally considered to require more steps for security:

> *I think yes, the more requests the better. But on the other hand, I became very lazy because the process was too complicated.* [Respondent 4]

PPP plays a greater role in influencing user trust in the app. Although keeping personal data secure often involves a time-consuming process, users preferred simpler and faster verification processes and security settings.

Hypothesis 3: PEU and Functional Factors

In contrast with the study by Alnawas et al [6], this study showed that PEU has no significant effect on functional factors (hypothesis 3). This hypothesis was rejected because users are often not proficient in using the information system provided by the app. This finding is supported by respondents' answers regarding the main obstacles they encountered when accessing the app, namely, time limitations and a busy practice schedule, which were mentioned by 69.4% (361/520) of the respondents. According to Smith [57], PEU also has no effect on the assessment of system performance because users find it difficult to locate information that matches their preferences. This is further supported by the work by Rakhmadian et al [58], who found that users become unhappy if they only have limited access to the information provided by the system. According to the findings of the interviews, users experienced constraints and limitations in feature functionality when using mobile learning. First, users found it difficult to navigate the account verification and registration processes. Second, users often felt that notifications that suddenly appeared during the use of the app were somewhat annoying:

> *Sudden updates are often confusing.* [Respondent 10]

Third, users found it difficult to begin working within the app because there were often sudden account log-ins and exits that forced them to re-enter their medical ID number:

> *I have trouble quite often. Suddenly, go out and go in again. Suddenly, log out and log in to the account again.* [Respondent 12]

Hypothesis 4: HIQ and Functional Factors

The results of this study showed that HIQ has an influence on functional factors (hypothesis 4). This is also supported by the work by Elsotouhy et al [9], who stated that information quality, which is the result of a comparison between expectations and user perceptions of the presented information, is an indicator of functional factors. The quality of the information also determines the adequacy, relevance, thoroughness, and timeliness of user interpretation [59]. In a study of user stickiness regarding mobile news apps, it was found that the information quality affects the satisfaction and attachment of app users [29].

According to the interviewees, the existence of reliable references and sources can minimize misunderstandings or the spread of hoaxes:

> *...minimizing misunderstandings of hoax information because the source of information is not detailed and clear.* [Respondent 2]

Interview respondents were also happy with the information presented via D2D because it included references, clear titles, trusted speakers, and rewards such as certificates after completing a webinar. However, some interviewees mentioned shortcomings regarding the content presented on the app (eg, the health journal content was not as complete as on the journal's website: "...less complete than those available on the journal website" [Respondent 1]). In addition, it was also rare to find health articles written by physicians on the app:

> *Unfortunately in D2D, not all doctors are authors. Unfortunately, there are no results of doctors' research in D2D.* [Respondent 5]

This is in line with the work by Yang et al [53], who noted that the quality of the information affects system performance and user stickiness.

Hypothesis 5: Gamification and Experiential Factors

This study found that gamification has an influence on experiential factors (hypothesis 5). This result is in line with the work by Cheng [13], who described how gamification plays a role in the learning process and significantly affects the user experience, thus impacting emotional attachment to an app (EARQ). Gamification features of web-based learning include reward elements and leaderboards to encourage learners to participate [24]. In research related to stickiness in retail apps, gamification is an important element in attracting users and retaining app use [7]. Another study found that gamification has a high influence on experiential factors in mobile commerce apps [32].

The results of the interviews showed that gamification made using the app more fun:

> *It feels like a learning app with a gamification feature makes it fun and exciting. The collection of program points or SKP points is one of the motivations for me to access webinars or learn apps.* [Respondent 1]

However, it was unfortunate that the gamification in the D2D app focused more on the SKP point collection program. This did not incentivize users with medical student or dental student profiles to join the app program because they had no need to obtain SKP points:

> *Never made, because there is no need yet.* [Respondent 3]

In addition, physicians who still have internship status even though they have obtained a registration certificate as a physician cannot participate in the program:

> *Currently, there is no need to use SKP points because they are still using Internship [points]*. [Respondent 9]

Even so, gamification is an interesting feature of this mobile learning app because it presents a dashboard that is significantly more informative than that of its competitors:

> *I think this is one of the main factors for using D2D. Unlike the next app, which only tracks the number of SKP points—it cannot collect SKP points.* [Respondent 6]

Hypothesis 6: Aesthetics and Experiential Factors

This study found that aesthetics have no significant effect on experiential factors (hypothesis 6). This result is in contrast to those of Alnawas et al [6], who found a significant relationship between aesthetics and experiential factors in retail apps. Zhou [59], whose results also contrasted with those of Alnawas et al [6], noted that design trends are dynamic, which affects users' aesthetic preferences regarding a website's (also dynamic) appearance. This is also supported by the work by Chen et al [8]—an increased number of target users of an app can affect subjective judgments related to the appearance of the app interface. As demonstrated by the demographic results, this mobile learning app has a diverse target user pool that includes general practitioners, specialists, dentists, medical students, dental students, coassistants, and residents. Thus, hypothesis 6 was rejected due to the variety of app user personas, which leads to many subjective judgments related to the visual assessment of the app.

Most interviewees stated that, when making decisions about using a product, the visual appearance or aesthetic aspect of the product is an important factor that is based on subjective interest:

> *I think it's very important, if I am not attracted to it from the beginning, I'm lazy to open it again.* [Respondent 5]

However, the interviews also showed that the color selection in the mobile learning app is still too striking:

> *...don't use the color of the ring.* [Respondent 2]

In addition, it was found that the text displayed by the mobile learning app was too small, so it was difficult to read:

> *The selection of the size of the text is too small.* [Respondent 15]

Moreover, some users suggested emphasizing the app's superior features, namely, medical discussions, so that users are more interested in participating in them:

> *If possible, the superior features that are emphasized in the medical discussion so that it can be more interactive between users.* [Respondent 12]

Other users suggested adding animation elements, such as a mascot, to create a lively atmosphere on the app:

> *My suggestion is adding a mascot or other animation to make it interesting.* [Respondent 5]

Finally, users indicated that they expected periodic design improvements to prevent them from getting bored with the appearance:

> *...The design is also always updated so that you don't get bored.* [Respondent 15]

The interviews indicated that there are many possible improvements to the appearance of the app that the development team can consider. This reinforces the statement by Zhou [59] that it is necessary to continuously improve the appearance of an app. The appearance of the mobile learning interface should also be tailored to user preferences [17]. Recommendations for mobile learning service providers include evaluating the app design using more comprehensive methods such as heuristic evaluation or user experience design.

Hypothesis 7: PE and Experiential Factors

PE had no significant effect on experiential factors in this study (hypothesis 7). The interviews showed that users did not feel much enjoyment during the learning process on the D2D app; they accessed the app with clear goals in mind, such as obtaining the required SKP points or accessing the latest health information:

> *It's normal. Enjoy it, I access the app to read discussions, literature, forums.* [Respondent 9]

Due to the similarity of the features and the flow of the process of using the app, users can become accustomed to or feel bored by it:

> *It's actually normal because almost all learning apps have a similar flow, so it's just normal to me.* [Respondent 7]

The interviews revealed that users have different perceptions when using mobile learning apps and when using social media apps. They tend to use mobile learning apps for the purpose of learning, whereas social media apps are used for entertainment. Finally, enjoyment related to using the app was found to be influenced by the interactions between users within the app:

> *In addition, there are still few doctors, and the webinar is very interesting. But the interaction between fellow doctors is lacking.* [Respondent 12]

The interviews showed that interaction between users is still quite minimal, especially via the discussion feature. To further

encourage users to learn, the development team might consider including gamification within the app [13].

Hypothesis 8: PHI and Experiential Factors

PHI had no significant effect on experiential factors in this study (hypothesis 8). These results are not in accordance with those of either Alnawas et al [6] or Hsu and Tang [7]; the latter stated that personalization significantly affects experiential factors. Information personalization is important because it connects the app user with the information provided on the app in accordance with the needs and interests of the user [29].

On the basis of the interviews, if the app's content is in accordance with users' needs and preferences, it will have an impact on the intensity of use:

> *From me, it is certain, because it is according to my needs, and I will continue to access it.* [Respondent 2]

In addition, users with student profiles noted having difficulty finding health information, journals, and webinars aimed at students:

> *I think it is still lacking because I want D2D to be able to categorize the material. If it can be divided by topic according to the station, it can be a children's station, so I can use it more comfortably.* [Respondent 4]

The interviews highlight the disappointment of respondents regarding the suboptimal implementation of personalization. Currently, D2D stores only users' work profiles, and the app has not optimized content personalization. Finally, it is important to remember that, when implementing personalization, it is necessary to consider the user's privacy and ensure that previous approval is granted [60].

Hypothesis 9: Functional Factors and CARQ

This study showed that functional factors do not have a significant influence on CARQ (hypothesis 9). This is in contrast to the work by Alnawas et al [6]. According to Zhampeissova et al [61], offering too many features or too much information via mobile learning increases the user's cognitive process load. Therefore, even if the system's ability to present information is very good, too much variety of that information can distract users from learning [18]. Wilmer et al [62] explained that users' attention when accessing mobile apps is generally diverted by additional features, resulting in a distraction from the users' original purpose. Wardaszko and Podgórski [63] stated that the success of cognitive processes is not entirely influenced by the success of the implementation of mobile learning. This is because each individual has different learning styles and cognitive needs, so a given app functionality may be beneficial for certain groups but not for others [63].

In the context of mobile learning, the interviews showed that users accessed the app with a clear purpose of obtaining the latest medical information:

> *I still need to find a reliable place or platform that can support my learning needs.* [Respondent 7]

However, users felt that the material provided to support their cognitive needs was unsuitable:

> *...But unfortunately, it is still rare for information for dentists.* [Respondent 3]

In addition, it is known that users obtain information in various ways, such as by reading various platforms, by preparing for an examination, and through videos:

> *To be honest, I cannot learn at this time from only this one platform, because my type of learning has to read as much as possible to understand better.* [Respondent 7]

Finally, the interviews made clear that users focus more on utilitarian benefits that can be felt directly, such as features that support learning, the collection of SKP points, the quality of the articles and medical information, and the speed of information processing:

> *Personally, I prioritize the usefulness of the app over the user experience.* [Respondent 14]

The interviews support that the functionality of the app does not have much effect on the cognitive process because what the user focuses on is the usefulness of the information to support the learning process. This is in line with the work by Cheng [13], who argued that information personalization is a very important part of directing users toward progress in the learning process.

Hypothesis 10: Functional Factors and EARQ

This study showed that functional factors affect EARQ (hypothesis 10). Previous research has shown that functional factors affect the emotions felt by users when using retail apps [6]. Furthermore, if the service provider focuses on achieving ease of operation and high performance, this affects the user's emotions and attitude toward using the app [47]. Chen et al [8] explained that technological capabilities, which implicitly include system performance, have a significant impact on the level of satisfaction with the use of technology in the web-based learning process. Good app performance capabilities, such as information processing speed, can make users feel satisfied and happy because they feel that the app is useful [44]. On the basis of the interviews, the emotions felt by users due to the quality of the app's functionality significantly affected their decisions regarding continued app use:

> *...The better the functionality of the app, the more it will meet my expectations, which has an impact on me being happy and likely to keep the app.* [Respondent 15]

Hypothesis 11: Experiential Factors and CARQ

Experiential factors affected CARQ in this study (hypothesis 11). In the context of mobile commerce, user experience has a significant influence on the level of ease of use when operating the app features [8]. Experiential factors are also related to the ease of information collection, which is one of the utilitarian benefits [6]. Dastane and Haba [17] explained that experiential factors can influence a user's perception of the benefits of the available features either directly or indirectly. Interviewees agreed that the appearance of the app affects the user's level of

trust and their perception of the process of obtaining information:

> *On the other hand, a good user experience can also make it easier for users to use their features and find information more easily.* [Respondent 10]

The experience provided by the app also affects the user's decision to continue using it:

> *Of course I will move to another one also if it looks bad.* [Respondent 9]

User considerations in the context of mobile learning still include utilitarianism and the experience provided:

> *But still in terms of performance, you also have to look at the performance of the app, at least not very slow, it's still okay for me to use it.* [Respondent 11]

One of the insights obtained from the interviews is that, if the app provides a suboptimal user experience but the features have a high benefit value, users will still use the platform. However, if a competitor is able to present similar features with a better user experience, users are more likely to switch to them.

Hypothesis 12: Experiential Factors and EARQ

Experiential factors were found to affect EARQ in this study (hypothesis 12). Alnawas et al [6] explained that user experience has a significant influence on the relationship between app quality and the emotional judgment of retail app users. Experiential factors also influence the users' emotions because they provide hedonistic benefits [43]. Improving the user experience is one way to shape the interaction between users and apps; user responses to interacting with an app or service can include emotions, perceptions, preferences, behaviors, and enjoyment [7]. The user experience then affects pleasure during purchasing activities and can encourage users to engage with the app [45]. The interviews showed that a good user experience on the app affects feelings of happiness when using it. Users will also be more interested in using the app if the experience is memorable:

> *But on the other hand, they will be more interested if the app has good coloring and a cute design.* [Respondent 4]

In addition, users will also have increased trust in apps that have a good design:

> *The experience on the app also affects me, so I have more trust in the app.* [Respondent 7]

Hypothesis 13: CARQ and Stickiness in Mobile Learning

This study found that there is a relationship between CARQ and stickiness in mobile learning in health care (hypothesis 13). This result is in accordance with previous research showing that CARQ has a significant influence on stickiness in retail apps [6]. In the context of mobile commerce apps, the results of previous research have shown that the cognitive relationship with apps affects user experience, such as satisfaction levels, and this satisfaction affects the user's desire to use apps in the future [34,64]. The cognitive relationship dimension includes the ease of searching for information [64] and the ease of efficiently purchasing goods [34]. In addition, CARQ has been proven to have a significant effect on learning persistence in the context of the web-based learning process using massive open online courses [13,65].

The interviews showed that one of the drivers of using the app continuously was users' desire to obtain the latest medical information:

> *In my opinion, if from me, the cognitive desire that I usually feel for the learning app is to get new information in the world of health.* [Respondent 14]

The ability of features to meet user needs also greatly affected the satisfaction that users felt and the intensity of their app use:

> *So if the app I consider not to provide benefits for me, yes, I also rarely have access intensity.* [Respondent 13]

However, there were also external factors that affected the intensity of app use, such as busy schedules or poor mobile network quality.

Hypothesis 14: EARQ and Stickiness in Mobile Learning

Finally, this study discovered a relationship between EARQ and stickiness in mobile learning (hypothesis 14). In retail app research, it is known that EARQ has a significant effect on user stickiness [6]. The user's decision to use an app is influenced by the level of emotional satisfaction that they feel [66]. Similar findings were obtained in the study by Souiden et al [67]. In understanding the decision-making process of retail app users, it is important to consider how users feel when shopping in both offline and online stores. In the context of mobile learning, learners' emotional involvement can positively influence their intention to continue using web-based courses [21]. Zolkepli et al [44] stated that the emotions felt when using an app are affective responses that can be positive or negative and affect the intensity of app use. The emotions felt during app use were found to be related to the user's desire to continue to use the app regularly:

> *The feeling I felt while using the D2D app for learning is happy, because the health information provided is quite complete in the app, so it makes me want to continue using the app.* [Respondent 15]

The sense of pleasure obtained from the experience encouraged users to become more willing to use the app:

> *I think maybe yes because I'm already happy, so I should at least be willing to use the app.* [Respondent 7]

Discussion

Principal Findings

This study showed that the cognitive and emotional connections with an app affect user stickiness in mobile learning. The user's EARQ is positively influenced by the performance of the app (functional factors) and the user experience (experiential factors), whereas the CARQ is influenced by user experience factors (experiential factors). App performance can affect emotional assessments triggered by feelings that arise when using apps (hedonic benefits), resulting in subjective

assessments. However, the experience that users have when using the app (experiential factors) has an effect on their emotions or feelings. In addition, user experience has an impact on cognitive assessments related to the ease of app use.

This study provides an overview of the influence of the relationship between app relationship quality, namely, EARQ and CARQ, and stickiness in mobile learning. In this study, the organismic variables are EARQ and CARQ, which represent the process of user assessment of app relationship quality [6,13]. EARQ was found to have a greater effect on stickiness in mobile learning apps than CARQ. The emotions felt during the use of an app were found to be related to the user's desire to continue using the app regularly, and the sense of pleasure obtained from the experience encouraged users to become more dependent on using the app. This is relevant to the results of previous research where, in retail apps, users were more inclined to build affective or emotional-quality relationships than cognitive-quality relationships [6,21]. The results of this study are relevant to the work by Alnawas et al [6] on retail apps, in which the relationship between CARQ and EARQ was found to affect user stickiness. Previous mobile learning research has produced similar results—cognitive and emotional connections were proven to affect users' desire to continue using apps in the web-based learning process [21]. However, this study showed that functional factors have no effect on CARQ. This is because the app provides diverse information without factoring in users' information needs.

Mobile learning service providers can use the results of this study to continue to develop features that improve user stickiness in mobile learning. Service providers must pay attention to app quality based on cognitive and emotional aspects when developing features in mobile learning apps. In addition to the usefulness of features, the experience of using the mobile learning app is an important aspect in encouraging ongoing user dependence. Service providers can improve the quality of mobile learning app functionality by focusing on several important aspects, especially the processing speed and quality of the information presented. Processing speed is a critical factor that affects the user experience as responsive and fast apps can improve user satisfaction and efficiency. In addition, the information presented in the mobile learning app must be accurate, relevant, and easily accessible to users. Quality information will help mobile learning app users make better decisions and increase trust in the app. Thus, improvements in these two aspects are expected to encourage user stickiness and ensure sustainable use of mobile learning apps.

Mobile learning service providers can improve the app's security in the context of personal data protection. To accomplish this, they can present the terms and conditions of the app and privacy policy in a clear and easy-to-understand manner. The purpose of this step is to show the seriousness of the mobile learning service provider regarding safeguarding users' personal data and building trust with users. In addition, mobile learning service providers can educate users on how to protect their personal data. This education can include instructions not to provide information to unknown parties and on maintaining the confidentiality of personal data.

Furthermore, mobile learning service providers can implement personalized content recommendations and health information tailored to mobile learning users' interests and information needs. This recommendation feature can be realized by using data on the user's profession. However, it is important to note that mobile learning user data are confidential and should be protected under the data privacy policy. Important components to consider in the implementation of this feature include effective communication, transparency, engagement, and an attractive value proposition. To increase user enjoyment and interaction during the learning process, mobile learning service providers can improve the gamification features on their apps. This can be accomplished by adopting several game elements, such as missions, rewards, dashboard rankings, and leveling. In addition, gamification provides a good user experience by supporting the learning process and increasing learning persistence.

Because most D2D app users are physicians, there are several external factors that affect the duration of app access. On the basis of the questionnaire, the biggest obstacle was limited time due to a full practice schedule (361/1120, 32.23%). This was also supported by most respondents providing more than one answer related to the location of their practice, the most common being physicians practicing in clinics (267/649, 41.1%), hospitals (117/649, 18%), and private practices (106/649, 16.3%).

Strength, Limitations, and Future Work

This study enriches previous studies on mobile learning for the Indonesian context using the stimulus-organism-response theory. Another strength of this study is the involvement of medical students, general practitioners, and specialist physicians. However, a limitation of this study is that the research respondents were primarily general practitioners and mobile learning users with an age range of 20 to 30 years; thus, future work should add more specialist respondents. On the basis of the interviews, further research can identify other factors that affect user stickiness, namely, perceived behavioral control and social influence as there were many obstacles users experienced such as busy activity schedules or network quality constraints. In addition, based on the interviews, an in-depth analysis is needed to determine whether encouragement from one's social environment also affects the desire to continue using an app.

Future work can examine whether there is an influence of moderation factors such as the age, gender, and occupation of mobile learning app users on the relationship between app relationship quality and user stickiness in mobile learning. Future research could also include comparative results based on the age and occupation of mobile learning app users. Further research can explore related factors that affect the app use experience by examining users' perception of usefulness after learning to use the app.

Acknowledgments

We want to convey our gratitude to the Faculty of Computer Science University of Indonesia for the internal grant year 2025.

Data Availability

The data are available upon request from the corresponding authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Research instruments and qualitative results.
[DOCX File , 58 KB - mededu_v11i1e63827_app1.docx]

References

1. Hailemariam RT, Nigatu AM, Melaku MS. Medical students' knowledge and attitude towards tele-education and associated factors at University of Gondar, Ethiopia, 2022: mixed method. BMC Med Educ 2023 Aug 22;23(1):599 [FREE Full text] [doi: 10.1186/s12909-023-04579-5] [Medline: 37608276]
2. Wu WH, Jim Wu Y, Chen C, Kao H, Lin C, Huang S. Review of trends from mobile learning studies: a meta-analysis. Comput Educ 2012 Sep;59(2):817-827. [doi: 10.1016/j.compedu.2012.03.016]
3. Statistik Telekomunikasi Indonesia. Badan Pusat Statistik Indonesia. 2020. URL: https://www.bps.go.id/id/publication/2021/10/11/e03aca1e6ae93396ee660328/statistik-telekomunikasi-indonesia-2020.html [accessed 2025-05-29]
4. Hidayat A. Jumlah Dokter di Indonesia Naik pada 2022, Tertinggi dalam 5 Tahun. databoks. URL: https://databoks.katadata.co.id/datapublish/2023/03/03/jumlah-dokter-di-indonesia-naik-pada-2022-tertinggi-dalam-5-tahun [accessed 2025-05-29]
5. Annur MC. Tak Merata, Mayoritas Dokter di Indonesia Masih Berpusat di Jawa. In 7 Maret. databoks. URL: https://databoks.katadata.co.id/datapublish/2022/03/07/tak-merata-mayoritas-dokter-di-indonesia-masih-berpusat-di-jawa [accessed 2025-05-29]
6. Alnawas I, Al Khateeb A, El Hedhli K. The effects of app-related factors on app stickiness: the role of cognitive and emotional app relationship quality. J Retail Consum Serv 2023 Nov;75:103412. [doi: 10.1016/j.jretconser.2023.103412]
7. Hsu TH, Tang JW. Development of hierarchical structure and analytical model of key factors for mobile app stickiness. J Innov Knowl 2020 Jan;5(1):68-79. [doi: 10.1016/j.jik.2019.01.006]
8. Chen YH, Chien SH, Wu JJ, Tsai PY. Impact of signals and experience on trust and trusting behavior. Cyberpsychol Behav Soc Netw 2010 Oct;13(5):539-546. [doi: 10.1089/cyber.2009.0188] [Medline: 20950178]
9. Elsotouhy MM, Ghonim MA, Alasker TH, Khashan MA. Investigating health and fitness app users' stickiness, WOM, and continuance intention using S-O-R model: the moderating role of health consciousness. Int J Hum Comput Interact 2022 Nov 03;40(5):1235-1250. [doi: 10.1080/10447318.2022.2135813]
10. 21st century health care challenges: a connected health approach. Deloitte Konsultan Indonesia. URL: https://www2.deloitte.com/id/en/ [accessed 2025-05-29]
11. Are you a college student or a recent fresh graduate? GUE Ecosystem. 2018. URL: https://gueecosystem.com/en/ [accessed 2025-05-29]
12. Halim F, Rebecca YHL, Ngada LM, Antonio F, Susana HF. Antecedents of intention to adopt mobile health (mhealth) application for physicians. J. Ilm Manaj Bisnis Inov Univ. Sam Ratulangi 2024;9(2):533-553 [FREE Full text]
13. Cheng YM. What makes learners enhance learning outcomes in MOOCs? Exploring the roles of gamification and personalization. Interact Technol Smart Educ 2023 Aug 21;21(2):308-330. [doi: 10.1108/itse-05-2023-0097]
14. Roca JC, Chiu CM, Martínez FJ. Understanding e-learning continuance intention: an extension of the Technology Acceptance Model. Int J Hum Comput Stud 2006 Aug;64(8):683-696. [doi: 10.1016/j.ijhcs.2006.01.003]
15. Ho XH, Nguyen DP, Cheng JM, Le AN. Customer engagement in the context of retail mobile apps: a contingency model integrating spatial presence experience and its drivers. J Retail Consum Serv 2022 May;66:102950. [doi: 10.1016/j.jretconser.2022.102950]
16. Aparicio M, Oliveira T, Bacao F, Painho M. Gamification: a key determinant of massive open online course (MOOC) success. Inf Manag 2019 Jan;56(1):39-54. [doi: 10.1016/j.im.2018.06.003]
17. Dastane O, Haba HF. What drives mobile MOOC's continuous intention? A theory of perceived value perspective. Int J Inf Learn Technol 2023 Mar 09;40(2):148-163. [doi: 10.1108/ijilt-04-2022-0087]
18. Lu Y, Wang B, Lu Y. Understanding key drivers of MOOC satisfaction and continuance intention to use. J Electron Commer Res 2019;20(2):105-117 [FREE Full text]
19. Mehrabian A, Russell JA. Approach to Environmental Psychology. Cambridge, MA: MIT Press; 1974.

20. Kim MJ, Lee CK, Jung T. Exploring consumer behavior in virtual reality tourism using an extended stimulus-organism-response model. J Travel Res 2018 Dec 26;59(1):69-89. [doi: 10.1177/0047287518818915]
21. Molinillo S, Aguilar-Illescas R, Anaya-Sánchez R, Carvajal-Trujillo E. The customer retail app experience: implications for customer loyalty. J Retail Consum Serv 2022 Mar;65:102842. [doi: 10.1016/j.jretconser.2021.102842]
22. Li CY, Fang YH. Toward better purchase decision-performance: linking person-environment fit to explorative and exploitative use of branded applications. Electron Commer Res Appl 2021 Jul;48:101063. [doi: 10.1016/j.elerap.2021.101063]
23. Baek TH, Yoo CY. Branded app usability: conceptualization, measurement, and prediction of consumer loyalty. J Advert 2018 Feb 15;47(1):70-82 [FREE Full text] [doi: 10.1080/00913367.2017.1405755]
24. Zhang Y, Rong X, Shu M, Chen Q. Identification of key influencing factors of user experience of mobile reading app in China based on the fuzzy-DEMATEL model. Math Probl Eng 2021 Jun 11;2021:1-12. [doi: 10.1155/2021/2847646]
25. Stocchi L, Pourazad N, Michaelidou N. Identification of two decision‐making paths underpinning the continued use of branded apps. Psychol Mark 2020 Jun 07;37(10):1362-1377. [doi: 10.1002/mar.21385]
26. Sharma N. A digital cohort analysis of consumers' mobile banking app experience. Int J Consum Stud 2023 Sep 19;48(1):1-19. [doi: 10.1111/ijcs.12989]
27. Bandara R, Fernando M, Akter S. Addressing privacy predicaments in the digital marketplace: a power‐relations perspective. Int J Consum Stud 2020 Mar 16;44(5):423-434. [doi: 10.1111/ijcs.12576]
28. Atulkar S, Singh AK. Role of psychological and technological attributes on customer conversion to use food ordering apps. Int J Retail Distrib Manag 2021 Apr 07;49(10):1430-1446 [FREE Full text] [doi: 10.1108/ijrdm-09-2020-0349]
29. Li M, Luo C. Exploring the influencing factors of user stickiness in a Chinese mobile news application. In: ICMECG '20: Proceedings of the 7th International Conference on Management of e-Commerce and e-Government. 2020 Presented at: ICMECG '20; July 1-3, 2020; Jeju Island, Republic of Korea p. 119-124 URL: https://dl.acm.org/doi/10.1145/3409891.3409909 [doi: 10.1145/3409891.3409909]
30. Yang D, Li C. Design of gamification theory in tourism application: take the application "travel in Zhenjiang" for example. In: Proceedings of the 2020 International Conference on Innovation Design and Digital Technology. 2020 Presented at: ICIDDT '20; December 5-6, 2020; Zhenjing, China p. 290-294 URL: https://ieeexplore.ieee.org/document/9522521 [doi: 10.1109/iciddt52279.2020.00059]
31. Ortega-Arranz A, Bote-Lorenzo ML, Asensio-Pérez JI, Martínez-Monés A, Gómez-Sánchez E, Dimitriadis Y. To reward and beyond: analyzing the effect of reward-based strategies in a MOOC. Comput Educ 2019 Dec;142:103639 [FREE Full text] [doi: 10.1016/j.compedu.2019.103639]
32. Djohan SA, Handhana D, Castafiore VB, Hendriana E. Can gamification stimulate customers to repurchase in the e-marketplace? The mediation effect of customer experience and engagement. Bp Int Res Critic Inst J Humanit Soc Sci 2022;5(1):4781-4796.
33. Huang YC, Chang LL, Yu CP, Chen JS. Examining an extended technology acceptance model with experience construct on hotel consumers' adoption of mobile applications. J Hosp Mark Manag 2019 Mar 06;28(8):957-980 [FREE Full text] [doi: 10.1080/19368623.2019.1580172]
34. Rose S, Clark M, Samouel P, Hair N. Online customer experience in e-retailing: an empirical model of antecedents and outcomes. Journal of Retailing 2012 Jun;88(2):308-322. [doi: 10.1016/j.jretai.2012.03.001]
35. Zarour M, Alharbi M. User experience framework that combines aspects, dimensions, and measurement methods. Cogent Eng 2018 Jan 08;4(1):1421006. [doi: 10.1080/23311916.2017.1421006]
36. Kang JW, Namkung Y. The role of personalization on continuance intention in food service mobile apps. Int J Contemp Hosp Manag 2019 Feb 11;31(2):734-752. [doi: 10.1108/ijchm-12-2017-0783]
37. Singh R, Söderlund M. Extending the experience construct: an examination of online grocery shopping. Eur J Mark 2020 Feb 05;54(10):2419-2446. [doi: 10.1108/ejm-06-2019-0536]
38. Pappas IO, Kourouthanassis PE, Giannakos MN, Lekakos G. The interplay of online shopping motivations and experiential factors on personalized e-commerce: a complexity theory approach. Telemat Inform 2017 Aug;34(5):730-742. [doi: 10.1016/j.tele.2016.08.021]
39. Benlian A. Web personalization cues and their differential effects on user assessments of website value. J Manag Inf Syst 2015 Jul 06;32(1):225-260 [FREE Full text] [doi: 10.1080/07421222.2015.1029394]
40. Bock DE, Mangus SM, Folse JA. The road to customer loyalty paved with service customization. J Bus Res 2016 Oct;69(10):3923-3932. [doi: 10.1016/j.jbusres.2016.06.002]
41. Lee SA, Lee J. Enhancing customers' brand loyalty via branded hotel apps. J Qual Assur Hosp Tour 2018 Oct 29;20(3):339-361. [doi: 10.1080/1528008x.2018.1537819]
42. Wu L. Factors of continually using branded mobile apps: the central role of app engagement. Int J Internet Mark Advert 2015;9(4):303. [doi: 10.1504/ijima.2015.072884]
43. Al‐Nabhani K, Wilson A, McLean G. Examining consumers' continuous usage of multichannel retailers' mobile applications. Psychol Mark 2021 Aug 30;39(1):168-195 [FREE Full text] [doi: 10.1002/mar.21585]
44. Zolkepli IA, Mukhiar SN, Tan C. Mobile consumer behaviour on apps usage: the effects of perceived values, rating, and cost. J Mark Commun 2020 Apr 03;27(6):571-593. [doi: 10.1080/13527266.2020.1749108]

45. van Noort G, van Reijmersdal EA. Branded apps: explaining effects of brands' mobile phone applications on brand responses. J Interact Mark 2022 Jan 31;45(1):16-26. [doi: 10.1016/j.intmar.2018.05.003]
46. Shahid S, Islam JU, Malik S, Hasan U. Examining consumer experience in using m-banking apps: a study of its antecedents and outcomes. J Retail Consum Serv 2022 Mar;65:102870. [doi: 10.1016/j.jretconser.2021.102870]
47. Fernandes T, Proença J. Reassessing relationships in consumer markets: emotion, cognition, and consumer relationship intention. J Relat Mark 2013 Jan;12(1):41-58. [doi: 10.1080/15332667.2013.763719]
48. Korosec-Serfaty M, Riedl R, Sénécal S, Léger PM. Attentional and behavioral disengagement as coping responses to technostress and financial stress: an experiment based on psychophysiological, perceptual, and behavioral data. Front Neurosci 2022 Jul 12;16:883431 [FREE Full text] [doi: 10.3389/fnins.2022.883431] [Medline: 35903805]
49. Fernandes G, O'Sullivan D. Project management practices in major university-industry R and D collaboration programs-a case study. J Technol Transf 2023 Mar 11;48(1):361-391 [FREE Full text] [doi: 10.1007/s10961-021-09915-9] [Medline: 35291661]
50. Tran T, Taylor DG, Wen C. Value co-creation through branded apps: enhancing perceived quality and brand loyalty. J Res Interact Mark 2022 Oct 14;17(4):562-580. [doi: 10.1108/jrim-04-2022-0128]
51. Nyffenegger B, Krohmer H, Hoyer WD, Malär L. Service brand relationship quality: hot or cold? J Serv Res 2014 Aug 26;18(1):90-106. [doi: 10.1177/1094670514547580]
52. Hsu CL, Chen MC. How does gamification improve user experience? An empirical investigation on the antecedences and consequences of user experience and its mediating role. Technol Forecast Soc Change 2018 Jul;132:118-129 [FREE Full text] [doi: 10.1016/j.techfore.2018.01.023]
53. Yang R, Wibowo S, Mubarak S, Rahamathulla M. Managing students' attitude, learning engagement, and stickiness towards e-learning post-COVID-19 in Australian universities: a perceived qualities perspective. J Mark High Educ 2023 May 02;34(2):1146-1177. [doi: 10.1080/08841241.2023.2204466]
54. Hair Jr JF, da Gabriel ML, Patel VK. AMOS covariance-based structural equation modeling (CB-SEM): Guidelines on its application as a marketing research tool. Rev Bras Mark 2014 May 23;13(2):44-55. [doi: 10.5585/remark.v13i2.2718]
55. Fornell C, Larcker DF. Evaluating structural equation models with unobservable variables and measurement error. J Mark Res 1981 Feb;18(1):39. [doi: 10.2307/3151312]
56. Morey T, Burt A, Moorman C, Redman TC. Beyond unicorns: educating, classifying, and certifying business data scientists. Harvard Business Review. URL: https://hdsr.mitpress.mit.edu/pub/t37qjoi7/release/4 [accessed 2025-05-29]
57. Smith TJ. Senior citizens and e-commerce websites: the role of perceived usefulness, perceived ease of use, and web site usability. InformingSciJ 2008;11:59-83. [doi: 10.28945/440]
58. Rakhmadian M, Sefaverdiana PV, Rahman N. Analisis Persepsi Kemanfaatan dan Persepsi Kemudahan Penggunaan Terhadap Penggunaan Sistem Informasi Akademik. Indonesian J Comput Inf Technol 2019 Nov 15;4(2):155-161. [doi: 10.31294/ijcit.v4i2.5833]
59. Zhou T. An empirical examination of continuance intention of mobile payment services. Decis Support Syst 2013 Jan;54(2):1085-1091. [doi: 10.1016/j.dss.2012.10.034]
60. Ampadu S, Jiang Y, Debrah E, Antwi CO, Amankwa E, Gyamfi SA, et al. Online personalized recommended product quality and e-impulse buying: a conditional mediation analysis. J Retail Consum Serv 2022 Jan;64:102789. [doi: 10.1016/j.jretconser.2021.102789]
61. Zhampeissova K, Gura A, Vanina E, Egorova Z. Academic performance and cognitive load in mobile learning. Int J Interact Mob Technol 2020 Dec 22;14(21):78. [doi: 10.3991/ijim.v14i21.18439]
62. Wilmer HH, Sherman LE, Chein JM. Smartphones and cognition: a review of research exploring the links between mobile technology habits and cognitive functioning. Front Psychol 2017;8:605 [FREE Full text] [doi: 10.3389/fpsyg.2017.00605] [Medline: 28487665]
63. Wardaszko M, Podgórski B. Mobile learning game effectiveness in cognitive learning by adults: a comparative study. Simul Gaming 2017 Apr 21;48(4):435-454. [doi: 10.1177/1046878117704350]
64. Barari M, Ross M, Surachartkumtonkun J. Negative and positive customer shopping experience in an online context. J Retail Consum Serv 2020 Mar;53:101985 [FREE Full text] [doi: 10.1016/j.jretconser.2019.101985]
65. Joo YJ, Lim KY, Kim EK. Online university students' satisfaction and persistence: examining perceived level of presence, usefulness and ease of use as predictors in a structural model. Comput Educ 2011 Sep;57(2):1654-1664. [doi: 10.1016/j.compedu.2011.02.008]
66. Martin J, Mortimer G, Andrews L. Re-examining online customer experience to include purchase frequency and perceived risk. J Retail Consum Serv 2015 Jul;25:81-95. [doi: 10.1016/j.jretconser.2015.03.008]
67. Souiden N, Ladhari R, Chiadmi NE. New trends in retailing and services. J Retail Consum Serv 2019 Sep;50:286-288. [doi: 10.1016/j.jretconser.2018.07.023]

Abbreviations

AVE: average variance extracted
CARQ: cognitive app relationship quality

CME: continuing medical education
D2D: Doctor-to-Doctor
EARQ: emotional app relationship quality
GOF: goodness of fit
HIPS: health information processing speed
HIQ: health information quality
PE: perceived enjoyment
PEU: perceived ease of use
PHI: personalized health information
PPP: perceived privacy protection
SKP: satuan kredit profesional (professional credit unit)

Edited by B Lesselroth; submitted 01.07.24; peer-reviewed by D Palazuelos, I Ogueji; comments to author 31.01.25; revised version received 05.02.25; accepted 13.05.25; published 13.06.25.

Please cite as:
Nurwardani S, Handayani PW
Health Workers' Perspectives on Mobile Health Care Learning Stickiness: Mixed Methods Study
JMIR Med Educ 2025;11:e63827
URL: https://mededu.jmir.org/2025/1/e63827
doi:10.2196/63827
PMID:40512533

Original Paper

Training Gaps in Digital Skills for the Cancer Health Care Workforce Based on Insights From Clinical Professionals, Nonclinical Professionals, and Patients and Caregivers: Qualitative Study

David Liñares[1,2,3], PhD; Theologia Tsitsi[4], MA, MSc, PhD; Noemí López-Rey[2,3,5], MD, PhD; Wilfredo Guanipa-Sierra[3,6], MD, PhD; Susana Aldecoa-Landesa[2,3,6], MD; Carme Carrión[7], Prof Dr; Daniela Cabutto[7], MSc; Deborah Moreno-Alonso[8], MD, PhD; Clara Madrid-Alejos[8], MSc; Andreas Charalambous[4,9], PhD; Ana Clavería[2,10], MD, PhD

[1]Galician Agency for Health Technology Assessment, Avalia-T. Galician Agency for Health Knowledge Management, Santiago, Spain

[2]I-Saúde Group, Galicia Sur Health Research Institute (IIS Galicia Sur), Servicio Galego de Saúde-Universidade de Vigo, Vigo, Spain

[3]Research Network in Chronicity, Primary Care and Health Promotion (Red de Investigación en Cronicidad, Atención Primaria y Promoción de la Salud), Zaragoza, Spain

[4]Department of Nursing, Cyprus University of Technology, Limassol, Cyprus

[5]Health Center CIS-Milagrosa (Servicio Galego de Saúde), Lugo, Spain

[6]Health Center Beiramar (Servicio Galego de Saúde), Vigo, Spain

[7]eHealth Lab Research Group, School of Health Sciences, Universitat Oberta de Catalunya (UOC)., Barcelona, Spain

[8]e-Oncologia and Cancer Education Research Unit, Institut Català d'Oncologia (ICO), Barcelona, Spain

[9]Department of Nursing, University of Turku, Turku, Finland

[10]Área de Vigo, Servicio Galego de Saúde, Vigo, Spain

Corresponding Author:
Noemí López-Rey, MD, PhD
Health Center CIS-Milagrosa (Servicio Galego de Saúde)
Río Ser, 7
Lugo, 27004
Spain
Phone: 34 982 14 99 00
Email: noemilopezrey@gmail.com

Abstract

Background: The integration of digital technologies is becoming increasingly essential in cancer care. However, limited digital health literacy among clinical and nonclinical cancer health care professionals poses significant challenges to effective implementation and sustainability over time. To address this, the European Union is prioritizing the development of targeted digital skills training programs for cancer care providers, the TRANSiTION project among them. A crucial initial step in this effort is conducting a comprehensive gap analysis to identify specific training needs.

Objective: The aim of this work is to identify training gaps and prioritize the digital skill development needs in the oncology health care workforce.

Methods: An importance-performance analysis (IPA) was conducted following a survey that assessed the performance and importance of 7 digital skills: information, communication, content creation, safety, eHealth problem-solving, ethics, and patient empowerment.

Results: A total of 67 participants from 11 European countries completed the study: 38 clinical professionals (CP), 16 nonclinical professionals (NCP), and 13 patients or caregivers (PC). CP acknowledged the need for a comprehensive training program that includes all 7 digital skills. Digital patient empowerment and safety skills emerge as the highest priorities for both CP and NCP. Conversely, NCP assigned a lower priority to digital content creation skills, and PC assigned a lower priority to digital information and ethical skills. The IPA also revealed discrepancies in digital communication skills across groups (H=6.50; P=.04).

Conclusions: The study showcased the pressing need for comprehensive digital skill training for cancer health care professionals across diverse backgrounds and health care systems in Europe, tailored to their occupation and care setting. Incorporating PC

perspectives ensures a balanced approach to addressing these training gaps. These findings provide a valuable knowledge base for designing digital skills training programs, promoting a holistic approach that integrates the perspectives of the various stakeholders involved in digital cancer care.

(JMIR Med Educ 2025;11:e78490) doi:10.2196/78490

KEYWORDS

health literacy; professional competence; health personnel; telemedicine; oncology nursing

Introduction

Background

Cancer is the second leading cause of premature mortality and morbidity worldwide [1]. In the European Union, nearly 4.7 million new cases of cancer and 2.1 million cancer-related deaths occur each year [2]. According to the European Commission, the urgency to address cancer control and outcomes is a significant political challenge, as reflected in Europe's Beating Cancer Plan [3], with cancer being one of the 5 missions included in the Horizon Europe program.

Health literacy (HL), defined as the individual capacity to access, understand, evaluate, and apply health information to make informed health decisions [4]**,** is widely recognized as a critical factor in effective cancer care [5,6]. With the increasing integration of digital technologies, such as symptom monitoring platforms, treatment adherence tools, telehealth, and mobile apps, in oncology, HL has evolved to encompass the digital environment, giving rise to the concept of digital health literacy (DHL) [7-11]. DHL has been defined in various ways, reflecting the evolving nature of health information environments. Broadly, DHL refers to the ability to seek, find, understand, and appraise health information from electronic sources and to apply this knowledge to solve a health problem [12,13]. However, there is ongoing debate about its key attributes, particularly the relative weight of technical skills, critical thinking, health knowledge, and digital engagement, in shaping a comprehensive definition [14,15]. In response, recent studies have focused on 4 major areas: (1) conceptualizing and measuring DHL; (2) identifying and addressing the digital divide; (3) exploring the factors that influence DHL development; and (4) examining the health outcomes associated with DHL levels [16]. Regarding the latter, DHL enhances access to and quality of health care to the extent of being considered a "super determinant" of health, a factor with a profound impact across various health outcomes [17,18].

In cancer care, DHL is particularly relevant, enabling health care professionals and patients to benefit from digital innovations such as electronic health records, patient portals, symptom tracking tools, and remote care services [19]. Low levels of DHL have been associated with poorer clinical outcomes, including reduced overall survival among patients with cancer [20-22]. Limited DHL not only hinders patients and caregivers (PC) but also poses challenges for health care professionals, potentially impeding the effective adoption of digital health solutions in clinical practice [23]. The Towards European Health Data Space (TEHDAS) project highlights significant disparities in the health system infrastructures across European countries, with not all of them being adequately equipped to ensure effective management of digital health [24]. Still, the main barriers to implementing digital health strategies in health care organizations are not technical issues (such as infrastructure or connectivity). Instead, they are rooted in gaps in digital skills among professionals and patients, concerns about data security and confidentiality in digital environments, and limited time availability [25-28].

As a result of these challenges, experts in the field emphasize the need to develop flexible and easily accessible training programs, such as online modules and hands-on learning approaches, supported by appropriate incentives to engage and retain the oncology workforce [29]. Nevertheless, the results of DigiCanTRain, a European-cofunded project under the EU4Health Programme (2021-2027) of approximately €1.98 million (US $2.32 million), highlight significant challenges in the implementation of digital skills training programs for cancer professionals across 25 EU countries. These challenges include a lack of coordination between national and international organizations in promoting training initiatives, as well as limited access to continuous accreditation mechanisms that ensure the quality and consistency of educational content [30]. In response to these gaps, DigiCanTRain aims to design, pilot, and evaluate a comprehensive digital skills training curriculum for both clinical and nonclinical oncology professionals, with the goal of enhancing the adoption of eHealth technologies and fostering more person-centered, efficient, and resilient cancer care.

Despite these initiatives, international continuing education programs fail to identify the specific digital skills required by health care professionals [31]. Even if DigComp 2.2: The Digital Competence Framework for Citizens provides a reference framework for the global population on existing digital competencies [32], it does not include specific competencies oriented to health care professionals in cancer care, such as ethical or patient empowerment skills [33,34]. Moreover, there are no validated and widely used measurement tools available to assess eHealth competencies [35,36]. Therefore, the review by Tinmaz et al [37] highlighted the need to create updated digital frameworks for different work settings, professional categories, and contexts. In the context of cancer care, the study by Leena et al [34] revealed that the digital skills of health care professionals are multifaceted. Consequently, the authors indicated that it is imperative that these skills be subjected to a process of assessment to facilitate the provision of training that is based on the actual learning needs of the professionals in question.

In view of this, the TRANSiTION project [38] was cofunded at 80% by the European Union with a total budget of €2,299,541.28 (US $2,690,371). The project aims to design an advanced training program for both clinical professionals (CP)

and nonclinical professionals (NCP) involved in cancer care, equipping them with essential digital skills to enhance the efficiency and effectiveness of information exchange with patients and other health care providers. TRANSiTION brings together an interdisciplinary consortium of 24 partners from 14 member states, all with extensive experience in the development, evaluation, and successful implementation of continuing professional development and training programs in oncology.

Theoretical Framework

According to the European Commission, the development of a digital skills training program should be preceded by a thorough analysis of training needs and existing gaps [39]. A needs analysis is a systematic process used to identify discrepancies between the current and ideal states of an organization or service [40].

One widely adopted and intuitive method for conducting such an analysis is importance-performance analysis (IPA). Originally developed by Martilla and James [41], IPA provides a visual and analytical framework to support strategic decision-making by comparing the importance of specific attributes to their perceived performance. It has been extensively applied across diverse sectors, including IT services, marketing, banking, tourism, and sports [42]. More recently, IPA has been used in the assessment of training needs [43,44], process improvement [45], and evaluation of health care services [46].

IPA is based on a 2D grid that plots attributes according to their mean scores on 2 axes: (1) importance, which is the value or relevance assigned to the attribute by users, and (2) performance, which is the perceived effectiveness or quality of that attribute.

The axes of the IPA grid are determined by the overall mean scores of importance and performance across all attributes assessed. The position of each attribute within the grid reflects its scores on these 2 dimensions. This allows for a relative comparison, enabling the identification of attributes that deviate from the general trend.

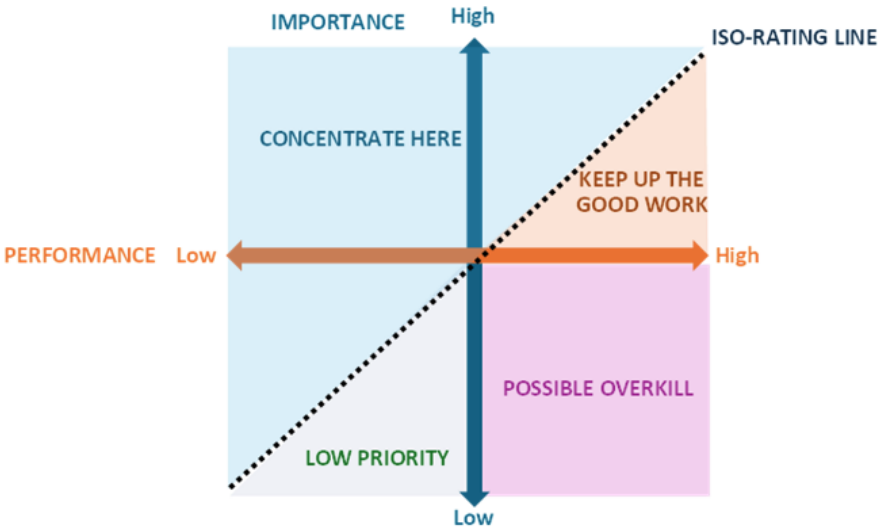
The resulting matrix is divided into 4 quadrants, each linked to distinct action strategies [47]: Quadrant I—Focus here: high importance, low performance; these are critical areas in need of immediate improvement. Quadrant II—Keep up the good work: high importance, high performance; these are strengths to be preserved. Quadrant III—Low priority: low importance, low performance; these areas require minimal attention as they are not strategically significant. Quadrant IV—Possible overkill: low importance, high performance; these attributes may be receiving more resources than necessary.

To enhance the discriminative power of the analysis, this study used a modified version of the IPA proposed by Abalo et al [48]. This adaptation addresses the common issue of attribute saturation, whereby attributes tend to receive uniformly high ratings, limiting the ability to differentiate among them. The modified approach uses ordinal rankings instead of mean scores, improving the identification of priority areas, particularly relevant in health care settings [49] and training needs assessments [50].

Based on this approach, an IPA chart was generated using discrepancy scores from 3 key stakeholder groups: CP, NCP, and PC. CP were defined as members of health care organizations providing direct cancer care (eg, oncologists, radiotherapists, oncology nurses, family physicians, and community nurses). NCP included individuals performing administrative or managerial tasks related to cancer care, regardless of professional category. Caregivers were defined as those who provide physical, emotional, practical, and, in some cases, medical support to individuals diagnosed with cancer. These may be family members, close friends, or other persons designated by the patient who play a critical role across all stages of treatment and recovery. A patient with cancer was defined as any individual at any stage of the disease, including those undergoing active treatment, in remission (with favorable progression and no ongoing treatment), considered cured but undergoing regular follow-up, cured without active medical surveillance, or experiencing a relapse.

The position of each attribute was analyzed in relation to the diagonal and the corresponding quadrants of the IPA grid (see Figure 1).

Figure 1. Representation of the alternative version of the importance-performance analysis (IPA) grid.



The current literature highlights a significant gap in digital skills among health care professionals [51]. Furthermore, although it is widely recognized that digital skills in health care are inherently multifaceted, a standardized framework to delineate and prioritize the most essential skills remains absent. Furthermore, we expected not only training gaps but also differences between the health professionals included in the study.

Consequently, the aim of this study was to identify training gaps and prioritize the digital skill development needs in the oncology health care workforce through the application of IPA.

Methods

Study Design

A selective methodology was used. It involved conducting an online survey among a group of experts selected within the consortium of the European project. Given the exploratory nature of the study and the specific expertise required from participants, a convenience sampling approach was used to facilitate recruitment across diverse stakeholder groups and countries. This strategy was deemed appropriate due to the practical constraints of accessing individuals with relevant professional or lived experience in cancer care across several member states, all within a limited time frame.

The inclusion criteria required participants to be CP, NCP, or PC from member states of the TRANSiTION consortium, with the ability to comprehend and respond in English. Moreover, participants were required to have the capacity to provide informed consent. Recruitment was supported by consortium partners who identified and invited suitable individuals based on their direct involvement in cancer care or their experience as patients or caregivers. This pragmatic sampling approach enabled the collection of meaningful insights while ensuring feasibility in a multinational context.

Instrument

An ad hoc online questionnaire, available exclusively in English, was developed by the research team. Although several validated instruments exist in the fields of DHL and eHealth for the general population [12,15], they did not adequately address the specific objectives or context of our study. The TRANSiTION project aimed to design a training course tailored to professionals involved in cancer care. Therefore, beyond assessing general skills, it was crucial to identify the specific learning needs, expectations, and contextual factors affecting this particular group. Given these requirements, the development of a customized instrument was deemed necessary to ensure the relevance, specificity, and practical utility of the collected data for informing the design of the training intervention.

The structure and items of the questionnaire were based on *DigComp 2.2: The Digital Competence Framework for Citizens* [32] and *Measuring What Matters: The Patient-Reported Indicator Surveys* [52]. For the development of the items, the contents of the *Core Curriculum in eHealth* from the EU×US initiatives were reviewed [53], as well as the document *Mapping Health Data Management Systems through Country Visits: Development, Needs, and Expectations of the EHDS* by TEHDAS [24].

The final selection of questionnaire items was reached by consensus in an online focus group conducted in May 2023. During this session, it was decided to include 2 additional and independent sections (ie, patient empowerment and ethics) informed by the prior work and practical experience of project partners who had previously identified core digital health skills as applied in professional practice [54].

The questionnaire was organized into 3 thematic sections. The first section included the information sheet and the consent form for participants. The second section addressed sociodemographic variables. The third section focused on assessing the need for training of 7 key digital skills: information, communication, content creation, safety, problem-solving, ethics, and patient empowerment. Moreover, 7-point Likert scale items were used to evaluate performance (with 1 representing “Very bad” and 7 representing “Very good”) and importance (with 1 representing “Not important” and 7 representing “Highly important”). The questionnaire items for CP and NCP were identical. Additionally, to make the results more representative, the third section asked about the performance of their colleagues within their organization from the respondent’s perspective, aiming to shift the focus away from self-assessment of personal skills. Similarly, the items on the importance of digital skills referred to how crucial these skills are for cancer care. In contrast, patients were asked about their perception of the digital skills shown by the cancer professionals attending to them and about the importance of these skills for their cancer care—or, in the case of caregivers, for the care of the patient. The information letters and questionnaires can be found in Multimedia Appendix 1.

Procedure

The questionnaire was administered as a closed, invitation-only survey, accessible exclusively to individuals invited by TRANSiTION consortium partners. It was fully compatible with mobile phones, tablets, and computers across all major operating systems. To prevent multiple entries from the same individual, participants were authenticated through their email address prior to receiving single-use access to the survey platform.

Recruitment was facilitated by the TRANSiTION consortium, which disseminated information about the study to eligible participants. The consortium consisted of 24 European reference partners in the field of cancer care, including research institutes, universities, hospitals, oncology centers, and patient organizations [38].

The survey was piloted by 8 members of the Spanish partners of the TRANSiTION consortium during June 2023 and July 2023. Based on the pilot, minor adjustments were made to improve the wording of certain items. On average, it was estimated that participants required approximately 15 minutes to complete the online questionnaire. Additionally, it was suggested to include the item “Have you received prior training in digital competencies/skills?” with a dichotomous response option of “Yes” or “No.”

Participants were recruited during July 2023 and August 2023. Data were collected and stored using an online questionnaire implemented through the eDelphi website [55] in September 2023. To ensure procedural standardization, daily monitoring of the data collection process was conducted, allowing for the immediate resolution of any questions or technical issues that arose.

Data Analysis

Data analysis included descriptive statistics, reported as mean and SD. Internal consistency of the items assessing digital skill domains was evaluated using Cronbach α. Normality was assessed using the Shapiro-Wilk test. In cases where the assumption of normality was not met, nonparametric tests were applied. Specifically, group comparisons of importance and performance scores across CP, NCP, and PC were conducted using 1-way ANOVA or, when appropriate, the Kruskal-Wallis test. All analyses were performed using SPSS version 25.0 (IBM Corp).

Ethical Considerations

The study was reviewed and approved by the Pontevedra-Vigo-Ourense Research Ethics Committee (reference: 2023/309).

Before the survey, they were informed about the study's objectives, reminded of the voluntary nature of their participation, and asked to provide informed consent.

The online questionnaire platform provides a confidential, single-use access system to the questionnaire for each participant, ensuring the confidentiality and anonymity of responses. No financial compensation was provided to the participants.

Results

Demographics

Initially, 152 participants expressed interest in taking part in the study. Of these, 33 were excluded for not meeting the inclusion criteria. Specifically, 24 participants who had registered as CP or NCP were in fact students without professional experience, and 9 PC were excluded due to insufficient English proficiency. Additionally, 28 participants (12 CP, 10 NCP, 3 patients, and 3 caregivers) accepted the informed consent but did not begin the questionnaire. A further 24 participants (8 CP, 10 NCP, 3 patients, and 3 caregivers) declined to provide informed consent.

A total of 67 participants completed the study: 38 CP, 16 NCP, and 13 PC. All participants who completed the survey responded to all questionnaire items. Of the total, 50 participants were women (50/67, 75%), participants were primarily aged between 31 years and 45 years (27/67, 40%), and most resided in municipalities with populations greater than 100,000 (44/67, 66%). Table 1 presents the main sociodemographic characteristics of the sample, stratified by group.

Table 1. Sociodemographic data of the participants, by group.

Characteristic	CP[a] (n=38)	NCP[b] (n=16)	PC[c] (n=13)
Gender (female), n (%)	29 (76)	12 (75)	9 (69)
Age (years), n (%)			
18-30	10 (26)	5 (31)	0 (0)
31-45	14 (37)	7 (44)	6 (46)
46-60	11 (29)	3 (19)	5 (39)
≥61	3 (8)	1 (6)	2 (15)
Population of the resident municipality, n (%)			
<50,000	7 (18)	2 (13)	5 (39)
50,000-100,000	6 (16)	1 (6)	2 (15)
>100,000	25 (66)	13 (81)	6 (46)

[a]CP: clinical professionals.

[b]NCP: nonclinical professionals.

[c]PC: patients and caregivers.

The participants were from 11 European countries: Belgium (3 CP, 5 NCP, and 2 PC), Bulgaria (10 CP and 2 NCP), Croatia (4 CP, 1 NCP, and 3 PC), Cyprus (6 CP), Greece (1 CP, 1 NCP, and 1 PC), Italy (1 CP, 1 NCP, and 2 PC), Lithuania (1 CP and 1 NCP), Poland (1 PC), Portugal (3 CP, 1 NCP, and 2 PC), Slovenia (1 CP, 1 NCP, and 1 PC), and Spain (8 CP, 3 NCP, and 1 PC).

The professions of the health care professionals were diverse. Among CP (n=38), 16 (42%) were oncologists, 12 (32%) were oncology nurses, 6 (16%) were clinical researchers, and 4 (11%) worked in other clinical professions related to cancer care. Additionally, 19 (50%) worked in public organizations, 12 (32%) worked in subsidized private organizations, 4 (11%) worked in nonsubsidized private organizations, and 3 (8%) preferred not to specify. Among NCP (n=16), 5 (31%) were clinical data managers, 4 (25%) were part of the administrative staff related to cancer care, 3 (19%) worked in health care service management, and 4 (25%) were in other nonclinical professions related to cancer care. Furthermore, 8 (50%) worked in public organizations, 7 (44%) worked in subsidized private

organizations, and 1 (6%) preferred not to specify. All PC (n=13) were users of public health care services.

Internal Consistency

As shown in Table 2, the internal consistency of the digital skills questionnaire was high across all domains (items B1 through B7 in Multimedia Appendix 1) and participant groups. At the global level, Cronbach α values ranged from 0.92 (information) to 0.97 (ethics), indicating excellent reliability. When analyzed by subgroup, responses from CP and PC consistently showed very high internal consistency, with α values greater than 0.85 in all domains. Responses from NCP also showed acceptable to excellent consistency, though slightly lower in the domains of eHealth problem-solving (α=0.86) and ethics (α=0.90). These results support the internal reliability of the instrument across different respondent profiles.

Table 2. Internal consistency of the digital skill domains.

Digital skills	Global, Cronbach α	Clinical professionals, Cronbach α	Nonclinical professionals, Cronbach α	Patients and caregivers, Cronbach α
Information	0.92	0.95	0.91	0.85
Communication	0.93	0.95	0.89	0.90
Content creation	0.95	0.95	0.92	0.95
Safety	0.96	0.96	0.95	0.97
eHealth problem-solving	0.94	0.97	0.86	0.92
Ethics	0.97	0.97	0.90	0.97
Patient empowerment	0.96	0.96	0.90	0.98

Performance and Importance Scores

This section presents the results from the third part of the questionnaire, which focused on 7 core digital competencies relevant to cancer care professionals. Additional findings related to perceived training needs are available in Multimedia Appendix 2. The 7 competencies assessed were information, communication, content creation, safety, eHealth problem-solving, ethics, and patient empowerment.

Information is defined as the ability to search, evaluate, and manage digital health information effectively. Communication refers to the capacity to interact, share, and collaborate using digital tools in health care contexts. Content creation is understood as the skill to produce, edit, and adapt digital content appropriately for clinical use. Safety encompasses data protection, privacy, and cybersecurity practices. Problem-solving is the ability to identify and resolve technical or digital challenges in the care process. Ethics is related to the understanding and application of ethical principles such as confidentiality, consent, and digital equity. Patient empowerment is defined as the ability to support patients with using digital tools to actively participate in their care.

Table 3 shows the importance and performance results for CP, NCP, and PC. As can be seen, all performance and importance scores reached notably high values, reflecting that oncology health care professionals perceive themselves as having strong digital skills, and, at the same time, that digital literacy is considered important in cancer care. Moreover, 22 of 38 CP (58%) reported having received prior training in digital skills, a figure that reached 11 of 16 NCP (69%).

Table 3. Importance and performance for clinical professionals (CP; n=38), nonclinical professionals (NCP; n=16), and patients and caregivers (PC; n=13).

Digital skills by group	Performance, mean (SD)[a]	Importance, mean (SD)[b]
Information		
CP	6.18 (0.98)	6.34 (1.34)
NCP	5.88 (0.96)	5.94 (1.23)
PC	5.85 (2.07)	5.23 (1.59)
Communication		
CP	6.03 (0.97)	6.45 (1.29)
NCP	5.94 (0.85)	5.69 (1.08)
PC	5.23 (1.78)	5.31 (1.32)
Content creation		
CP	5.37 (1.58)	5.79 (1.36)
NCP	5.56 (1.03)	5.44 (1.03)
PC	4.15 (2.03)	4.15 (2.03)
Safety		
CP	5.21 (1.49)	6.08 (1.56)
NCP	5.25 (1.06)	5.81 (1.05)
PC	4.69 (1.84)	4.62 (2.10)
eHealth problem-solving		
CP	5.53 (1.31)	6.18 (1.37)
NCP	5.56 (1.09)	5.81 (1.17)
PC	4.38 (1.80)	4.38 (1.55)
Ethics		
CP	5.61 (1.28)	6.03 (1.50)
NCP	5.50 (1.15)	5.75 (1.06)
PC	4.92 (2.10)	4.38 (1.85)
Patient empowerment		
CP	5.50 (1.45)	6.32 (1.36)
NCP	5.56 (1.09)	6.00 (1.15)
PC	5.08 (2.11)	5.23 (1.36)

[a]Overall mean (SD): 5.38 (1.47).

[b]Overall mean (SD): 5.57 (1.40).

The digital skills that showed the highest performance among CP were information (mean 6.18, SD 0.98) and communication (mean 6.03, SD 0.97). Similarly, these same digital skills, but in reverse order, were considered the most important for cancer care (mean 6.45, SD 1.29; mean 6.34, SD 1.34, respectively). Additionally, patient empowerment digital skills achieved a similar level of importance (mean 6.32, SD 1.36). In contrast, the digital skill with the lowest performance and importance scores was content creation (mean 5.37, SD 1.58).

Digital communication (mean 5.94, SD 0.85) and information (mean 5.88, SD 0.96) skills exhibited the highest performance scores among NCP. In contrast, the skills considered most important for cancer care were those related to patient empowerment (mean 6.00, SD 1.15). In comparison, the digital skills of content creation and eHealth problem-solving demonstrated the lowest performance scores (mean 5.56, SD 1.03; mean 5.56, SD 1.09, respectively), with the former also showing the lowest importance score (mean 5.44, SD 1.03).

For PC, the digital skills in which they thought oncology health care professionals perform best were communication (mean 5.23) and information (mean 5.85). The digital skills considered most important were communication (mean 5.31, SD 1.32) and patient empowerment (mean 5.23, SD 1.36). In contrast, the lowest performance and importance scores were observed for the digital skills related to content creation (both: mean 4.15, SD 2.03).

Discrepancy Analysis

Table 4 presents the results of the discrepancy analysis, defined as the mean difference between performance and importance for each of the digital skills. As shown, for CP, information digital skills exhibited the most positive discrepancy (mean difference 0.62). Conversely, safety (mean difference –0.87) and patient empowerment (mean difference –0.82) digital skills showed the most negative results. Regarding NCP, communication digital skills displayed the most positive discrepancy (mean difference 0.25), while safety digital skills had the most negative discrepancy (mean difference –0.56). For PC, the highest positive discrepancy was observed for information digital skills (mean difference 0.62), whereas the most negative discrepancy pertained to patient empowerment digital skills (mean difference –0.15).

Table 4. Discrepancy for clinical professionals (CP; n=38), nonclinical professionals (NCP; n=16), and patients and caregivers (PC; n=13).

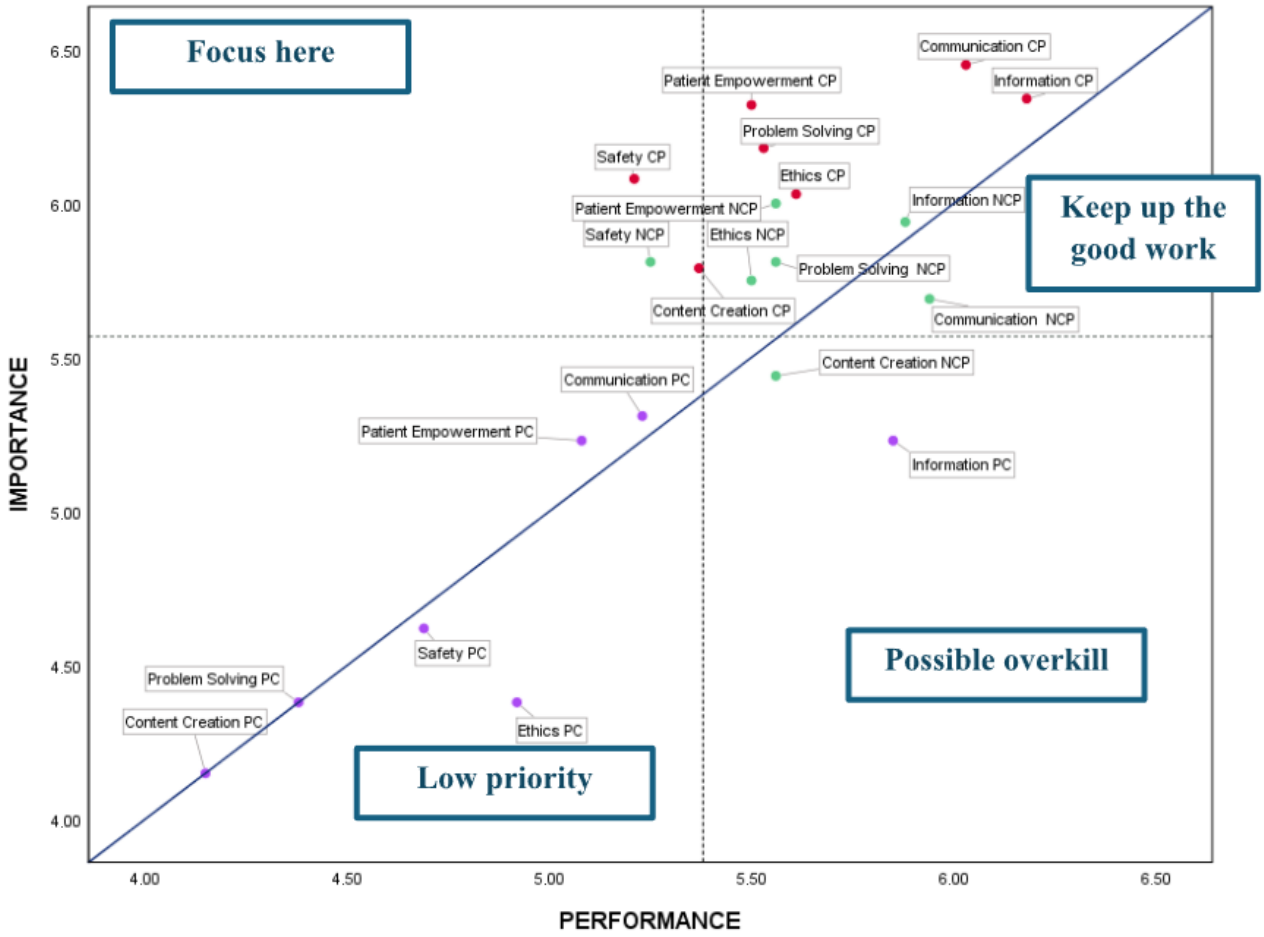
Digital skills, by group	Discrepancy, mean difference	*F* (*df*)	*H*
Information		1.13 (2)	3.58
CP	–0.16		
NCP	–0.06		
PC	0.62		
Communication		1.25 (2)	6.50*
CP	–0.42		
NCP	0.25		
PC	–0.08		
Content creation		0.65 (2)	1.26
CP	–0.42		
NCP	0.12		
PC	0		
Safety		1.10 (2)	2.89
CP	–0.87		
NCP	–0.56		
PC	0.07		
eHealth problem-solving		0.85 (2)	1.86
CP	–0.65		
NCP	–0.25		
PC	0		
Ethics		1.24 (2)	2.19
CP	–0.42		
NCP	–0.25		
PC	0.54		
Patient empowerment		0.70 (2)	2.76
CP	–0.82		
NCP	–0.44		
PC	–0.15		

Notably, despite the small sample size, statistically significant differences were identified in the discrepancies associated with communication digital skills. Specifically, the discrepancy was positive for CP, negative for NCP, and nearly neutral for PC (–0.42 vs 0.25 vs –0.08; H= 6.50, P=.04.

IPA Chart and Key Findings

As a result of these discrepancies, the IPA chart was developed (Figure 2). The axes of the graph were formed using the mean scores of performance (mean 5.38) and importance (mean 5.57), as shown in Table 2. Results were segmented for CP (in red), NCP (in green), and PC (in purple). This segmentation allowed for the identification of high-priority training needs, as well as areas where training is of low priority or may even represent a misallocation of resources. Two key findings emerged from the overall analysis.

Figure 2. Importance-performance analysis (IPA) graph comparing clinical professionals (CP), nonclinical professionals (NCP), and patients and caregivers (PC).



First, both CP and NCP perceived a greater need for training compared with the skills PC attributed to them. This is evident, as most elements related to health care professionals are positioned above the diagonal, with some at a significant distance. In contrast, digital skills for PC are located on or near the diagonal, with a few even below it.

Second, all digital skills for CP fell within the "Focus here" quadrant, indicating that, although their prioritization may vary, all these skills require targeted training for this group. When evaluating each element individually, the IPA chart confirmed that patient empowerment and, particularly, security skills are the furthest from the diagonal and within the "Focus here" quadrant for both CP and NCP. Consequently, these represent the most critical digital skills requiring training.

For NCP, security and patient empowerment skills were similarly the top training priorities. Notably, communication skills met the expected importance levels (located in the "Keep up the good work" quadrant), whereas content creation skills were deemed redundant (falling into the "Possible overkill" quadrant).

In the case of PC, ethical skills fell into the "Low priority" quadrant, while information skills are in the "Possible overkill" quadrant. The remaining digital skills clustered near the diagonal, indicating a moderate level of alignment between their perceived importance and performance.

Discussion

Principal Findings

This study identified key training gaps and digital skill development priorities within the oncology health care workforce by applying IPA to a diverse panel of experts. The panel included 3 stakeholder profiles: CP, such as physicians and nurses; NCP, including managers, educators, and researchers; and PC.

A major finding was the assessment of 7 digital skills (information, communication, content creation, safety, eHealth problem-solving, ethical, and patient empowerment) and the specific training needs associated with each stakeholder group.

Notably, the prioritization of these competences differed across groups. For CP, the analysis revealed that all digital skills require further training, with safety and patient empowerment emerging as the highest priorities. NCP showed a comparable pattern, although they assigned less importance to content creation. None of the competencies was considered a low priority in this group.

In contrast, PC viewed information digital skills as having high performance but relatively low importance. The remaining competencies were positioned in the low performance–low importance quadrant, though communication and patient

empowerment were perceived as the most important among them.

These findings underscore the need for tailored digital training strategies that reflect the differing perceptions and priorities of each stakeholder group.

Comparison With Other Studies

Several previous studies have highlighted the widespread lack of digital literacy among health care professionals [56], which has led to the development of initiatives aimed at addressing this issue, particularly in the oncology field [38,57]. As Foadi and Varghese [58] suggested, although CP demonstrate a high level of proficiency with using specialized software for their daily tasks, they do not receive sufficient training on the basic principles of digital systems, which are essential for thriving in an ever-evolving digital health care environment [59]. In fact, Ramachandran et al [60] emphasized that, although health care professionals should demonstrate a general orientation toward digital skills, opportunities are created for specialized digital skills, particularly for the safe and equitable use of new technologies. Regarding cancer care, the study by Barbosa et al [61] identified digital safety skills as one of the 6 key domains for therapeutic radiographers and radiation therapists. Concerning patient empowerment, Navarro Martínez et al [62] emphasized a worrisome trend: Nurses, especially younger ones, exhibit limited or negligible use of technology to empower patients.

Incorporating the experience of health care service users is a central principle for the transition to digital health care with a patient-centered approach [63], making it essential to understand what PC expect regarding the digital skills of their oncology professionals. eHealth policies should be designed to consider the diverse perspectives of health professionals, patients, and caregivers. Furthermore, it is essential to bridge the digital divide for patients with cancer who have low digital literacy, enabling them to effectively use digital platforms designed by professionals [64]. In this regard, it is noteworthy that this study highlights that NCP perceived training in digital content creation skills as a potential misallocation of resources. Therefore, although these digital skills are important for enhancing cancer care, they should not be prioritized in training programs for this professional profile. The results suggest that enhancing digital information skills training within the oncology workforce may not be a pressing priority, as indicated by feedback from PC. However, this does not imply that patients with cancer are uninterested in receiving information on all aspects of the disease [65] nor that they are unwilling to embrace the potential of telecommunications to access relevant information [66]. In fact, the most necessary digital health function among patients with cancer and caregivers is information and education on symptom management following cancer treatment [67], and the use of digital health technology can be experienced as a person guiding them during their cancer treatment [68]. However, as previous studies noted, the interest of patients with chronic diseases in receiving health information through digital modalities is often hindered by educational and age-related gaps. In many cases, this must be preceded by social and instrumental support from health promoters [69].

This study has also yielded a significant finding when comparing results across different groups: CP, NCP, and PC. Digital communication skills were rated as highly important by all 3 groups. This is consistent with the fact that social support through digital tools provides significant benefits for both patients and caregivers during cancer treatment [70]. However, when analyzing discrepancies between CP, NCP, and PC, we found that the discrepancies between performance and importance were statistically significant. The results indicate that digital communication skills training is a priority for CP, somewhat unnecessary for NCP, and considered almost neutral by PC. First, NCP have limited direct interactions with patients, as their primary responsibilities center on the management and administration of services, which may account for the observed differences between CP and NCP. Accordingly, communication skills training for cancer care professionals who interact directly with patients has become a critical focus, using structured checklists to systematically assess oncologists' behaviors during specialized doctor-patient consultations [71]. Second, the review by Henry et al [72] emphasized that digital communication skills are essential for CP to perform telehealth tasks, which may explain the priorities for training in this group. Third, although patients with cancer consider online platforms a preferred option for cancer follow-up consultations and delivering good news, they are not seen as suitable for initial visits or discussing bad news [73]. Therefore, PC may demonstrate a less favorable stance toward digital communication skills than CP. For instance, when oncology nurses and surgeons do not mention electronic patient-reported outcome measures, patients also refrain from discussing them [74].

Finally, it is worth mentioning that, in comparison, both CP and NCP are more critical of their digital skills training needs than PC. Overall, health care professionals recognize that they are not achieving a level of performance that aligns with the importance of these skills for cancer care. This finding, in which service providers perceive a worse outcome than users, has been observed in other studies using IPA [47]. This apparent "halo effect," where PC may rate professionals' digital skills more positively than professionals rate themselves, could reflect a high level of trust in health care providers. This is particularly relevant in the context of implementing new technologies and addressing the urgent need for digital skills training, as it suggests that end users may be more receptive to digital innovation than expected, potentially easing the path for adoption and integration of new tools in cancer care.

Strengths and Limitations

This study has several strengths. It sought input from a sample of experts purposely selected by the TRANSiTION project consortium [38]. The partners belong to 14 countries from different geographic regions, with different types of organizations and occupational profiles. The partners have the personal email addresses of their members. Therefore, this ensured variability in the professional profiles of the participants, as well as different geogeographical backgrounds; additionally, patient involvement in the long-term improvement of HL is an essential requirement [25], making the inclusion of their perspective in this gap analysis a notable strength. Furthermore,

although the use of the IPA method is well-documented in the literature, to the best of our knowledge, it has not been applied to evaluate training needs in digital skills for health care professionals, specifically in cancer care. Therefore, this study represents an innovative approach that could serve as a methodological foundation for future research. In addition, the survey probes aspects of cancer care that are not included in more general questionnaires. Moreover, the internal reliability of the instrument across different respondent profiles supports the incorporation of these skills in training design or scale development.

This study also has several limitations. First, the questionnaire did not rely on previously validated scales to assess digital skills. However, there is currently no gold standard for measuring digital skills among health care professionals, and the instrument was developed based on established digital competence frameworks to ensure alignment with the study's specific objectives. Moreover, the aim of the project was not to validate a scale but to collect meaningful data to inform the design of a massive open online course (MOOC) [38]. Although participants from several EU countries were included, the use of a nonprobabilistic, purposive sampling strategy through consortium partners limits the representativeness of the sample and the generalizability of the findings beyond the context of the TRANSiTION network. Additionally, as the survey was available only in English, 9 potential participants were excluded due to insufficient language proficiency, which may have introduced selection bias. The administration of the survey during the summer months also contributed to a relatively low response rate within a limited population. Moreover, the use of self-reported data introduces the possibility of response bias; although prior research supports the validity of self-report measures [75], the results should be interpreted with appropriate caution. Finally, although IPA is a widely used tool to guide priority setting in training and service improvement [43,44,46,76], it has certain methodological limitations [77]. The technique relies heavily on mean values to allocate items to quadrants, which may not fully capture the distributional characteristics of the data. In addition, the cut-off points, typically overall means, are somewhat arbitrary, which can affect the robustness and interpretability of the results. This study did not include a sensitivity analysis to explore the impact of alternative quadrant definitions, which could have helped mitigate this limitation. This decision was primarily influenced by the exploratory nature of the research and the limited sample size. Future research should address these aspects to improve the reliability of IPA-based prioritization.

Impact on Organizations and Health

There is a growing body of literature that demonstrates the pivotal role of HL for achieving better health outcomes and higher quality of care [78]. Crucially, HL is modifiable, and improving HL is increasingly recognized as a way of improving outcomes, including in Europe's Beating Cancer Plan. Therefore, the concept has rapidly gained an emerging strategic role in several governing bodies and cancer organizations, although more comprehensive implementation of interventions and strategies is still needed.

The information that cancer patients need to know for their diagnosis and treatment is indeed complicated. It includes a new language of health terminology, understanding consents for complex treatments and procedures, attending appointments at the right time and place, and seeking help appropriately and in a timely manner. Equally, citizens should have the necessary skills to interpret information and make appropriate decisions for cancer screening (eg, high-risk group of citizens) and prevention (eg, lifestyle behaviors). Competencies in HL and communication can significantly contribute to reducing barriers related to HL and to improving the quality of health care and health outcomes for patients [78]. However, studies have shown that health professionals tend to overestimate the HL of patients and citizens and lack adequate competence to compensate for it. Therefore, preparing staff to respond to patients' HL is seen as a responsibility of health care organizations, which should be incorporated into their training programs.

Cancer care systems need to adapt to technological advancements by providing online health materials that are evidence-based, quality-controlled, reliable, and both culturally and linguistically appropriate. Therefore, the incorporation and management of digital technologies that facilitate interactions between health care professionals and patients seem essential in current and future training programs. It should be noted that the digital skills of CP involved in cancer care are multifaceted, and all of them are essential for providing high-quality cancer care [34]. Therefore, the findings of this study support the need to implement comprehensive training programs for CP that address the main digital skills cited in the literature [37].

It is worth mentioning that, although patient empowerment is a vague concept, it has been increasingly applied in cancer care over the past decade [33]. The accumulated evidence suggests that shared decision-making and the use of interactive digital tools lead to positive outcomes for cancer patients [79,80].

In addition, the future European Health Data Space, which aims to provide a coherent, reliable, and efficient system for the exchange and reuse of health data in research, innovation, policy-making, and regulation, will require a greater mastery of digital security skills for both CP and NCP [81].

The importance given by the European Union to digital skills training has already been outlined. Based on expert input, our consortium believes it is essential that countries incorporate DHL training at the earliest stages of education for health professionals and health managers and develop programs focusing on digital skills in oncology [29]—not only in the university setting but also in continuing education—including hands-on training through internships, clinical rotations, and simulation exercises and promoting interdisciplinary collaboration.

The extension and adaptation of competencies to the health environment represent an example to follow. The involvement so far of 60 countries worldwide and 1300 participants will facilitate this impact, supported by multiple meetings with European stakeholders [38] and diffusion through scientific societies and partners' universities and national health services.

Finally, it points to research gaps for new scientific projects in the fields of social sciences and citizen science.

Conclusions

This study conducted a gap analysis using the IPA to assess the digital skills of health care professionals in oncology and identified areas where further training is needed. The results highlight the necessity of developing comprehensive training programs for CP. Additionally, it underscores the critical importance of digital safety and patient empowerment skills for both PC and NCP. The study incorporates the perspectives of PC, who prioritize different training needs for health care professionals, placing comparatively less emphasis on digital information and ethical skills. These findings provide a knowledge base for designing training programs and eHealth policies, promoting a holistic approach that integrates the perspectives of the various stakeholders involved in digital cancer care.

Policy Summary

Health care professionals acknowledge that their digital skills require enhancement across all areas. Specifically, doctors and nurses need additional training in digital problem-solving, communication, and, above all, skills linked to patient safety and empowerment. These priorities align with the perspectives of PC, who emphasize the critical need for health care providers to strengthen their digital communication and patient empowerment capabilities.

Acknowledgments

This study has been elaborated in the framework of the European TRANSiTION project (GA 101101261) cofunded by the European Union. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them. The authors also extend their gratitude to all collaborators and participants who contributed to the study.

During the preparation of this work, the authors used DeepL Write in order to improve readability and language. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Clinical professional and nonclinical professional survey.
[DOCX File , 68 KB - mededu_v11i1e78490_app1.docx]

Multimedia Appendix 2
Additional analysis of training needs.
[DOCX File , 146 KB - mededu_v11i1e78490_app2.docx]

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018 Nov;68(6):394-424 [FREE Full text] [doi: 10.3322/caac.21492] [Medline: 30207593]
2. Dyba T, Randi G, Bray F, Martos C, Giusti F, Nicholson N, et al. The European cancer burden in 2020: incidence and mortality estimates for 40 countries and 25 major cancers. Eur J Cancer 2021 Nov;157:308-347 [FREE Full text] [doi: 10.1016/j.ejca.2021.07.039] [Medline: 34560371]
3. Communication from the commission to the european parliament and the council: Europe's Beating Cancer Plan. European Commission. 2021 Mar 02. URL: https://eur-lex.europa.eu/resource.html?uri=cellar:8dec84ce-66df-11eb-aeb5-01aa75ed71a1.0002.02/DOC_1&format=PDF [accessed 2024-12-26]
4. Ishikawa H, Kiuchi T. Association of health literacy levels between family members. Front Public Health 2019;7:169 [FREE Full text] [doi: 10.3389/fpubh.2019.00169] [Medline: 31275918]
5. Holden CE, Wheelwright S, Harle A, Wagland R. The role of health literacy in cancer care: a mixed studies systematic review. PLoS One 2021;16(11):e0259815 [FREE Full text] [doi: 10.1371/journal.pone.0259815] [Medline: 34767562]
6. Papadakos JK, Hasan SM, Barnsley J, Berta W, Fazelzad R, Papadakos CJ, et al. Health literacy and cancer self-management behaviors: a scoping review. Cancer 2018 Nov 01;124(21):4202-4210 [FREE Full text] [doi: 10.1002/cncr.31733] [Medline: 30264856]
7. Golinelli D, Boetto E, Carullo G, Nuzzolese AG, Landini MP, Fantini MP. Adoption of digital technologies in health care during the COVID-19 pandemic: systematic review of early scientific literature. J Med Internet Res 2020 Nov 06;22(11):e22280 [FREE Full text] [doi: 10.2196/22280] [Medline: 33079693]

8. Li Q, Lu Y, Hao Y, Zhao Y, Qi X, Qiao J. Adaptive digital and non-digital self-management in permanent enterostomy patients: a qualitative study based on the Chronic Illness Trajectory framework. Eur J Oncol Nurs 2025 Feb;74:102733 [FREE Full text] [doi: 10.1016/j.ejon.2024.102733] [Medline: 39637689]
9. Marthick M, McGregor D, Alison J, Cheema B, Dhillon H, Shaw T. Supportive care interventions for people with cancer assisted by digital technology: systematic review. J Med Internet Res 2021 Oct 29;23(10):e24722 [FREE Full text] [doi: 10.2196/24722] [Medline: 34714246]
10. Papachristou N, Kotronoulas G, Dikaios N, Allison SJ, Eleftherochorinou H, Rai T, et al. Digital transformation of cancer care in the era of big data, artificial intelligence and data-driven interventions: navigating the field. Semin Oncol Nurs 2023 Jun;39(3):151433. [doi: 10.1016/j.soncn.2023.151433] [Medline: 37137770]
11. Singh S, Fletcher GG, Yao X, Sussman J. Virtual care in patients with cancer: a systematic review. Curr Oncol 2021 Sep 08;28(5):3488-3506 [FREE Full text] [doi: 10.3390/curroncol28050301] [Medline: 34590602]
12. Norman CD, Skinner HA. eHEALS: the eHealth literacy scale. J Med Internet Res 2006 Nov 14;8(4):e27 [FREE Full text] [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
13. Thorup CB, Uitto M, Butler-Henderson K, Wamala-Andersson S, Hoffrén-Mikkola M, Schack Thoft D, et al. Choosing the best digital health literacy measure for research: mixed methods study. J Med Internet Res 2025 Apr 08;27:e59807 [FREE Full text] [doi: 10.2196/59807] [Medline: 40198098]
14. Gilstad H. Toward a comprehensive model of eHealth literacy. 2014 Presented at: Practical Aspects of Health Informatics; May 19-20, 2014; Trondheim, Norway URL: https://ceur-ws.org/Vol-1251 [doi: 10.13140/2.1.4569.0247]
15. van der Vaart R, Drossaert C. Development of the digital health literacy instrument: measuring a broad spectrum of health 1.0 and health 2.0 skills. J Med Internet Res 2017 Jan 24;19(1):e27 [FREE Full text] [doi: 10.2196/jmir.6709] [Medline: 28119275]
16. Yang K, Hu Y, Qi H. Digital health literacy: bibliometric analysis. J Med Internet Res 2022 Jul 06;24(7):e35816 [FREE Full text] [doi: 10.2196/35816] [Medline: 35793141]
17. Arias López MDP, Ong BA, Borrat Frigola X, Fernández AL, Hicklent RS, Obeles AJT, et al. Digital literacy as a new determinant of health: a scoping review. PLOS Digit Health 2023 Oct;2(10):e0000279 [FREE Full text] [doi: 10.1371/journal.pdig.0000279] [Medline: 37824584]
18. van Kessel R, Wong BLH, Clemens T, Brand H. Digital health literacy as a super determinant of health: more than simply the sum of its parts. Internet Interv 2022 Mar;27:100500 [FREE Full text] [doi: 10.1016/j.invent.2022.100500] [Medline: 35242586]
19. Melhem S, Nabhani-Gebara S, Kayyali R, Almomani H, Alrashdan Y, Elian M, et al. The Role of Digital Literacy, Health Literacy, and Information-Seeking Behaviour in Cancer Care: Empowering Survivors through Knowledge and Action. In: Fusi J, Scarfò G, Franzoni F, editors. Health Promotion - From Knowledge to Action. London, England: IntechOpen Limited; 2025.
20. Heudel PE, Delrieu L, Dumas E, Crochet H, Hodroj K, Charrier I, et al. Impact of limited e-Health literacy on the overall survival of patients with cancer. JCO Clin Cancer Inform 2022 Feb;6:e2100174 [FREE Full text] [doi: 10.1200/CCI.21.00174] [Medline: 35213209]
21. Pimentel-Parra GA, Soto-Ruiz MN, San Martín-Rodríguez L, Escalada-Hernández P, García-Vivar C. Effectiveness of digital health on the quality of life of long-term breast cancer survivors: a systematic review. Semin Oncol Nurs 2023 Aug;39(4):151418 [FREE Full text] [doi: 10.1016/j.soncn.2023.151418] [Medline: 37045645]
22. Verma R, Saldanha C, Ellis U, Sattar S, Haase KR. eHealth literacy among older adults living with cancer and their caregivers: a scoping review. J Geriatr Oncol 2022 Jun;13(5):555-562. [doi: 10.1016/j.jgo.2021.11.008] [Medline: 34810146]
23. Kemp E, Trigg J, Beatty L, Christensen C, Dhillon HM, Maeder A, et al. Health literacy, digital health literacy and the implementation of digital health technologies in cancer care: the need for a strategic approach. Health Promot J Austr 2021 Feb;32 Suppl 1:104-114. [doi: 10.1002/hpja.387] [Medline: 32681656]
24. Abboud L, Cosgrove S, Kesisoglou I. Mapping health data management systems through country visits: development, needs and expectations of the EHDS 1 0 Document info 0. Towards European Health Data Space (TEHDAS). 2023. URL: https://tehdas.eu/tehdas1/app/uploads/2023/04/tehdas-mapping-health-data-management-systems-through-country-visits.pdf [accessed 2025-09-05]
25. Hasannejadasl H, Roumen C, Smit Y, Dekker A, Fijten R. Health literacy and eHealth: challenges and strategies. JCO Clin Cancer Inform 2022 Sep;6:e2200005. [doi: 10.1200/CCI.22.00005] [Medline: 36194843]
26. Montero Delgado J, Merino Alonso F, Monte Boquet E, Ávila de Tomás J, Cepeda Díez J. Competencias digitales clave de los profesionales sanitarios. Educación Médica 2020 Sep;21(5):338-344 [FREE Full text] [doi: 10.1016/j.edumed.2019.02.010]
27. Schreiweis B, Pobiruchin M, Strotbaum V, Suleder J, Wiesner M, Bergh B. Barriers and facilitators to the implementation of eHealth services: systematic literature analysis. J Med Internet Res 2019 Nov 22;21(11):e14197 [FREE Full text] [doi: 10.2196/14197] [Medline: 31755869]
28. Charalambous A. Digital transformation in healthcare: have we gone off the rails? Asia Pac J Oncol Nurs 2024 May;11(5):100481 [FREE Full text] [doi: 10.1016/j.apjon.2024.100481] [Medline: 38774536]

29. Protogiros D, Cloconi C, Tsitsi T, Nicolaidou I, Kyriacou E, Couespel N, et al. Achieving digital transformation in cancer care across Europe: practical recommendations from the TRANSiTION project. J Cancer Policy 2025 Jun;44:100584 [FREE Full text] [doi: 10.1016/j.jcpo.2025.100584] [Medline: 40210108]
30. Kaihlanen A, Virtanen L, Kainiemi E, Sulosaari V, Heponiemi T. Continuing education in digital skills for healthcare professionals - mapping of the current situation in EU Member States. Int J Health Policy Manag 2024;13:8309. [doi: 10.34172/ijhpm.8309] [Medline: 39099482]
31. Tischendorf T, Hasseler M, Schaal T, Ruppert S, Marchwacka M, Heitmann-Möller A, et al. Developing digital competencies of nursing professionals in continuing education and training - a scoping review. Front Med (Lausanne) 2024;11:1358398 [FREE Full text] [doi: 10.3389/fmed.2024.1358398] [Medline: 38947234]
32. DigComp 2.2: The Digital Competence Framework for Citizens. European Commission. 2022 Mar 17. URL: https://pact-for-skills.ec.europa.eu/community-resources/publications-and-documents/digcomp-22-digital-competence-framework-citizens_en [accessed 2025-09-05]
33. Kim SH, Choe YH, Kim DH. Patient empowerment in cancer care: a scoping review. Cancer Nurs 2024;47(6):471-483. [doi: 10.1097/NCC.0000000000001228] [Medline: 36907924]
34. Tuominen L, Poraharju J, Carrion C, Leeni L, Leino-Kilpi H, Moretó S, et al. Digital skills of health care professionals in cancer care: a systematic review. Digit Health 2024;10:20552076241240907 [FREE Full text] [doi: 10.1177/20552076241240907] [Medline: 38528966]
35. Longhini J, Rossettini G, Palese A. Digital health competencies among health care professionals: systematic review. J Med Internet Res 2022 Aug 18;24(8):e36414 [FREE Full text] [doi: 10.2196/36414] [Medline: 35980735]
36. Mainz A, Nitsche J, Weirauch V, Meister S. Measuring the digital competence of health professionals: scoping review. JMIR Med Educ 2024 Mar 29;10:e55737 [FREE Full text] [doi: 10.2196/55737] [Medline: 38551628]
37. Tinmaz H, Lee Y, Fanea-Ivanovici M, Baber H. A systematic review on digital literacy. Smart Learn Environ 2022;9(1):21 [FREE Full text] [doi: 10.1186/s40561-022-00204-y] [Medline: 40478098]
38. TRANSiTION. European Cancer Organisation. URL: https://www.europeancancer.org/eu-projects/impact/resource/transition [accessed 2024-12-20]
39. Digitalisation in education and training: Checklist for successful projects and initiatives. European Union. 2022. URL: https://op.europa.eu/en/publication-detail/-/publication/5362d258-0967-11ed-b11c-01aa75ed71a1/language-en [accessed 2025-09-05]
40. Kim S, Ji Y. Gap Analysis. In: The International Encyclopedia of Strategic Communication. Hoboken, NJ: Wiley; 2018.
41. Martilla JA, James J. Importance-performance analysis. Journal of Marketing 1977 Jan 01;41(1):77-79 [FREE Full text] [doi: 10.1177/002224297704100112]
42. Feng M, Mangan J, Wong C, Xu M, Lalwani C. Investigating the different approaches to importance–performance analysis. The Service Industries Journal 2014 Jun 30;34(12):1021-1041 [FREE Full text] [doi: 10.1080/02642069.2014.915949]
43. Ahmed H. How to use importance-performance analysis (IPA)-based SWOT analysis as a new quantitative methodology for developing actual strategic plans in universities. SN Soc Sci 2021 Jan 11;1(1):32 [FREE Full text] [doi: 10.1007/s43545-020-00039-9]
44. Cladera M. An application of importance-performance analysis to students' evaluation of teaching. Educ Asse Eval Acc 2020 Oct 20;33(4):701-715 [FREE Full text] [doi: 10.1007/s11092-020-09338-4]
45. Shin J, Kim N. Importance-performance analysis of patient-safety nursing in the operating room: a cross-sectional study. Risk Manag Healthc Policy 2024;17:715-725 [FREE Full text] [doi: 10.2147/RMHP.S450340] [Medline: 38559872]
46. Zarei E, Bagheri A, Daneshkohan A, Khodakarim S. Patients' views on service quality in selected Iranian hospitals: an importance-performance analysis. Shiraz E-Med J 2020 May 12;21(9):A [FREE Full text] [doi: 10.5812/semj.97938]
47. Serrano-Gómez V, García-García O, Rial-Boubeta A. Using importance–performance analysis (IPA) to improve golf club management: the gap between users and managers' perceptions. Sustainability 2023 Apr 26;15(9):7189 [FREE Full text] [doi: 10.3390/su15097189]
48. Abalo J, Varela J, Manzano V. Importance values for importance–performance analysis: a formula for spreading out values derived from preference rankings. Journal of Business Research 2007 Feb;60(2):115-121 [FREE Full text] [doi: 10.1016/j.jbusres.2006.10.009]
49. Izadi A, Jahani Y, Rafiei S, Masoud A, Vali L. Evaluating health service quality: using importance performance analysis. Int J Health Care Qual Assur 2017 Aug 14;30(7):656-663. [doi: 10.1108/IJHCQA-02-2017-0030] [Medline: 28809594]
50. Martínez-Izaguirre M, Yániz-Álvarez-de-Eulate C, Villardón-Gallego L. Aplicación de un análisis de importancia y realizaciónde competencias para la identificación de prioridades enla formación docente. Revista de Educación 2021;393:97-128 [FREE Full text] [doi: 10.4438/1988-592X-RE-2021-393-487]
51. Linares D, Charalambous A. Addressing digital skills gaps in cancer care: a call for targeted training programs. Eur J Oncol Nurs 2025 Apr;75:102842. [doi: 10.1016/j.ejon.2025.102842] [Medline: 40010018]
52. Patient-reported indicators for assessing health system performance: Measuring what matters: the patient-reported indicator surveys. OECD. 2021. URL: https://health.ec.europa.eu/system/files/2021-02/pagoda_en_0.pdf [accessed 2025-09-05]
53. Värri A. Foundational eHealth Curriculum for the Healthcare Workforce. Tampere University. 2018. URL: https://cris.tuni.fi/ws/portalfiles/portal/15782480/MIE_FoundCur_25_April2018.pdf(accessed [accessed 2025-09-05]

54. Segur i Ferrer J. Les competències bàsiques dels professionals de la salut necessàries per incorporar la Salut Digital en la pràctica professional. Treball Final de Màster. Máster Universitari en Salut Digital 2022:1-327. [doi: 10.62727/dsalut.scs/11592]
55. eDelphi.org. URL: https://www.edelphi.org [accessed 2025-09-05]
56. Reixach E, Andrés E, Sallent Ribes J, Gea-Sánchez M, Àvila López A, Cruañas B, et al. Measuring the digital skills of Catalan health care professionals as a key step toward a strategic training plan: digital competence test validation study. J Med Internet Res 2022 Nov 30;24(11):e38347 [FREE Full text] [doi: 10.2196/38347] [Medline: 36449330]
57. DigiCanTrain. Turku University of Applied Sciences. URL: https://digicantrain.turkuamk.fi [accessed 2024-12-20]
58. Foadi N, Varghese J. Digital competence - a key competence for todays and future physicians. J Eur CME 2022;11(1):2015200 [FREE Full text] [doi: 10.1080/21614083.2021.2015200] [Medline: 34992949]
59. Charalambous A, Dodlek N. Big data, machine learning, and artificial intelligence to advance cancer care: opportunities and challenges. Semin Oncol Nurs 2023 Jun;39(3):151429. [doi: 10.1016/j.soncn.2023.151429] [Medline: 37085405]
60. Ramachandran S, Chang H, Worthington C, Kushniruk A, Ibáñez-Carrasco F, Davies H, et al. Digital competencies and training approaches to enhance the capacity of practitioners to support the digital transformation of public health: rapid review of current recommendations. JMIR Public Health Surveill 2024 Sep 09;10:e52798 [FREE Full text] [doi: 10.2196/52798] [Medline: 39248660]
61. Barbosa B, Bravo I, Oliveira C, Antunes L, Couto JG, McFadden S, et al. Digital skills of therapeutic radiographers/radiation therapists - document analysis for a European educational curriculum. Radiography (Lond) 2022 Nov;28(4):955-963 [FREE Full text] [doi: 10.1016/j.radi.2022.06.017] [Medline: 35842952]
62. Navarro Martínez O, Igual García J, Traver Salcedo V. Estimating patient empowerment and nurses' use of digital strategies: eSurvey study. Int J Environ Res Public Health 2021 Sep 18;18(18):1 [FREE Full text] [doi: 10.3390/ijerph18189844] [Medline: 34574766]
63. Gilbert RM. Reimagining digital healthcare with a patient-centric approach: the role of user experience (UX) research. Front Digit Health 2022;4:899976 [FREE Full text] [doi: 10.3389/fdgth.2022.899976] [Medline: 36016600]
64. Melhem SJ, Nabhani-Gebara S, Kayyali R. Digital trends, digital literacy, and e-Health engagement predictors of breast and colorectal cancer survivors: a population-based cross-sectional survey. Int J Environ Res Public Health 2023 Jan 13;20(2):1 [FREE Full text] [doi: 10.3390/ijerph20021472] [Medline: 36674237]
65. Ellis EM, Varner A. Unpacking cancer patients' preferences for information about their care. J Psychosoc Oncol 2018;36(1):1-18 [FREE Full text] [doi: 10.1080/07347332.2017.1357666] [Medline: 28786762]
66. Tashkandi E, BaAbdullah M, Zeeneldin A, AlAbdulwahab A, Elemam O, Elsamany S, et al. Optimizing the communication with cancer patients during the COVID-19 pandemic: patient perspectives. Patient Prefer Adherence 2020;14:1205-1212 [FREE Full text] [doi: 10.2147/PPA.S263022] [Medline: 32764893]
67. Yoo S, Sung JH, Lee K, Hong B, Oh EG, Kim SH, et al. The needs for digital health and eHeath literacy of cancer patients, caregivers, and healthcare providers: a multicenter, descriptive correlational study. Eur J Oncol Nurs 2024 Jun;70:102581. [doi: 10.1016/j.ejon.2024.102581] [Medline: 38749385]
68. Darley A, Furlong E, Maguire R, McCann L, Coughlan B. Relationship and attachment to digital health technology during cancer treatment. Semin Oncol Nurs 2024 Apr;40(2):151587 [FREE Full text] [doi: 10.1016/j.soncn.2024.151587] [Medline: 38342642]
69. Gordon NP, Crouch E. Digital information technology use and patient preferences for internet-based health education modalities: cross-sectional survey study of middle-aged and older adults with chronic health conditions. JMIR Aging 2019 Apr 04;2(1):e12243 [FREE Full text] [doi: 10.2196/12243] [Medline: 31518291]
70. Katsaros D, Hawthorne J, Patel J, Pothier K, Aungst T, Franzese C. Optimizing social support in oncology with digital platforms. JMIR Cancer 2022 Jun 24;8(2):e36258 [FREE Full text] [doi: 10.2196/36258] [Medline: 35749161]
71. Stubenrauch S, Schneid E, Wünsch A, Helmes A, Bertz H, Fritzsche K, et al. Development and evaluation of a checklist assessing communication skills of oncologists: the COM-ON-Checklist. J Eval Clin Pract 2012 Apr;18(2):225-230. [doi: 10.1111/j.1365-2753.2010.01556.x] [Medline: 21029271]
72. Henry BW, Block DE, Ciesla JR, McGowan BA, Vozenilek JA. Clinician behaviors in telehealth care delivery: a systematic review. Adv Health Sci Educ Theory Pract 2017 Oct;22(4):869-888. [doi: 10.1007/s10459-016-9717-2] [Medline: 27696102]
73. Kumar D, Gordon N, Zamani C, Sheehan T, Martin E, Egorova O, et al. Cancer patients' preferences and perceptions of advantages and disadvantages of telehealth visits during the COVID-19 pandemic. JCO Clin Cancer Inform 2023 Sep;7:e2300040. [doi: 10.1200/CCI.23.00040] [Medline: 37656925]
74. Thestrup Hansen S, Jørgensen L, Schmidt V, Gebhard Ørsted L, Piil K. Empowered or challenged? The dual impact of condition-specific electronic patient-reported outcome measures in the person-centred care of women with breast cancer: a qualitative study. Eur J Oncol Nurs 2024 Dec;73:102712. [doi: 10.1016/j.ejon.2024.102712] [Medline: 39486313]
75. Lance CE. More statistical and methodological myths and urban legends. Organizational Research Methods 2010 Dec 30;14(2):279-286 [FREE Full text] [doi: 10.1177/1094428110391814]
76. Han T, Wei Q, Wang R, Cai Y, Zhu H, Chen J, et al. Service quality and patient satisfaction of internet hospitals in China: cross-sectional evaluation with the Service Quality Questionnaire. J Med Internet Res 2024 Nov 08;26:e55140 [FREE Full text] [doi: 10.2196/55140] [Medline: 39514849]

77. Azzopardi E, Nash R. A critical evaluation of importance–performance analysis. Tourism Management 2013 Apr;35:222-233 [FREE Full text] [doi: 10.1016/j.tourman.2012.07.007]
78. Kaper MS, Winter AFD, Bevilacqua R, Giammarchi C, McCusker A, Sixsmith J, et al. Positive outcomes of a comprehensive health literacy communication training for health professionals in three European countries: a multi-centre pre-post intervention study. Int J Environ Res Public Health 2019 Oct 15;16(20):1 [FREE Full text] [doi: 10.3390/ijerph16203923] [Medline: 31619010]
79. Makarov DV, Feuer Z, Ciprut S, Lopez NM, Fagerlin A, Shedlin M, et al. Randomized trial of community health worker-led decision coaching to promote shared decision-making for prostate cancer screening among Black male patients and their providers. Trials 2021 Feb 10;22(1):128 [FREE Full text] [doi: 10.1186/s13063-021-05064-4] [Medline: 33568208]
80. Tuominen L, Leino-Kilpi H, Poraharju J, Cabutto D, Carrion C, Lehtiö L, et al. Interactive digital tools to support empowerment of people with cancer: a systematic literature review. Support Care Cancer 2024 May 31;32(6):396 [FREE Full text] [doi: 10.1007/s00520-024-08545-9] [Medline: 38816629]
81. Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. European Union. 2022. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0197 [accessed 2024-12-20]

Abbreviations

CP: clinical professionals
DHL: digital health literacy
HL: health literacy
IPA: importance-performance analysis
MOOC: massive open online course
NCP: nonclinical professionals
PC: patients and caregivers
TEHDAS: Towards European Health Data Space

Edited by B Lesselroth; submitted 03.06.25; peer-reviewed by C Pollard, O Akinsola; comments to author 18.06.25; revised version received 27.07.25; accepted 15.08.25; published 08.10.25.

Please cite as:
Liñares D, Tsitsi T, López-Rey N, Guanipa-Sierra W, Aldecoa-Landesa S, Carrión C, Cabutto D, Moreno-Alonso D, Madrid-Alejos C, Charalambous A, Clavería A
Training Gaps in Digital Skills for the Cancer Health Care Workforce Based on Insights From Clinical Professionals, Nonclinical Professionals, and Patients and Caregivers: Qualitative Study
JMIR Med Educ 2025;11:e78490
URL: https://mededu.jmir.org/2025/1/e78490
doi:10.2196/78490
PMID:

Original Paper

Impact of Learner Autonomy on the Performance in Voluntary Online Cardiac Auscultation Courses: Prospective Self-Controlled Study

Yudong Fang[1]; Ligang Fang[2], MD; Xue Lin[2], MD

[1]Capital Medical University, Beijing, China

[2]Peking Union Medical College Hospital, Beijing, China

Corresponding Author:
Xue Lin, MD
Peking Union Medical College Hospital
1 Shuaifuyuan, Dongcheng District
Beijing, 100730
China
Phone: 86 10 69155068
Fax: 86 10 69155068
Email: linxuepumch@qq.com

Abstract

Background: Learner autonomy—the ability to self-direct and regulate learning—is a key determinant of success in online education, yet its quantifiable impact in voluntary noncredit courses remains unclear. Understanding how autonomy translates into measurable behaviors and outcomes in clinical skills training may inform more effective online learning design and learning outcomes.

Objective: This study aims to quantify the association between behavioral indicators of learner autonomy and performance in a voluntary noncredit online cardiac auscultation course.

Methods: We conducted a prospective, self-controlled, single-center study. A total of 199 registrants (n=122 physicians and n=77 medical students) were recruited via WeChat and attended four weekly 2-hour synchronous sessions using authentic patient heart sound recordings with imaging-based explanations. The primary outcome was the final posttraining quiz score (0-100); training effectiveness was assessed by the pre- to posttraining score change. The autonomy indicators were full participation (attendance at all four sessions), in-class engagement (number of responses to brief content-aligned prompts posed approximately every 10-15 minutes; responses recorded for participation monitoring only), and postclass review (frequency/duration of reviewing recordings and materials). Analyses included Wilcoxon signed rank tests, χ^2 tests, multivariable linear regression, and receiver operating characteristic profiling of “excellent learners” (top 10% improvement).

Results: Of the 199 registrants, 146 (73.4%) attended ≥1 session and 46 (23.1%) completed all sessions. Median test scores improved from 40 (IQR 20-50) to 70 (IQR 50-83; P<.001). Intrinsic motivation was associated with full participation (χ^2_1=4.03; P=.045). In multivariable models, full participation (unstandardized B=41.55, 95% CI 24.43-58.67; standardized β=0.60; P<.001) and in-class engagement (B=4.79 per additional response, 95% CI 3.05-6.45; β=0.70; P<.001) independently predicted higher final scores (adjusted R^2=0.48). Receiver operating characteristic profiling indicated that greater postclass review (recordings/materials) led to learners achieving excellent performance.

Conclusions: In this voluntary online clinical skills course, showing up consistently, engaging during class, and reviewing after class—practical expressions of learner autonomy—were key correlates of short-term performance. These behaviors may be encouraged through simple, feasible course designs such as clear expectation setting, periodic interactive prompts, and structured review opportunities, which warrant prospective evaluation in future studies.

(JMIR Med Educ 2025;11:e78363) doi:10.2196/78363

KEYWORDS

learner autonomy; online learning; clinical skills; medical education; student engagement

Introduction

Online education has become an essential component of health professional learning and continuing professional development, especially since the COVID-19 pandemic [1-3]. Among factors determining online learning success, learner autonomy—the capacity to self-direct and regulate one's learning—has been recognized as fundamental in distance education theory [4,5] and self-determination theory [6-8]. Autonomous learners set personal goals, regulate study behaviors, and reflect on their progress, achieving deeper and more durable learning [9]. However, despite extensive theoretical discussion, quantitative evidence linking measurable expressions of autonomy—such as participation, attention, and review behaviors—to concrete learning outcomes in clinical skills training remains scarce [10,11]. Understanding this relationship may help learners and educators make more strategic use of limited study time and improve learning efficiency.

For medical professionals, online education generally follows two distinct patterns. The first consists of institution-driven structured courses provided by medical schools or training organizations through online platforms, often with mandatory participation and formal assessments [12,13]. The second is learner-driven and self-directed, where participants independently choose courses aligned with their interests or professional needs, often without credit or compulsory evaluation [7,14]. Such voluntary online courses place high demands on learners' autonomy, time management, and self-discipline. Capturing behavioral indicators of autonomy in freely accessible nonmandatory settings and examining their relationship with learning outcomes can yield insights into continuing medical education.

Cardiac auscultation remains a crucial yet challenging diagnostic skill in clinical practice [15-17]. Traditional bedside teaching is limited by the scarcity of authentic cardiac sounds and the difficulty of linking acoustic findings to underlying pathophysiology. These challenges make auscultation training well suited to online formats that can provide repeated exposure to verified recordings and integrated visual explanations. In line with this rationale, recent studies have shown that carefully designed digital modules can complement traditional instruction and improve diagnostic competence [18].

Based on these premises, we developed a voluntary noncredit online cardiac auscultation course that used authentic patient recordings, multimodal imaging, and interactive components. To examine the learner side of online success, we quantified autonomy through observable behaviors—full participation, active in-class engagement, and postclass review frequency—and evaluated how these dimensions predicted learning outcomes. This study aimed to provide quantitative evidence on how learner autonomy influences achievement in voluntary online clinical skills training.

A preliminary version of this study was shared as a preprint on Research Square [19].

Methods

Learner Recruitment

This was a prospective, self-controlled, single-center study that recruited doctors and medical students interested in free cardiac auscultation training through multiple channels within our hospital and affiliated medical school. The recruitment campaign was coordinated by the hospital's Department of Medical Education and disseminated via several official WeChat groups, including those for undergraduate medical students, staff physicians, and visiting trainees. In addition, printed recruitment posters containing a QR code for registration were displayed in various teaching and clinical areas of the hospital and medical school.

Interested individuals could scan the QR code to access an online registration form. The form collected basic demographic information, participants' prior experience with cardiac auscultation, and their motivation for enrolling in the course. Enrollment was open until the maximum capacity of the online classroom was reached, resulting in a total of 199 registrants on a first-come, first-served basis.

Ethical Considerations

This study was approved by the Institutional Review Board of Peking Union Medical College Hospital (I-23PJ1679). During registration, all prospective participants were presented with an electronic informed consent form that described the study purpose and procedures, explained that the online teaching sessions would be recorded, and noted that anonymized data might be used for research analysis. Only those who read and electronically signed the consent form were able to complete registration and participate in the course. Participation was entirely voluntary, and no financial or material compensation was provided.

Teaching Process

Teaching Contents

The cardiac sounds used in this training were authentic recordings collected from real patients over the past decade. When patients presented with abnormal heart sounds, simultaneous phonocardiographic recordings were obtained using an electronic stethoscope, and echocardiographic videos were captured to ensure precise correlation between auscultatory findings and underlying pathophysiology. Each recording was processed using acoustic analysis software to remove distortion and verify signal integrity, and all recordings were independently reviewed and validated by experienced cardiologists to confirm diagnostic accuracy.

Through this long-term systematic collection, we established a curated library of several hundred authentic cardiac sound recordings representing a broad spectrum of pathological findings. A total of 80 representative sounds were selected from this database, together with more than 30 clinical cases, 5 animations, and 100 echocardiographic clips, to form the course materials and cover the essential elements of cardiac auscultation relevant to internal medicine and clinical diagnostics.

Teaching Setting

All training sessions were conducted on an online teaching platform (Plaso, PLASO Network Technologies Co, Ltd) and consisted of four 2-hour live-streamed sessions held weekly. The platform supported real-time teacher-student interaction, and participants used in-ear headphones to ensure optimal sound quality.

Participants completed a 10-item cardiac sound identification test both before and after the course. Each correct answer was awarded 10 points (maximum score of 100 per test). The pre- and posttests used different sets of heart sounds that together covered all essential auscultatory findings.

During each live session, the instructor posed brief content-aligned questions approximately every 10-15 minutes. Items were randomly drawn from a preestablished question bank covering clinically relevant cardiac sounds and key conceptual checkpoints. Typical examples included "Which of the following four murmurs best represents a patent ductus arteriosus?" and "Which of the following murmurs is the most common systolic murmur?" Learners submitted responses through the platform; answers were recorded only to confirm participation and attention, not to evaluate correctness or to influence grading.

After each class, video recordings and supplementary review materials were made available on the platform for 1 month to facilitate voluntary review.

Definition of Key Variables

Motivation Classification

Learning motivation was categorized as intrinsic or extrinsic according to self-determination theory, in which intrinsic motivation refers to engaging in an activity for inherent interest or enjoyment, whereas extrinsic motivation involves participating to obtain external rewards or avoid negative consequences [20]. Motivation type was determined using a single self-report item included on the WeChat recruitment page. For physicians, the question read "Are you attending this training simply because you want to master cardiac auscultation, rather than for work or promotion needs?" This wording was chosen because, in current clinical practice, cardiac auscultation is no longer indispensable for diagnosis—most structural cardiac abnormalities can be identified by echocardiography. Therefore, physicians who voluntarily return to study auscultation, often years after graduation, generally do so out of strong personal interest and professional self-expectation rather than institutional or promotion requirements. Selecting this option was classified as intrinsic motivation, whereas choosing "because it is required for work or promotion requirements" represented extrinsic motivation. For medical students, the item was simplified to two options on the registration page: "interested in learning about heart sounds" (intrinsic motivation) or "because it is needed for work or exams" (extrinsic motivation).

Participation in training was defined as attending at least one session or reviewing the provided postclass materials at least once. Full participation was defined as attending all four sessions.

Other Variables

Participant engagement during sessions was measured by the duration of attendance recorded by the online system and the frequency of responses to randomly posed questions.

Postclass review activities were assessed based on the frequency and duration of video reviews and the number of supplementary materials reviewed.

Total learning time included cumulative time spent attending live classes and reviewing class materials.

Training effectiveness was determined by comparing pre- and posttraining quiz scores. Learning outcomes were indicated by the final quiz scores. Participants whose score improvement exceeded the 90th percentile of all participants were classified as excellent learners.

Statistical Analysis

Continuous variables were assessed for normality using the Shapiro-Wilk test. Normally distributed data were presented as means (SDs), while nonnormally distributed data were expressed as medians (IQRs) or median (minimum, maximum). Categorical variables were presented as frequencies and percentages. Nonnormally distributed data were analyzed using the Wilcoxon signed rank test. Differences in categorical variables were analyzed using either the χ^2 test or Fisher exact test, as appropriate.

Spearman correlation coefficients and multivariate linear regression analyses were used to identify factors influencing final quiz scores. Variables with Spearman correlation coefficients (P<.05) were included in the multivariate regression model. Collinearity diagnostics were performed, and variables inducing collinearity were excluded. Factors associated with excellent learners were evaluated using receiver operating characteristic (ROC) curve analysis.

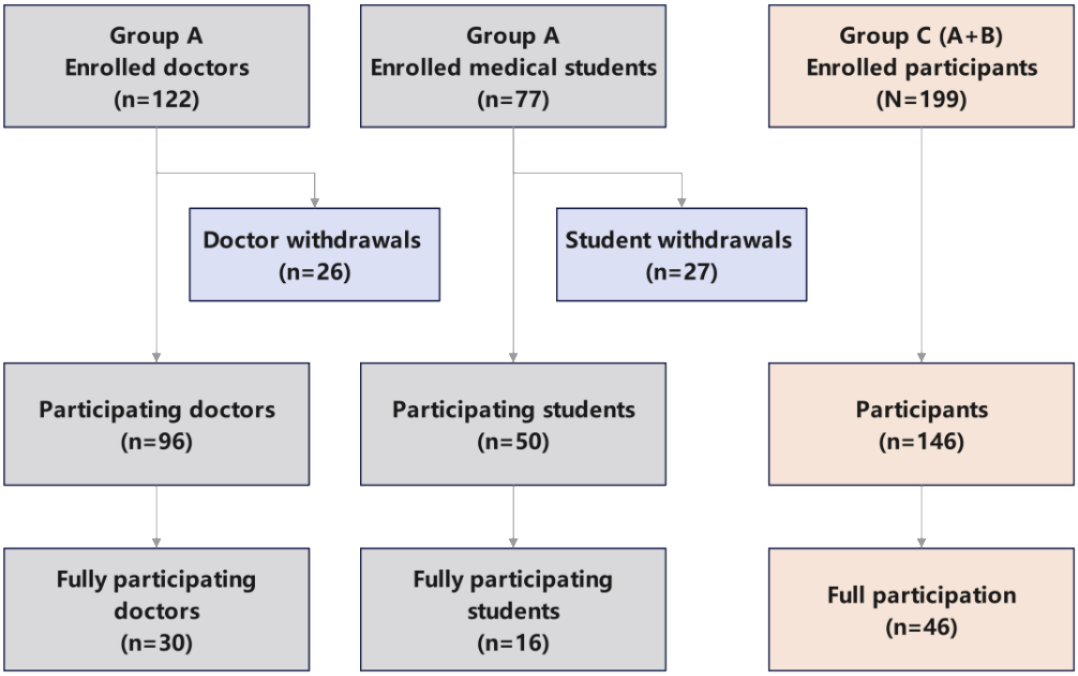
All statistical analyses were performed using SPSS Statistics for Windows, Version 23 (IBM Corp) and GraphPad Prism, Version 10.1.2 (GraphPad Software). Statistical significance was defined as P<.05.

Results

Overview

A total of 199 individuals voluntarily registered for the training, including 122 doctors and 77 medical students. The recruitment of doctors was completed within 2 days, whereas student recruitment required about a month. All registrants indicated that they had not yet mastered the skill of cardiac auscultation. The doctors were significantly older than the students and had substantially more years of prior exposure to auscultation learning. From registration to actual participation, a higher proportion of doctors attended the training compared with medical students (96/122, 78.7% vs 50/77, 64.9%; χ^2_1=4.57; P=.03). Overall, 73.4% (146/199) of registrants participated in the training, corresponding to a 26.6% (53/199) dropout rate. However, only 23.1% (46/199) of all registrants attended all four sessions (Figure 1), with no significant difference in full participation between doctors and students.

Figure 1. Participant flow diagram from registration to full course completion.



For participation motivation, doctors were more driven by intrinsic motivation than medical students (89/122, 73.0% vs 15/77, 19.5%; χ^2_1=54.10; *P*<.001). Intrinsic motivation was significantly correlated with age (*r*=0.394; *P*=.004) and total study time (*r*=0.145; *P*=.04), but not with the number of in-class random questions answered (*r*=0.153; *P*=.11). A χ^2 test further indicated that intrinsic motivation was associated with full participation (χ^2_1=4.03; *P*=.045). A significant increase in posttraining scores suggested that the program was effective (Table 1).

Table 1. Characteristics of participants in online heart auscultation training.

Variable[a]	Total (N=199)	Doctors (n=122)	Medical students (n=77)	*P* value
Age (years), median (IQR)	26 (23-31)	29 (26-35)	22 (20-25)	<.001
Female sex, n (%)	136 (68.3)	91 (74.6)	49 (63.6)	.26
Years of learning heart sound auscultation, median (IQR)	5 (3-8)	7 (5-11)	2 (1-4)	<.001
Training participants, n (%)[b]	146 (73.4)	96 (78.7)	50 (64.9)	.03
Participants who attended all 4 sessions, n (%)	46 (23.1)	30 (24.6)	16 (20.8)	.61
Motivation				<.001
Intrinsic motivation, n (%)	104 (52.3)	89 (73.0)	15 (19.5)	
Extrinsic motivation, n (%)	95 (47.7)	33 (27.0)	62 (80.5)	
Pretraining score, median (IQR)	40 (20-50)	40 (20-50)	30 (20-50)	.47
Posttraining score, median (IQR)	70 (50-83)	70 (50-90)	50 (50-75)	.29
Individual score change, median (IQR)	30 (10-45)	35 (10-50)	30 (0-40)	.66

[a]Continuous variables compared using the Mann-Whitney *U* test; categorical variables compared using the χ^2 test or Fisher exact test, as appropriate. Two-sided *P*<.05 indicates statistical significance.

[b]Defined as participants attending ≥1 session or reviewing postclass materials at least once.

During the four sessions, a total of 16 random questions were posed, and 11 sets of review materials were distributed online after each class. The system automatically recorded the number of in-class responses and the frequency of postclass access to review materials. Of the 146 training participants, 49 (33.6%) participants watched the lecture videos after class; however, only 10 (6.8%) viewed all four full recordings. The lecture material was reviewed by 111 (76.0%) participants, but only 20 (13.7%) accessed all available materials.

The duration of postclass video viewing varied substantially among participants. Age was significantly and positively correlated with total study time (*r*=0.366; *P*<.001), time spent reviewing recorded videos (*r*=0.330; *P*<.001), and the number of review materials accessed (*r*=0.355; *P*<.001; Table 2).

Table 2. Assessment results for classroom attendance and postclass review. Data are limited to training attendees only.

Variable[a]	Total	Doctors	Medical students	*P* value
Live class participations (n), median (IQR)	2 (1-4)	2 (1-4)	2.5 (1-4)	.41
Total duration of live class participation (min), median (IQR)	191 (84-384)	183 (79-375)	210 (84-404)	.77
Random questions in class (n), median (IQR)	5 (1-16)	5 (1-14)	5 (3-16)	.23
Course materials reviewed after class (n), median (IQR)	1 (0-3)	1 (0-5)	0 (0-1)	.01
Duration of watching lecture videos (min), median (IQR)	0 (0-578)	0 (0-578)	0 (0-169)	.01
Total study time (min), median (IQR)	202 (86-422)	192 (86-424)	210 (88-416)	.89

[a]Continuous variables compared using the Mann-Whitney *U* test. Two-sided *P*<.05 indicates statistical significance.

Analysis of Factors Affecting Training Scores

We performed univariate correlation analyses to examine the relationships between multiple factors—including learning motivation, overall participation, number of class attendances, duration of class participation, frequency of in-class responses to random questions, number of postclass material reviews, number of times watching lecture videos, and total study time—and the final scores. Significant correlations were observed between the final scores and full participation (r=0.351; P=.02), frequency of in-class responses (r=0.431; P=.004), and number of postclass material reviews (r=0.345; P=.03).

Age, participant type (doctor or student), years of prior auscultation study, motivation for participation, and time spent attending live sessions were not significantly correlated with final scores. Further multivariable linear regression analysis showed that only full participation and frequency of in-class responses remained independently associated with final scores, whereas postclass material review frequency did not significantly affect performance (Table 3).

Table 3. Multivariate linear regression results for factors associated with final auscultation scores (adjusted R^2=0.483).

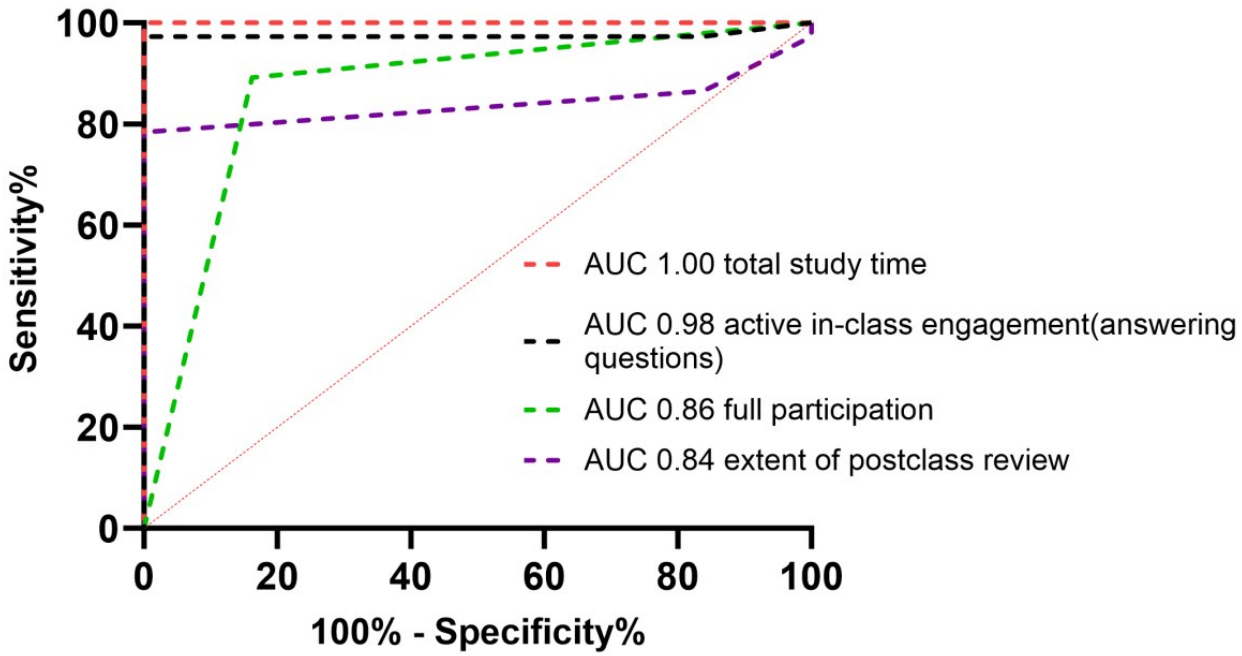
Regression variables	B (95% CI)	β	*P* value
Constant	–21.789 (–50.855 to 7.277)	—[a]	.13
Full participation	41.547 (24.426 to 58.667)	0.602	<.001
Times of answering random questions in classes	4.794 (3.054 to 6.445)	0.695	<.001

[a]Not applicable.

Factors Influencing Becoming an Excellent Learner

Six participants increased their scores by more than 60 points after the training, placing them in the top 10% of all participants and defining them as excellent learners. ROC curve analysis suggests that total study time, actively answering questions in class, full participation, and the extent of postclass review were all significantly related to achieving excellent learner status (Figure 2).

Figure 2. Receiver operating characteristic curve analysis of factors related to excellent performance. AUC: area under the curve.



Discussion

Principal Findings

In this voluntary noncredit online cardiac auscultation course, cohort-level test scores improved from pre- to posttraining, indicating overall instructional effectiveness. Nevertheless, only 23% of registrants completed all four sessions, underscoring the difficulty of sustaining participation in voluntary settings. Within the cohort, intrinsic motivation was positively associated with full participation, and both full participation and more frequent content-aligned in-class responses were independently associated with higher final scores. Postclass review was not an independent predictor in multivariable models, yet it consistently characterized learners who achieved excellent performance in ROC analyses.

Interpretation and Comparison With Prior Work

Taken together, the data suggest a plausible pathway for success in voluntary online learning: autonomous motivation→sustained participation→active content-aligned engagement→higher performance. This pathway is consistent with the self-determination theory and with contemporary applications of autonomy-supportive design in health professional education, which emphasize persistence and self-regulation in learning tasks [20,21]. Methodologically, our study contributes by operationalizing autonomy through objective behavioral traces—attendance, periodic in-class responses, and review—rather than relying solely on self-report, which aligns with recent work mapping measurable dimensions of student engagement in health professional education [22].

Low completion in our cohort mirrors patterns reported in large-scale voluntary online courses, where completion is often modest; for example, completion of massive open online courses frequently clusters around 10% in some contexts. A recent systematic review concluded that dropout is multifactorial and more closely related to course attributes, learner behaviors, and motivation/self-regulation than to content alone, reinforcing the need for proactive planning to sustain participation in voluntary courses [11]. In practical terms, clear expectation setting before enrollment—regarding time commitment, learning objectives, and assessment schedule—and simple scheduling reminders have been reported to improve learner persistence in online education [11] and may help participants maintain full engagement in voluntary courses.

Beyond attendance, maintaining focused attention during sessions appears salient. Empirical work shows that off-task device use is more frequent online than face-to-face and is associated with perceived distraction, which can erode engagement [23]. In medical education contexts, survey evidence and expert opinion further suggest that attention tends to decline after approximately 10-15 minutes unless refreshed through interactive elements. Accordingly, periodic content-aligned checks (brief questions or polls), as implemented in our study, are advisable in synchronous sessions [24].

Finally, postclass review may play a selective role in mastery. Although postclass review was not an independent predictor in our adjusted model, high performers engaged in postclass review more often. This pattern is consonant with meta-analytic evidence that spaced or retrieval-based reviews improve long-term retention and, in some cases, clinical behavior change relative to massed learning [25]. In health professional education, spaced e-learning has been shown to enhance knowledge retention in basic life support compared with single-session learning [26]. Future implementations could explore embedding simple, low-burden review structures—such as a 48-hour recap and a 1-week revisit—and providing automated reminders to prompt learners to consolidate key material.

Notably in our cohort, the associations between participation/engagement and performance were not explained by measured demographic characteristics (age, years of exposure to auscultation, workplace), suggesting that the behavioral pathway outlined above could be applicable across subgroups in similar settings.

Limitations

This study has several limitations. It was conducted at a single center with voluntary participation, which may limit generalizability and introduce self-selection bias. The primary outcome was a short-term test score, and long-term retention was not assessed. Although the pre- and posttests followed the same blueprint, using different item sets may have introduced measurement variability. Motivation was assessed by a single unvalidated item, and the voluntary single-center sample entails self-selection; hence, residual confounding and limited generalizability remain.

Conclusions and Broader Implications

Learners who consistently attended all sessions, stayed focused during class, and reviewed material after class achieved the best results in this voluntary online clinical skills course. These behaviors represent practical expressions of learner autonomy and can be strengthened through simple habits—committing to scheduled sessions, engaging actively with class tasks, and revisiting key materials on a spaced schedule. For busy health professionals, such disciplined learning routines may transform limited study time into durable competence.

Beyond this specific course, these findings highlight the importance of quantifying and supporting learner autonomy as a measurable construct in online medical education. Incorporating behavioral analytics to monitor participation, engagement, and review patterns may provide educators with actionable insights to design more autonomy-supportive learning environments. Future research should investigate how these learner-driven behaviors can be further supported through course design and motivation-enhancing strategies to sustain long-term engagement and performance across diverse educational settings.

Acknowledgments

This project is supported by the Educational Reform Project of Peking Union Medical College, Chinese Academy of Medical Sciences, under grants 2023zlgl032 and 2021zlgc0110.

ChatGPT (OpenAI) was used for language polishing, and the authors take full responsibility for all revisions.

Data Availability

The datasets generated and analyzed during this study are not publicly available because they contain identifiable personal information and course platform activity records in Chinese. Deidentified data may be made available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: XL, YF
Methodology: XL
Investigation: YF
Data curation: XL
Formal analysis: XL
Validation: LF
Writing – original draft: XL
Writing – review & editing: XL, YF, LF
Supervision: YF
Project administration: YF
All authors had full access to all data in the study. XL took responsibility for the integrity of the data and accuracy of the data analysis. All authors approved the final version of the manuscript and agree to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

References

1. Goldberg LR, Crocombe LA. Advances in medical education and practice: role of massive open online courses. Adv Med Educ Pract 2017;8:603-609 [FREE Full text] [doi: 10.2147/AMEP.S115321] [Medline: 28860891]
2. Zhu H, Xu J, Wang P, Bian J, Zhao Z, Liu H, et al. The irreplaceable role of medical massive open online courses in China during the COVID-19 pandemic. BMC Med Educ 2023 May 09;23(1):323 [FREE Full text] [doi: 10.1186/s12909-023-04315-z] [Medline: 37158861]

3. O'Doherty D, Dromey M, Lougheed J, Hannigan A, Last J, McGrath D. Barriers and solutions to online learning in medical education - an integrative review. BMC Med Educ 2018 Jun 07;18(1):130 [FREE Full text] [doi: 10.1186/s12909-018-1240-0] [Medline: 29880045]
4. Moore MG. The theory of transactional distance. In: Moore MG, editor. Handbook of Distance Education. New York, NY: Routledge; 2013.
5. MacNeill H, Masters K, Nemethy K, Correia R. Online learning in health professions education. Part 1: teaching and learning in online environments: AMEE Guide No. 161. Med Teach 2024 Jan;46(1):4-17 [FREE Full text] [doi: 10.1080/0142159X.2023.2197135] [Medline: 37094079]
6. Learner autonomy. Wikipedia. 2023. URL: https://en.wikipedia.org/w/index.php?title=Learner_autonomy&oldid=1178564267 [accessed 2024-06-23]
7. Gupta DK, Chaudhuri A, Gaine D. A systematic review of self-directed learning in medical education in undergraduate medical students. Curr Med Issues 2025 Mar;23(1):61. [doi: 10.4103/cmi.cmi_96_24]
8. Knowles MS. Self-Directed Learning: A Guide for Learners and Teachers. New York, NY: Association Press; 1975.
9. Learner autonomy. ScienceDirect. URL: https://www.sciencedirect.com/topics/psychology/learner-autonomy [accessed 2024-06-23]
10. Gupta N, Ali K, Jiang D, Fink T, Du X. Beyond autonomy: unpacking self-regulated and self-directed learning through the lens of learner agency- a scoping review. BMC Med Educ 2024 Dec 23;24(1):1519. [doi: 10.1186/s12909-024-06476-x] [Medline: 39716158]
11. Huang H, Jew L, Qi D. Take a MOOC and then drop: a systematic review of MOOC engagement pattern and dropout factor. Heliyon 2023 Apr;9(4):e15220 [FREE Full text] [doi: 10.1016/j.heliyon.2023.e15220] [Medline: 37123924]
12. George PP, Papachristou N, Belisario JM, Wang W, Wark PA, Cotic Z, et al. Online eLearning for undergraduates in health professions: a systematic review of the impact on knowledge, skills, attitudes and satisfaction. J Glob Health 2014 Jun;4(1):010406 [FREE Full text] [doi: 10.7189/jogh.04.010406] [Medline: 24976965]
13. Regmi K, Jones L. A systematic review of the factors - enablers and barriers - affecting e-learning in health sciences education. BMC Med Educ 2020 Mar 30;20(1):91 [FREE Full text] [doi: 10.1186/s12909-020-02007-6] [Medline: 32228560]
14. Berger-Estilita J, Krista L, Gogollari A, Schmitz F, Elfering A, Guttormsen S. Self-directed learning in health professions: a mixed-methods systematic review of the literature. PLoS One 2025;20(5):e0320530 [FREE Full text] [doi: 10.1371/journal.pone.0320530] [Medline: 40315189]
15. Owen SJ, Wong K. Cardiac auscultation via simulation: a survey of the approach of UK medical schools. BMC Res Notes 2015 Sep 10;8:427 [FREE Full text] [doi: 10.1186/s13104-015-1419-y] [Medline: 26358413]
16. Martínez G, Guarda E, Baeza R, Garayar B, Chamorro G, Casanegra P. A heart sound simulator as an effective aid in teaching cardiac auscultation to medical students and internal medicine residents. Rev Esp Cardiol (Engl Ed) 2012 Dec;65(12):1135-1136. [doi: 10.1016/j.recesp.2012.03.022] [Medline: 22796032]
17. Vukanovic-Criley JM, Criley S, Warde CM, Boker JR, Guevara-Matheus L, Churchill WH, et al. Competency in cardiac examination skills in medical students, trainees, physicians, and faculty: a multicenter study. Arch Intern Med 2006 Mar 27;166(6):610-616. [doi: 10.1001/archinte.166.6.610] [Medline: 16567598]
18. McGee RG, Wark S, Mwangi F, Drovandi A, Alele F, Malau-Aduli BS, ACHIEVE Collaboration. Digital learning of clinical skills and its impact on medical students' academic performance: a systematic review. BMC Med Educ 2024 Dec 18;24(1):1477 [FREE Full text] [doi: 10.1186/s12909-024-06471-2] [Medline: 39696150]
19. Fang Y, Fang L, Zhu W, Lin X. The impact of learner autonomy on the performance in voluntary online cardiac auscultation courses. Research Square Preprint posted online on September 2, 2024. [doi: 10.21203/rs.3.rs-4758934/v1]
20. Ryan RM, Deci EL. Self-Determination Theory Basic Psychological Needs in Motivation, Development, and Wellness. New York, NY: Guilford Publications; 2018.
21. Kusurkar RA. 32 Self-determination theory in health professions education research and practice. In: Ryan RM, editor. The Oxford Handbook of Self-Determination Theory. Oxford, England: Oxford University Press; 2023.
22. Kassab SE, Al-Eraky M, El-Sayed W, Hamdy H, Schmidt H. Measurement of student engagement in health professions education: a review of literature. BMC Med Educ 2023 May 20;23(1):354 [FREE Full text] [doi: 10.1186/s12909-023-04344-8] [Medline: 37210491]
23. Aivaz KA, Teodorescu D. College students' distractions from learning caused by multitasking in online vs. face-to-face classes: a case study at a public university in Romania. Int J Environ Res Public Health 2022 Sep 06;19(18):11188 [FREE Full text] [doi: 10.3390/ijerph191811188] [Medline: 36141459]
24. Çakmakkaya ÖS, Meydanlı EG, Kafadar AM, Demirci MS, Süzer Ö, Ar MC, et al. Factors affecting medical students' satisfaction with online learning: a regression analysis of a survey. BMC Med Educ 2024 Jan 03;24(1):11 [FREE Full text] [doi: 10.1186/s12909-023-04995-7] [Medline: 38172870]
25. Martinengo L, Ng MSP, Ng TDR, Ang Y, Jabir AI, Kyaw BM, et al. Spaced digital education for health professionals: systematic review and meta-analysis. J Med Internet Res 2024 Oct 10;26:e57760 [FREE Full text] [doi: 10.2196/57760] [Medline: 39388234]

26. Ranjbar F, Sharif-Nia H, Shiri M, Rahmatpour P. The effect of spaced E-Learning on knowledge of basic life support and satisfaction of nursing students: a quasi-experimental study. BMC Med Educ 2024 May 15;24(1):537 [FREE Full text] [doi: 10.1186/s12909-024-05533-9] [Medline: 38750506]

Abbreviations

ROC: receiver operating characteristic

Edited by A Stone, T Leung; submitted 01.06.25; peer-reviewed by T Gladman, C Pawlik; comments to author 10.10.25; revised version received 24.10.25; accepted 24.10.25; published 25.11.25.

Please cite as:
Fang Y, Fang L, Lin X
Impact of Learner Autonomy on the Performance in Voluntary Online Cardiac Auscultation Courses: Prospective Self-Controlled Study
JMIR Med Educ 2025;11:e78363
URL: https://mededu.jmir.org/2025/1/e78363
doi:10.2196/78363
PMID:

Original Paper

Student Satisfaction in Social Media–Based Learning Environments: Development, Validation, and Psychometric Evaluation of the CuSAERS (Questionnaire of Satisfaction With Educational Activities Performed on Social Media)

Roy La Touche[1,2,3*], PhD; Álvaro Reina-Varona[1,2*], MSc; Mónica Grande-Alonso[4,5*], PhD; José Vicente León-Hernández[1,2*], PhD; Joaquín Pardo-Montero[1,2,6*], PhD; Néstor Requejo-Salinas[1,2*], PhD; Raúl Ferrer-Peña[1,5,7,8*], PhD; Alba Paris-Alemany[2,3,9*], MD, PhD

[1]Departamento de Fisioterapia, Centro Superior de Estudios Universitarios (CSEU) La Salle, Universidad Autónoma de Madrid, Madrid, Spain

[2]Motion in Brains Research Group, Centro Superior de Estudios Universitarios (CSEU) La Salle, Universidad Autónoma de Madrid, Madrid, Spain

[3]Instituto de Dolor Craneofacial y Neuromusculoesquelético (INDCRAN), Madrid, Spain., Madrid, Spain

[4]Departamento de Cirugía, Ciencias Médicas y Sociales, Facultad de Medicina, Universidad de Alcalá, Alcalá de Henares, Spain., Alcalá de Henares, Spain

[5]Clinical-Teaching Research Group on Rehabilitation Sciences (INDOCLIN), CSEU La Salle, UAM, Madrid, Spain, Madrid, Spain

[6]Hospital La Paz Institute for Health Research (IdiPAZ), Madrid, Spain

[7]Cognitive Neuroscience, Pain and Rehabilitation Research Group (NECODOR), Faculty of Health Sciences, Rey Juan Carlos University, Alcorcón, Spain, Alcorcón, Spain

[8]Centro de Salud Entrevías, Gerencia Asistencial de Atención Primaria de la Comunidad de Madrid, Madrid, Spain

[9]Department of Basic Health Sciences, Universidad Rey Juan Carlos, Alcorcón, Spain., Alcorcón, Spain

[*]all authors contributed equally

Corresponding Author:
Raúl Ferrer-Peña, PhD
Departamento de Fisioterapia, Centro Superior de Estudios Universitarios (CSEU) La Salle, Universidad Autónoma de Madrid
Calle La Salle, 10
Madrid
Spain
Phone: 34 917401980
Email: drraulferrer@gmail.com

Abstract

Background: Social media platforms are increasingly integrated into higher education, enabling collaborative, student-centered learning. Yet, few instruments specifically measure students' satisfaction with these activities across platforms. A brief, valid tool is needed to evaluate perceived quality and guide instructional design in social media–based learning environments.

Objective: This study investigated the use of social media as educational tools in the university environment, with the aim of designing and validating the CuSAERS (Questionnaire of Satisfaction With Educational Activities Performed on Social Media).

Methods: Using a mixed and sequential methodology, we explored the perceptions of bachelor's and master's degree students in physiotherapy who participated in teaching activities through X (formerly Twitter) and Instagram. The first phase of the project identified key dimensions of satisfaction from the literature, expert interviews, and cognitive interviews. The second phase assessed the psychometric properties of the CuSAERS in a sample of 150 students, addressing construct validity, internal reliability, concurrent validity, and discriminant validity.

Results: Exploratory factor analysis supported a 3-factor structure—perception of learning, task satisfaction/environment, and self-realization—explaining 61.9% of the variance, with acceptable overall reliability. Concurrent validity was supported by moderate correlations with the Academic Satisfaction Scale. Master's students reported higher scores than bachelor's students.

Conclusions: CuSAERS provides preliminary evidence as a promising measure of student satisfaction with social media–based learning activities; its use should remain formative and cautious until confirmatory and invariance analyses are completed.

Trial Registration: No applicable.

(JMIR Med Educ 2025;11:e73805) doi:10.2196/73805

KEYWORDS

academic engagement; collaborative learning; digital learning; educational technology; learning environments; psychometric validation; social media; student satisfaction

Introduction

Social media refers to online resources designed to facilitate interaction and engagement among individuals [1]. Over the last decade, these platforms have transformed the way people communicate, learn, and collaborate, extending far beyond personal and social contexts into education [2-4]. Tools such as X (formerly Twitter), Instagram, Facebook, and YouTube now support innovative pedagogical strategies aligned with digitalization and global collaboration, promoting learning experiences that transcend geographic and temporal boundaries [5-7].

In health professions education, social media can promote collaborative learning, reflective discussion, and peer interaction, enhancing students' engagement and motivation [8-10]. Platforms like X facilitate professional dialogue and access to scientific information, while Instagram and YouTube enable visual learning and dissemination of clinical content [11,12]. However, drawbacks include information overload, variable content quality, and risks related to privacy or professionalism [13,14]. These limitations underscore the need for structured pedagogical design and critical evaluation of their educational use.

Despite the rapid adoption of social media in educational settings, systematic approaches to evaluate their educational impact remain limited. Existing measures often emphasize usability, frequency of use, or performance metrics while neglecting students' affective responses and satisfaction with learning experiences [15-18]. Yet satisfaction is closely linked to engagement, retention, and perceived learning quality, making it a key outcome for educational quality assurance [19-21].

Assessing satisfaction requires specific, psychometrically sound instruments that capture the attitudinal evaluation of students toward learning activities conducted on social media, beyond usability or behavioral engagement [22,23]. Such measurement tools can inform educators and institutions about the perceived quality and effectiveness of teaching innovations in digital environments and guide continuous improvement in educational design [24].

This study aimed to develop and provide an initial psychometric evaluation of the CuSAERS (Questionnaire of Satisfaction With Educational Activities Performed on Social Media), a tool designed to assess students' satisfaction with learning experiences delivered through social media. Specifically, we examined its internal consistency, factorial structure, and construct, concurrent, and discriminant validity across subgroups using different social media platforms.

Methods

Ethical Considerations

The university's ethics committee of La Salle Higher Center for University Studies determined that, given the design of the study, formal ethics committee review and approval were not required and therefore granted an exemption, authorizing the study to be conducted without ethics committee approval. All students signed an informed consent document before participating, which explained the objectives of the study, the confidentiality of the data, and the voluntary nature of their participation. The database was encoded using an alphanumeric code, and no personal data was recorded. No compensation was provided to the participants.

Study Design

This study used an exploratory sequential mixed design, combining qualitative and quantitative methods for the development, construction, and validation of the CuSAERS. The first qualitative phase, which included a literature review, semistructured interviews, expert content validation, cognitive interviews, and a pilot test, has been previously published [25]. This first phase had been aimed at identifying the dimensions of satisfaction perceived by students with regard to educational activities carried out on social media, such as interaction with the teacher, the quality of the content, and collaboration between colleagues.

The second phase focused on the psychometric evaluation of the CuSAERS. During this stage, the questionnaire was administered to a representative sample of bachelor's and master's students to assess its psychometric properties, including validity and reliability.

Participants

The sample selected on a nonprobabilistic basis consisted of 160 bachelor's and master's degree students in physiotherapy from a Spanish university. Of these, 110 were bachelor's students and 50 were master's students. All participants were aged 18 years or older and enrolled in official academic programs. As an inclusion criterion, participants had to have previously participated in educational activities developed on social media, specifically on X or Instagram.

Teaching Innovation Activities

The master's students participated in an academic activity designed on Instagram. This activity consisted of the critical analysis of therapeutic exercises published in video format on this platform. Students watched the videos, identified technical errors, proposed modifications, and justified their proposals based on scientific evidence. This task not only fostered autonomous and critical learning but also promoted the use of social media as a professional learning tool. The students'

responses and observations were subsequently discussed in a digital forum supervised by the teacher, who acted as moderator and interlocutor, offering specific and targeted feedback.

The bachelor's students participated in an activity developed on X. This activity included a virtual debate on aspects related to chronic pain. Each student had to perform prior research on the assigned topic and present their position in the form of X threads, using technical but accessible language. The debates were enriched with bibliographical references and the integration of specific hashtags to facilitate the monitoring of the discussions. The teacher's involvement focused on moderating the debate, posing critical questions to deepen the arguments, and providing direct feedback on the content and quality of the interactions.

Both activities were designed to maximize students' active participation, integrating the use of social media as an innovative pedagogical resource. The strategies used sought to foster critical, technical, and communicative skills in digital contexts, thus contributing to more dynamic and meaningful learning.

Procedure

Phase 1: Development of CuSAERS

The first phase was described in a previous study and consisted of a literature review and interviews with experts in educational methodology, social media, and rehabilitation, as well as with physiotherapy students [25].

Phase 2: Psychometric Evaluation

In the second phase, the CuSAERS questionnaire was administered to a sample of 150 students, of whom 90 were bachelor's and 60 were master's students. The students completed the questionnaire online, and a cross-sectional design was used to assess the psychometric properties of the instrument.

Data Analysis

Descriptive Statistics

The data analysis was performed exclusively using JAMOVI (version 2.6; The Jamovi Project) software, which allowed all statistical evaluations of the study to be performed. Descriptive statistics were used to summarize the categorical variables, expressed in absolute and relative frequencies, and the continuous variables, which were reported in terms of means, SDs, and 95% CIs.

Normality Assessment

The normality of the data was comprehensively assessed, using both statistical tests and graphical methods.

The Kolmogorov–Smirnov test was applied to determine whether the distributions of the variables differed significantly from a normal distribution. This test was complemented by a visual analysis using quantile–quantile (Q–Q) plots and histograms, which allowed the alignment of the data with the theoretical normal curve to be assessed.

In addition, skewness and kurtosis coefficients were calculated to assess the shape of the distributions. Skewness values close to 0 indicate symmetric distributions, whereas positive or negative values indicate skewing to the right or left, respectively. As for kurtosis, values close to 3 indicate mesokurtic distributions, whereas higher or lower values suggest leptokurtic or platykurtic distributions, respectively. These metrics provided a quantitative framework for interpreting the normality of the data distributions.

Lastly, additional analyses were performed to assess the influence of outliers on the distributions. They were identified and visually examined using boxplots, which allowed us to determine whether these outliers had a significant impact on the structure of the data.

Construct Validity

We conducted an exploratory factor analysis (EFA) using the minimum residual extraction method with oblimin rotation. Sampling adequacy was evaluated with the Kaiser-Meyer-Olkin (KMO) index and Bartlett's test of sphericity. The number of factors was determined using multiple criteria: Kaiser's rule (eigenvalue ≥1), scree plot, parallel analysis, and exploratory graph analysis. Factors were retained when they had ≥2 items with minimal cross-loadings.

The root-mean-square error of approximation (RMSEA) was calculated with a 90% CI. RMSEA values up to 0.08 are considered indicative of a reasonable fit to the data, with values closer to 0.05 or lower suggesting a good fit. In addition, the Tucker-Lewis Index (TLI) was calculated, with values close to or above 0.95 indicating an excellent fit. The Bayesian information criterion (BIC) was also assessed, in which lower (more negative) values are preferred, suggesting a model that better explains the data with fewer parameters. Lastly, the model fit was confirmed using the chi-square test. A nonsignificant chi-square value indicates that the observed and expected covariances of the model do not differ significantly, supporting an adequate model fit.

Item retention followed a priori thresholds to preserve content validity and parsimony: only items with primary loadings ≥0.40 on a single factor were kept; items with cross-loadings within 0.20 of the primary loading or with communality/uniqueness values indicating poor common variance were considered for removal [26]. The final solution and retained items were based on these empirical criteria in conjunction with theoretical interpretability.

Internal Consistency (Reliability)

Internal consistency was assessed using Cronbach α coefficient, with values above 0.70 considered adequate, and McDonald omega (ω) coefficient, which provides a complementary estimate of internal consistency. This approach allowed a robust assessment of the homogeneity of the items that make up the dimensions of the CuSAERS.

Concurrent Validity

Concurrent validity was assessed by calculating Pearson correlations between CuSAERS scores and the Academic Satisfaction Scale (ASS), used as a reference instrument. The ASS defines academic satisfaction as "the well-being and

enjoyment that students perceive in their experiences within the academic role" [27].

This instrument, composed of 7 items organized into a single factor, uses a 7-point Likert scale and has demonstrated high reliability (ordinal α=.92) and structural validity in Chilean university contexts [28].

In addition, previous experience in using social media, measured in months, was included as an additional variable to explore its relationship with CuSAERS scores. This information allowed us to assess whether the length of time spent using social media significantly influences the perception of academic satisfaction in activities performed in these environments.

The values of the correlations between the CuSAERS, ASS, and social media experience were interpreted according to the criteria of Schober et al [29]. The correlations were classified as follows: insignificant (0.00-0.10), weak (0.10-0.39), moderate (0.40-0.69), strong (0.70-0.89), and very strong (0.90-1.00).

Floor and Ceiling Effect

The presence of floor and ceiling effects was assessed by calculating the percentage of participants who obtained the lowest or highest possible scores on the questionnaire. An effect was considered significant if more than 15% of participants were at these extremes.

Discriminant Validity

The discriminant validity of the CuSAERS was explored by comparing groups of students with various levels of engagement in educational activities on social media platforms. Specifically, analyses were conducted to compare students who participated in activities on X versus those who used Instagram, examining differences in their satisfaction scores.

Additionally, students with prior experience using social media were compared with those without such experience to determine the impact of this variable on perceptions of educational satisfaction. A comparative analysis was also performed between physiotherapy master's students and bachelor's students to evaluate potential differences based on academic level.

The comparisons were conducted using the Mann–Whitney U test as a nonparametric method for independent samples, and effect sizes were calculated using rank-biserial correlation. Effect sizes were interpreted as small (r=0.10), moderate (r=0.30), and large (r=0.50). These analyses identified significant differences that support the ability of the CuSAERS to distinguish between groups with diverse educational characteristics and contexts.

Discriminant comparisons (eg, bachelor's vs master's degree; prior social media experience) are reported as exploratory contrasts with effect sizes and are not adjusted for covariates. In subsequent work we will conduct confirmatory factor analysis (CFA), multigroup measurement invariance (by degree level, gender, and age), cross-validation in an independent multisite sample, and covariate-adjusted models (eg, regression or standard error of means) to strengthen causal interpretability of group differences.

Results

Normality Analysis

The normality of the data was assessed through a comprehensive analysis that included both statistical tests and graphical methods, providing an in-depth understanding of the distribution of the items in the CuSAERS. The results of the Kolmogorov–Smirnov test indicated that the data significantly deviated from a normal distribution (P<.001), confirming that the variables did not meet the assumption of normality. This finding was consistent with skewness and kurtosis values, which suggested deviations from perfect symmetry and distributions far from the ideal mesokurtic shape. Most of the items showed negative skewness, indicating a tendency for participants to give higher scores on the satisfaction scale, while predominantly platykurtic kurtosis reflected lower concentration at the scale's extremes.

The visual analysis complemented these statistical evaluations. Q–Q plots and histograms demonstrated that the data did not align with the theoretical normal curve, confirming the presence of biases and lighter tails in the distribution of scores. Additionally, the influence of outliers was explored using boxplots, which revealed greater dispersion in specific items, such as Item 8 and Item 12. These items, which had the highest SDs, reflected the heterogeneity in the participants' perceptions regarding these specific aspects of the questionnaire.

Descriptive Analysis

The descriptive analysis of the items showed means and medians centered around values close to 3, reflecting a tendency toward neutral or moderately positive responses. SDs ranged between 0.752 and 1.046, demonstrating differences in the dispersion of responses. Items with the highest dispersion were Item 8 and Item 12, whereas Item 16 presented the lowest variability.

Regarding distribution, most items exhibited negative skewness, suggesting a slight tendency toward higher scores, except for Item 8, which showed skewness close to 0. Platykurtic kurtosis predominated, indicating lower concentration of responses at the extremes, except for some items, such as Item 3 and Item 4, which displayed more leptokurtic distributions.

In terms of internal consistency, the item-total correlations were heterogeneous, ranging from −0.001 (Item 16) to 0.539 (Item 10). Items 11 and 16 stood out for having near-zero or negative item-total correlations, suggesting a lower contribution to the instrument's consistency. Cronbach α and McDonald ω coefficients ranged between 0.645 and 0.716, and between 0.674 and 0.755, respectively, indicating moderate reliability overall. The highest reliability values were observed for Items 11 and 16, although these items had low item-total correlations, which might imply atypical behavior for these items (Table 1).

Table 1. Descriptive statistics and reliability coefficients upon item removals.

Item	Mean (SD)	Median (IQR)	Skewness	Kurtosis	Item 1: all correlation	Cronbach α	McDonald ω
Item 3	2.98 (0.797)	3.0 (3-3)	−0.948	1.01	0.481	0.653	0.683
Item 4	3.12 (0.764)	3.0 (3-4)	−0.977	1.31	0.455	0.657	0.685
Item 6	2.92 (0.821)	3.0 (3-3)	−0.897	0.7	0.296	0.679	0.729
Item 7	3.02 (0.846)	3.0 (3-4)	−0.552	−0.323	0.490	0.650	0.684
Item 8	2.49 (1.046)	3.0 (2-3)	−0.05	−1.18	0.287	0.683	0.732
Item 9	2.98 (0.831)	3.0 (2-4)	−0.352	−0.618	0.432	0.659	0.695
Item 10	3.19 (0.773)	3.0 (3-4)	−0.929	0.887	0.539	0.645	0.674
Item 11	3.18 (0.831)	3.0 (2-4)	−0.351	−1.47	0.029	0.715	0.752
Item 12	2.63 (1.05)	3.0 (2-3)	−0.204	−1.14	0.382	0.666	0.721
Item 13	3.04 (0.846)	3.0 (2-4)	−0.387	−0.78	0.274	0.682	0.731
Item 16	3.01 (0.752)	3.0 (2-4)	−0.021	−1.22	−0.001	0.716	0.755
Item 17	3.02 (0.935)	3.0 (3-4)	−0.786	−0.164	0.311	0.674	0.723

Factor Analysis (Construct Validity)

The data were suitable for factor analysis (KMO=0.754; Bartlett's χ^2_{45}=685; P<.001). An exploratory factor analysis with oblimin rotation supported a 3-factor solution explaining 61.9% of the variance (sum of squared loadings: 3.07, 1.56, 1.56; 30.7%, 15.6%, 15.6%). Interfactor correlations were small (r=0.06-0.20).

Factor 1 (perception of learning) showed high primary loadings for Item 10 (0.821), Item 4 (0.820), Item 3 (0.803), Item 7 (0.748), and Item 9 (0.695), with uniqueness ranging from 0.318 to 0.518 (Table 2).

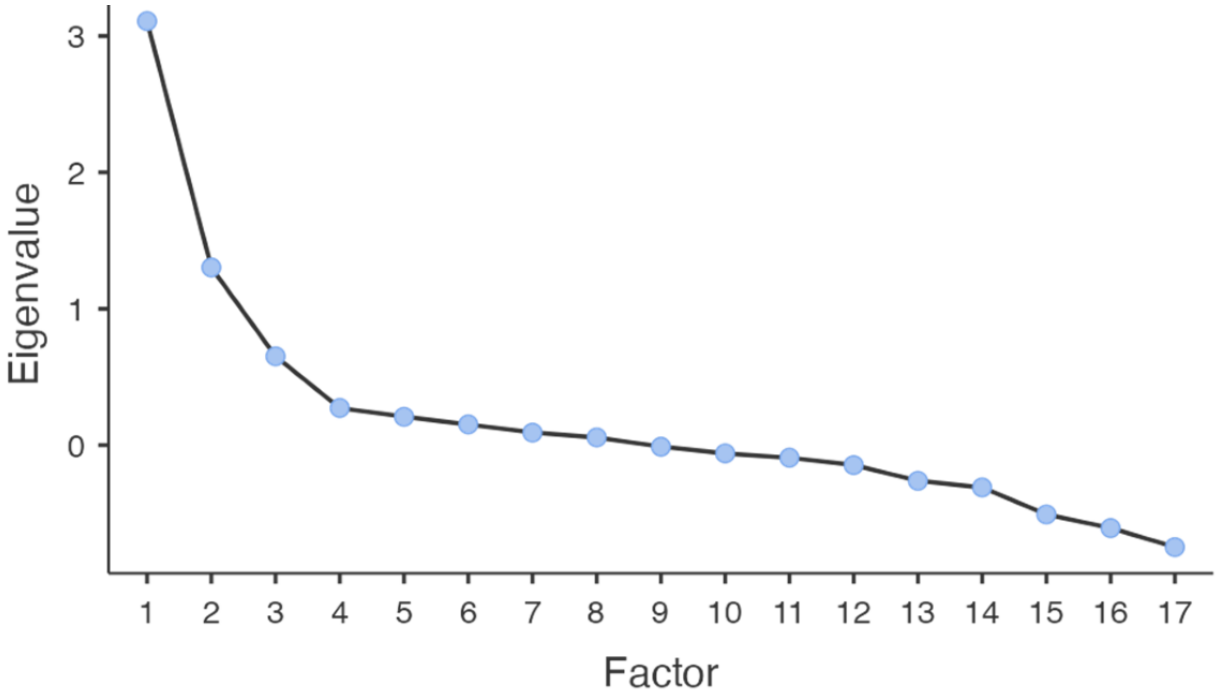
Table 2. Factor loadings of the items in relation to the factor solution. Extraction was performed using the minimum residuals method with oblimin rotation. Only primary loadings ≥0.40 are displayed. Factor labels follow the retained 3-factor solution. Factor 3 should be interpreted with caution because of a boundary estimate for Item 8. Factor 2, representing task satisfaction/environment, was defined by Item 13 (0.763), Item 6 (0.745), and Item 17 (0.626), with higher uniqueness for Item 17 (0.593). Factor 3, representing self-realization, was defined by Item 8 (1.002; boundary estimate) and Item 12 (0.722).

Item	Statement	Factor 1 (perception of learning)	Factor 2 (task satisfaction)	Factor 3 (self-realization)	Uniqueness
Item 10	Training activities on social media foster my reflection, synthesis, and reasoning. [Las actividades formativas en redes sociales fomentaron mi reflexión, síntesis y razonamiento.]	0.821	—[a]	—	0.318
Item 4	Educational activities on social media motivate me to ask questions and participate in discussions. [Las actividades educativas en redes sociales me motivan a hacer preguntas y participar en discusiones.]	0.820	—	—	0.330
Item 3	Educational activities on social media promote my participation. [Las actividades educativas en redes sociales promueven mi participación.]	0.803	—	—	0.334
Item 7	The use of social media in education has increased my interest in the course content. [El uso de redes sociales en la educación aumentó mi interés en los contenidos de la asignatura.]	0.748	—	—	0.431
Item 9	Using social media as a learning tool is beneficial. [Es positivo usar redes sociales como herramienta de aprendizaje.]	0.695	—	—	0.518
Item 13	The time spent on training activities on social media is well utilized. [El tiempo dedicado a actividades formativas en redes sociales está bien aprovechado.]	—	0.763	—	0.419
Item 6	My educational experience on social media makes me feel that it is a suitable environment to express my ideas. [Mi experiencia educativa en redes sociales me hace sentir que es un entorno adecuado para expresar mis ideas.]	—	0.745	—	0.427
Item 17	My experience indicates that social media is suitable for acquiring knowledge related to my field of study. [Mi experiencia indica que las redes sociales son adecuadas para adquirir conocimientos relacionados con mi carrera.]	—	0.626	—	0.593
Item 8	I am satisfied with my participation in educational activities conducted through social media. [Estoy satisfecho con mi participación en las actividades educativas desarrolladas con redes sociales.]	—	—	1.002	0.001
Item 12	I am satisfied with what I have learned in educational activities on social media. [Estoy satisfecho con lo que he aprendido en las actividades educativas en redes sociales.]	—	—	0.722	0.441

[a]Empty cells indicate loadings <0.40 on that factor.

Overall model fit was acceptable (RMSEA=0.058; 90% CI 0.00-0.100; TLI=0.961; χ^2_{18}=27.9; BIC=–63.4; *P*=.06). All retained items met the a priori loading threshold (≥0.40), and cross-loadings were minimal (Table 2). The scree plot and parallel analysis supported the retention of three factors: the first 2 observed eigenvalues were clearly above the simulated distribution, the third was approximately at the simulated threshold, and from the fourth onward, the observed eigenvalues fell below the simulated ones (Figure 1). We note that Item 8 presented a boundary (Heywood) estimate—an occurrence that can arise in small samples and oblique solutions—so this third factor should be considered provisional pending confirmatory testing and item redevelopment in an independent, larger sample.

Figure 1. Scree plot of the exploratory factor analysis of the CuSAERS (Questionnaire of Satisfaction With Educational Activities Performed on Social Media), showing the eigenvalues of the extracted factors; the "elbow" in the curve suggests retaining 4 factors, which explain the majority of the total variance.



Consistent with the scree plot and parallel analysis (Figure 1), we retained the 3-factor solution as the primary model (Table 3). The 2-factor alternative, although parsimonious, collapses conceptually distinct domains and does not map cleanly onto the theorized constructs (Table S1 in Multimedia Appendix 1). The 4-factor alternative introduces a psychometrically weak structure: after applying the a priori retention rule (primary loading ≥0.40), one factor is left with a single indicator (Table S2 in Multimedia Appendix 1), which is not recommended for latent constructs. Taken together, the 3-factor model offers the best balance between empirical support and theoretical interpretability.

Table 3. Fit measures for the comparison between the 3 models. The 3-factor solution is the primary model; the 2- and 4-factor solutions are provided for comparison.

Model factors	Variance (%)	RMSEA[a] (90% CI)	TLI[b]	BIC[c]	Chi-square (*df*)	*P* value
2	57.5	0.064 (<0.001 to 0.111)	0.963	−44.4	21.6 (13)	.06
3	61.9	0.058 (<0.001 to 0.1)	0.961	−63.4	27.9 (18)	.06
4	58.4	0.055 (<0.001 to 0.092)	0.948	−86	35.8 (24)	.06

[a]RMSEA: root-mean-square error of approximation.

[b]TLI: Tucker-Lewis Index.

[c]BIC: Bayesian information criterion.

Internal Consistency

Table 4 presents the descriptive statistics and reliability coefficients for the evaluated scales. The overall mean for the CuSAERS was 2.93 (SD 0.503), with moderate internal consistency coefficients for both Cronbach α=0.743 and McDonald ω=0.787. Among the subscales, perception of learning showed the highest internal consistency (α=0.882; ω=0.884) and a mean of 3.06 (SD 0.662), indicating a positive perception of learning.

Table 4. Descriptive statistics and reliability of the CuSAERS (Questionnaire of Satisfaction With Educational Activities Performed on Social Media) questionnaire and its subscales.

Scale	Mean (SD)	Cronbach α	McDonald ω
CuSAERS total	2.93 (0.503)	0.743	0.787
Perception of learning	3.06 (0.662)	0.882	0.884
Task satisfaction	2.99 (0.709)	0.749	0.756
Self-realization	2.56 (0.977)	0.849	0.849

Concurrent Validity

Concurrent validity was assessed using Spearman correlations between the CuSAERS scores and the ASS, as well as prior experience with social media (Table 5). The results showed that the total CuSAERS score had a moderate positive correlation with the ASS (ρ=0.59; *P*<.001), supporting its validity as an instrument to evaluate the perception of educational activities on social media. Additionally, significant positive correlations were observed between the ASS and the subscales perception of learning (ρ=0.44; *P*<.001), task satisfaction (ρ=0.31; *P*<.001), and self-realization (ρ=0.29; *P*<.01), indicating a consistent relationship between academic satisfaction and the dimensions evaluated by the CuSAERS.

Table 5. Correlation matrix.

Variable	ASS[a]	Social media experience	CuSAERS[b] total	Perception of learning	Task satisfaction	Self-realization
ASS	—[c]					
Social media experience	0.17[d]	—				
CuSAERS total	0.59[e]	0.27[e]	—			
Perception of learning	0.44[e]	0.20[f]	0.72[e]	—		
Task satisfaction	0.31[e]	0.14	0.54[e]	0.09	—	
Self-realization	0.29[f]	0.23[f]	0.56[f]	0.10	0.19[d]	—

[a]ASS: Academic Satisfaction Scale.

[b]CuSAERS: Questionnaire of Satisfaction With Educational Activities Performed on Social Media.

[c]Not applicable.

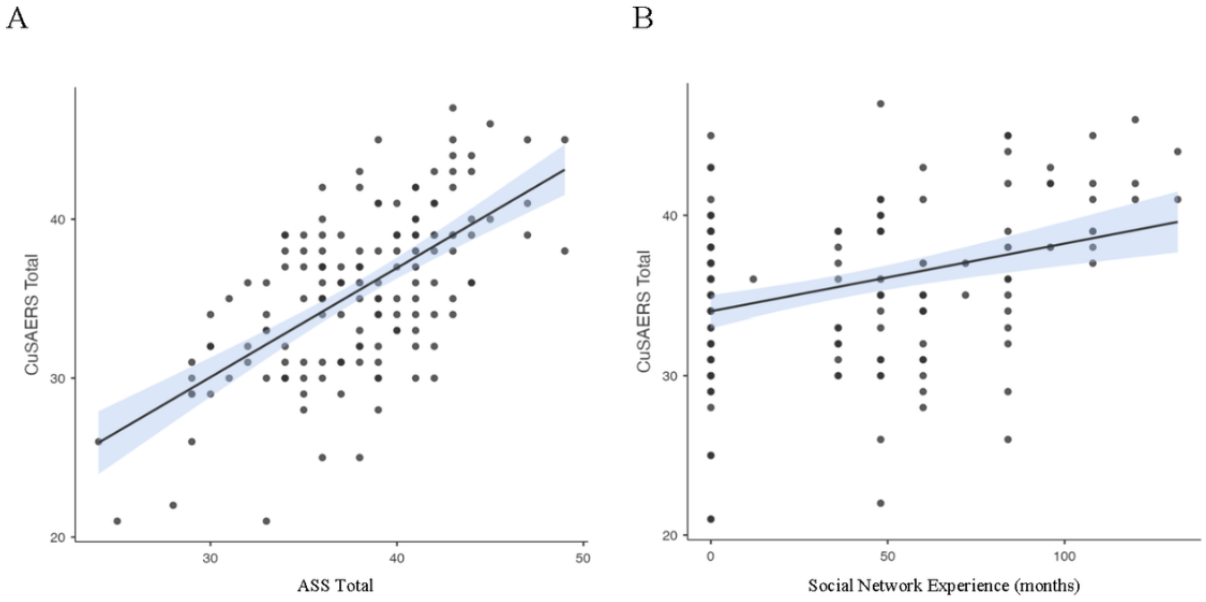
[d]*P*<.05.

[e]*P*<.001.

[f]*P*<.01.

On the other hand, prior experience with social media showed a weak positive correlation with the CuSAERS total score (ρ=0.27, *P*<.001), as well as with perception of learning (ρ=0.20; *P*<.01) and self-realization (ρ=0.23; *P*<.01). However, no significant correlations were observed with the task satisfaction subscale (ρ=0.14; *P*>.05). These results suggest that, although prior experience with social media has a positive relationship with overall perception and some specific aspects evaluated by the CuSAERS, its influence on dimensions such as motivation is limited (Figure 2).

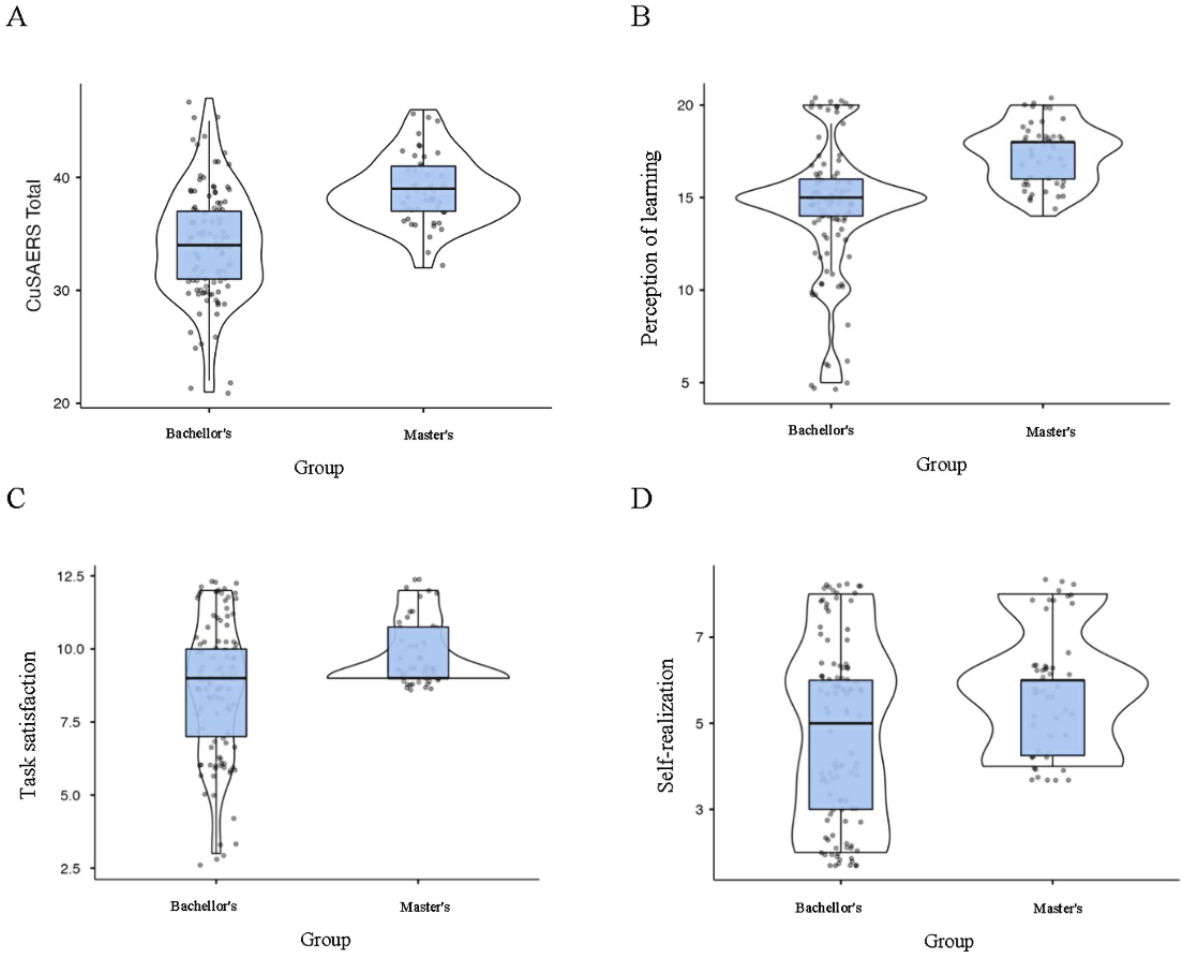
Figure 2. Correlations between the total CuSAERS (Questionnaire of Satisfaction With Educational Activities Performed on Social Media) score and external variables, showing (A) the relationship with Academic Satisfaction Scale (ASS) and (B) the relationship with experience with social media (measured in months).



Discriminant Validity

The discriminant validity of the CuSAERS was evaluated by comparing groups of students based on their academic level (bachelor's or master's). These comparisons were conducted using the nonparametric Mann-Whitney *U* test, and the effect size was calculated using rank-biserial correlation. The statistical results and graphical inspection through box and violin plots reveal clear differences between the 2 groups, supporting the CuSAERS's ability to discriminate between students with different educational levels (Figure 3).

Figure 3. Comparative box and violin plots between bachelor's and master's groups in the CuSAERS (Questionnaire of Satisfaction With Educational Activities Performed on Social Media) subscales and total score, showing (A) total CuSAERS scores, (B) the perception of learning subscale, (C) the task satisfaction subscale, and (D) comparison of the self-realization subscale.



For the total CuSAERS score (Figure 3A), master's students had significantly higher scores (U=1113; P<.001), with a median of 39 (IQR 37-41), compared with the bachelor's group, whose median was 34 (IQR 31-37). Box and violin plots showed greater dispersion in the bachelor's group, whereas scores for the master's group were more compact and centered on higher values.

The perception of learning subscale (Figure 3B) also showed significant differences (U=1064; r=0.613; P<.001), with higher scores and less dispersed distributions in the master's group (median 18, IQR 16-18). The plots reveal greater variability in the bachelor's group, with individual cases at considerably low levels.

For the task satisfaction (Figure 3C) and self-realization (Figure 3D) subscales, significant differences were found (U=2006; r=0.271; P=.005). For task satisfaction, the master's group showed higher scores with less dispersion (median 9, IQR 9-10.8), whereas for self-realization, the master's group's scores were also higher (median 6, IQR 4.25-6), in contrast to the greater dispersion and lower values observed in the bachelor's group (Table 6).

Table 6. Comparison of CuSAERS scores by academic level.

Variable	Bachelor's degree (physiotherapy; n=110)			Master's degree (physiotherapy; n=50)			Mann-Whitney *U* test	*P* value	Effect size
	Mean (SD)	Median (IQR)	SEM[a]	Mean (SD)	Median (IQR)	SEM			
CuSAERS Total	34.10 (4.96)	34 (31-37)	0.47	38.96 (3)	39 (37-41)	0.42	1113	<.001	0.595
Perception of learning	14.86 (3.46)	15 (14-16)	0.33	17.34 (1.62)	18 (16-18)	0.23	1064	<.001	0.613
Task satisfaction	8.62 (2.37)	9 (7-10)	0.22	9.78 (0.15)	9 (9-10.8)	0.15	2006	.005	0.271
Self-realization	4.81 (2.08)	5 (3-6)	0.19	5.80 (1.44)	6 (4.25-6)	0.20	2006	.005	0.271

[a]SEM: standard error of means.

The results also indicated differences between students with and without prior experience in social media. The group with experience obtained higher scores across all subscales and in the total CuSAERS score, although the differences were only significant for the total score (U=2514; r=0.2045; P=.03). For the perception of learning and self-realization, although statistical significance was not reached (P>.05), trends toward higher scores were observed in the group with experience (Table 7).

Table 7. Comparison of CuSAERS (Questionnaire of Satisfaction With Educational Activities Performed on Social Media) scores by participants' experience with social media.

Variable	Experience with social media (n=89)			No experience with social media (n=71)			Mann-Whitney *U* test	*P* value	Effect size
	Mean (SD)	Median (IQR)	SEM[a]	Mean (SD)	Median (IQR)	SEM			
CuSAERS Total	36.42 (4.96)	37 (33-39)	0.53	34.55 (4.77)	35 (31-38)	0.57	2514	.03	0.205
Perception of learning	15.65 (2.97)	15 (15-17)	0.31	14.85 (3.67)	15 (14-17.5)	0.44	2794	.20	0.116
Task satisfaction	9.06 (2.31)	9 (8-11)	0.25	8.89 (1.89)	9 (8-10)	0.22	2920	.40	0.076
Self-realization	5.35 (2.09)	6 (4-7)	0.22	4.83 (1.74)	4 (4-6)	0.21	2639	.07	0.165

[a]SEM: standard error of means.

Results for the motivation facet—considered at the content-validation stage but not retained by the final EFA—are provided in Table S3 in Multimedia Appendix 1 for completeness. These results suggest that familiarity with the use of social media might positively influence overall perceptions of educational activities, although its impact on specific subdimensions is more limited.

Discussion

Overview

The present findings support CuSAERS as a preliminary instrument for assessing student satisfaction in social media–mediated learning environments. In its 3-factor configuration—perception of learning, task satisfaction/environment, and self-realization—the scale shows acceptable reliability at the total-score level and coherent relations with external criteria, while some facets remain candidates for refinement. Construct validity, analyzed through EFA, supported a 3-factor solution that aligns with prior conceptualizations of satisfaction in digital educational settings [18,19] and with the multidimensional notion of academic satisfaction [30-32].

Among the 3 identified factors, perception of learning and self-realization showed the strongest relations with external criteria, consistent with previous findings highlighting the importance of acquiring new skills and the sense of achievement in digital learning experiences [5,7]. These components moderately correlated with the ASS, supporting the concurrent validity of the CuSAERS. In line with research linking satisfaction to performance and retention, students who perceived greater learning and personal growth also tended to report higher overall academic satisfaction [33,34]. Socioemotional elements and social support are key in online contexts; recent work shows that emotional support and effective communication through social media increase student satisfaction and motivation [35].

Psychometrically, internal consistency was acceptable overall. The perception of learning dimension exhibited the highest internal consistency, reflecting its stability. Although motivation emerged at the content-validation stage as a relevant facet, the present exploratory analyses did not retain a stable motivation factor. This facet remains a target for item redevelopment and confirmatory evaluation in future studies. Notably, recent studies suggest that visually rich platforms (eg, Instagram) can enhance motivation, whereas text-centric platforms (eg, X) may require higher digital literacy for comparable satisfaction [36,37]. These

considerations are congruent with standard scale-development practice, where less stable content is refined after initial validation [26].

Discriminant analyses revealed significant differences between bachelor's and master's students, with higher scores among the latter on most subscales. Explanations may include greater academic and professional experience and stronger self-management skills among master's students [3,23].

By contrast, bachelor's students engaged in an X-based debate on chronic pain, a context in which lower academic maturity and less experience with educational uses of social media may temper satisfaction [2,9,15,38-41]. Importantly, the bachelor group's scores were moderately positive, indicating overall acceptance of social media–based activities, while calling for replication in other health disciplines.

Regarding platform differences (X vs Instagram), platform use was not randomized and tasks differed in nature; therefore, any platform contrasts are descriptive and noncausal. Prior literature suggests that Instagram's audiovisual format may favor motivation and participation [8,42,43], whereas X is powerful for debate but may demand greater digital literacy [11,16].

The moderate correlation between CuSAERS scores and ASS scores further supports the instrument's concurrent validity, indicating that satisfaction with social media activities is related to overall academic satisfaction [27,28]. Prior experience with social media showed weak-to-moderate associations with satisfaction, suggesting that technological familiarity helps but does not solely determine satisfaction; interaction quality, content relevance, social presence, and the teacher's facilitating role also matter [13,14,19,21,36,44].

Taken together, the findings support CuSAERS as a preliminary instrument with acceptable reliability at the total-scale level and a provisional self-realization subscale. The scale can inform formative pedagogical reflections, whereas policy or high-stakes decisions should await confirmatory replication and measurement-invariance testing.

Limitations and Future Perspectives

This study used a nonprobabilistic, single-institution sample, which limits generalizability [45]. Future work should include larger, multisite samples spanning diverse programs and contexts [46,47]. Platform use was not randomized: students in the master's degree used Instagram and students in the bachelor's degree used X, and activities differed, so platform effects are confounded by academic level and task; future studies will randomize platform assignment or use within-subject counterbalanced designs with covariate-adjusted models.

Next steps will subject the 3-factor structure to CFA, test configural, metric, and scalar measurement invariance (degree level, gender, age), and conduct cross-validation in an independent, multi-institutional sample [48]. The motivation facet will undergo item redevelopment prior to confirmatory testing. Beyond psychometrics, it will be useful to relate CuSAERS to outcomes such as academic performance, motivation types, peer and teacher interaction, and participation in learning communities [49].

Conclusions

This study provides initial evidence that the CuSAERS is a promising, early-stage instrument for assessing student satisfaction with learning activities delivered over social media. In its 3-factor configuration—perception of learning, task satisfaction/environment, and self-realization—the scale showed acceptable reliability at the total-score level and coherent associations with an established academic satisfaction measure. Group contrasts by academic level were observed, but platform-related differences should be interpreted cautiously given the nonrandomized design.

CuSAERS should be regarded as an early-stage measure with promising properties. Its routine use will benefit from CFA, measurement-invariance testing, and multisite replication, alongside ongoing item refinement. Within the broader trend toward digitization in higher education, tools such as CuSAERS can inform formative pedagogical decisions and help tailor social media–based activities to student needs, while definitive applications should await confirmatory evidence.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Suuplementary tables containing descriptive statistics and reliability coefficients upon item removal, two-factor EFA solution, and four-factor EFA solution.
[DOCX File , 26 KB - mededu_v11i1e73805_app1.docx]

References

1. Aichner T, Grünfelder M, Maurer O, Jegeni D. Twenty-five years of social media: a review of social media applications and definitions from 1994 to 2019. Cyberpsychol Behav Soc Netw 2021;24(4):215-222 [FREE Full text] [doi: 10.1089/cyber.2020.0134] [Medline: 33847527]
2. Wong SY, Tee WJ. The effectiveness and impact of social media approach on students' learning performances. In: Redesigning Learning for Greater Social Impact. Singapore: Springer; 2018:355-363.
3. Ansari JAN, Khan NA. Exploring the role of social media in collaborative learning the new domain of learning. Smart Learn Environ 2020;7(1):1-16. [doi: 10.1186/s40561-020-00118-7]

4. Abendaño ART, Quimada RT, Coloquit LMP. The effectiveness and utilization of social media as academic medium in the UNC College of Education. Int J Res Educ 2022;2(2):142-154. [doi: 10.26877/IJRE.V2I2.12053]
5. Gikas J, Grant MM. Internet High Educ 2013;19:18-26. [doi: 10.1016/j.iheduc.2013.06.002]
6. Dahlstrom E. ECAR Study of Undergraduate Students and Information Technology, 2012 (Research Report).: EDUCAUSE Center for Applied Research; 2012. URL: https://library.educause.edu/~/media/files/library/2012/9/ers1208.pdf?la=en [accessed 2025-11-30]
7. Cavus N, Ibrahim D. m‐Learning: an experiment in using SMS to support learning new English language words. Brit J Educ Tech 2009;40(1):78-91. [doi: 10.1111/j.1467-8535.2007.00801.x]
8. Otsuka E, Wallac S, Chiu D. Design and evaluation of a Twitter hashtag recommendation system. 2014 Presented at: ACM International Conference on Web Search and Data Mining; February 24-28, 2014; New York p. 330-333.
9. Bolderston A, Meeking K, Snaith B, Watson J, Westerink A, Woznitza N. Five years of #MedRadJClub: an impact evaluation of an established Twitter journal club. J Med Radiat Sci 2022;69(2):165-173 [FREE Full text] [doi: 10.1002/jmrs.569] [Medline: 35143706]
10. Lozano Díaz A, González Moreno MJ, Cuenca Piqueras C. Youtube como recurso didáctico en la Universidad. Edmetic 2020;9(2):159-180. [doi: 10.21071/edmetic.v9i2.12051]
11. Moran M, Seaman J, Tinti-Kane H. Teaching, Learning, and Sharing: How Today's Higher Education Faculty Use Social Media. Boston: Pearson Learning Solutions; 2011.
12. Toland R. Facebook as a learning tool. Perspec Learn 2013;14(1):1-5 [FREE Full text]
13. Cao Y, Ajjan H, Hong P. Using social media applications for educational outcomes in college teaching: a structural equation analysis. Brit J Educ Technol 2013;44(4):581-593. [doi: 10.1111/bjet.12066]
14. Madden M, Zickuhr K. 65% of online adults use social networking sites. Pew Internet & American Life Project. URL: https://www.pewresearch.org/internet/2011/08/26/65-of-online-adults-use-social-networking-sites/ [accessed 2025-11-30]
15. Rutherford C. Using online social media to support preservice student engagement. MERLOT Journal of Online Learning and Teaching. 2010. URL: https://jolt.merlot.org/vol6no4/rutherford_1210.pdf? [accessed 2025-11-30]
16. Voorn RJ, Kommers PA. Social media and higher education: introversion and collaborative learning from the student's perspective. Int J Social Media Interact Learn Environ 2013;1(1):59. [doi: 10.1504/ijsmile.2013.051650]
17. Zhu C. Student satisfaction, performance, and knowledge construction in online collaborative learning. Educ Technol Soc 2012;15(1):127-136 [FREE Full text]
18. Zhang Y, Lin C. Effects of community of inquiry, learning presence and mentor presence on K‐12 online learning outcomes. J Comput Assist Learn 2021;37(3):782-796. [doi: 10.1111/jcal.12523]
19. Richardson JC, Maeda Y, Lv J, Caskurlu S. Social presence in relation to students' satisfaction and learning in the online environment: a meta-analysis. Comput Human Behav 2017;71:402-417. [doi: 10.1016/j.chb.2017.02.001]
20. Gunawardena CN, Zittle FJ. Social presence as a predictor of satisfaction within a computer‐mediated conferencing environment. Am J Distance Educ 1997;11(3):8-26. [doi: 10.1080/08923649709526970]
21. Cobb SC. Social presence, satisfaction, and perceived learning of RN-to-BSN students in web-based nursing courses. Nurs Educ Perspect 2011;32(2):115-119. [doi: 10.5480/1536-5026-32.2.115] [Medline: 21667794]
22. Armah JK, Bervell B, Bonsu NO. Modelling the role of learner presence within the community of inquiry framework to determine online course satisfaction in distance education. Heliyon 2023;9(5):e15803 [FREE Full text] [doi: 10.1016/j.heliyon.2023.e15803] [Medline: 37180887]
23. Tolsgaard MG, Kulasegaram KM, Ringsted CV. Collaborative learning of clinical skills in health professions education: the why, how, when and for whom. Med Educ 2016;50(1):69-78. [doi: 10.1111/medu.12814] [Medline: 26695467]
24. Giroux CM, Moreau KA. A qualitative exploration of the teaching- and learning-related content nursing students share to social media. Can J Nurs Res 2022;54(3):304-312 [FREE Full text] [doi: 10.1177/08445621211053113] [Medline: 34755574]
25. La Touche R, Paris Alemany A, Grande Alonso M. Construcción y validación de contenido del cuestionario de satisfacción de actividades educativas realizadas en redes sociales (CuSAERS). In: Serrano Villalobos O, Velasco Furlong L, Arcos Rodríguez R, editors. Avances para la innovación docente en salud y comunicación. 1st Edition. Madrid: Dykinson; 2023:978-984.
26. Costello AB, Osborne JW. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. Pract Assess Res Eval 2005;10(7):1-9 [FREE Full text]
27. Medrano LA, Fernández Liporace M, Pérez E. Computerized assessment system for academic satisfaction (ASAS) for first-year university student. Electron J Res Educ Psychol 2017;12(2):541-562. [doi: 10.25115/ejrep.33.13131]
28. Vergara-Morales J, Del Valle M, Diaz A, Perez MV. Adaptation of the Academic Satisfaction Scale in Chilean university students. Psicologia Educativa 2018;24(2):99-106. [doi: 10.5093/psed2018a15]
29. Schober P, Boer C, Schwarte LA. Correlation coefficients: appropriate use and interpretation. Anesth Analg 2018;126(5):1763-1768. [doi: 10.1213/ANE.0000000000002864] [Medline: 29481436]
30. Hanssen TES, Solvoll G. The importance of university facilities for student satisfaction at a Norwegian University. Facilities 2015;33:744-759. [doi: 10.1108/F-11-2014-0081]
31. Jereb E, Jerebic J, Urh M. Revising the importance of factors pertaining to student satisfaction in higher education. Organizacija 2018;51(4):271-285. [doi: 10.2478/orga-2018-0020]

32. Nastasić A, Banjević K, Gardašević D. Student satisfaction as a performance indicator of higher education institution. Mednarodno Inovativno Poslovanje = Journal of Innovative Business and Management 2019;11(2):67-76. [doi: 10.32015/jibm/2019-11-2-8]
33. Duque LC. A framework for analysing higher education performance: students' satisfaction, perceived learning outcomes, and dropout intentions. Total Qual Manage Bus Excell 2013;25(1-2):1-21. [doi: 10.1080/14783363.2013.807677]
34. Mihanović Z, Batinić A, Pavičić J. The link between students' satisfaction with faculty, overall students' satisfaction with student life and student performances. Rev Innov Compet 2016;2(1):37-60. [doi: 10.32728/ric.2016.21/3]
35. Zalazar-Jaime MF, Moretti LS, García-Batista ZE, Medrano LA. Evaluation of an academic satisfaction model in E-learning education contexts. Interact Learn Environ 2023;31(7):4687-4697. [doi: 10.1080/10494820.2021.1979047]
36. Peña-Acuña B, Alfonso-Jaramillo JF. Instagram and YouTube, visual culture and university education: a systematic review. VISUAL REVIEW: Int Visual Cult Rev 2024;16(2):53-66. [doi: 10.62161/revvisual.v16.5206]
37. Callaghan N, Bower M. Learning through social networking sites – the critical role of the teacher. Educ Media Int 2012;49(1):1-17. [doi: 10.1080/09523987.2012.662621]
38. Joseph MA, Natarajan J, Seshan V, Roach EJ, Omari OA, Karkada S. Effects of Twitter use on academic performance and satisfaction in a pathophysiology course among Omani nursing students: a quasi-experimental study. BMC Nurs 2023;22(1):439 [FREE Full text] [doi: 10.1186/s12912-023-01609-x] [Medline: 37990319]
39. Sinclair W, McLoughlin M, Warne T. To Twitter to woo: harnessing the power of social media (SoMe) in nurse education to enhance the student's experience. Nurse Educ Pract 2015;15(6):507-511. [doi: 10.1016/j.nepr.2015.06.002] [Medline: 26119057]
40. Gao F, Li L. Predicting educators' use of Twitter for professional learning and development. Educ Inf Technol 2019;24(4):2311-2327. [doi: 10.1007/s10639-019-09872-9]
41. Scott BT, Harmeyer D, Wu SF. Utilizing Twitter and #hashtags toward enhancing student learning in an online course environment. Int J Distance Educ Technol 2014;12(3):75-83. [doi: 10.4018/ijdet.2014070106]
42. Obeso M, Pérez-Pérez M, García-Piqueres G, Serrano-Bedia A. Enhancing students' learning outcomes through smartphones: a case study of using Instagram in higher management education. Int J Manag Educ 2023;21(3):100885. [doi: 10.1016/j.ijme.2023.100885]
43. Morais S, Pereira T, Raposo R, Gouveia T. Uses, perceptions and impacts of Instagram: a study with young higher education students. Proceedings of the European Conference on Social Media 2024;11(1):206-215. [doi: 10.34190/ECSM.11.1.2211]
44. Saura J, Palacios-Marqués D, Iturricha-Fernández A. Ethical design in social media: assessing the main performance measurements of user online behavior modification. J Bus Res 2021;129:271-281. [doi: 10.1016/j.jbusres.2021.03.001]
45. MacCallum RC, Widaman KF, Zhang S, Hong S. Sample size in factor analysis. Psychol Methods 1999;4(1):84-99. [doi: 10.1037/1082-989X.4.1.84]
46. Brown TA. Confirmatory Factor Analysis for Applied Research. 2nd Edition. New York: The Guilford Press; 2015.
47. Byrne BM. Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming. 2nd Edition. New York: Routledge/Taylor & Francis Group; 2010.
48. Kline RB. Principles and Practice of Structural Equation Modeling. 4th Edition. New York: The Guilford Press; 2016.
49. Deci EL, Ryan RM. The "what" and "why" of goal pursuits: human needs and the self-determination of behavior. Psychol Inq 2000;11(4):227-268. [doi: 10.1207/s15327965pli1104_01]

Abbreviations

ASS: Academic Satisfaction Scale
BIC: Bayesian information criterion
CFA: confirmatory factor analysis
CuSAERS: Questionnaire of Satisfaction With Educational Activities Performed on Social Media
EFA: exploratory factor analysis
KMO: Kaiser-Meyer-Olkin
Q–Q: quantile–quantile
RMSEA: root-mean-square error of approximation
TLI: Tucker-Lewis Index

Edited by S Tsuei; submitted 12.03.25; peer-reviewed by S Santilli, Y Assefa, F Cuenca Martinez; comments to author 31.07.25; revised version received 27.10.25; accepted 27.10.25; published 19.12.25.

Please cite as:
La Touche R, Reina-Varona Á, Grande-Alonso M, León-Hernández JV, Pardo-Montero J, Requejo-Salinas N, Ferrer-Peña R, Paris-Alemany A
Student Satisfaction in Social Media–Based Learning Environments: Development, Validation, and Psychometric Evaluation of the CuSAERS (Questionnaire of Satisfaction With Educational Activities Performed on Social Media)
JMIR Med Educ 2025;11:e73805
URL: https://mededu.jmir.org/2025/1/e73805
doi:10.2196/73805
PMID:

Original Paper

Factors Influencing Educators' Perspectives on Accepting Extended Reality in Health Care Education: Qualitative Study

Zuheir Khlaif[1], PhD; Nisreen Salama[2], PhD; Bilal Hamamra[1], PhD; Allam Mousa[3], PhD

[1]Faculty of Humanities and Educational Sciences, An-Najah National University, Nablus, Occupied Palestinian Territory

[2]Faculty of Nursing, Arab American University, Jenin, Occupied Palestinian Territory

[3]Faculty of Engineering, An Najah National University, Nablus, Occupied Palestinian Territory

Corresponding Author:
Zuheir Khlaif, PhD
Faculty of Humanities and Educational Sciences
An-Najah National University
Old Campus Street
Nablus, PS358
Occupied Palestinian Territory
Phone: 970 0592754908
Email: zkhlaif@najah.edu

Abstract

Background: Palestinian higher education institutions face limitations in providing interactive practical training for medical education. Extended reality (XR), which encompasses virtual reality and augmented reality, is increasingly recognized for addressing these challenges by offering immersive learning experiences.

Objective: This study investigates the factors influencing the acceptance and adoption of XR in health care education within Palestinian universities, exploring its potential to transform traditional teaching methods.

Methods: A qualitative approach was used in this study to collect data through semistructured interviews and artifacts from the participants. The participants of the study were 25 faculty members from 2 large Palestinian universities who teach in the field of medical sciences.

Results: Three primary categories—external, internal, and design-related factors—emerged as pivotal in influencing XR adoption. Professional development, technical support, and infrastructure were key external enablers. Internally, prior experience with digital tools and positive attitudes had a significant impact on the adoption of XR. Design factors, including ease of use and interactivity, played a crucial role but also posed challenges for less tech-savvy educators. Despite barriers such as cost and technical issues, XR demonstrated notable benefits, including enhanced learning outcomes, improved knowledge retention, and the ability to simulate complex medical scenarios.

Conclusions: XR technologies offer transformative potential for health care education in Palestine. By addressing challenges and leveraging XR's strengths, educational institutions can foster innovation and improve student engagement and skill acquisition. The study contributes to the theoretical understanding of technology acceptance in education by identifying the interplay of external, internal, and design factors. Practically, it emphasizes strategic investments in infrastructure, professional training, and institutional policies to optimize XR integration.

(JMIR Med Educ 2025;11:e65042) doi:10.2196/65042

KEYWORDS

extended reality (XR); health care education; educational technology; Sustainable Development Goals (SDGs); Palestine

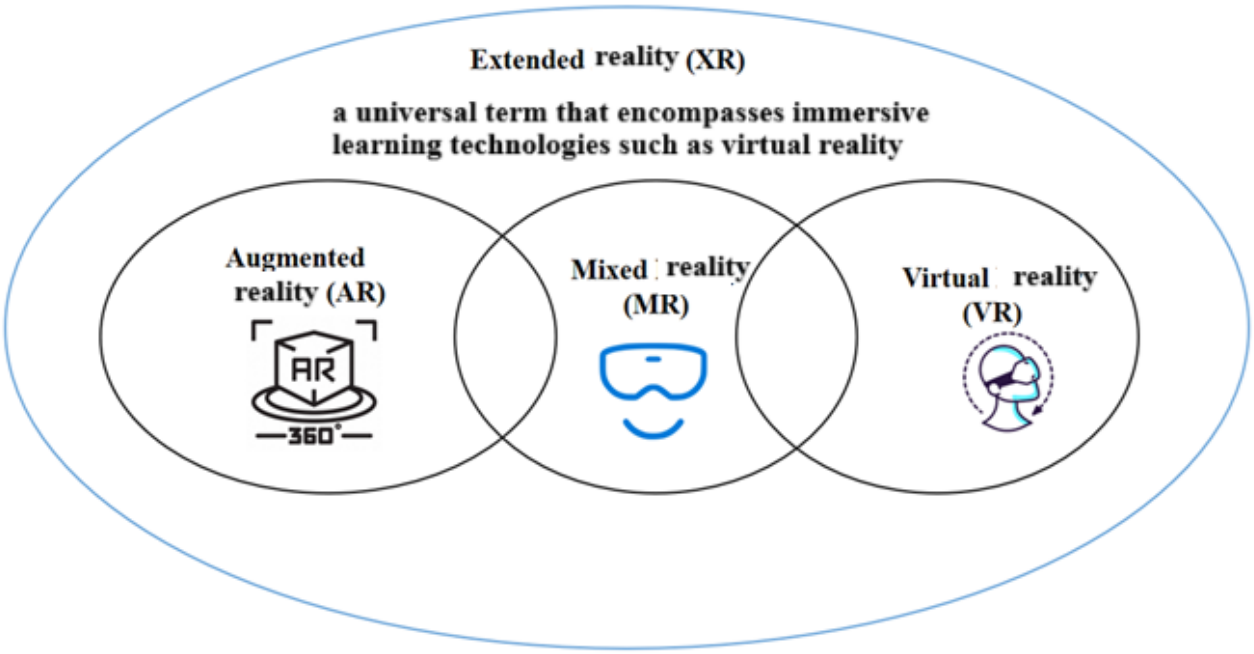
Introduction

Background

The integration of emerging technologies such as artificial intelligence and XR has significantly enhanced learners' engagement and interaction with educational environments [1]. XR, an umbrella term encompassing virtual reality (VR), augmented reality (AR), and mixed reality (MR), facilitates human-machine interaction through computer-generated content [2]. VR provides a fully immersive digital environment, AR overlays virtual objects onto real-world settings, and MR blends physical and digital elements into a seamless interactive

experience [2]. By incorporating visual, auditory, and interactive elements, XR enhances real-world learning by immersing students in digital environments that facilitate deeper engagement with educational content [3]. As shown in Figure 1, XR encompasses AR, MR, and VR, creating a spectrum of immersive experiences. Figure 1, developed by the authors, shows how XR encompasses AR, MR, and VR.

Figure 1. Illustration showing how extended reality (XR) encompasses augmented reality (AR), mixed reality (MR), and virtual reality (VR), as developed by the authors.



XR technologies are transforming medical education, particularly in resource-constrained settings, by providing cost-effective, risk-free learning environments. Head-mounted devices enable students to practice medical skills repeatedly without compromising patient safety. A review of 27 studies found XR-based training to be highly effective in surgery and anatomy, although more large-scale research is needed to fully assess its impact [4]. While some studies, such as Behmadi et al [5], found no statistically significant improvements in knowledge retention when comparing XR-based learning with traditional methods, they reported enhanced student engagement and satisfaction. This suggests that XR should be integrated alongside conventional teaching methods to cater to diverse learning styles and develop critical decision-making skills. Additionally, XR reduces dependence on costly medical equipment and provides students with virtual laboratories to conduct experiments, mitigating financial and logistical barriers [3,6]. By enabling hands-on practice in safe, controlled environments, XR improves learning accessibility and instructional quality [6].

In Palestinian higher education, XR adoption remains in its early stages, particularly in health care disciplines such as medicine and nursing. Its integration is closely tied to broader digital transformation policies, which aim to modernize educational practices and align with global trends in technology-enhanced learning. However, the transition faces barriers such as limited resources, inadequate infrastructure, and the need for faculty capacity-building. Despite these challenges, Palestinian universities are increasingly prioritizing XR within their strategic frameworks, reflecting a commitment to leveraging emerging technologies to enhance learning outcomes.

This study contributes to the growing research on XR adoption by examining Palestinian higher education institutions, a unique and underrepresented context. Unlike studies from technologically advanced regions, where XR adoption is often supported by substantial funding and infrastructure, this research highlights the policy-related, infrastructural, and faculty-specific challenges of implementing XR in resource-limited settings. While previous studies in Latin America and Southeast Asia have explored XR's role in medical education, they have often overlooked institutional policies, faculty readiness, and regional constraints. By addressing these factors, this study provides a comparative perspective on XR adoption in Palestinian universities and offers global insights into key determinants for sustainable technology integration in medical education.

Literature Review

The Unified Theory of Acceptance and Use of Technology Model in XR Adoption for Health Care Education

Overview

VR and AR are highly adaptable technologies that use various systems, setups, and content types, ranging from immersive and dynamic to nonimmersive and static environments. These technologies are characterized by 3 key elements: immersion, presence, and interaction [7,8]. Immersion depends on the

technological medium, such as head-mounted displays, concave or 3D projections, or interactive videos where users engage as protagonists. Presence and interaction, by contrast, relate to an individual's perception of connectedness within the virtual environment and their ability to act upon it and receive feedback [7,8]. These elements are crucial in defining the effectiveness and adoption of XR technologies, particularly in educational contexts such as health care training.

The Unified Theory of Acceptance and Use of Technology (UTAUT), developed by Venkatesh et al [9], serves as the theoretical framework for examining the adoption of XR in health care education. Over time, various theories and models have been proposed to understand the factors influencing the acceptance of new technologies. One of the foundational models in this field is the Technology Acceptance Model (TAM), developed by Davis [10] and based on the Theory of Reasoned Action. The TAM has been widely used to investigate the adoption of emerging technologies, including XR in surgical training, artificial intelligence adoption, and mobile learning [11-13]. However, the TAM has been critiqued for its limited predictive accuracy, as it fails to account for technology acceptance in nearly 40% of cases [14].

To address these limitations, Venkatesh et al [9] developed the UTAUT, which integrates multiple previous models, including the Theory of Reasoned Action, to provide a more comprehensive framework for understanding technology adoption. The UTAUT identifies 4 key constructs that shape an individual's intention to use, as well as their actual usage behavior of a new technology: Performance Expectancy, Effort Expectancy, Social Influence, and Facilitating Conditions.

Performance Expectancy

In the context of XR adoption in health care education, performance expectancy refers to the extent to which faculty members perceive XR as beneficial for both educators and students. XR facilitates immersive learning, enhances knowledge retention, and allows for realistic simulations of medical procedures without risks to patients. Faculty members recognize its potential to improve teaching effectiveness, help achieve learning objectives, and develop students' practical skills in a safe environment.

Effort Expectancy

Effort expectancy concerns the perceived ease of use of XR technology in teaching and learning environments. Faculty members' willingness to integrate XR depends on how intuitive and user-friendly the technology is. Instructors with prior experience using digital tools often find XR more accessible, while others may require training and technical support to overcome usability challenges.

Social Influence

The role of social influence in XR adoption is significant, as faculty members' decisions are shaped by the expectations of colleagues, mentors, and students. In this study, social influence extends beyond professional networks to include university policies, peer recommendations, and online technology communities that encourage XR integration.

Facilitating Conditions

The successful adoption of XR technology in health care education depends heavily on facilitating conditions, including institutional support, infrastructure, and technical assistance. Universities must provide adequate resources, professional development programs, and robust digital infrastructure to support faculty members' continued use of XR. Without sufficient technical support and access to XR-compatible hardware, faculty adoption may be hindered.

Virtual Reality in Health Care

VR technology has revolutionized health care education by offering immersive, interactive learning experiences. It creates realistic simulations that allow students to practice procedures and decision-making without risks [15]. For instance, the Stanford Virtual Heart Project uses VR to help students understand cardiac anatomy through 3D models. Furthermore, VR allows for detailed exploration of the human body, aiding in the comprehension of anatomical structures. The HoloAnatomy program at Case Western Reserve University uses Microsoft HoloLens for holographic dissections, enhancing anatomy learning and critical thinking [15].

VR provides simulation-based training for medical procedures, enhancing skills and confidence before actual patient care. Apps such as Surgical Theater aid in understanding and practicing surgical procedures [16]. VR also enables collaborative learning through merged MR and mixed XR platforms, enhancing teamwork and communication skills by combining the virtual and physical worlds [15]. VR makes learning more engaging and accessible by bringing medical simulations to remote or underresourced areas, ensuring that high-quality medical education is available to a broader audience [17].

XR Adoption in Education

Researchers have reported that XR is the future of learning and teaching in higher education institutions worldwide [6]. XR promotes teamwork through collaboration in VR environments [6]. Moreover, using XR in education transforms the learning process from traditional methods to active learning, where learners engage directly in learning activities [18]. In addition, XR enhances personalized learning by creating a customized learning environment tailored to the learner's needs and abilities [19]. Educators can design digital content to simulate immersive and interactive environments, making XR particularly suitable for teaching high-risk procedures, using expensive equipment, and conducting practical training activities that require additional resources, such as those in medicine or medical sciences [20].

Other studies have confirmed that using XR in learning fosters creativity, innovation, and design among learners [21,22]. Aguayo and Eames [23] reported that virtual agents in XR facilitated language learning and teaching by making complex knowledge more accessible. Moreover, these technologies offer immersive and interactive simulations that replicate complex real-world scenarios [24]. Consequently, using XR in education enhances the real-world experience by incorporating sounds, videos, and graphics into the learning environment, enabling learners to interact more effectively [3].

Benefits and Challenges

VR and AR offer numerous benefits over conventional therapy, including cost reduction, fewer hospital visits for immobile patients, user-friendly experiences, and improved patient safety. They also facilitate data collection in research settings and reduce surgical errors in training [25,26]. However, many studies indicate that determining the benefits of VR and AR in health care is challenging due to their small sample sizes, heterogeneous nature, and lack of proper controls. Pain relief mechanisms remain debated, and implementation is technically complex and expensive, with lower acceptance among older adults [7,27].

One of the main advantages of using XR is its ability to provide practical training without the need for expensive physical equipment [28]. For example, medical students can practice surgical procedures in a virtual environment, reducing the need for costly medical equipment and minimizing the risks associated with real-life practice on patients [29]. The implementation of virtual laboratories through XR not only makes education more accessible but also enhances the learning experience by offering hands-on practice in a safe and controlled environment [6].

Factors That Influence the Continued Intention to Use XR in Health Care Sciences

A previous study explored the potential of VR and AR technologies, particularly XR, in enhancing patient care, medical education, and presurgical planning [30]. It highlights the potential of integrating XR into medical education to provide immersive, interactive experiences.

Burian et al [31] highlighted that XR technologies provide more effective training compared with traditional methods, particularly for novices. Chuah [32] conducted a study involving 45 relevant studies to assess and analyze user acceptance of XR technology from multiple theories, disciplines, and perspectives. Wearable XR technology is influenced by various factors, including cost, technical and performance issues, hardware size, sensory inputs, content quality, cognitive impacts, user satisfaction, attitudes, intentions to visit tourism sites, expectation confirmation, personality traits, knowledge transfer, support needs, presence, boundary considerations, consumer characteristics, spatial awareness, control, participation, effectiveness, familiarity, innovativeness, value perceptions, decision comfort, spatial understanding, cognitive load, virtual embodiment perception, sense of presence, health and privacy concerns, and psychological and physical risks. Curran et al [33] stated that XR technologies offer portability, standardization, replicability, accessibility, and the ability to function without heavy manikin parts. They can be widely distributed without the need for a live instructor, thereby increasing learner engagement and enhancing spatial representation.

Kluge et al [34] noted that despite limited experience with XR technology, staff and students at the University of Newcastle view it as a standard tool for teaching. They aim to develop a sustainable implementation framework within 5 years. VR, AR, and MR technologies are disrupting medical education by offering immersive experiences and alleviating traditional learning constraints. XR technologies, particularly in emergency medicine, enable remote clinical skill development, even amidst the challenges posed by COVID-19 [33]. Li and Keskitalo [35] emphasized that XR technology is commonly used in health care education for safe treatments, communication, and decision-making. It supports 5 cognitive-processing dimensions: remembering, understanding, applying, analyzing, and evaluating. AR, VR, and MR can positively impact medical education. To effectively implement XR, it is important to consider existing resources and Bloom's taxonomy, and select the most suitable technology. Optimizing and expanding XR utilization are crucial for promoting deeper learning in health care. There is a significant gap in research regarding the factors influencing the continued use of XR in medical science, emphasizing the need for further studies in this area.

Methods

Study Design

We used a qualitative approach to explore faculty members' lived experiences with using XR in medical education at 2 Palestinian universities. This approach allows researchers to gain insights into the phenomena from practitioners who are using XR in teaching [36]. We gave participants the opportunity to share their experiences with using XR in their courses. The 2 universities involved are The Arab American University-Palestine and An-Najah National University. Both universities have established digital transformation centers focused on immersive technologies to enhance teaching and learning. The Arab American University-Palestine has developed a VR laboratory within its nursing and medicine departments, using cutting-edge technology to improve educational experiences. Similarly, An-Najah National University has created VR laboratories for its Departments of Dental Medicine and Medicine, along with a general XR center that serves both the university and the surrounding community. These laboratories provide training in XR technologies for faculty and students, supporting the integration of immersive tools into the curriculum.

Participants

This study involved 25 faculty members (18 males and 7 females) from diverse disciplines within medical sciences, representing 2 universities in the West Bank of Palestine. All participants had a minimum of 1 year of experience using XR in their teaching of undergraduate medical sciences courses, ensuring their expertise and relevance to the study's focus. Participants contributed by engaging in semistructured interviews and providing examples of their own work and that of their students. Table 1 presents the demographic characteristics of the participants, highlighting their diversity and alignment with the study's objectives.

Table 1. Demographic information about the participants in the study.

Variable and category	Frequency, n
Gender	
Male	18
Female	7
Education level	
Doctorate	19
Master's degree	6
Teaching experience	
5 years or less	5
6-10 years	12
11 years or more	8
Medical sciences field	
Nursing	7
Human medicine	8
Pharmacy	4
Dentistry	6

Recruitment of Participants and Justification of Sample Size

This study used purposive sampling, a qualitative method that selects participants based on predefined criteria relevant to the research objectives [37]. Faculty members, aligned with the study's focus on XR adoption in medical education, were recruited through official invitations sent by the deans of the 2 Palestinian universities.

Eligible participants were required to meet several inclusion criteria: at least one year of experience using XR in teaching, an academic position within the Faculty of Medical Sciences, and a minimum of 3 years of higher education teaching experience. Their expertise was further validated through prior research or publications on XR, active involvement in curriculum development, and proficiency with XR tools, demonstrated through certifications, training, or workshops. Additionally, student evaluations and feedback on XR-based teaching were considered.

To enhance transferability, the study included faculty from the fields of nursing, human medicine, pharmacy, and dentistry, ensuring representation across disciplines with varying levels of reliance on XR-based learning. Participants ranged from early-career faculty with 5 or fewer years of teaching experience to senior faculty with over 10 years, providing a range of professional perspectives. The sample also encompassed both experienced and novice XR users, drawn from universities at different stages of XR adoption—one with an established XR laboratory and structured training programs, and the other in an early adoption phase—offering a comprehensive view of institutional challenges and implementation strategies.

The sample size of 25 faculty members was determined based on data saturation, a key principle in qualitative research. Saturation is reached when further data collection no longer generates new themes or insights [38]. In this study, saturation was achieved by the 22nd interview, as recurring patterns related to institutional barriers, faculty perceptions, and XR adoption challenges emerged. The remaining interviews were conducted to reinforce the robustness of these themes. Prior research [9,36] on technology adoption in higher education suggests that qualitative studies with 15-30 participants provide sufficient depth to capture rich, context-specific insights. Given the exploratory nature of this study, the selected sample size was deemed appropriate for gathering in-depth faculty perspectives on XR integration in medical education.

We acknowledge that the small sample size may limit the generalizability of the findings. However, this study was designed to explore educators' lived experiences and provide deep insights into XR adoption in medical education, rather than aiming for broad generalizability. The qualitative methodology, supported by data triangulation through interviews and artifacts, enhances the trustworthiness and validity of the findings despite the limited sample size. To mitigate this limitation, participants were purposively selected from diverse medical disciplines and institutions to ensure a variety of perspectives. Rigorous thematic analysis techniques were used, validated by multiple coders with high interrater reliability. Additionally, the findings were cross-referenced with existing literature to contextualize and strengthen the conclusions.

While future studies with larger samples are necessary to further validate these findings, this study provides valuable foundational insights into XR adoption in medical education. These findings can serve as a reference for similar educational contexts, aiding institutions and educators in navigating the complexities of XR implementation.

Study Context

The context of this study involved faculty members from the Faculty of Healthcare Sciences at 2 major universities in the West Bank of Palestine. Both universities have a clear vision and policy to integrate emerging technologies, such as XR, into medical education and research. To facilitate the adoption of XR in teaching, 4 training workshops were organized at each university to equip faculty members with the knowledge and skills necessary to understand and apply XR in their practices. The length of each training session varied depending on the topic, generally ranging from 2 to 7 days. These sessions provided teachers with the opportunity to create and develop learning objects using XR. Various topics were covered, including an introduction to VR, AR, and XR; designing lessons using objects on the platform; and creating avatars for lessons, among others.

Research Instrument

A semistructured interview was the primary research instrument for data collection. We developed an interview protocol to guide the process. The protocol (Multimedia Appendix 1) consisted of 2 sections. The first section introduced the study to the participants and confirmed the confidentiality of their responses. Participants were informed that the interview would be recorded, provided they agreed, and that they could withdraw at any time. They were also asked to sign a consent form before the interview was recorded. The second section contained interview questions developed from a literature review, aligned with the research questions. Another data source consisted of artifacts provided by the participants during the interviews or submitted via email, illustrating how XR was used in medical education.

Data Collection

The researchers sent invitations to the nominated participants to schedule semistructured interviews at their convenience. Participants were given the flexibility to choose the time and location of the interviews. Each interview, lasting 30-45 minutes, was conducted individually and recorded. The interview protocol began by asking participants to discuss their general experiences with technology, followed by specific inquiries about their experiences with XR, allowing them to share their stories and journeys. Follow-up questions were asked to delve deeper into their experiences, especially regarding student collaboration on projects. For example, when one participant mentioned that students learn more through exploration, a follow-up question was, “Can you provide more details about what exploring entails and the role of XR in facilitating that exploration?” The interviews were designed to capture as much detailed information as possible from the participants’ experiences. In addition to the interviews, participants provided various artifacts, including samples of student work and activities implemented using XR. These artifacts served as valuable secondary data sources.

Trustworthiness

To ensure trustworthiness and methodological rigor, this study followed established qualitative research principles, emphasizing credibility, confirmability, dependability, and transferability. Credibility was enhanced through the use of multiple data collection methods, including semistructured interviews, participant-submitted artifacts, and institutional policy documents. Member checking was conducted by sharing transcribed interviews and preliminary themes with participants, enabling them to verify the accuracy of interpretations and provide clarifications where necessary. Additionally, peer debriefing was used, where external researchers reviewed the coding process and emerging themes to refine definitions, minimize bias, and ensure analytical coherence.

Conformability was maintained by carefully documenting the research process and prioritizing participants’ statements to minimize researcher bias. The interview protocol was developed based on the research questions, a pilot study, and expert reviews, ensuring its relevance and alignment with the study’s objectives. Dependability was reinforced through rigorous coding procedures, with data being independently coded 3 times, achieving 92% interrater reliability. This process ensured consistency in theme identification and analysis. Transferability was supported by purposive sampling, which selected participants from diverse academic disciplines and career stages to capture a broad range of perspectives on XR adoption. Additionally, data triangulation—incorporating interviews, observational notes, and artifacts from faculty members—provided a comprehensive understanding of the factors influencing XR integration in medical education. By incorporating these robust qualitative validation strategies, the study strengthens its credibility, validity, and applicability, ensuring that the findings accurately reflect participants’ experiences while addressing potential biases in sample selection and data interpretation.

Ethical Considerations

This research was approved by the institutional review board committee at An-Najah National University under reference Med. April.2024/18. The study adhered to strict ethical guidelines to ensure participant confidentiality, informed consent, and data protection. Before participation, all participants were provided with detailed information about the study’s objectives, procedures, and their rights as participants. Participants were explicitly informed that their participation was entirely voluntary, and they could withdraw from the study at any time without facing any consequences or needing to provide an explanation. To document consent, participants signed an informed consent form that outlined the scope of the study, the types of data to be collected, and how the data would be used solely for research purposes. The consent form also detailed the measures taken to protect their identities and ensure the confidentiality of their responses.

In terms of data protection, stringent protocols were followed to safeguard participant information. Personal data, including contact details, were securely stored on a password-protected and encrypted computer. Both physical and digital access to these data was restricted to authorized researchers only. Furthermore, any identifying information was anonymized during the data analysis process to further protect participants’ privacy. Interview data were stored in encrypted files, and backup copies were securely maintained to prevent data loss.

Additionally, the researchers communicated the steps taken to comply with ethical research standards, including adherence to the principles outlined by Ngozwana [39]. These procedures ensured that participants felt confident their contributions were protected and respected throughout the research process. By incorporating these comprehensive protocols, the study's transparency and ethical rigor were enhanced.

Data Analysis

To address the research questions, inductive thematic analysis was used, following the 6-step methodology proposed by Braun and Clarke [38]. The research process involved conducting 10.5 hours of recorded interviews. Before data analysis, the researchers transcribed the audio recordings and shared the text files with participants for validation. Participants were given the opportunity to amend, rewrite, or supplement the content as needed. Once the files were returned, no further modifications were expected. The thematic analysis was carried out in 6 phases, as outlined by Braun and Clarke [38], and these phases guided the data analysis process in this study.

The researchers began their analysis by organizing the individual transcript files, labeling them as F1, F2, and so on up to F25. The initial phase involved familiarizing themselves with the data. The researchers, who had conducted the interviews, carefully read through each transcript while simultaneously listening to the corresponding audio files. This process of immersion allowed the researchers to take detailed notes in the margins of the transcripts, ensuring a thorough understanding of the data. In the second phase, coding, the researchers developed labels for significant features of the data that were relevant to the research questions. As they meticulously read through the transcripts line-by-line, they created a coding book to document these labels. This phase focused on identifying key concepts or ideas that were central to the research, such as how faculty members experience XR in teaching medical concepts and in transferring medical knowledge and skills to their students. Next, the researchers moved to the theme development phase, where they identified patterns related to the research questions. They grouped similar and contrasting codes into coherent themes, resulting in a structured coding book (Table 2 and Multimedia Appendix 2). In the fourth phase, reviewing themes, the researchers examined these themes across all data, using inductive analysis to refine and combine them as needed. This iterative process of theme development was essential for generating meaningful insights from the data.

The fifth phase involved defining and naming themes based on the research questions, literature review, and theoretical framework. This phase integrated both inductive and deductive approaches, which are common in qualitative data analysis. Finally, in the reporting phase, the researchers documented and presented the analysis results, supporting the findings with direct quotations from the participants' interviews. This approach helped provide concrete evidence and context for the identified themes.

The data analysis process began with a thorough review of each transcript to familiarize the researchers with the material. During this phase, ideas and concepts were identified as units of analysis, and a coding process was developed. A coding book was created to assist with data coding. Ideas and concepts sharing similar characteristics were organized into emerging subthemes, which were then grouped under primary themes, informed by prior research findings. The research questions guided the analysis process. Table 2 illustrates the inductive thematic analysis approach used to develop the coding book for each research question.

Table 2. An overview of the themes and data sources (interviews and artifacts).

Major themes and sub-themes	Findings from interviews	Findings from artifacts
External Factors		
Professional Development	Faculty attended training sessions on XR[a], viewed as beneficial for skill-building and knowledge sharing.	Training materials and lesson plans demonstrated faculty engagement with XR-based teaching.
Technical Support	Faculty emphasized the need for technical support, linking it to continued XR use.	Technical support records showed assistance provided for troubleshooting XR-related issues.
Infrastructure	Challenges related to limited XR devices, weak Wi-Fi, and lack of educational resources.	Limited XR-enabled classrooms and laboratories were documented in institutional reports.
Social Influence	Positive influence from colleagues, social media, and university policies encouraged XR adoption.	Shared lesson plans and peer-reviewed materials indicated knowledge exchange among faculty.
XR Features		
Ease of Use	Some faculty found XR easy to use, while others struggled with the complexity of designing activities.	User guides and simplified XR tools were developed to address faculty concerns about complexity.
Interactivity	Interactivity was valued for facilitating learning objectives and engagement.	Lesson artifacts included interactive 3D models and gamified simulations.
Imagination and Immersion	Imagination linked to visualization capabilities, enhancing the learning experience.	Students' assignments showed creative applications of XR for visualization.
Internal Factors		
Previous Experience With ICTs[b]	Prior experience with ICTs contributed to smoother XR adoption.	Faculty-created resources mirrored existing ICT teaching practices, aiding XR integration.
Digital Competencies	Digital literacy skills were essential for the effective use of XR in teaching.	Assessment rubrics reflected the need for digital competencies in grading XR-based tasks.
Attitudes Toward XR	Most faculty had positive attitudes, seeing XR as engaging; a minority found it too complex.	Students' reflections indicated excitement about XR use; some reported difficulty in self-directed learning.
Design Factors		
Design Challenges	Time constraints and lack of technical expertise made designing XR activities difficult.	Course syllabi showed attempts to integrate XR but highlighted gaps in structured activity design.

[a]XR: extended reality.

[b]ICT: information and communication technology.

Results

Key Factors

Participants identified key factors based on their experiences with XR in health care education. Faculty workload and responsibility were recognized as significant factors influencing the integration of XR into teaching practices. Additionally, experience with medical technology was found to be linked to the use of XR. Data analysis revealed various influencing factors, which were grouped into 3 categories: internal factors, design factors, and external factors, which include XR features (Table 2). Each category encompasses several specific factors, as detailed in the following sections. Table 2 provides an overview of the themes and data sources.

External Factors

Institutional Drivers of XR Adoption in Teaching

External factors related to higher education institution policies and readiness, such as professional development, technical support, infrastructure, social influence, and XR features, were identified as factors that could positively influence the acceptance and continued use of XR in teaching practices.

Professional Development

All participants confirmed their attendance at training sessions on using XR in their teaching. These sessions covered a range of topics, from recognizing the value of XR as an advanced technology to creating lessons and activities using existing platform assets. For example, one faculty member mentioned, "The training was helpful in various aspects, such as understanding the value of XR and learning how to use it in my class activities" [D1]. Additionally, some faculty members viewed the training sessions as opportunities to share their knowledge and receive feedback, enhancing their practices in medical education. One participant noted, "I shared the activities I designed from scratch to get feedback from my colleagues. It was a good chance to share experiences and learn from others" [P5].

Integrating VR into medical education, particularly in fields such as nursing, offers students the chance to learn in authentic, immersive environments and practice practical tasks through simulations. However, some participants raised concerns that

VR might not substantially enhance student learning, as they believed students could already visualize real-life situations without the need for VR. This underscores the importance of professional development programs in equipping educators with the strategies needed to design VR experiences that extend beyond imagination, offering unique, hands-on, and interactive learning opportunities that traditional methods cannot replicate.

Technical Support

Most participants emphasized the importance of technical support in ensuring the continued use of XR, as it helps minimize technical difficulties faced by faculty members in medical sciences. Technical support encompassed a range of services, from creating platform accounts to troubleshooting issues with platform assets and student access. One faculty member stated, "Technical support is essential to continue using XR as it's a new technology, and I had no previous experience with it" [D25]. Some faculty members associated the availability of technical support with saving time and being able to focus more on the quality of activities and assessments. However, a few reported a lack of technical support due to insufficient staffing at the XR center. Providing technical support also positively impacted students' timely completion of assignments and tasks using XR.

Infrastructure

Infrastructure plays a crucial role in the use and continued adoption of XR in medical education. Participants defined infrastructure in terms of the availability of suitable VR devices, strong Wi-Fi, and educational resources. One faculty member mentioned, "I confront challenges to find assets related to Nursing relevant to my teaching topic" [D24]. A few participants cited the lack of infrastructure as a significant challenge. For instance, one faculty member said, "I have 45 students in Human Medicine, and it's difficult to take them all to the computer lab to use the VR devices because there is only one device" [D5].

One of the novel findings of this study is its emphasis on the intersection of institutional policies, faculty readiness, and infrastructural limitations in shaping XR adoption in medical education. While faculty members acknowledged the pedagogical benefits of XR, they also highlighted significant challenges related to institutional support and funding constraints. Unlike institutions in high-income regions, where government and private sector investments facilitate the widespread adoption of XR, Palestinian universities rely primarily on limited internal budgets and external grants. Consequently, the lack of funding for VR-compatible hardware, insufficient training opportunities for faculty, and inadequate technical support staff emerged as critical barriers to adoption. This study underscores the importance of targeted policy interventions, including faculty incentives, resource-sharing initiatives, and digital transformation strategies, to address these systemic barriers and promote sustainable XR integration.

Social Influence

Positive Influence of Colleagues

When asked about the impact of colleagues on their use of XR, most participants reported a positive influence from both within and outside the university. This influence included sharing expertise, providing technical and instructional support, and exchanging lessons and learning objects on the XR platform. One participant stated, "It was challenging to design lessons using the XR platform, so I asked a colleague for help" [D20].

The Power of Social Media

Some participants reported being members of social media groups focused on advanced technology in engineering, such as Twitter groups. These communities helped them exchange ideas about designing VR activities and share lessons using 3D and 360-degree techniques. One participant shared a lesson about the human body on Facebook and received feedback to improve the lesson using advanced VR features. Another participant said, "I share my lessons and activities in the group and exchange ideas on using XR in teaching various topics" [D6].

XR Features

Impact of XR Features on Adoption in Medical Education

The XR features reported by the majority of participants included ease of use, imagination, interaction, and immersion, all of which could influence the use of XR in medical education.

Ease of Use

Most participants highlighted the importance of XR's ease of use in lesson activities and content presentation. Some linked the simplicity of designing activities and lessons on the platform to their intention to continue using it. One faculty member stated, "I liked using XR because it was easy to design activities to show hidden parts of the human body" [D4]. However, a few participants found XR complicated and challenging for designing lesson activities, which led them to stop using it, although they continued assigning XR-related tasks to students.

Interaction Feature

Many participants emphasized the role of interactive features in facilitating and achieving lesson objectives. One participant said, "interactivity is important for me and my students because it enables activities that are otherwise impossible" [D11]. They also highlighted the importance of designing interaction types between students and learning activities, as well as interactions among students.

Imagination and Immersion

All participants confirmed the significance of imagination in medical sciences education, which could lead to greater immersion in class activities. One faculty member reported, "My students used XR to virtually perform a surgery" [D3]. Many participants linked imagination to visualization features that attract faculty members to use XR in assignments and activities.

Internal Factors

Role of Digital Competencies and Experience in XR Adoption

Internal factors included previous experience with information and communication technologies and digital competencies.

Previous Experience With Information and Communication Technologies

The majority of respondents indicated that previous experience with information and communication technologies and mobile technology was crucial for accepting and using XR in medical education. Experience with smartphones also facilitated their use of mobile VR for course instruction and activities.

Digital Competencies

In this study, digital competencies refer to the knowledge, skills, and attitudes related to XR. Most interviewees reported that their digital competencies were crucial for continuing to use XR. One faculty member stated, "My experience and knowledge are important for using XR in medical education." [D9].

Design Factors

Design factors refer specifically to the pedagogical and instructional aspects of XR integration, particularly in creating activities, materials, and assessments in health care. This process often requires the application of instructional design principles. Most participants indicated that their primary difficulty was designing course-related activities. Many felt they lacked the technological expertise required, and some reported insufficient time due to commitments at private hospitals and clinics.

Attitudes Toward Using XR in Medical Education

Positive Attitudes

Many participants expressed positive attitudes toward using XR in medical education, attributing these attitudes to features such as interaction, visualization, and immersion. Some mentioned that the simplicity of XR saved both time and effort. For instance, one faculty member noted, "My students were excited to use XR." [D4].

Negative Attitudes

A minority of participants, fewer than one-third, reported obstacles that negatively influenced their attitudes toward using XR. These challenges were related to the complexity of XR usage.

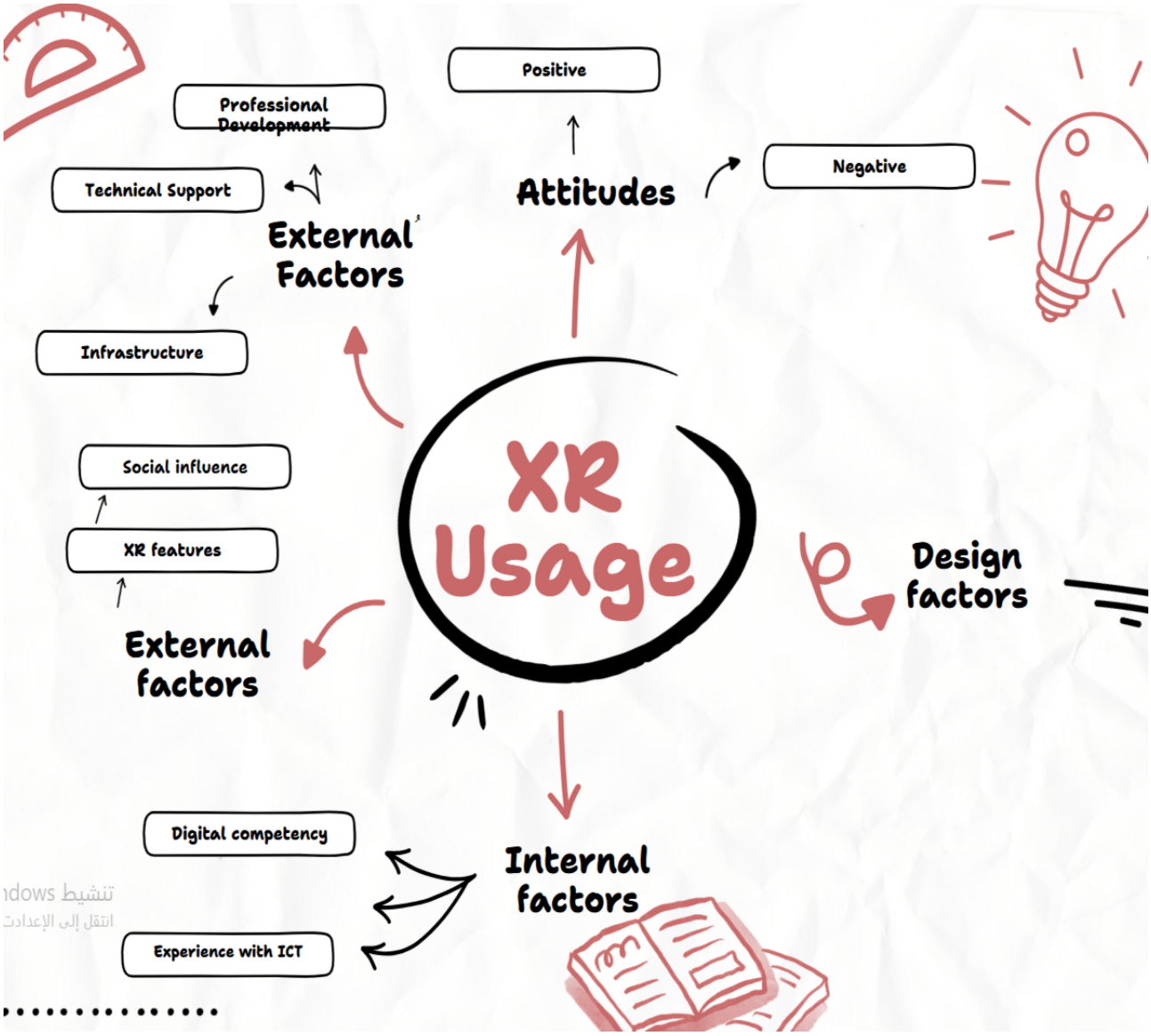
Socioeconomic Factors and XR Adoption

Socioeconomic factors significantly influence XR adoption in Palestinian higher education, particularly in infrastructure investment, faculty training, and institutional support. The high costs of XR hardware, software, and maintenance, coupled with limited government funding and restricted access to grants, present a major challenge. Many universities struggle to scale XR beyond pilot initiatives, limiting widespread faculty adoption. A key barrier is faculty training and professional development. While some institutions offer workshops, many educators lack consistent training and technical support, resulting in uneven adoption across disciplines. Comparisons with developing regions, such as Latin America and Southeast Asia, reveal similar constraints, while well-funded institutions in South Korea and Germany address these challenges through government investment, faculty incentives, and public-private partnerships.

To bridge this digital divide, Palestinian universities must increase funding, establish training programs, and explore resource-sharing models across institutions. Addressing these socioeconomic barriers will ensure sustainable XR integration, enabling faculty to effectively utilize immersive technologies in medical education. Future research should explore scalable funding models and institutional collaborations to support long-term XR adoption. In conclusion, we visualize the factors influencing faculty members' use of XR in medical education.

Figure 2 visualizes the factors influencing faculty members' use of XR in medical education.

Figure 2. Visualization of the factors influencing faculty members to use extended reality (XR) in medical education.



Discussion

Principal Findings

The findings of this study highlight the complex factors influencing XR adoption in medical education, categorized into external, internal, and design-related elements. Professional development emerged as a key enabler, with faculty members who participated in XR training reporting increased confidence and capability in integrating the technology into their teaching. These sessions provided both technical knowledge and collaborative learning environments, aligning with prior research emphasizing the role of continuous professional development in technology adoption [40]. However, technical support and infrastructure remain critical challenges. While access to reliable support enhances faculty engagement with XR, inconsistent availability of assistance and limited institutional investment in technical staff hinder seamless integration. Similarly, infrastructure gaps—such as limited access to XR devices, inadequate internet connectivity, and insufficient educational resources—remain substantial barriers, especially in resource-constrained settings [41]. These findings align with studies in other developing regions, where high costs and inadequate infrastructure are primary obstacles to XR adoption [42].

Although this study applies the UTAUT model to analyze XR adoption, the findings suggest that policy and institutional support are crucial facilitating conditions not explicitly accounted for in the framework. Additionally, socioeconomic constraints, including funding limitations and digital infrastructure challenges, significantly influence adoption behaviors [43]. To extend the theoretical framework, we propose a modified model that integrates 2 additional dimensions: Pedagogical Readiness, encompassing faculty training, instructional design capabilities, and institutional encouragement for XR use, and Technical and Logistical Support, emphasizing the role of digital infrastructure, maintenance, and technical assistance. These modifications offer a more contextualized perspective on XR adoption in developing regions, reinforcing the need for localized implementation strategies [44].

The role of social influence in XR adoption extends beyond institutional policies and peer encouragement. This study found

that faculty skepticism, generational differences in adoption, and student perceptions significantly influence XR use. While peer influence and institutional endorsement encourage adoption, some senior faculty members expressed skepticism about XR, fearing that it might disrupt traditional pedagogical methods rather than complement them. These concerns align with prior research on faculty resistance to emerging technologies [32]. By contrast, younger faculty members demonstrated greater openness, reflecting trends observed in broader educational technology adoption studies [30]. Additionally, students' positive engagement with XR significantly influenced faculty willingness to integrate the technology, reinforcing prior findings that student enthusiasm can drive faculty adoption [33]. However, some educators expressed concerns that XR might encourage passive rather than active learning, highlighting the need for interactive and problem-solving–oriented XR applications to maximize its educational impact [33].

To overcome financial barriers to XR adoption in Palestinian universities, alternative and sustainable funding models are essential. While current efforts often rely on short-term external grants, more resilient approaches—such as public-private partnerships, collaborations with technology firms, and the use of open-source XR platforms—could help support long-term implementation and scalability [45]. Although the return on investment in XR may not be immediately measurable in financial terms, it can be demonstrated through improvements in student performance, engagement, and retention. These outcomes contribute to institutional sustainability by reducing dropout rates and enhancing overall learning effectiveness [41].

Lastly, design-related challenges, particularly the complexity of XR tools and time constraints for faculty, emerged as barriers to effective integration. While many faculty members appreciated the interactive and immersive capabilities of XR, others found content creation and instructional design challenging, highlighting the need for user-friendly design tools and targeted training [46]. Digital competencies were also found to be a critical factor, with faculty members possessing stronger digital skills demonstrating greater ease in XR adoption. This underscores the importance of developing digital literacy as a core competency in medical education [47].

Overall, this study emphasizes the need for a holistic approach to XR adoption, integrating technical, economic, and pedagogical strategies. In comparison to universities in Latin America and Southeast Asia, where national digital education strategies and structured funding initiatives have facilitated XR adoption, Palestinian institutions require policy-driven interventions and regional partnerships to develop scalable, sustainable funding models [48]. Addressing these economic and infrastructural constraints will be essential to ensure that XR can be effectively integrated into medical education in underresourced contexts.

Theoretical and Practical Implications

The study on integrating XR in Palestinian health care education highlights key theoretical and practical implications. Theoretically, it advances technology acceptance models by identifying factors influencing XR adoption, including institutional policies, social influences, digital competencies, and attitudes. It also emphasizes the need for robust infrastructure and professional development to support technology integration.

Practically, effectively implementing XR in medical education requires a well-structured, phased approach to fully realize its transformative potential. The first priority for institutions should be establishing robust foundational infrastructure, including high-quality hardware and software solutions that are scalable and adaptable to evolving needs. This requires substantial initial investments, not only in technology but also in developing technical support systems to address challenges such as high costs, operational complexities, and the demands of maintaining cutting-edge solutions. Along with infrastructure development, it is essential to provide comprehensive training for educators and students, focusing on the digital literacy skills needed to use XR effectively. Clear guidelines should also be developed to ensure consistent, meaningful integration of XR into the curriculum.

Higher education institutions should also consider designing and adopting performance indicators specifically tailored to measure the success and impact of XR implementation. These indicators could include metrics such as student engagement levels, improvements in skill acquisition, and the cost-effectiveness of XR solutions. By establishing these benchmarks, institutions can monitor progress and identify areas for improvement, ensuring a data-driven approach to XR adoption.

Early adoption strategies should emphasize piloting risk-free, immersive simulations that allow educators and students to explore XR's capabilities in a controlled environment. These pilot programs serve to demonstrate the tangible value of XR, helping to build confidence among stakeholders and secure their buy-in for broader implementation. However, institutions should be cautious of potential pitfalls. For instance, underestimating the need for ongoing technical support can lead to system failures and diminished user satisfaction. Similarly, neglecting to align XR initiatives with specific, well-defined learning outcomes can result in unfocused or ineffective use of the technology. Finally, failing to allocate adequate resources for regular maintenance and updates may jeopardize the long-term sustainability of XR programs.

Addressing Technical and Economic Barriers to XR Adoption

Sustainable integration of XR in medical education requires a strategic focus on viable funding models, cost-effectiveness, and long-term impact. In resource-constrained settings, such as Palestinian universities, advancing XR implementation depends less on reiterating existing challenges and more on identifying innovative, context-sensitive solutions. Strategic partnerships—with private technology firms, medical institutions, and international funding bodies—can facilitate access to sponsored XR hardware, software, and training. These collaborations support the co-development of immersive learning programs and enable cost-sharing arrangements that reduce the financial burden on institutions.

Adopting open-source XR platforms also presents a promising avenue for sustainable integration. These tools offer flexibility in content creation and deployment without the high costs associated with proprietary systems, making them particularly suitable for universities with limited budgets. Beyond initial implementation, institutions must assess XR's return on investment through educational outcomes rather than direct financial metrics. Improvements in student engagement, knowledge retention, and academic performance are strong indicators of XR's value and can contribute to institutional sustainability by reducing dropout rates and enhancing graduate readiness.

To ensure scalability and impact, universities should adopt data-informed strategies, including cost-benefit analyses, pilot programs, and scalable deployment models. Aligning financial planning with pedagogical goals ensures that XR technologies are integrated not only as innovative teaching tools but also as sustainable investments in the future of medical education. A methodical, forward-looking approach enables institutions to transform economic limitations into opportunities for creative problem-solving and long-term growth.

Limitations

The study acknowledges several limitations. First, the research was conducted during the initial stages of XR adoption in Palestinian higher education, which may limit the generalizability of the findings. The small sample size and the focus on specific institutions further constrain the applicability of the results to other contexts. Additionally, the high upfront costs and technical challenges associated with XR technologies may pose barriers that were not fully explored due to the limited scope of the study. Finally, the study relies on self-reported data from participants, which could introduce bias or inaccuracies in the findings.

Future Research

Future research should focus on scaling the study to larger populations across multiple universities to provide a more comprehensive understanding of XR's applicability and effectiveness in diverse educational contexts. Expanding research across different institutions, disciplines, and settings will offer broader insights into how XR can be integrated into various pedagogical frameworks.

Additionally, longitudinal studies are essential to track XR adoption over time. These studies would assess the long-term impact of XR on educational outcomes, skill retention, learner engagement, instructor effectiveness, and curriculum integration. Examining how faculty and students interact with XR technologies over extended periods will help identify patterns of adoption, sustained challenges, and evolving best practices. This approach will also contribute to understanding the long-term sustainability of XR implementation in higher education.

Further research should also investigate the technical and pedagogical challenges associated with XR adoption. Identifying these challenges could lead to detailed, actionable guidelines that institutions can use to optimize XR deployment strategies. Beyond health care education, exploring XR's potential in fields such as engineering, humanities, and business would provide insights into its broader applicability. Moreover, examining how XR interacts with emerging technologies such as artificial intelligence, machine learning, and data analytics may reveal innovative ways to enhance teaching and learning experiences.

Another critical area for future research is the development of performance indicators to measure the success of XR adoption. These indicators should assess learning outcomes, user satisfaction, cost-effectiveness, and scalability, providing institutions with data-driven benchmarks to evaluate and refine their XR initiatives.

Finally, addressing the digital divide in XR adoption is crucial, particularly in developing regions. Investigating how educational institutions can ensure equitable access to XR technologies for students from varied socioeconomic backgrounds will help create inclusive and accessible learning environments. This research will be instrumental in bridging technological disparities and promoting digital equity in higher education.

Conclusion

XR technologies have the potential to revolutionize health care education by providing immersive learning experiences that enhance practical skills and knowledge retention. This study highlights several key factors for the successful adoption of XR in medical education, including professional development, adequate infrastructure, robust technical support, positive social influence, and user-friendly design. Strategic investments in these areas are vital to overcoming initial barriers and aligning XR adoption with the Sustainable Development Goals of quality education and good health. By addressing these complex factors, educational institutions can create an environment conducive to the successful integration of XR technology, ultimately improving teaching practices and student learning outcomes in medical and nursing programs, particularly in Palestine.

The findings of this study emphasize that successful XR adoption in medical education requires more than just technological availability—it demands strong institutional policies, sustained funding mechanisms, and structured faculty development programs. Higher education institutions must move beyond pilot initiatives and develop long-term strategies for integrating XR into curricula, supported by clear guidelines, resource-sharing models, and institutional incentives. Additionally, regional collaborations among universities in developing contexts could facilitate knowledge exchange and infrastructure sharing, reducing the financial burden on individual institutions. Future research should further explore **scalable policy interventions** that enable sustainable XR adoption, particularly in resource-constrained environments where technology-enhanced learning can play a crucial role in addressing educational inequalities.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview protocol.
[DOCX File , 15 KB - mededu_v11i1e65042_app1.docx]

Multimedia Appendix 2
A sample of the inductive thematic analysis used in this study.
[DOCX File , 17 KB - mededu_v11i1e65042_app2.docx]

References

1. Doolani S, Wessels C, Kanal V, Sevastopoulos C, Jaiswal A, Nambiappan H, et al. A review of extended reality (XR) technologies for manufacturing training. Technologies 2020 Dec 10;8(4):77. [doi: 10.3390/technologies8040077]
2. Morimoto T, Kobayashi T, Hirata H, Otani K, Sugimoto M, Tsukamoto M, et al. XR (extended reality: virtual reality, augmented reality, mixed reality) technology in spine medicine: status quo and quo vadis. J Clin Med 2022 Jan 17;11(2):470 [FREE Full text] [doi: 10.3390/jcm11020470] [Medline: 35054164]
3. Catbas FN, Luleci F, Zakaria M, Bagci U, LaViola JJ, Cruz-Neira C, et al. Extended Reality (XR) for Condition Assessment of Civil Engineering Structures: A Literature Review. Sensors (Basel) 2022 Dec 06;22(23):9560 [FREE Full text] [doi: 10.3390/s22239560] [Medline: 36502261]
4. Barteit S, Lanfermann L, Bärnighausen T, Neuhann F, Beiersmann C. Augmented, Mixed, and Virtual Reality-Based Head-Mounted Devices for Medical Education: Systematic Review. JMIR Serious Games 2021 Jul 08;9(3):e29080 [FREE Full text] [doi: 10.2196/29080] [Medline: 34255668]
5. Behmadi S. Virtual reality-based medical education versus lecture-based method in teaching start triage lessons in emergency medical students: Virtual reality in medical education. Journal of Advances in Medical Education & Professionalism 2022 Jul 31;10(1):1-6. [doi: 10.1164/ajrccm-conference.2023.207.1_meetingabstracts.a1676]
6. Alam A, Mohanty A. Foundation for the future of higher education or 'misplaced optimism'? Being human in the age of artificial intelligence. Innovations in Intelligent Computing and Communication 2023 Jan 1:17-29.
7. Xiong J, Hsiang E, He Z, Zhan T, Wu S. Augmented reality and virtual reality displays: emerging technologies and future perspectives. Light Sci Appl 2021 Oct 25;10(1):216 [FREE Full text] [doi: 10.1038/s41377-021-00658-8] [Medline: 34697292]
8. Bargeri S. , Scalea, S. , Agosta, F., Banfi, G., Corbetta, D., Filippi, M., ... & Gianola, S. Effectiveness and safety of virtual reality rehabilitation after stroke: an overview of systematic reviews.EClinicalMedicine. 2023 :64.
9. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Quarterly 2003;27(3):425. [doi: 10.2307/30036540]
10. Davis F. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly 1989 Sep;13(3):319. [doi: 10.2307/249008]
11. Khlaif ZN, Sanmugam M, Hattab MK, Bensalem E, Ayyoub A, Sharma RC, et al. Mobile technology features and technostress in mandatory online teaching during the COVID-19 crisis. Heliyon 2023 Aug;9(8):e19069 [FREE Full text] [doi: 10.1016/j.heliyon.2023.e19069] [Medline: 37636397]
12. Park Y, Tak YW, Kim I, Lee HJ, Lee JB, Lee JW, et al. User Experience and Extended Technology Acceptance Model in Commercial Health Care App Usage Among Patients With Cancer: Mixed Methods Study. J Med Internet Res 2024 Dec 18;26:e55176 [FREE Full text] [doi: 10.2196/55176] [Medline: 39693615]
13. Wang X, Young GW, Iqbal MZ, Guckin CM. The potential of extended reality in Rural Education's future – perspectives from rural educators. Educ Inf Technol 2023 Sep 04;29(7):8987-9011. [doi: 10.1007/s10639-023-12169-7]
14. Bach M. P.., Understanding Determinants of Management Simulation Games Adoption in Higher Educational Institutions Using an Integrated Technology Acceptance Model/Technology?Organisation?Environment Modeleducator Perspective. Information 2025;16(1):E. [doi: https://doi.org/10.3390/info16010045]
15. Skalidis I. , Muller, O. , & Fournier, S. CardioVerse: The cardiovascular medicine in the era of Metaverse.Trends in cardiovascular medicine,2023 ;33(8):471. [doi: https://doi.org/10.1016/j.tcm.2022.05.004]
16. Tang Y. M.., A systematic review of immersive technology applications for medical practice and education-trends, application areas, recipients, teaching contents, evaluation methods, and performance. Educational Research Review 2022;35:100429. [doi: https://doi.org/10.1016/j.edurev.2021.100429]
17. Suh I. , T. McKinney, and K. -C. Siu. Current perspective of metaverse application in medical education, research and patient care. in Virtual Worlds. 2023. MDPI . [doi: https://doi.org/10.3390/virtualworlds2020007]
18. Garlinska M. The influence of emerging technologies on distance education. Electronics 2023;12(7):1550. [doi: https://doi.org/10.3390/electronics12071550]

19. Marougkas A. How personalized and effective is immersive virtual reality in education? A systematic literature review for the last decade. Multimedia Tools and Applications 2024;83(6):18185-18233. [doi: https://doi.org/10.1007/s11042-023-15986-7]
20. Lee A. T., R.K. Ramasamy, and A. Subbarao, Understanding Psychosocial Barriers to Healthcare Technology Adoption: A Review of TAM Technology Acceptance Model and Unified Theory of Acceptance and Use of Technology and UTAUT Frameworks. Healthcare 2025;13(3):A. [doi: https://doi.org/10.3390/healthcare13030250]
21. Kuleto V. Extended reality in higher education, a responsible innovation approach for generation y and generation z. Sustainability 2021;13(21):11814. [doi: https://doi.org/10.3390/su132111814]
22. Serna-Mendiburu G. M. C.R. Guerra-Tamez. Shaping the future of creative education: the transformative power of VR in artdesign learning. in Frontiers in Education. 2024. Frontiers Media SA . [doi: https://doi.org/10.3389/feduc.2024.1388483]
23. Aguayo C, Eames C. Using mixed reality (XR) immersive learning to enhance environmental education. The Journal of Environmental Education 2023 Mar 16;54(1):58-71. [doi: 10.1080/00958964.2022.2152410]
24. González-Erena P. V., S. Fernández-Guinea, and P. Kourtesis, Cognitive Assessment and Training in Extended Reality: Multimodal Systems, Clinical Utility, and Current Challenges. Encyclopedia 2025;5(1):8. [doi: https://doi.org/10.3390/encyclopedia5010008]
25. Goudman L. , Jansen, J. , Billot, M., Vets, N., De Smedt, A., Roulaud, M., ... & Moens, M. Virtual reality applications in chronic pain management: systematic review and meta-analysis.JMIR Serious Games,2022. , e3 ;10(2):4402. [doi: https://doi.org/10.2196/34402]
26. Catania V. , Rundo, F. , Panerai, S., & Ferri, R. Virtual reality for the rehabilitation of acquired cognitive disorders: a narrative review.Bioengineering,2023 ;11(1):35. [doi: https://doi.org/10.3390/bioengineering11010035]
27. Samnakay S. , Bell, E. , Evans, D., Sommerfield, D., Sommerfield, A., Hauser, N., & von Ungern?Sternberg, B. S. Assessing the Use and Acceptability of Virtual Reality to Assist Coping in Children Undergoing Clinical Procedures.Journal for Specialists in Pediatric Nursing,2025. , e7 ;30(1):0002. [doi: https://doi.org/10.1111/jspn.70002]
28. de G. Adopting extended reality? A systematic review of manufacturing training and teaching applications. Journal of manufacturing systems 2023;71:645-663. [doi: https://doi.org/10.1016/j.jmsy.2023.10.016]
29. Mistry D, Brock CA, Lindsey T. The Present and Future of Virtual Reality in Medical Education: A Narrative Review. Cureus 2023 Dec;15(12):e51124 [FREE Full text] [doi: 10.7759/cureus.51124] [Medline: 38274907]
30. Arthur T, Melendez-Torres G, Harris D, Robinson S, Wilson M, Vine S. Extended Reality Interventions for Health and Procedural Anxiety: Panoramic Meta-Analysis Based on Overviews of Reviews. J Med Internet Res 2025 Jan 08;27:e58086 [FREE Full text] [doi: 10.2196/58086] [Medline: 39778203]
31. Burian B, Ebnali M, Robertson J, Musson D, Pozner C, Doyle T, et al. Using extended reality (XR) for medical training and real-time clinical support during deep space missions. Appl Ergon 2023 Jan;106:103902. [doi: 10.1016/j.apergo.2022.103902] [Medline: 36162274]
32. Chuah SHW. Wearable XR-technology: literature review, conceptual framework and future research directions. IJTMKT 2019;13(3/4):205. [doi: 10.1504/ijtmkt.2019.104586]
33. Curran VR, Xu X, Aydin MY, Meruvia-Pastor O. Use of extended reality in medical education: an integrative review. Med Sci Educ 2023 Feb 19;33(1):275-286 [FREE Full text] [doi: 10.1007/s40670-022-01698-4] [Medline: 36569366]
34. Kluge MG, Maltby S, Keynes A, Nalivaiko E, Evans DJR, Walker FR. Current state and general perceptions of the use of extended reality (XR) technology at the University of Newcastle: interviews and surveys from staff and students. Sage Open 2022 Apr 26;12(2):21582440221093348. [doi: 10.1177/21582440221093348]
35. Li N, Keskitalo T. Literature review - XR in medical and healthcare education. WHO. 2022. URL: https://ec.europa.eu/programmes/erasmus-plus/project-result-content/f91f0f50-2e07-4d8e-a459-eaf7e0d3ebe4/PrepaCare-IO2_Literature_Review_Publication.pdf [accessed 2025-04-28]
36. Creswell J. Qualitative Research Designs Selection and Implementation. The Counseling Psychologist 2007;35:236-264.
37. Campbell S. Purposive sampling: complex or simple? Research case examples. J Res Nurs 2020;25(8):652-661.
38. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
39. Ngozwana N. Ethical dilemmas in qualitative research methodology: researcher's reflections. Int J Educ Methodol 2018 Feb 15;4(1):19-28. [doi: 10.12973/ijem.4.1.19]
40. Jin Y. Generative AI in higher education: A global perspective of institutional adoption policies and guidelines. Computers and Education: Artificial Intelligence 2025;8:A. [doi: https://doi.org/10.1016/j.caeai.2024.100348]
41. Luo S. , D. Zou, and L. Kohnke, A systematic review of research on xReality (XR) in the English classroom: Trends, research areas, benefits, and challenges. Computers & Education: X Reality 2024;4:100049. [doi: https://doi.org/10.1016/j.cexr.2023.100049]
42. Dinh A. Perceptions about augmented reality in remote medical care: Interview study of emergency telemedicine providers. JMIR Formative Research 2023;7:e45211. [doi: https://doi.org/10.2196/45211]
43. Dingel J. Predictors of Health Care Practitioners? Intention to Use AI-Enabled Clinical Decision Support Systems: Meta-Analysis Based on the Unified Theory of Acceptance and Use of Technology. Journal of medical internet research 2024;26:e57224. [doi: https://doi.org/10.2196/57224]

44. Fugate J. , M. Tonsanger, and S. Macrine, Immersive Extended Reality (I-XR) in Medical and Nursing Education: A Systematic Review and Pedagogical Directives 2025 Jan 13:1-50. [doi: 10.20944/preprints202501.0928.v1]
45. Kelly S. , S.-A. Kaye, and O. Oviedo-Trespalacios, What factors contribute to the acceptance of artificial intelligence? A systematic review. Telematics and Informatics 2023;77:101925.
46. Meccawy M. Teachers? prospective attitudes towards the adoption of extended reality technologies in the classroom: Interests and concerns. Smart Learning Environments 2023;10(1):36. [doi: https://doi.org/10.1186/s40561-023-00256-8]
47. Mainz A. Measuring the Digital Competence of Health Professionals: Scoping Review. JMIR Medical Education 2024;10(1):e55737. [doi: https://doi.org/10.2196/55737]
48. Al-Adwan A. S.., Extending the technology acceptance model (TAM) to Predict University Students? intentions to use metaverse-based learning platforms. Education and Information Technologies 2023;28(11):15381-15413. [doi: https://doi.org/10.1007/s10639-023-11816-3]

Abbreviations

AR: augmented reality
MR: mixed reality
TAM: Technology Acceptance Model
UTAUT: Unified Theory of Acceptance and Use of Technology
VR: virtual reality
XR: extended reality

Edited by A Bahattab; submitted 03.08.24; peer-reviewed by M Smith, D Patel, T Moser; comments to author 01.01.25; revised version received 15.01.25; accepted 23.03.25; published 01.05.25.

Please cite as:
Khlaif Z, Salama N, Hamamra B, Mousa A
Factors Influencing Educators' Perspectives on Accepting Extended Reality in Health Care Education: Qualitative Study
JMIR Med Educ 2025;11:e65042
URL: https://mededu.jmir.org/2025/1/e65042
doi:10.2196/65042
PMID:40310667

Assessing ChatGPT's Capability as a New Age Standardized Patient: Qualitative Study

Joseph Cross[1*], PhD; Tarron Kayalackakom[2*], MD; Raymond E Robinson[3], MPH, MBA, MD, EdS; Andrea Vaughans[4], MD; Roopa Sebastian[4], MSc, PhD; Ricardo Hood[2], MD; Courtney Lewis[2], MD; Sumanth Devaraju[2], MD; Prasanna Honnavar[2], PhD; Sheetal Naik[2], MD; Jillwin Joseph[2], PhD; Nikhilesh Anand[5], MD; Abdalla Mohammed[6], BSc; Asjah Johnson[6], BSc; Eliran Cohen[6], BSc; Teniola Adeniji[6], BSc; Aisling Nnenna Nnaji[6], BSc; Julia Elizabeth George[6]

[1]Medical University of the Americas, PO Box 701, Charlestown, Saint Kitts and Nevis

[2]Department of Education Enhancement, College of Medicine, American University of Antigua, St Johns, Antigua and Barbuda

[3]Department of Health Informatics, School of Professional Studies, Northwestern University, Evanston, IL, United States

[4]Department of Biochemistry, Cell Biology and Genetics, College of Medicine, American University of Antigua, Basseterre, Antigua and Barbuda

[5]Department of Medical Education, School of Medicine, University of Texas Rio Grande Valley, Edinburgh, TX, United States

[6]School of Medicine, Xavier University, Orangestad, Aruba

[*]these authors contributed equally

Corresponding Author:
Joseph Cross, PhD
Medical University of the Americas, PO Box 701, Charlestown, Saint Kitts and Nevis

Abstract

Background: Standardized patients (SPs) have been crucial in medical education, offering realistic patient interactions to students. Despite their benefits, SP training is resource-intensive and access can be limited. Advances in artificial intelligence (AI), particularly with large language models such as ChatGPT, present new opportunities for virtual SPs, potentially addressing these limitations.

Objectives: This study aims to assess medical students' perceptions and experiences of using ChatGPT as an SP and to evaluate ChatGPT's effectiveness in performing as a virtual SP in a medical school setting.

Methods: This qualitative study, approved by the American University of Antigua Institutional Review Board, involved 9 students (5 females and 4 males, aged 22 - 48 years) from the American University of Antigua College of Medicine. Students were observed during a live role-play, interacting with ChatGPT as an SP using a predetermined prompt. A structured 15-question survey was administered before and after the interaction. Thematic analysis was conducted on the transcribed and coded responses, with inductive category formation.

Results: Thematic analysis identified key themes preinteraction including technology limitations (eg, prompt engineering difficulties), learning efficacy (eg, potential for personalized learning and reduced interview stress), verisimilitude (eg, absence of visual cues), and trust (eg, concerns about AI accuracy). Postinteraction, students noted improvements in prompt engineering, some alignment issues (eg, limited responses on sensitive topics), maintained learning efficacy (eg, convenience and repetition), and continued verisimilitude challenges (eg, lack of empathy and nonverbal cues). No significant trust issues were reported postinteraction. Despite some limitations, students found ChatGPT as a valuable supplement to traditional SPs, enhancing practice flexibility and diagnostic skills.

Conclusions: ChatGPT can effectively augment traditional SPs in medical education, offering accessible, flexible practice opportunities. However, it cannot fully replace human SPs due to limitations in verisimilitude and prompt engineering challenges. Integrating prompt engineering into medical curricula and continuous advancements in AI are recommended to enhance the use of virtual SPs.

(JMIR Med Educ 2025;11:e63353) doi:10.2196/63353

KEYWORDS

medical education; standardized patient; AI; ChatGPT; virtual patient; assessment; standardized patients; LLM; effectiveness; medical school; qualitative; flexibility; diagnostic

Introduction

Standardized patients (SPs) have been a cornerstone of medical education since the 1960s, offering students an immersive, real-world experience in a controlled environment. Studies have demonstrated that SP programs are superior for teaching consultation skills compared with traditional methods, with medical students trained using SPs showing increased confidence and competency compared with those trained through other modalities [1,2].

While SPs provide valuable opportunities for students to practice diagnostic and interpersonal skills under standardized conditions, several inherent challenges exist. The resource-intensive nature of SP programs has been a persistent issue, with significant costs associated with recruitment, training, and maintenance of an SP bank [1,3]. Additionally, questions have emerged about SPs' ability to adequately represent the nuances of real patient presentations.

These challenges are particularly pronounced in specific contexts. For instance, Caribbean medical schools face unique obstacles due to limited local health care infrastructure and varying access to clinical training resources. Many offshore institutions in countries such as Aruba and Antigua and Barbuda must rely on partnerships with local health care providers, often resulting in inconsistent access across student cohorts [4,5]. The COVID-19 pandemic exposed additional vulnerabilities in traditional SP programs. The discontinuation of the USMLE Step 2 Clinical Skills examination in 2022, for instance, highlighted the risks of relying solely on in-person SP encounters for assessment [5].

In the 21st century, virtual SPs have emerged. These are computer programs that simulate specific illnesses and respond to learner inputs [6]. They have become invaluable tools in both teaching and assessment. However, their development also requires significant resources, making it challenging for institutions without robust educational technology support departments [7].

As the field of artificial intelligence (AI) has advanced, the potential for its application in medical education has expanded. Large language models (LLMs), such as ChatGPT (OpenAI), have revolutionized natural language processing. These sophisticated neural networks, trained on vast amounts of web-based data, are adept at predicting subsequent words in a sequence [8]. ChatGPT, a chatbot based on the GPT-3.5 model, has an enormous 175 billion parameters and displays a remarkable capacity for understanding and reasoning, bordering on human-like proficiency [9]. Since its introduction in November 2022, sectors spanning from history to entertainment have rapidly adopted the LLM [10].

This advancement in AI has led to the development of virtual SP chatbots. A number of major educational material suppliers and specialized companies are offering chatbot SPs, based on LLMs capable of natural language interactions, for students to practice clinical skills. One example is Oscer, which can present more than 200 virtual patient conditions and boasts above 90% accuracy in symptomology [11]. Similarly, the University of Texas Medical Branch makes use of an AI agent termed Virti, which they use to conduct virtual Observed Structured Clinical Examinations with medical students [12]. Other publicly accessible sites offering virtual patients include Soma Lab [13] and Body Interact [14]. However, for this new generation of virtual patients there is again considerable time and resources required for the company or the institution to develop the program and train the LLM on specific datasets and student access can be limited by cost and locality [7].

The debut of ChatGPT sparked inquiries into its potential as an SP. Liu et. al [15] crafted 10 medical histories with ChatGPT, which were then vetted by experienced physicians. Their results highlighted ChatGPT's promise in clinical education, although some responses came across as robotic [15]. Suarez et.al [16] gathered dental student's feedback after interacting with an AI chatbot. The majority found the experience valuable, especially those who made a correct diagnosis. This underscores the potential of integrating AI into health sciences training [16].

Weidener and Fischer [17] emphasized the growing consensus on incorporating AI into medical education. Their study indicated the importance of both practical and technological skills for leveraging AI in medicine [17]. Similarly, Jowsey et. al [18] have recommended adoption of AI into medical education as a way of preparing future physicians for the reality of modern practice.

We were aware that SPs at our school, American University of Antigua (AUA), were in limited supply and had received feedback indicating that while SPs are effective, students would like greater access to them. In fact, some students had no access during their course, depending on their cohort.

One of our study's aims was to assess medical students' perceptions and experiences regarding the use of AI in medicine—specifically by examining their views before and after interacting with ChatGPT as an SP. A second aim was to evaluate whether ChatGPT can perform adequately as a virtual SP in a medical school setting. Guided by these aims, our investigation focused on the following research questions: (1) How do students perceive the effectiveness of ChatGPT compared with traditional SPs in medical training scenarios? (2) To what extent can ChatGPT function effectively as a virtual SP in medical education?

By addressing these questions, our study seeks to inform the potential integration of AI-driven virtual SPs into medical curricula, particularly in settings where access to traditional SPs is limited.

Methods

Ethical Considerations

This study was given expedited approval by the AUA Research committee (no. AUAIRBa23011). Eleven medical student volunteers enrolled in the MD course at AUA were recruited via a campus-wide email. Two participants were lost to follow-up, leaving a total of 9 participants. Students were 5 females and 4 males, aged 22-48 years, comprising students from both first and second years of the basic sciences course

section of the MD program. Participants were explicitly informed that their involvement in the research was completely voluntary. They were also assured that their responses would remain confidential and anonymous, and all participants signed informed consent agreements. All data were anonymized and no compensation was provided to participants.

Study Design

Students were given access to ChatGPT version 4.0 accounts, the most recent available at the time of the study. Students were observed during a live role-play, in which a student inputted a predetermined prompt, provided by the study authors, into ChatGPT. The prompt directed the LLM to present as a patient with a neurological condition (Figure 1).

The student, in the role of physician, then interviewed the ChatGPT and attempted to make a differential diagnosis (Figure 2).

Figure 1. Prompt used in ChatGPT role-play.

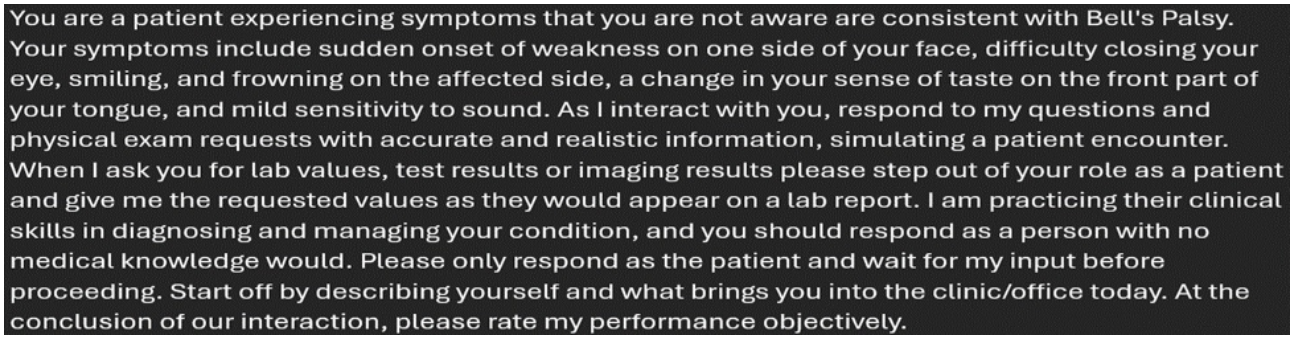
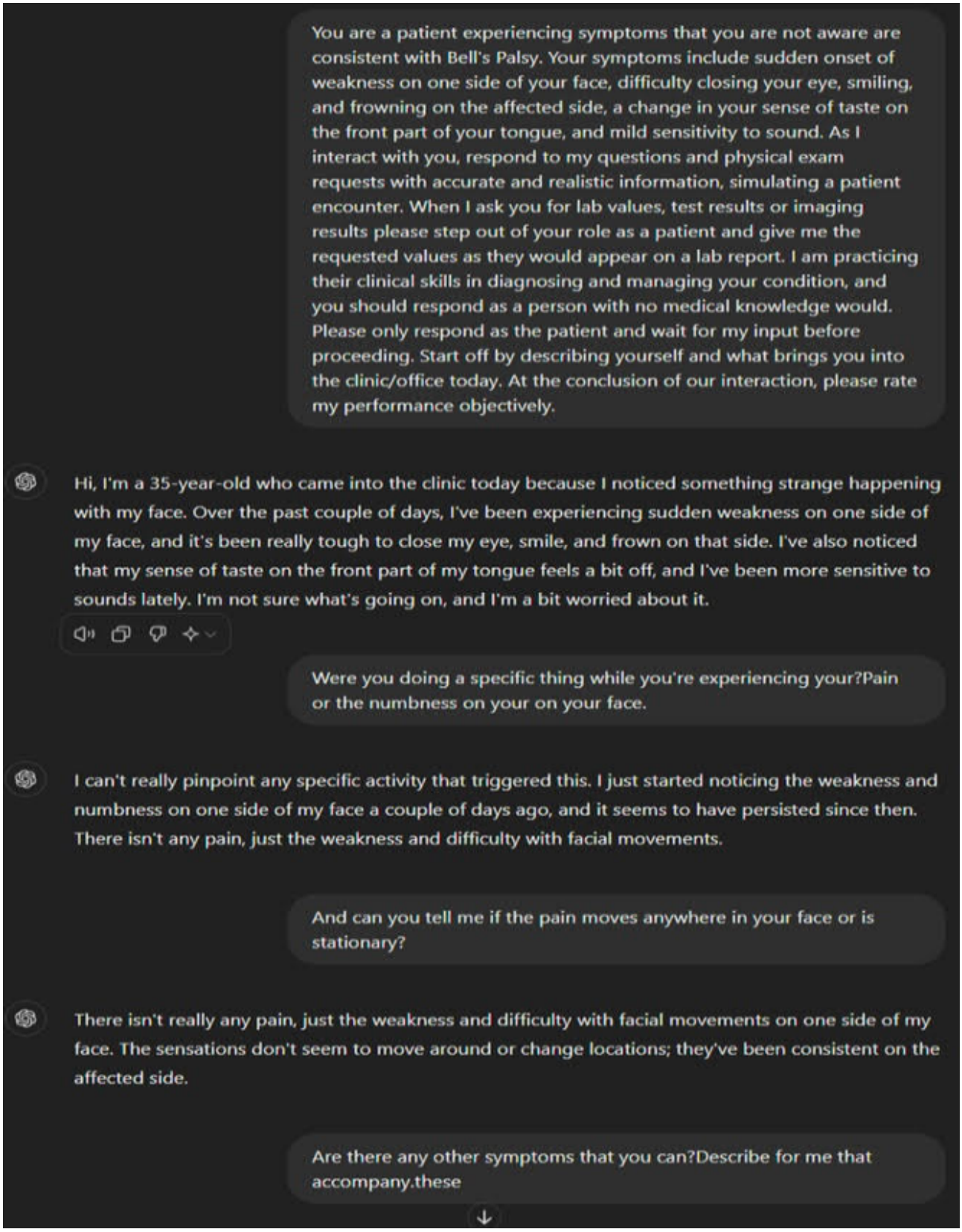


Figure 2. Screenshot example of ChatGPT standardized patient interaction.



Prompt Development

The development of the prompt for ChatGPT's simulated patient interaction underwent an iterative process prior to its use by students. This process involved a 6-member faculty team comprising both clinical and nonclinical faculty, ensuring a diverse range of perspectives and expertise. The faculty were tasked with using the prompt in simulated interactions with ChatGPT, assessing the following factors:

1. Consistency: ensuring the chatbot consistently adhered to the patient role and provided responses aligned with the illness script.
2. Accuracy: evaluating whether the responses were medically plausible and aligned with the provided case information.
3. Likelihood of misleading the SP: assessing whether the chatbot responses could inadvertently lead users to incorrect assumptions or conclusions.
4. Quality of output: reviewing the depth and appropriateness of responses to ensure a realistic and effective simulation experience.
5. Adherence to prompt instructions: verifying that ChatGPT's responses followed the specific behavioral and informational instructions embedded in the prompt.

Faculty provided detailed feedback based on their observations, leading to refinements in the prompt. Suggestions included adjustments to phrasing, additional clarifications to the illness script, and enhancements to behavioral instructions to minimize the potential for ChatGPT to deviate from the assigned patient role. This iterative process was instrumental in optimizing the prompt's effectiveness before deployment in the study.

Rationale for Clinical Case Selection

Bell palsy was chosen as the clinical condition for the simulation due to its relevance to the material being taught at the time. This alignment ensured that the scenario was both clinically pertinent and integrated with the participants' ongoing coursework in both basic sciences and clinical disciplines. The familiarity of the students with the foundational aspects of Bell palsy was intended to facilitate meaningful engagement with the simulated patient, allowing them to focus on the interaction and diagnostic questioning rather than struggling with unfamiliar content.

Purpose of the Evaluation

It is important to note that the primary goal of this study was not to evaluate the students' diagnostic accuracy. Instead, the focus was on assessing their perceptions of ChatGPT's performance as a simulated patient. This distinction was critical to the study design, as it allowed for an emphasis on the usability, realism, and educational value of AI-driven SP interactions without conflating these aspects with the participants' clinical competencies.

The role-play was conducted verbally, as a voice control extension added to the ChatGPT accounts allowed natural language conversation between the student and the LLM [19]. A structured questionnaire consisting of 15 open-ended questions was administered before and after interaction with ChatGPT in the role of an SP. Students were asked about specific elements of their interaction and interviews were conducted in person by faculty team members (Multimedia Appendix 1).

Participating students were introduced to the ethical considerations of using LLMs such as ChatGPT. This included training on the importance of deidentifying patient data, recognizing the limitations of AI, and understanding the potential biases inherent in AI responses, such as those related to gender or ethnicity. This ethical orientation aimed to ensure that students approached the interactions responsibly and with an awareness of the technology's constraints.

Thematic Analysis

The results of the students' group work were recorded, transcribed, and coded by 3 different authors (JC, TK, and RER). Following discussions in regular meetings, findings were summarized, and a category system consisting of main and subcategories, according to Mayring's [20] qualitative content analysis, was agreed upon. Selected text passages were used as quotations to illustrate each category. Inductive category formation, a qualitative research method used to analyze data by identifying patterns, themes, or categories that emerge directly from the data itself, without predefined hypotheses or coding frameworks, was used to analyze open-ended survey responses and interview transcripts.

To explore differences in prompt engineering techniques across academic levels, we asked students to describe how they approached questioning and refining their prompts during the postsession interviews. First-year students, who had less clinical exposure, were expected to rely more on general inquiry methods, while second-year students might leverage slightly more clinical insight. Recording these observations allowed us to compare prompt engineering strategies between these groups and understand how curriculum familiarity influenced interactions with the AI-driven simulated patient interactions.

Results

A total of 9 students participated (5 females and 4 males, aged 22 - 48 years) (Table 1). All students had had some prior experience with traditional SPs, with more senior students having had a greater number of encounters. This contextualizes their perceptions of ChatGPT as a supplement and provides a baseline for understanding the comparative effectiveness of the AI-based approach.

Table . Demographic data.

Characteristics	Participants (n=9), n
Age (years)[a]	
22 - 30	4
31 - 40	4
41+	1
Sex	
Male	4
Female	5
Semester	
1	0
2	7
3	1
4	1
5	0
Prior experience with SPs[b]	
Yes	9
No	0
Prior experience with AI[c]/ChatGPT	
Yes	6
No	3

[a]Mean age: 31.22 (SD 6.8) years

[b]SPs: standardized patients.

[c]AI: artificial intelligence.

The thematic analysis of student feedback prior to interaction with ChatGPT as an SP identified several key themes and subthemes (Table 2). Under the theme of technology limitations, students noted challenges with prompt engineering, such as difficulty in asking effective questions, because the AI could not role-play a physical examination. In terms of learning efficacy, students mentioned the potential for personalized learning materials, grammatical assistance, and the ability for repeated practice without the constraints of limited SP availability. Additionally, some students highlighted the potential for increased convenience, as they could practice as often and whenever they wanted. A potential reduction in SP interview stress was also seen as a benefit of increasing virtual practice. However, under the theme of verisimilitude (ie, the degree to which a simulation mirrors real-life scenarios, including the subtle behaviors and interactions that contribute to a convincing experience), students expressed concerns about the absence of visual cues and rapport, which are important in real patient interactions. Finally, trust issues were raised regarding the accuracy of the LLMs output.

Table . Thematic analysis of student feedback preinteraction with ChatGPT standardized patient.

Themes and subthemes	Representative quotations
Theme 1. Technology limitations	
Prompt engineering	"The challenges might be just asking the right questions, because it's an AI, you can't ask them to do physical examinations."
Theme 2. Learning efficacy	
Personalized learning materials	"Triple checking work and not only getting the right answer, but getting explanations for the right answer and then why the wrong answer is wrong."
Grammatical assistance	"It would be helpful because English is not my first language."
Repetition	"There's usually 10 medical students to one patient, and sometimes you're fighting over each other to get the interview, so this allows us to get more repetitions."
Depth of medical knowledge	"The sky's the limit with regards to what we can practice."
Interview stress or anxiety	"It will kind of be a bit more stress free because you know you're talking to a computer rather than an actual patient."
Convenience	"Be able to practice it as much as I want, as often as I want and any time I want."
Theme 3. Verisimilitude	
Absence of visual cues	"You have to figure out ways to ask the question without the visual cues."
Absence of rapport or empathy	"Building the communication and the relationship with your patient is important."
Theme 4. Trust	
Inaccurate output	"One incident was in the small group activity, where it gave us the wrong answer."

Following interaction with ChatGPT, the thematic analysis of student feedback revealed some changes in perceptions (Table 3). While technology limitations were still noted, students mentioned that they had learnt to improve the output from ChatGPT by tailoring prompts. They also reported alignment issues, such as ChatGPT not providing information on sensitive topics such as patient sexual history. Learning efficacy remained a significant theme, with students appreciating the convenience and repetition benefits. They found the ability to practice history taking without stress and receive feedback useful for skill development. However, verisimilitude issues persisted as a theme, with students noting the absence of visual and tonal cues, and the lack of rapport and empathy, all of which impacted the effectiveness of the patient interview and the ability to make a diagnosis. Some students experienced information overload, feeling that ChatGPT provided more information than a real patient would.

Table . Thematic analysis of student feedback postinteraction with ChatGPT standardized patient.

Themes and subthemes	Representative quotations
Theme 1. Technology limitations	
Prompt engineering	"You could put in the prompt that you want to tailor the responses you want to get back."
Alignment	"When I asked like about sexual history, they were not able to give information."
Theme 2. Learning efficacy	
Convenience	"Having ChatGPT to practice history whenever we want, I think that's the improvement."
Repetition	"You are able to have a lot more repetitions than you are in lab."
Interview stress or anxiety	"Since it's a computer, it's not real. I had less anxiety."
Feedback	"I can ask 'hey, how did you think I did?'"
Skills development	"It highlighted the importance of on-the-spot thinking and memory recall in a medical scenario."
Overall enhanced learning	"It's going to make you sharper. You know, you're probably going to be ahead of your peers, you're going to be able to answer a patient in a better, more detailed manner. Give them a better treatment or care plan."
Theme 3. Verisimilitude	
Absence of visual cues	"For the standardized patient you physically see them. You can see if they're in pain, they don't have to explain where they are in pain."
Absence of tonal cues	"ChatGPT had the same tone, even if it was saying something sad."
Absence of rapport or empathy	"It takes away the personal connection between the doctor and the patient."
Information overload	"It felt like it was offering more information than a regular patient would."

To provide broader context, we compared ChatGPT with some other virtual SP platforms or platforms that could provide this function (Table 4). The comparison highlights the unique strengths and weaknesses of ChatGPT identified in this study in comparison with other platforms, including Claude AI (another chatbot often ranking near the top of benchmarking tables), Body Interact, Oscer AI, and Soma Lab [13,14,21-23]. Both ChatGPT and Claude AI offer flexibility and unlimited practice but are limited by uncurated outputs and reliance on prompt engineering. Oscer AI and Som Lab provide curated clinical cases with tailored feedback, yet their visual representation and interactivity vary, with Soma Lab integrating natural conversational voice modes. Body Interact enhances verisimilitude through patient avatars and curated cases but lacks voice interaction. Cost structures range from free access for basic use to subscription-based models for advanced features.

Table . Comparison of various platforms able to function as standardized patients.

Platform	Technology limitations	Learning efficacy	Verisimilitude	Model cost
ChatGPT	Requires effective prompt engineering; uncurated outputs	Offers flexibility and unlimited practice	Limited visual and tonal cues; natural conversational voice mode	Free and subscription-based options
Claude AI	Requires effective prompt engineering; uncurated outputs	Offers flexibility and unlimited practice	Limited visual and tonal cues; limited voice interaction	Free and subscription-based options
Body Interact	Requires effective prompt engineering; curated clinical cases	Facilitates skill development through realistic scenarios	Patient avatars; lacks voice interaction	Subscription or licensing fees
Oscer AI	Requires effective prompt engineering; wide range of curated clinical cases	Focus on history-taking skills with limited versatility	Limited visual cues; voice interaction possible	Free. Subscription for full access
Soma Lab	Requires effective prompt engineering; wide range of curated clinical cases	Counseling-focused; supports repeated practice with tailored feedback	Static patient avatars; natural conversational voice mode	Variable costs based on usage and features

Discussion

Principal Findings

This study investigated the use of ChatGPT as an SP by qualitative analysis of students' responses to a questionnaire, preinteraction and postinteraction, with ChatGPT performing the role of SP. In terms of diagnostic skill development, our conclusions were drawn from a combination of faculty observations and student self-report. Faculty members who observed the sessions noted that students demonstrated more structured reasoning and improved question formulation after repetitive practice with ChatGPT. In postsession interviews, students themselves expressed feeling more confident and organized in their clinical reasoning steps. This alignment between external observation and self-assessment suggests that the interaction with ChatGPT, although lacking nonverbal cues and certain realistic elements, still provides a valuable platform for honing diagnostic interviewing skills. Thematic analysis provided insights into student perceptions. Major themes identified were technology limitations, learning efficacy, and verisimilitude. Our results suggest that the current version of ChatGPT (ChatGPT version 4.0 at the time of this study) can function effectively as an augmentation to traditional SPs but cannot fully substitute for SPs. These results are broadly in line with those of other studies using LLMs in the role of SP [24-29].

The technological limitations of LLMs in the context of SP exercises were both anticipated and confirmed in our study. The subtheme of prompt engineering was particularly important. Students were made aware of the importance of correctly worded prompts before the exercise, and we found that the faculty-provided prompt, developed through a trial and error process, proved effective in this regard.

The significance of prompt engineering when using LLMs as virtual SPs, or in developing related materials, is also supported by other studies [28,30-33]. It has been suggested that prompt engineering could be incorporated into medical curricula through, for example, hands-on workshops, simulation-based learning, and courses on AI in health care [28,30-32].

The postinteraction interviews also revealed an additional subtheme of alignment. Alignment refers to the problem of ensuring that AI acts in accordance with human intentions and human values [34]. Students noted that the LLM did not provide a response when asked about a patient's sexual history, a standard question in any medical consultation. Ensuring that ChatGPT does not output material which could be considered offensive under societal norms is a component of alignment [35]. However, our results demonstrate an "alignment tax," in that the model becomes less useful due to constraints imposed by the alignment. The development of LLMs designed specifically for medical education may overcome this issue [36].

Learning efficacy was also a major theme identified in this study. Important subthemes in this category were repetition and convenience. Students noted the benefits of having access to ChatGPT for practice at any time or place and having virtually unlimited ability to repeat the exercises. As mentioned earlier, access to SPs is limited in many medical schools [15]. The ability to augment this shortfall with a virtual SP may be a positive option for many medical students and medical schools.

Interestingly, some students expressed that they experienced considerable anxiety as much as a day before they were scheduled to interact with an SP, although they were aware that the SP was not a real patient. The ability to practice with an LLM such as ChatGPT was seen as beneficial, because students could develop questioning techniques to a point where even during the session with a real SP they could still perform well.

Some differences between preinteraction and postinteraction in terms of subthemes were evident under the major theme of learning efficacy. Before the exercise students were focused more on anticipated or previous experiences in using LLMs for personalized learning materials, for example, developing mnemonics, practice questions, or flash cards. This reflects the experience of other medical students [37]. Responses following the exercise were focused on diagnostic patient interaction skills. This is to be expected as students now had actual experience of ChatGPT in this role and knew that this was to be the focus of our study.

Verisimilitude was a major theme in both preinteraction and postinteraction responses. All students mentioned this as a limiting factor. Absence of facial cues, changes in tone, or body language and an inability to develop rapport were all seen as drawbacks of the virtual SP. Some students also mentioned that this impacted their role as physician. For example, a student physician leaning into the patient to show interest, or other types of body language, was redundant in the exercise. Other studies have also highlighted that the output from ChatGPT cannot replicate the true stimuli a physician relies on in a patient visit [28,31,38,39]. We note that virtual patients are developing rapidly, so issues with verisimilitude may be overcome in future, although it may take some time before ChatGPT, specifically, is able to incorporate a visual or physical layer.

Trust as a theme was evident in preinterview responses but had disappeared in postinterview responses. We note that our faculty team, consisting of clinicians and PhD-qualified members, did not notice any "hallucinations" in output, despite multiple repetitions of the exercise. Yanagita et al [40] recently found that high-quality illness scripts, used for improving medical student's clinical reasoning, could be generated by ChatGPT with relatively few errors. Magalhaes et al [25] also found that a majority of students trusted ChatGPT's output. Nevertheless, even a single error in ChatGPT output, given multiple health care providers may receive the same output, could affect many patients. It is therefore imperative that the veracity of AI output be thoroughly tested before it is fully integrated into health care and medical education settings [28].

Other subthemes for learning efficacy evident postinteraction were feedback and information overload. Our prompt included a direction for ChatGPT to provide feedback on how students could improve their performance. We note that it was necessary to revise the prompt several times during the study, as initially it provided only positive feedback, which did not help in identifying areas for improvement. Responses under the information overload subtheme suggested that students found that the LLM tended to provide more information in regard to

a given question than perhaps a real patient or SP would. This presumably related to the depth of medical knowledge of the LLM but should be considered in further iterations of this exercise. It may be possible to refine the prompt to reduce this effect.

Table 4 compares various platforms able to be used as SPs in medical education, highlighting strengths and limitations across technology, learning efficacy, verisimilitude, and cost. ChatGPT and Claude AI offer affordable, flexible options for unlimited practice but face challenges with uncurated outputs and limited realism in visual and tonal cues. In contrast, platforms such as Body Interact and Soma Lab provide curated cases and interactive features, although often at a higher cost. These findings reinforce that while ChatGPT is a valuable and accessible tool for augmenting SP training, it cannot fully replicate the nuances of human SPs. Addressing limitations such as effective prompt engineering and enhancing realism through improved visual and auditory features could significantly improve its use.

It is possible that the use of ChatGPT as a virtual SP may influence trainees' sensitivity toward patients through the absence of the genuine human interaction students may have with SPs and real patients [41]. The rapid evolution of AI technologies is addressing these gaps to an extent. For instance, the advanced voice mode (AVM) in newer versions of ChatGPT incorporates natural speech patterns and emotional intonations, which may help simulate more realistic interactions. While AI cannot yet replicate the full nuances of real patient encounters, it serves as a consistent and flexible supplementary tool for medical training. Future advancements in AI capabilities may further enhance their ability to foster empathy and connection, thereby reducing potential concerns around decreased sensitivity in trainees.

A number of recent studies have confirmed the use of ChatGPT, or similar LLMs, as virtual SPs [28,29,42]. Similarly to our study, these studies have highlighted ChatGPT's potential to reduce resource constraints and improve accessibility in medical training while offering immersive, flexible practice opportunities. At the same time, limitations created by a lack of verisimilitude were also noted.

Both the necessity and challenges of integrating AI, including LLMs, into medical curricula have also been widely acknowledged [43-47]. Addressing inequities in AI models derived from biased training data is crucial, as these can perpetuate disparities in patient care. Strategies to ensure fairness and equitable outcomes, such as transparency in algorithmic design, have been emphasized in recent studies [45,48]. Additionally, resource allocation, faculty training, and the development of tailored content for medical applications add layers of complexity to curricula integration [46,48]. To move forward, curricula must incorporate foundational AI competencies, including ethical considerations, algorithmic fairness, and practical skills such as prompt engineering. Embedding these competencies into existing core courses, rather than as electives, will ensure comprehensive and equitable learning opportunities [43,44,46,48].

To effectively integrate AI into medical curricula, assessments should be designed to balance the use of AI tools while maintaining the integrity of evaluation processes [44]. Educators should implement secure examination protocols, such as locked-down computers and stricter proctoring, to prevent misuse of AI during assessments. However, assessments can also creatively incorporate AI by engaging students in critiquing AI-generated responses or using these tools to identify knowledge gaps and provide tailored feedback. Generative AI can enhance formative assessments by offering immediate and individualized feedback, guiding students' learning trajectories. We note that our results demonstrate the efficacy of this approach, with the virtual SP providing valuable insights to each student individually on how to improve their patient interactions.

Study Limitations

The small sample size, comprising only 9 participants from a single institution, and potential ascertainment bias, with tech-savvy volunteers possibly skewing results, limited the study's generalizability. This lack of diversity in the sample highlights the need for future studies to include larger and more diverse participant pools to enhance the robustness and generalizability of the findings. Our team is currently working on a multicenter, randomized controlled trial with a mixed methods approach. The study uses a convergent parallel mixed methods design and will span 8 months across multiple medical schools. It will use the new AVM of ChatGPT to simulate an SP. The AVM offers several advantages over the original voice mode, including reduced latency and an ability to inject emotion into its voice [29]. The study aims to draw conclusions based on robust statistical data comparing the average percentage improvement of the experimental group with the control groups on Observed Structured Clinical Examination scores, as well as qualitative data exploring students' learning and perceptions of the AI through focus groups.

Conclusions

This study found ChatGPT to be an effective supplement, although not a full replacement, to traditional SPs. Students and faculty appreciated its potential, noting benefits such as flexible practice times, reduced stress, and improved diagnostic skills. Some shortcomings were noted, including the need for effective prompt engineering and the lack of nonverbal cues affecting realism. Despite these challenges, its reliability and convenience make it a valuable training tool.

Students' diagnostic skills were not formally assessed in this study. However, based on their self-reported perceptions and observations of their interactions with ChatGPT, it appears that the AI can be a valuable tool for practicing clinical reasoning and problem-solving skills. Future research could explore the impact of ChatGPT on students' diagnostic accuracy and clinical performance.

Overall, ChatGPT offers a significant adjunct to traditional SPs, providing accessible, flexible practice opportunities for medical students. The study underscores the importance of integrating prompt engineering into medical curricula and refining AI interactions for balanced information delivery. Continuous

advancements in virtual patient technology and AI capabilities, including improved verbal and auditory flow, are expected to further enhance ChatGPT's use in medical education. Future studies are planned with a larger sample size and using the recently released ChatGPT version 4.o1 with AVM.

Acknowledgments

The authors thank Richard Millis, PhD, for support and guidance during this study. JC was affiliated with the American University of Antigua at the time of the study and is currently affiliated with the Medical University of the Americas. NA was affiliated with the American University of Antigua at the time of the study and is currently affiliated with the School of Medicine, University of Texas. RER and SN were affiliated with the American University of Antigua at the time of the study and are currently unaffiliated.

Authors' Contributions

JC and TK contributed equally to this work. JC conceived the original idea and study design. JC, TK, RER, SD, AV, RH, PH, SN, RS, CL, NA, and JJ refined the study design and conducted the study activity. TK, RS, and RER conducted study participant interviews. AM, AJ, EC, TA, ANN, and JEG contributed to data organization, input, and analysis. JC, TK, and RER contributed to coding, theme construction, and writing original draft of paper. All authors contributed to the review and editing of the paper and approved the submitted version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview questions.
[DOCX File, 14 KB - mededu_v11i1e63353_app1.docx]

References

1. Cleland JA, Abe K, Rethans JJ. The use of simulated patients in medical education: AMEE Guide No 42. Med Teach 2009 Jun;31(6):477-486. [doi: 10.1080/01421590903002821] [Medline: 19811162]
2. Flanagan OL, Cummings KM. Standardized patients in medical education: a review of the literature. Cureus 2023 Jul;15(7):e42027. [doi: 10.7759/cureus.42027] [Medline: 37593270]
3. Mühling T, Schreiner V, Appel M, Leutritz T, König S. Comparing virtual reality-based and traditional physical Objective Structured Clinical Examination (OSCE) stations for clinical competency assessments: randomized controlled trial. J Med Internet Res 2025 Jan 10;27:e55066. [doi: 10.2196/55066] [Medline: 39793025]
4. Shankar PR, Dwivedi NR. Using standardized patients for teaching-learning and assessment in a Caribbean medical school. EIMJ 2015 Jun 10;7(2). [doi: 10.5959/eimj.v7i2.358]
5. Mladenovic J, van Zanten M, Pinsky WW. Evolution of educational commission for foreign medical graduates certification in the absence of the USMLE step 2 clinical skills examination. Acad Med 2023 Apr 1;98(4):444-447. [doi: 10.1097/ACM.0000000000005051] [Medline: 36538680]
6. Phanudulkitti C, Puengrung S, Meepong R, Vanderboll K, Farris KB, Vordenberg SE. A systematic review on the use of virtual patient and computer-based simulation for experiential pharmacy education. Explor Res Clin Soc Pharm 2023 Sep;11:100316. [doi: 10.1016/j.rcsop.2023.100316] [Medline: 37635840]
7. Huang G, Reynolds R, Candler C. Virtual patient simulation at US and Canadian medical schools. Acad Med 2007 May;82(5):446-451. [doi: 10.1097/ACM.0b013e31803e8a0a] [Medline: 17457063]
8. Bubeck S, Chandrasekaran V, Eldan R, et al. Sparks of artificial general intelligence: early experiments with GPT-4. arXiv. Preprint posted online on Mar 22, 2023. [doi: 10.48550/arXiv.2303.12712]
9. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
10. Temsah O, Khan SA, Chaiah Y, et al. Overview of early ChatGPT's presence in medical literature: insights from a hybrid literature review by ChatGPT and human experts. Cureus 2023 Apr;15(4):e37281. [doi: 10.7759/cureus.37281] [Medline: 37038381]
11. About Oscer. Oscer. 2024. URL: https://www.oscer.ai/about [accessed 2024-06-11]
12. Educational affairs - Office of Clinical Education. The University of Texas Medical Branch. 2024. URL: https://www.utmb.edu/som/som-educational-affairs/office-of-clinical-education-(oce)/clerkship---pediatrics/pedi-clerkship [accessed 2025-05-14]
13. Soma Lab. 2024. URL: https://counselor.somalab.ai/signup [accessed 2024-12-16]

14. Body Interact. URL: https://bodyinteract.com/ [accessed 2024-12-15]
15. Liu X, Wu C, Lai R, et al. ChatGPT: when the artificial intelligence meets standardized patients in clinical training. J Transl Med 2023 Jul 6;21(1):447. [doi: 10.1186/s12967-023-04314-0] [Medline: 37415217]
16. Suárez A, Adanero A, Díaz-Flores García V, Freire Y, Algar J. Using a virtual patient via an artificial intelligence chatbot to develop dental students' diagnostic skills. Int J Environ Res Public Health 2022 Jul 18;19(14):8735. [doi: 10.3390/ijerph19148735] [Medline: 35886584]
17. Weidener L, Fischer M. Artificial intelligence teaching as part of medical education: qualitative analysis of expert interviews. JMIR Med Educ 2023 Apr 24;9:e46428. [doi: 10.2196/46428] [Medline: 36946094]
18. Jowsey T, Stokes-Parish J, Singleton R, Todorovic M. Medical education empowered by generative artificial intelligence large language models. Trends Mol Med 2023 Dec;29(12):971-973. [doi: 10.1016/j.molmed.2023.08.012] [Medline: 37718142]
19. Voice control for ChatGPT x MIA AI. Chrome Web Store. 2024. URL: https://chromewebstore.google.com/detail/voice-control-for-chatgpt/eollffkcakegifhacjnlnegohfdlidhn?hl=en&pli=1 [accessed 2025-05-14]
20. Mayring P. Qualitative content analysis theoretical foundation, basic procedures and software solution. In: Approaches to Qualitative Research in Mathematics Education Advances in Mathematics Education 2015. [doi: 10.1007/978-94-017-9181-6_13]
21. Claude. URL: https://claude.ai/new [accessed 2024-12-15]
22. MMLU—wikipedia. Wikipedia Contributors. 2024. URL: https://en.wikipedia.org/wiki/MMLU [accessed 2024-12-16]
23. Learn clinical reasoning—free sign up. Oscer. URL: https://www.oscer.ai/dashboard/home [accessed 2024-12-15]
24. Kavadella A, Dias da Silva MA, Kaklamanos EG, Stamatopoulos V, Giannakopoulos K. Evaluation of ChatGPT's real-life implementation in undergraduate dental education: mixed methods study. JMIR Med Educ 2024 Jan 31;10(1):e51344. [doi: 10.2196/51344] [Medline: 38111256]
25. Magalhães Araujo S, Cruz-Correia R. Incorporating ChatGPT in medical informatics education: mixed methods study on student perceptions and experiential integration proposals. JMIR Preprints. Preprint posted online on Jul 27, 2023. [doi: 10.2196/preprints.51151]
26. Moldt JA, Festl-Wietek T, Madany Mamlouk A, Nieselt K, Fuhl W, Herrmann-Werner A. Chatbots for future docs: exploring medical students' attitudes and knowledge towards artificial intelligence and medical chatbots. Med Educ Online 2023 Dec;28(1):2182659. [doi: 10.1080/10872981.2023.2182659] [Medline: 36855245]
27. Sarangi PK, Panda BB, P S, Pattanayak D, Panda S, Mondal H. Exploring radiology postgraduate students' engagement with large language models for educational purposes: a study of knowledge, attitudes, and practices. Indian J Radiol Imaging 2025 Jan;35(1):35-42. [doi: 10.1055/s-0044-1788605] [Medline: 39697505]
28. Holderried F, Stegemann-Philipps C, Herschbach L, et al. A Generative Pretrained Transformer (GPT)-powered chatbot as a simulated patient to practice history taking: prospective, mixed methods study. JMIR Med Educ 2024 Jan 16;10(1):e53961. [doi: 10.2196/53961] [Medline: 38227363]
29. Barra FL, Costa A, Rodella G, Semeraro F, Carenzo L. Shaping the future of simulator interactions: the role of ChatGPT's advanced voice mode. Resuscitation 2024 Dec;205:110452. [doi: 10.1016/j.resuscitation.2024.110452] [Medline: 39617251]
30. Gray M, Baird A, Sawyer T, et al. Increasing realism and variety of virtual patient dialogues for prenatal counseling education through a novel application of chatgpt: exploratory observational study. JMIR Preprints. Preprint posted online on Jul 17, 2023. [doi: 10.2196/preprints.50705]
31. Webb JJ. Proof of concept: using ChatGPT to teach emergency physicians how to break bad news. Cureus 2023 May;15(5):e38755. [doi: 10.7759/cureus.38755] [Medline: 37303324]
32. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res 2023 Oct 4;25:e50638. [doi: 10.2196/50638] [Medline: 37792434]
33. Heston T, Khun C. Prompt engineering in medical education. Int Med Educ 2023 Aug 31;2(3):198-205. [doi: 10.3390/ime2030019]
34. Aligning language models to follow instructions. Open AI. 2024. URL: https://openai.com/index/instruction-following [accessed 2025-05-04]
35. Our approach to AI safety. Open AI. 2024. URL: https://openai.com/index/our-approach-to-ai-safety [accessed 2025-05-04]
36. Tu T, Palepu A, Schaekermann M, et al. Towards conversational diagnostic AI. arXiv. Preprint posted online on Jan 10, 2024. [doi: 10.48550/arXiv.2401.05654]
37. Hashemi-Sabet F, Leung FH. Equity-driven MCAT prep: a ChatGPT advantage. Can Med Educ J 2024 May;15(2):105-106. [doi: 10.36834/cmej.78169] [Medline: 38827918]
38. Musallam E, Alhaj Ali A, Alkhafaji M. OpenAI's ChatGPT clinical simulation. Nurse Educ 2024;49(6):E361-E362. [doi: 10.1097/NNE.0000000000001657]
39. Safranek CW, Sidamon-Eristoff AE, Gilson A, Chartash D. The role of large language models in medical education: applications and implications. JMIR Med Educ 2023 Aug 14;9:e50945. [doi: 10.2196/50945] [Medline: 37578830]
40. Yanagita Y, Yokokawa D, Fukuzawa F, Uchida S, Uehara T, Ikusaka M. Expert assessment of ChatGPT's ability to generate illness scripts: an evaluative study. BMC Med Educ 2024 May 15;24(1):536. [doi: 10.1186/s12909-024-05534-8] [Medline: 38750546]

41. Wu Y, Zheng Y, Feng B, Yang Y, Kang K, Zhao A. Embracing ChatGPT for medical education: exploring its impact on doctors and medical students. JMIR Med Educ 2024 Apr 10;10:e52483. [doi: 10.2196/52483] [Medline: 38598263]
42. Jiang Y, Fu X, Wang J, et al. Enhancing medical education with chatbots: a randomized controlled trial on standardized patients for colorectal cancer. BMC Med Educ 2024 Dec 20;24(1):1511. [doi: 10.1186/s12909-024-06530-8] [Medline: 39707245]
43. Hersh W. Artificial Intelligence: Implications for Health Professions Education: AMIA Academic Forum; 2024.
44. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. Acad Med 2024 Jan 1;99(1):22-27. [doi: 10.1097/ACM.0000000000005439] [Medline: 37651677]
45. Ray PP, Majumder P. The potential of ChatGPT to transform healthcare and address ethical challenges in artificial intelligence-driven medicine. J Clin Neurol 2023 Sep;19(5):509-511. [doi: 10.3988/jcn.2023.0158] [Medline: 37635433]
46. Xu T, Weng H, Liu F, et al. Current status of ChatGPT use in medical education: potentials, challenges, and strategies. J Med Internet Res 2024 Aug 28;26:e57896. [doi: 10.2196/57896] [Medline: 39196640]
47. McCoy LG, Nagaraj S, Morgado F, Harish V, Das S, Celi LA. What do medical students actually need to know about artificial intelligence? NPJ Digit Med 2020;3:86. [doi: 10.1038/s41746-020-0294-7] [Medline: 32577533]
48. Gupta N, Khatri K, Malik Y, et al. Exploring prospects, hurdles, and road ahead for generative artificial intelligence in orthopedic education and training. BMC Med Educ 2024 Dec 28;24(1):1544. [doi: 10.1186/s12909-024-06592-8] [Medline: 39732679]

Abbreviations

AI: artificial intelligence
AUA: American University of Antigua
AVM: advanced voice mode
LLM: large language model
MUA: Medical University of the Americas
SP: standardized patient

Edited by D Chartash; submitted 20.06.24; peer-reviewed by PK Sarangi, R Saifi, SM Araujo; revised version received 18.03.25; accepted 18.03.25; published 20.05.25.

Please cite as:
Cross J, Kayalackakom T, Robinson RE, Vaughans A, Sebastian R, Hood R, Lewis C, Devaraju S, Honnavar P, Naik S, Joseph J, Anand N, Mohammed A, Johnson A, Cohen E, Adeniji T, Nnenna Nnaji A, George JE
Assessing ChatGPT's Capability as a New Age Standardized Patient: Qualitative Study
JMIR Med Educ 2025;11:e63353
URL: https://mededu.jmir.org/2025/1/e63353
doi:10.2196/63353

Original Paper

Virtual Reality Simulation for Undergraduate Nursing Students for Care of Patients With Infectious Diseases: Mixed Methods Study

Wen Chang[1,2], PhD; Chun-Chih Lin[3,4], PhD; Julia Crilly[5,6], PhD; Hui-Ling Lee[1], MSN; Li-Chin Chen[1,2,4], MSN; Chin-Yen Han[1,4], PhD

[1]Department of Nursing, Chang Gung University of Science and Technology, Taoyuan City, Taiwan

[2]Nursing Management Department of Administration Center, Chang Gung Medical Foundation, Taoyuan City, Taiwan

[3]Department of Nursing, Chang Gung University of Science and Technology, Chiayi County, Taiwan

[4]Department of Nursing, New Taipei Municipal TuCheng Hospital (Built and Operated by Chang Gung Medical Foundation), New Taipei City, Taiwan

[5]Department of Emergency Medicine, Gold Coast Health, Southport, Australia

[6]School of Nursing and Midwifery, Griffith University, Southport, Australia

Corresponding Author:
Chin-Yen Han, PhD
Department of Nursing
Chang Gung University of Science and Technology
261
Wenhua 1st Road, Guishan Dist.
Taoyuan City, 33303
Taiwan
Phone: 886 3 2118999 ext 3417
Fax: 886 3 2118866
Email: cyhan@mail.cgust.edu.tw

Abstract

Background: Virtual reality simulation (VRS) teaching offers nursing students a safe, immersive learning environment with immediate feedback, enhancing learning outcomes. Before the COVID-19 pandemic, nursing students had limited training and opportunities to care for patients in isolation units with infectious diseases. However, the pandemic highlighted the ongoing global priority of providing care for patients with infectious diseases.

Objective: This study aims to (1) examine the effectiveness of VRS in preparing nursing students to care for patients with infectious diseases by assessing its impact on their theoretical knowledge, learning motivation, and attitudes; and (2) evaluate their experiences with VRS.

Methods: This 2-phased mixed methods study recruited third-year undergraduate nursing students enrolled in the Integrated Emergency and Critical Care course at a university in Taiwan. Phase 1 used a quasi-experimental design to address objective 1 by comparing the learning outcomes of students in the VRS teaching program (experimental group) with those in the traditional teaching program (control group). Tools included an infection control written test, the Instructional Materials Motivation Survey, and a learning attitude questionnaire. The experimental group participated in a VRS lesson titled “Caring for a Patient with COVID-19 in the Negative Pressure Unit” as part of the infection control unit. In phase 2, semistructured interviews were conducted to address objective 2, exploring students’ learning experiences.

Results: A total of 107 students participated in phase 1, and 18 students participated in phase 2. Both the VRS and control groups showed significant improvements in theoretical knowledge scores (for the VRS group t_{46}=–7.47; P<.001, for the control group t_{59}=–4.04; P<.001). However, compared with the control group, the VRS group achieved significantly higher theoretical knowledge scores ($t_{98.13}$=2.70; P=.008) and greater learning attention (t_{105}=2.30; P=.02) at T1. Additionally, the VRS group demonstrated a statistically significant higher regression coefficient for learning confidence compared with the control group (β=.29; P=.03). The students’ learning experiences in the VRS group were categorized into 4 themes: Applying Professional Knowledge to Patient Care, Enhancing Infection Control Skills, Demonstrating Patient Care Confidence, and Engaging in Real Clinical Cases. The core theme identified was Strengthening Clinical Patient Care Competencies.

Conclusions: The findings suggest that VRS teaching significantly enhanced undergraduate nursing students’ infection control knowledge, learning attention, and confidence. Qualitative insights reinforced the quantitative results, highlighting the holistic

benefits of VRS teaching in nursing education, including improved learning outcomes. The positive impact on student motivation and attitudes indicates a potentially transformative approach to nursing education, particularly in the post–COVID-19 era, where digital and remote learning tools play an increasingly vital role.

(JMIR Med Educ 2025;11:e64780) doi:10.2196/64780

KEYWORDS

virtual reality; infection control; learning motivation; learning attitudes; nursing education

Introduction

Background

Virtual reality (VR) has emerged as an innovative teaching strategy in nursing education. VR technology leverages simulated scenarios to overcome time and space limitations, offering students opportunities to learn in safe, realistic settings and receive immediate feedback [1]. VR simulation (VRS) teaching strategies enhance learning motivation, student immersion, knowledge and skill acquisition, confidence [2,3], active participation, and learning effectiveness [4-6]. The goal of undergraduate nursing education is to prepare students for clinical practice, making it essential to strengthen their professional competencies and attitudes. Integrating information technology into nursing education enhances students' learning outcomes. Nursing education should align with the broader clinical practice environment, incorporating technology to support students in developing their competencies [7].

The COVID-19 pandemic has profoundly impacted nursing curricula and teaching worldwide. In emergency and critical care, university-level nursing curricula must reflect clinical environments. Emphasizing situated learning enhances students' abilities and confidence in providing emergency patient care [8,9]. Before the pandemic, nursing students rarely had opportunities to care for patients with infectious diseases in isolation units. However, the demand for care related to infectious diseases remains a global priority [9]. Strengthening courses on infectious diseases can help students develop positive attitudes toward clinical practice [10]. Updating infectious disease courses with more practical experiences can further support nursing students in developing positive attitudes when caring for patients with infectious diseases during clinical practice.

Learning theories related to VR teaching include constructivism, situated learning, and experiential learning. In VR learning, learners actively absorb information and construct new knowledge [11]. Situated learning theory emphasizes real-world interactions and activities in authentic contexts, transforming these experiences into applicable knowledge [12]. VR offers an interactive virtual environment, using visual effects to present abstract problems and providing opportunities for active manipulation and repeated practice [11]. Experiential learning theory posits that learning is the transformation of experience, with knowledge creation emerging from interactions, conflicts, and problem-solving between individuals and their environment. This theory highlights the potential of immersive technology to provide meaningful experiences [12]. Compared with other teaching methods, VR teaching is easy to use, generates positive and active learning experiences [13], and enhances learning outcomes, including improvements in knowledge, skills, and clinical decision-making [14,15]. Engagement in VR environments provides students with experiences closely aligned with clinical practice, boosting their motivation and attitudes and leading to better educational outcomes [14,15].

Motivation and attitude play a significant role in influencing learning outcomes. Enhanced motivation strengthens active learning and improves results [12,16]. Studies have shown a positive correlation between motivation and learning outcomes, making learning easier and fostering proactive engagement [17,18]. Keller's Attention, Relevance, Confidence, and Satisfaction (ARCS) model of motivation incorporates a learning motivation scale to assess motivational aspects within a course [12,16]. Designing courses with integrated motivational models can inspire learners, enhance motivation, and increase classroom engagement [19,20]. In nursing education, particularly in emergency and critical care courses, VR can address the limitations of clinical settings and traditional teaching methods caused by resource constraints [21-24]. VR stimulates learners' motivation, promotes active participation, and enhances learning outcomes [25,26]. By incorporating VR teaching, courses can more closely align with clinical practice, providing students with a solid foundation in professional knowledge and skills.

Even long after the pandemic, there will remain a global need for the care of patients with infectious diseases. However, opportunities for students to participate in the actual care of such patients in isolation units remain limited. To date, little attention has been given in the literature to identifying educational strategies that address this gap in developing nursing students' professional knowledge and skills. This mixed methods study was guided by 2 research questions: (1) What is the effectiveness of VRS teaching on nursing students' theoretical knowledge, learning motivation, and attitudes toward the care of patients with infectious diseases? and (2) What are the learning experiences of nursing students in a VRS program? Our a priori hypothesis was that VRS teaching would significantly improve nursing students' infection control theoretical knowledge, learning motivation, and attitudes regarding the care of patients with infectious diseases.

Objectives

This study had 2 objectives: (1) to evaluate the effectiveness of VRS teaching on nursing students' theoretical knowledge, learning motivation, and attitudes toward the care of patients with infectious diseases, and (2) to explore their learning experiences in a VRS program designed for this target population.

Methods

Study Design

This study used a 2-phased mixed methods approach to comprehensively evaluate a VRS teaching program on the care of patients with infectious diseases, which was part of the infection control unit within the Integrated Emergency and Critical Care course. Phase 1 utilized a quantitative study design to assess the learning effectiveness of the VRS teaching method, while phase 2 used qualitative phenomenography to explore students' experiences and perceptions of the program.

Phase 1: Outcomes of the VRS Program on Students' Knowledge, Learning Motivation, and Attitudes to the Care of Patients With Infectious Diseases

Overview

A quasi-experimental design was used to compare learning outcomes—knowledge, motivation, and attitude—between students in the VRS teaching program (experimental group) and those in the traditional teaching course (control group). Data were collected from August 2022 to July 2023.

Participants

This study used convenience sampling and was conducted at a clinical competence center at a university in Taiwan. Third-year undergraduate nursing students enrolled in the Integrated Emergency and Critical Care course were eligible to participate. One class of students was assigned to the experimental group, and another to the control group. The Integrated Emergency and Critical Care course is an elective offered in both the first and second semesters. Researchers used a random selection process to assign students in the infection control unit to the VRS program in the first semester and to traditional teaching in the second semester. The participating school provided a 2-week add/drop period, during which members of the research team gave in-class briefings about the study, and students were free to choose whether to participate in the experimental group. The selection criteria for the experimental group were (1) aged ≥20 years, (2) enrolled in the Integrated Emergency and Critical Care course, and (3) willing to participate in this study. Students with a history of epilepsy were excluded from the VRS. Sample size estimation was conducted using G*Power software version 3.1 [27]. Following Cohen's rule [28], 2 groups were included, with a medium effect size of f=0.25, a correlation of 0.5, a power of 0.8, and an α value of .05, resulting in a required sample size of ≥86, with ≥43 participants per group. A total of 47 students were recruited for the experimental group. None of the students in the experimental group refused to participate in the VRS program. All participants were taught by the same instructor, and the course content was consistent across both groups.

Instruments

Infection Control Written Test

Previous research has shown that VRS teaching can enhance the development of both knowledge and practical skills in undergraduate nursing students, with outcomes effectively assessed using a written test [29]. In this study, the infection control knowledge assessment involved a written test administered to students before (T0) and after (T1) the infection control lesson. The test consisted of 10 questions aligned with the learning objectives of the infection control unit. These included single- and multiple-choice items covering both theoretical knowledge and practical skills, such as donning and doffing personal protective equipment (PPE). The test addressed the same key infection control techniques for all students, aiming to evaluate their baseline abilities and the changes in knowledge following the lesson. To better capture postlearning changes and minimize the influence of memory recall on the posttest results, the order of the questions was adjusted, and some questions were modified. The test items were reviewed by the course instructor and clinical experts (senior emergency nurses) to ensure content validity.

Instructional Materials Motivation Survey

The Instructional Materials Motivation Survey (IMMS) is a self-reported questionnaire administered before (T0) and after (T1) the infection control lesson. Designed primarily to evaluate students' motivation in learning a course [12], the IMMS comprises 36 items distributed across 4 subscales based on the ARCS motivation model: Attention (12 items), Relevance (9 items), Confidence (9 items), and Satisfaction (6 items). Each item is rated on a 5-point Likert scale, with higher scores indicating greater learning motivation. The original IMMS scale has demonstrated high reliability, with Cronbach α values ranging from 0.81 to 0.96 [12]. In this study, the IMMS exhibited excellent reliability, with a Cronbach α of 0.94.

Learning Attitude Questionnaire

A learning attitude questionnaire was administered at T0 and T1. This 20-item self-reported questionnaire was developed by several members of the research team to assess students' attitudes toward caring for patients with infectious diseases and their participation in the infection control unit. Items were rated on a 5-point Likert scale, where 1 indicates "strongly disagree," 2 "disagree," 3 "neutral," 4 "agree," and 5 "strongly agree." Higher scores reflected a more positive learning attitude. The questionnaire demonstrated strong validity and reliability, with an average Content Validity Index of 0.9 and a Cronbach α of 0.955.

VRS Lesson Plan for Caring for a Patient With COVID-19 in the Negative Pressure Unit

In the infection control unit, VRS teaching was implemented for the experimental group. The VRS scenario, developed by several research team members with VR training certification, depicted a real clinical case of a febrile patient visiting the emergency department for triage, later confirmed to have COVID-19, and subsequently admitted to a negative pressure isolation unit (Figure 1). The teaching content emphasized a nurse's role in providing care within a negative pressure isolation room, including the proper techniques for donning and doffing PPE. The lesson's learning objectives were for students to differentiate care for patients with infectious diseases, correctly don and doff PPE, and provide appropriate patient care. The VR Oculus Quest equipment, including a headset and controllers, was supplied by the School of Nursing of the participating university. The VRS lesson plan was reviewed by the course instructor and clinical experts (senior emergency

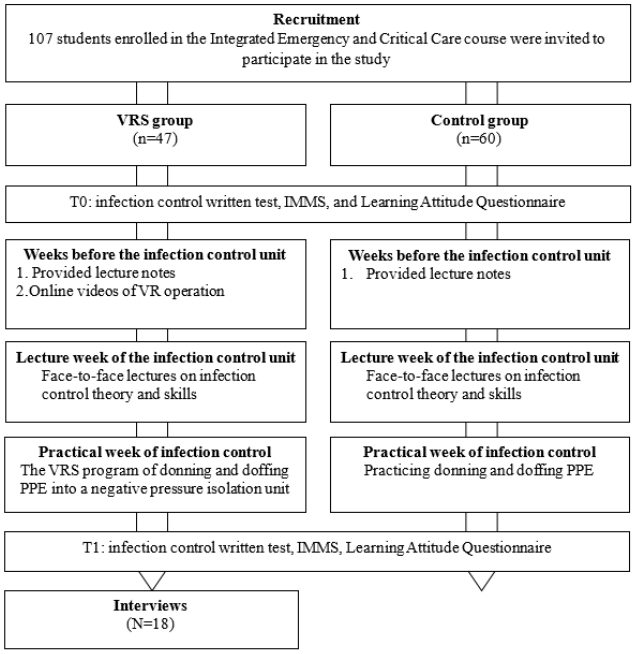
nurses) to ensure content validity. The lesson utilized VR technology to deliver immersive visual effects and interactive scenarios, aiming to enhance students' awareness and provide opportunities for practical exercises, thereby improving learning outcomes [30]. The assessment included the standard procedure for applying PPE, such as N95 masks, goggles, hair caps, and gloves. Upon completing the assessment, students received immediate feedback, with the computer screen highlighting missed items. This direct feedback was intended to reinforce learning effectiveness [1].

Figure 1. Screen captures of virtual reality simulation videos.

Procedure

To maintain neutrality, none of the research team members were involved in teaching either the experimental or control group. Instead, several team members focused on designing the VRS program and creating a VR system operation video to help students become familiar with operating the VR system. During the experimental group's class, research team members were available to address any technical issues that participants encountered. They also met with the unit instructors before the start of the unit to ensure consistency in teaching between the 2 groups and alignment in the course delivery process. The study protocol is illustrated in Figure 2.

Figure 2. Study protocol process. IMMS: Instructional Materials Motivation Survey; PPE: personal protective equipment; VR: virtual reality; VRS: virtual reality simulation.



The infection control unit spanned 2 weeks and included 1 lecture (100 minutes) and 1 practical session (100 minutes). Both the VRS group and the control group received lecture notes before the start of the course. Additionally, the VRS group was provided with prerecorded online VR videos demonstrating donning and doffing PPE in a negative pressure isolation unit, as well as instructions on operating the VR system. For the VRS group, the first week consisted of a 100-minute lesson on

infection control theory and skills, while the second week involved 100 minutes of VRS instruction. The VRS class was divided into 5 groups, each comprising 8-10 students, who worked collaboratively on the drills. The lesson began with an introduction to VR system operation (5-10 minutes), followed by group VRS scenario drills (30-40 minutes). Each student then executed their part of the VRS lesson, which lasted approximately 6-8 minutes. Upon completion, the system provided feedback, serving as the students' learning outcomes. Group members first discussed the session among themselves, followed by a 10-minute instructor-led debriefing session. During this session, students were encouraged to ask questions and share their reflections on the VRS program execution. Feedback and reflection were incorporated to help students consolidate their learning and transform it into meaningful learning outcomes. The groups then switched roles and conducted a second round of drills and discussions for another 30-40 minutes. A pretest (T0) and posttest (T1) on infection scenario cases were conducted to evaluate the students' learning outcomes, motivation, attitudes, and knowledge related to the infection control unit. For the control group, a traditional teaching strategy was used. During the first week, theoretical lectures were delivered, accompanied by a video to aid students in understanding the process. Lecture notes, identical to those provided to the VRS group, were distributed before class and included a video link demonstrating the standard PPE procedure. In the second week, students were divided into 6 groups of 9-11 members to practice donning and doffing PPE. Instructors provided individualized guidance to correct mistakes. Within the same groups, students evaluated and discussed the PPE practice. Although an instructor-led debriefing session was planned for the control group, it was postponed to the following week due to the large number of students and time constraints.

Data collected for this study were individually coded and entered into a computer for analysis using SPSS version 22.0 (IBM Corp.). Descriptive statistics, including frequency, percentage, mean, and SD, were calculated. Inferential statistics, such as independent *t* tests, paired *t* tests, and generalized estimating equations, were applied. Results with a *P* value of <.05 were considered statistically significant.

Phase 2: The Students' Learning Experiences of the VRS in Caring for Patients With Infectious Diseases

Overview

Phase 2 utilized qualitative phenomenography to explore students' experiences and perceptions of the VRS program. The key concepts in phenomenography are "phenomenon" and "experience." This methodology aims to identify the shared and generalized aspects of participants' thoughts or concepts regarding their experiences of a specific phenomenon, with a focus on describing their understanding of these experiences [31]. In this study, phenomenography was applied to understand how learners organize and structure the content they acquire during the learning process [31]. Interviews with students were analyzed to uncover their learning experiences and outcomes, with the goal of providing evidence to support the ongoing improvement of educational programs.

Participants

Students in the experimental group who met the following inclusion criteria were recruited: (1) aged 20 years or older, (2) enrolled in the Emergency and Critical Care course and participating in VRS teaching, and (3) consenting to participate in and have interviews recorded.

Procedure

Participants took part in in-depth semistructured interviews. These interviews facilitate meaningful conversations, providing a detailed understanding of complex issues [31]. In a phenomenographic study, interview questions need to be as open-ended as possible to accurately capture the participants' thoughts. The interview guide is provided in Multimedia Appendix 1. The in-depth interviews in this study contributed to understanding nursing students' experiences with VRS teaching. Each eligible participant received a consent form outlining the study's purpose, the voluntary nature of participation, and the confidentiality of their data. Participants completed the consent forms, and suitable interview times were scheduled. The interviews were conducted by a single researcher (CYH) who had prior experience in qualitative research gained during doctoral studies, had served as a principal investigator on research projects, and had published several qualitative research articles. CYH was not involved in teaching this subject and was unfamiliar with the participants in the experimental group. During the interviews, CYH utilized interview skills to encourage participants to articulate their VR learning experiences. The interviews were audio-recorded and lasted between 42 and 62 minutes. A sample size of 18 students was sufficient to generate rich data and achieve saturation.

Qualitative Analysis and Trustworthiness

Following each interview, the same researcher (CYH) transcribed the audio recordings verbatim to ensure detailed documentation and analyzed the interview data. Data analysis followed the 7 steps of phenomenographic analysis: familiarization, condensation, comparison, grouping, articulating, labeling, and contrasting [32]. The trustworthiness of the research findings was established using Lincoln and Guba's [33] criteria of credibility, transferability, dependability, and confirmability. In terms of credibility, phenomenographic research emphasizes the precise description of each stage of the study process, the application of the researcher's ideas to the phenomena, the careful formulation of interview questions and processes, and the thorough analysis and presentation of conclusions. Peer debriefing, which involves collaborative data analysis to explore diverse interpretations, enhances data interpretation and credibility, contributing to the development of credible research outcomes. Transferability is supported by providing in-depth data that represent a comprehensive view of the research, highlighting its relevance and context. Dependability is ensured by supporting categorizations with excerpted interview content, illustrating the similarities and differences among participants in relation to the phenomenon, and confirming the logical connection between the collected data and the phenomena captured by the descriptive categorization. Confirmability is established by documenting the interviewer's feelings and thoughts during the interview

process, thereby creating an audit trail. The data analysis is thoroughly described, with detailed records of decisions made and strategies adopted during concept formation. These reflections on theoretical and methodological aspects further contribute to the audit trail and the confirmability of the findings [34].

Ethical Consideration

Ethical approval was obtained from the Institutional Review Board of Chang Gung Medical Foundation (approval number 202002386B0). Potential participants were fully informed about the nature and purpose of the study, emphasizing that participation was entirely voluntary and that they had the right to withdraw from the study at any time. They were explicitly assured that their academic results would not be affected by their decision to participate or not. Participants were also guaranteed that their data would remain confidential and that they would not be identifiable in any reports. All participants provided written informed consent, and none of the students withdrew from the study.

Results

Phase 1: Outcomes of the VRS Program on Students' Knowledge, Learning Motivation, and Attitudes to the Care of Patients With Infectious Diseases

Overview

Participants in phase 1 consisted of 107 third-year undergraduate students: 47 in the experimental group, who received VRS teaching, and 60 in the control group, who participated in traditional practical sessions on donning and doffing isolation gowns. As shown in Table 1, the majority of participants (99/107, 92.5%) were female, with an average age of 21.14 (SD 0.69) years.

Table 1. Demographic characteristics of participants in phase 1 (N=107).

Participant demographics	Study groups		Total participants (N=107)
	Virtual reality simulation (n=47)	Control (n=60)	
Gender, n (%)			
Male	2 (4.3)	6 (10)	8 (7.5)
Female	45 (95.7)	54 (90)	99 (92.5)
Age (years), mean (SD)	N/A[a]	N/A	21.14 (0.69)
Male	22.00 (1.41)	20.50 (0.55)	20.88 (0.99)
Female	21.40 (0.54)	20.96 (0.70)	21.16 (0.67)

[a]N/A: not applicable.

Effectiveness of VRS in Infection Control Theoretical Knowledge

Pre- and posttest assessments of infection control knowledge were conducted at T0 and T1, with a maximum score of 10 points. The combined results of all participants showed an average pretest knowledge score of 7.58 (SD 1.13) and a posttest knowledge score of 8.58 (SD 1.16), indicating improved knowledge after course completion (t_{106}=–7.08; P<.001). Both the VRS and control groups demonstrated significant improvements in theoretical knowledge scores (for the VRS group t_{46}=–.747; P<.001 and for the control group t_{58}=–4.04; P<.001). However, the VRS group achieved significantly higher posttest scores compared with the control group ($t_{98.13}$=2.70; P=.008), suggesting that VRS teaching was more effective in enhancing students' knowledge (Table 2 and Figure 3).

Figure 3. Change in infection control knowledge in the two groups. T0: pretest; T1: posttest; **: *P*<.01.

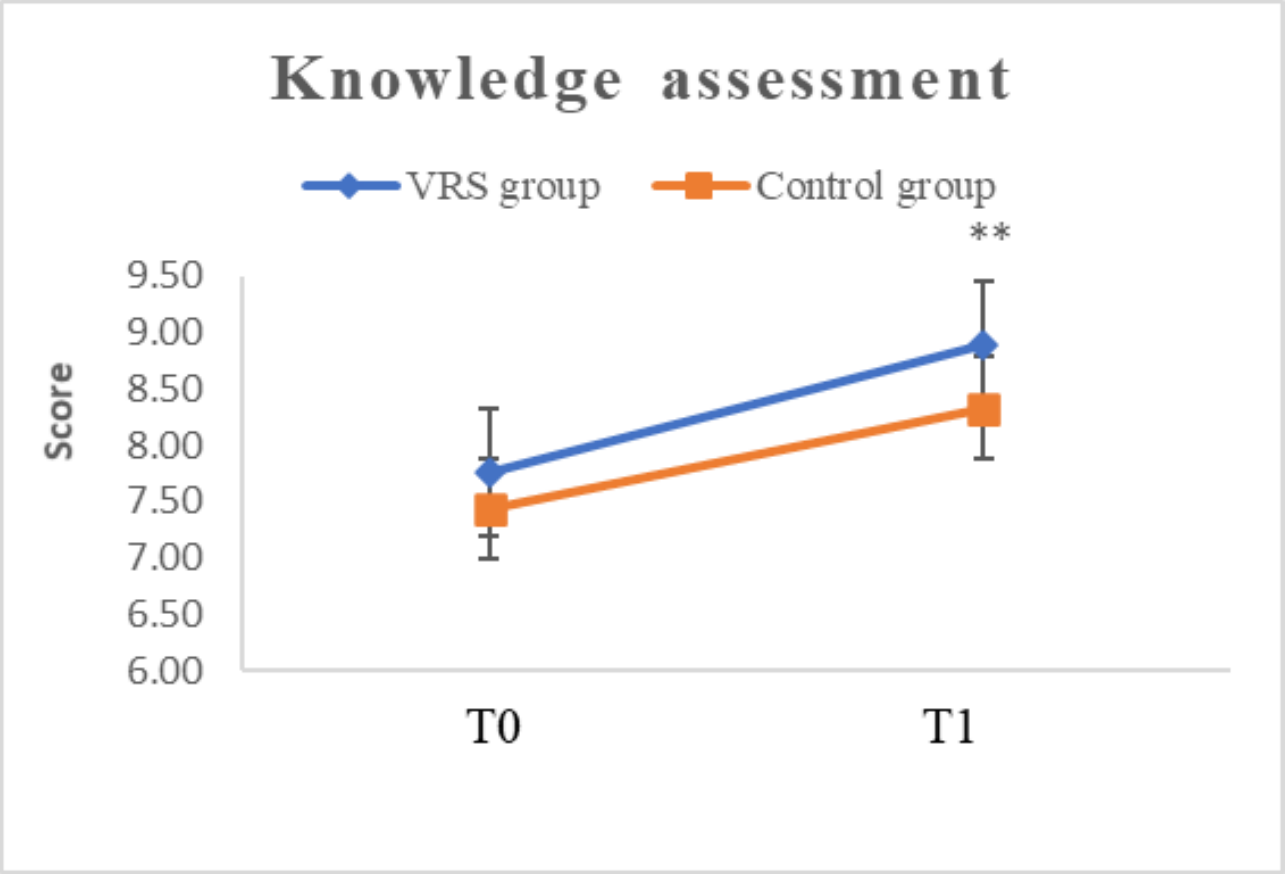


Table 2. Comparison of theoretical knowledge scores in the 2 groups (N=107).

Variable	T0[a], mean (SD)	T1[b], mean (SD)	*t* test (*df*)	*P* value
Knowledge assessment				
Virtual reality simulation	7.77 (1.05)	8.89 (0.79)	–7.47 (46)	<.001
Control	7.43 (1.18)	8.33 (1.34)	–4.04 (58)	<.001
t test (*df*)	1.52 (105)	2.70 (98.13)	N/A[c]	N/A
P value	.13	.008[d]	N/A	N/A
Total	7.58 (1.13)	8.58 (1.16)	–7.08 (106)	<.001

[c]N/A: not applicable.

Effectiveness of VRS on Learning Motivation

The learning motivation of all students increased slightly from T0 (mean 3.84, SD 0.47) to T1 (mean 3.94, SD 0.40) (t_{106}=–3.10; *P*=.002), with no significant differences between the groups at either T0 ($t_{76.50}$=0.09; *P*=.93) or T1 ($t_{80.95}$=1.43; *P*=.16). At T0, except for the Confidence dimension—which was lower in the VRS group compared with the control group ($t_{78.53}$=–2.12; *P*=.04)—the other dimensions of the ARCS model (ie, Attention, Relevance, and Satisfaction) did not differ significantly between the groups. At T1, only the Attention dimension differed significantly between the VRS and control groups, being higher in the VRS group (t_{105}=2.30; *P*=.02), as shown in Table 3.

Table 3. Comparison of learning outcomes between the 2 groups at different time points (N=107).

Variable and groups	T0[a]			T1[b]		
	Mean (SD)	*t* test (*df*)	*P* value	Mean (SD)	*t* test (*df*)	*P* value
Motivation		0.09 (76.50)	.93		1.43 (80.95)	.16
Virtual reality simulation	3.84 (0.57)			4.01 (0.46)		
Control	3.83 (0.38)			3.89 (0.34)		
Attention		0.91 (105)	.36		2.30 (105)	.02[c]
Virtual reality simulation	3.84 (0.67)			4.03 (0.52)		
Control	3.74 (0.48)			3.82 (0.42)		
Relevance		0.76 (82.99)	.45		0.88 (105)	.38
Virtual reality simulation	4.15 (0.54)			4.21 (0.48)		
Control	4.07 (0.41)			4.14 (0.37)		
Confidence		–2.12 (78.53)	.04[c]		0.86 (105)	.39
Virtual reality simulation	3.45 (0.58)			3.65 (0.49)		
Control	3.66 (0.40)			3.58 (0.43)		
Satisfaction		0.51 (78.30)	.61		0.48 (105)	.63
Virtual reality simulation	3.98 (0.64)			4.18 (0.51)		
Control	3.92 (0.44)			4.13 (0.50)		
Attitude		–0.60 (105)	.55		0.03 (105)	.98
Virtual reality simulation	4.09 (0.57)			4.34 (0.52)		
Control	4.15 (0.50)			4.34 (0.56)		

[a]T0: pretest.

[b]T1: posttest.

[c]*P*<.05.

To address potential bias stemming from the differences in the Confidence dimension between the VRS and control groups at T0, the Generalized Estimating Equations model was applied to analyze and compare changes in both groups throughout the study period and to evaluate the outcomes of the VRS intervention. For the VRS group, the regression coefficients for the Confidence dimension were significant (β=.29; *P*=.03), with positive parameter estimates compared with the control group. This finding indicates that the VRS intervention enhanced students' learning confidence.

Effectiveness of VRS on Learning Attitude

The learning attitude score increased slightly in the VRS group from T0 (mean 4.09, SD 0.57) to T1 (mean 4.34, SD 0.52) and in the control group from T0 (mean 4.15, SD 0.50) to T1 (mean 4.34, SD 0.56). However, no significant differences were observed between the groups at either T0 or T1, as shown in Table 3.

Phase 2: The Students' Learning Experiences of the VRS in Caring for Patients With Infectious Diseases

Overview

In phase 2 of this study, 18 students from the VRS group who had expressed willingness to be interviewed were recruited for qualitative interviews. All interview participants were female. Data analysis followed the phenomenographic steps of familiarization, condensation, comparison, grouping, articulating, labeling, and contrasting [32]. Each theme elicited from the participants' pool of meaning represented a concept of their learning experiences associated with engaging in the VRS program. The core theme captured the relationship between each theme and participants' overall understanding of their VRS learning experiences. The students' learning experiences were categorized into 4 themes: (1) Application of Professional Knowledge to Patient Care, (2) Enhanced Infection Control Skills, (3) Demonstrated Confidence in Patient Care, and (4) Participation in Real Clinical Cases. The core theme was identified as Strengthening Clinical Patient Care Competencies.

Theme 1: Application of Professional Knowledge to Patient Care

The students described how they applied their theoretical knowledge of infection control during the VRS teaching process, particularly regarding the various factors that need to be considered when entering and exiting the negative pressure isolation unit. Through the feedback and debriefing provided by the VRS teaching, they were able to reflect on the content of their infection control learning, thereby deepening their professional knowledge in this area.

Theme 2: Enhanced Infection Control Skills

The participants shared their experiences of using VRS to practice the care skills learned in the infection control unit. Through hands-on practice and observing their classmates, they noted improvements in their skills. They also reported that the course's practical and interactive scenarios enhanced their learning interest, which translated into positive learning outcomes.

Theme 3: Participation in Real Clinical Cases

Most of the students in the integrated care course for Emergency and Critical Care expressed a desire to intern or work in emergency departments or intensive care units. However, during the pandemic, many hospitals' critical care units halted the acceptance of nursing interns or prevented students from participating in the care of patients with infectious diseases during their ward internships. The participants reported that the realistic, scenario-based case studies in the VRS enabled them to practice clinical skills that would otherwise have been unavailable, thereby bridging the gap between theory and practice. The experience of providing care for simulated patients in a context closely resembling clinical settings enhanced their learning experience.

Theme 4: Demonstrated Confidence in Patient Care

During the interviews, participants shared that they had been concerned about their ability to provide effective clinical care for patients with infectious diseases due to the impact of the pandemic. However, after participating in the VRS course, they reported a boost in their confidence in providing this type of care.

Core Theme: Strengthening Clinical Patient Care Competencies

The core theme that emerged from the analysis of the qualitative data on students' learning experiences in the VRS teaching program was the strengthening of their clinical care competencies. The VRS learning program allowed students to apply the professional knowledge and skills they had learned in the course to carefully design patient scenarios. Through VRS practical exercises, students improved the skills required to care for patients with infectious diseases. By engaging with clinical cases and performing learning tasks in a realistic setting, they gained greater confidence in caring for patients with infectious diseases. In other words, this approach of connecting learning experiences enhanced their clinical care competence, better preparing them for the future care of patients with infectious diseases.

Discussion

Principal Findings

The results of this study showed significant improvements in infection control knowledge scores in both groups, with the VRS group achieving higher scores, highlighting the effectiveness of VRS teaching in enhancing theoretical knowledge. The VRS group also achieved a higher attention score at T1 compared with the control group. Additionally, the VRS intervention enhanced students' learning confidence. Students' reflections on their learning experiences and perceptions of the VRS teaching emphasized the following themes: Application of Professional Knowledge to Patient Care, Enhanced Infection Control Skills, Demonstrated Confidence in Patient Care, Participation in Real Clinical Cases, and Strengthening Clinical Patient Care Competencies.

This study made every effort to control variables to ensure consistency in learning content and teaching quality between the 2 groups, including the use of the same teaching materials and the same instructor for both groups. The VRS provided immediate feedback, allowing students to actively engage in the care process within a VR patient scenario, which contributed to enhanced learning outcomes in the VRS group. However, due to the larger number of students in the control group, their debriefing session was delayed. Future studies could investigate the impact of debriefing timing on the effectiveness of VRS-based teaching.

Comparison With Prior Work: Effectiveness of VRS in Infection Control Knowledge

The results of this study demonstrated significant improvement in infection control knowledge in both groups after the learning process. The VRS group achieved significantly higher scores on the infection control written test compared with the control group at T1, indicating that VRS teaching was more effective in enhancing students' theoretical knowledge. This finding aligns with previous research. Systematic reviews and meta-analyses on VR in nursing education have demonstrated its effectiveness in improving knowledge [5,35]. Another review reported a moderate effect size (g=0.47) for VR teaching in knowledge acquisition [29]. Additionally, this review noted that subgroup analysis showed VR training involving multiple self-practice sessions of less than 30 minutes was effective in imparting procedural knowledge to undergraduate nursing students [29]. This finding is consistent with the results of our study, where each student engaged in VRS learning for 6-8 minutes, with the option for continued practice for those wishing to further develop their skills. An integrative review also concluded that VRS teaching is effective in enhancing the acquisition of clinical skills and knowledge [36]. An extensive review of 29 randomized controlled trials involving 2722 students found that VR, augmented reality, and mixed reality were as effective as traditional methods in enhancing knowledge, highlighting their potential role in preclinical education [37]. Similarly, a German study on teaching tracheal suction skills observed no statistical differences among various teaching methods in terms of knowledge and skill improvement, suggesting that VR can serve as a supplementary resource to existing learning strategies, supporting students in preparing for clinical practice [23].

Impact of VRS on Learning Motivation and Attitude

A systematic review and meta-analysis of 26 studies found no significant impact of VR on nursing students' motivation and cognitive load compared with traditional teaching methods [38]. This aligns with the findings of our study, which also showed no significant difference in learning motivation. However, other studies have reported higher motivation and satisfaction with VR, though it may also increase cognitive load [39].

Additionally, several studies have shown that VR positively impacts learners by enhancing attention and motivation, building self-efficacy, and reinforcing learning confidence and performance [5,6,21,25]. A Taiwanese study comparing traditional and VR teaching on nasogastric tube feeding found nonsignificant higher scores in the VR group, which demonstrated greater motivation and satisfaction but also experienced a higher cognitive load [39]. These findings highlight the need to carefully consider cognitive load in future course designs [39].

A South Korean study reported higher neonatal resuscitation knowledge, motivation, problem-solving skills, and confidence in the VR group compared with the control groups, along with lower anxiety levels [26]. An integrative review on VR teaching for emergency patients also revealed increased confidence in handling emergencies [25]. Although some studies have reported no significant differences in anxiety and confidence [3,40], further research is needed to determine VR's impact on learning confidence and stress. A Chinese study on disaster nursing courses found significant improvements in preparedness, confidence, and performance in the experimental group, highlighting VR's potential as a cost-effective simulation method [41]. Technical issues with VR were noted as disadvantages, which may explain the lower precourse confidence in the VRS group compared with the control group. Ensuring that students are familiar with the VR system before the course begins may help improve their confidence [1,27,42,43].

Student's Learning Experiences of VRS

Analysis of the qualitative data obtained in this study revealed 4 themes in students' experiences and perceptions of VR learning: Application of Professional Knowledge to Patient Care, Enhanced Infection Control Skills, Demonstrated Confidence in Patient Care, and Participation in Real Clinical Cases. The core theme identified was the Strengthening Clinical Patient Care Competencies. Similarly, previous qualitative studies have demonstrated a positive impact of VR learning on knowledge [23,24,43], skills [21,23,24], confidence [23], and engagement [4]. Some of these studies focused on VR learning as a tool, while others examined the characteristics of the VR environment [21,23,43]. By contrast, our study applied a phenomenographic methodology to explore students' experiences and perceptions of VR learning, linking these to various outcomes. As a result, our findings provide unique insights into students' conceptions of VR learning.

Strengths and Limitations

The strength of this study lies in its combination of quantitative and qualitative methods, providing a comprehensive understanding of the effectiveness of VRS teaching. However, some potential limitations and weaknesses should be considered. Although the sample size was adequate, it may not fully represent the diversity of nursing students, as it was drawn from only 1 university. The course in which the VRS program was applied was an elective unit offered in both the first and second semesters, with a maximum enrollment of 60 students per class. The number of students enrolled in each class was beyond the researchers' control, resulting in an imbalance in the number of students between the 2 groups. It is recommended that future studies compare the effectiveness of VRS using groups of equal size and a larger number of participants. This study did not conduct a formal survey on potential side effects among students in the experimental group. However, during the VRS session, research team members and the instructor periodically checked in with the students. None of the students reported any discomfort that required them to pause or stop the activity. The control group engaged in practical exercises for donning and doffing PPE, which differs from the traditional nursing classroom teaching methods used in previous studies. Future research is needed to build on our findings and develop a more detailed understanding of the effectiveness of VRS programs.

Conclusions

This study highlights the effectiveness of VRS teaching in enhancing infection control knowledge, learning motivation, attitudes, and course satisfaction among undergraduate nursing students. By combining insights from qualitative data with quantitative information, we have provided a holistic understanding of the potential role of VRS in nursing education. Despite its limitations, this study opens avenues for future research and presents a compelling case for the broader implementation of VR in nursing education curricula. Future studies should consider longitudinal designs to evaluate the long-term impacts of VRS teaching on nursing education. Additionally, expanding the participant pool to include a more diverse range of students could yield more generalizable results. The findings have significant implications for nursing education, suggesting that VRS teaching can effectively enhance learning outcomes, particularly in areas that require high levels of practical knowledge and skills. The positive impact on student motivation and attitudes also points to a potentially transformative shift in how nursing education can be delivered, especially in a post–COVID-19 era, where digital and remote learning tools are becoming increasingly important.

Use of Generative Artificial Intelligence

During the preparation of this work, the authors used ChatGPT (OpenAI) to enhance the clarity of the content. After using ChatGPT, the authors reviewed and edited the content as needed and took full responsibility for the content of the published article.

Acknowledgments

This research was funded by the National Science and Technology Council (grants MOST 110-2511-H-255-013 and MOST 111-2410-H-255-010). The authors thank those students who participated in this research.

Data Availability

The data that support the findings of this study are available on reasonable request from the corresponding author.

Authors' Contributions

The conception and design of the study were carried out by WC, CCL, and CYH, while data acquisition and collection were performed by WC and CYH. The analysis and interpretation of data involved WC, CCL, LCC, and CYH. WC, CCL, and CYH drafted the manuscript, and critical revisions were made by WC, CCL, JC, HLL, LCC, and CYH. The final approval of the manuscript was provided by WC, CCL, JC, HLL, LCC, and CYH. Administrative, technical, or material support was provided by WC, CCL, and CYH, with resources provided by HLL and LCC. Supervision and validation were undertaken by CCL and CYH, and funding for the study was obtained by CYH.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide.
[PDF File (Adobe PDF File), 169 KB - mededu_v11i1e64780_app1.pdf]

References

1. Shorey S, Ng ED. The use of virtual reality simulation among nursing students and registered nurses: a systematic review. Nurse Educ Today 2021 Mar;98:104662. [doi: 10.1016/j.nedt.2020.104662] [Medline: 33203545]
2. Butt A, Kardong-Edgren S, Ellertson A. Using game-based virtual reality with haptics for skill acquisition. Clinical Simulation in Nursing 2018 Mar;16:25-32 [FREE Full text] [doi: 10.1016/j.ecns.2017.09.010]
3. Günay İ, Zaybak A. Comparison of the effectiveness of a virtual simulator with a plastic arm model in teaching intravenous catheter insertion skills. Comput Inform Nurs 2018;36(2):98-105. [doi: 10.1097/cin.0000000000000405]
4. Helle N, Vikman MD, Dahl-Michelsen T, Lie SS. Health care and social work students' experiences with a virtual reality simulation learning activity: qualitative study. JMIR Med Educ 2023 Sep 20;9:e49372 [FREE Full text] [doi: 10.2196/49372] [Medline: 37728988]
5. Park S, Shin HJ, Kwak H, Lee HJ. Effects of immersive technology-based education for undergraduate nursing students: systematic review and meta-analysis using the Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) approach. J Med Internet Res 2024 Jul 24;26:e57566 [FREE Full text] [doi: 10.2196/57566] [Medline: 38978483]
6. Yoon H. Effects of immersive straight catheterization virtual reality simulation on skills, confidence, and flow state in nursing students. Comput Inform Nurs 2024;42(12):872-887. [doi: 10.1097/cin.0000000000001141]
7. Abou Hashish EA, Alnajjar H. Digital proficiency: assessing knowledge, attitudes, and skills in digital transformation, health literacy, and artificial intelligence among university nursing students. BMC Med Educ 2024 May 07;24(1):508. [doi: 10.1186/s12909-024-05482-3] [Medline: 38715005]
8. Andreoli KG, Musser LA. Challenges confronting the future of emergency nursing. J Emerg Nurs 2020 Sep;46(5):573-578. [doi: 10.1016/j.jen.2020.04.006] [Medline: 32828478]
9. Jones T, Curtis K, Shaban RZ. Academic and professional characteristics of Australian graduate emergency nursing programs: a national study. Australas Emerg Care 2020 Sep;23(3):173-180. [doi: 10.1016/j.auec.2020.02.003] [Medline: 32115399]
10. Provenzano S, Santangelo O, Armetta F, Pesco G, Allegro A, Lampasona M, et al. COVID-19 infection: comparing the knowledge, attitude and practices in a sample of nursing students. Acta Biomed 2020 Nov 30;91(12-S):e2020001 [FREE Full text] [doi: 10.23750/abm.v91i12-S.10252] [Medline: 33263338]
11. Mallek F, Mazhar T, Faisal Abbas Shah S, Ghadi Y, Hamam H. A review on cultivating effective learning: synthesizing educational theories and virtual reality for enhanced educational experiences. PeerJ Comput Sci 2024;10:e2000 [FREE Full text] [doi: 10.7717/peerj-cs.2000] [Medline: 38855256]
12. Keller J. Motivational Design for Learning and Performance: The ARCS Model Approach. New York, NY: Springer Nature; 2010.
13. Thompson D, Thompson A, McConnell K. Nursing students' engagement and experiences with virtual reality in an undergraduate bioscience course. Int J Nurs Educ Scholarsh 2020 Sep 16;17(1):20190081 [FREE Full text] [doi: 10.1515/ijnes-2019-0081] [Medline: 32941161]
14. Fealy S, Jones D, Hutton A, Graham K, McNeill L, Sweet L, et al. The integration of immersive virtual reality in tertiary nursing and midwifery education: a scoping review. Nurse Educ Today 2019 Aug;79:14-19. [doi: 10.1016/j.nedt.2019.05.002] [Medline: 31078869]

15. Rourke S. How does virtual reality simulation compare to simulated practice in the acquisition of clinical psychomotor skills for pre-registration student nurses? A systematic review. Int J Nurs Stud 2020 Feb;102:103466. [doi: 10.1016/j.ijnurstu.2019.103466] [Medline: 31783192]
16. Keller JM. Motivation, learning, and technology: applying the ARCS-V motivation model. Particip Educ Res 2016;3(2):1-13. [doi: 10.17275/per.16.06.3.2]
17. Killam LA, Timmermans KE, Shapiro SJ. Motivation and engagement of nursing students in 2 gamified courses. Nurse Educ 2021 Aug 6;46(6):E173-E178. [doi: 10.1097/nne.0000000000001065]
18. Wu H, Li S, Zheng J, Guo J. Medical students' motivation and academic performance: the mediating roles of self-efficacy and learning engagement. Med Educ Online 2020 Dec;25(1):1742964 [FREE Full text] [doi: 10.1080/10872981.2020.1742964] [Medline: 32180537]
19. Mohamed Mohamed Bayoumy H, Alsayed S. Investigating relationship of perceived learning engagement, motivation, and academic performance among nursing students: a multisite study. AMEP 2021 Apr;Volume 12:351-369. [doi: 10.2147/amep.s272745]
20. Kowitlawakul Y, Tan JJM, Suebnukarn S, Nguyen HD, Poo DCC, Chai J, et al. Utilizing educational technology in enhancing undergraduate nursing students' engagement and motivation: A scoping review. J Prof Nurs 2022;42:262-275. [doi: 10.1016/j.profnurs.2022.07.015] [Medline: 36150870]
21. Chang YM, Lai CL. Exploring the experiences of nursing students in using immersive virtual reality to learn nursing skills. Nurse Educ Today 2021 Feb;97:104670. [doi: 10.1016/j.nedt.2020.104670] [Medline: 33264739]
22. Mäkinen H, Haavisto E, Havola S, Koivisto J. Graduating nursing students' user experiences of the immersive virtual reality simulation in learning - a qualitative descriptive study. Nurs Open 2023 May;10(5):3210-3219 [FREE Full text] [doi: 10.1002/nop2.1571] [Medline: 36598872]
23. Plotzky C, Loessl B, Kuhnert B, Friedrich N, Kugler C, König P, et al. My hands are running away - learning a complex nursing skill via virtual reality simulation: a randomised mixed methods study. BMC Nurs 2023 Jun 27;22(1):222. [doi: 10.1186/s12912-023-01384-9] [Medline: 37370124]
24. Saab M, McCarthy M, O'Mahony B. Virtual reality simulation in nursing and midwifery education: a usability study. Comput Inform Nurs 2023;41(10):815-824. [doi: 10.1097/cin.0000000000001010]
25. Wood J, Ebert L, Duff J. Implementation methods of virtual reality simulation and the impact on confidence and stress when learning patient resuscitation: an integrative review. Clin Simul Nursing 2022;66:5-17 [FREE Full text] [doi: 10.1016/j.ecns.2022.02.006]
26. Yang S, Oh Y. The effects of neonatal resuscitation gamification program using immersive virtual reality: a quasi-experimental study. Nurse Educ Today 2022 Oct;117:105464 [FREE Full text] [doi: 10.1016/j.nedt.2022.105464] [Medline: 35914345]
27. Faul F, Erdfelder E, Lang A, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007 May;39(2):175-191 [FREE Full text] [doi: 10.3758/bf03193146] [Medline: 17695343]
28. Cohen J. Statistical Power Analysis for the Behavioral Sciences (2nd Edition). Hillside, NJ: Lawrence Erlbaum Associates; 1998.
29. Woon A, Mok W, Chieng Y, Zhang H, Ramos P, Mustadi H, et al. Effectiveness of virtual reality training in improving knowledge among nursing students: a systematic review, meta-analysis and meta-regression. Nurse Educ Today 2021 Mar;98:104655 [FREE Full text] [doi: 10.1016/j.nedt.2020.104655] [Medline: 33303246]
30. George R, Titus SK. Benefits and barriers of using virtual reality in teaching undergraduate nursing students. Nurse Educ 2024;49(5):E244-E249. [doi: 10.1097/nne.0000000000001660]
31. Marton F. Phenomenography-A research approach to investigating different understanding of reality. J Thought 1986;21(3):28-49.
32. Dahlgren L, Fallsberg M. Phenomenography as a qualitative approach in social pharmacy research. J Soc Admin Pharm 1991;8:150-156. [doi: 10.4135/9781412963909.n316]
33. Lincoln YS, Guba EG. Naturalistic Inquiry. Thousand Oaks, CA: Sage; 1985.
34. Sjöström B, Dahlgren LO. Applying phenomenography in nursing research. J Adv Nurs 2002 Nov;40(3):339-345. [doi: 10.1046/j.1365-2648.2002.02375.x] [Medline: 12383185]
35. Chen F, Leng Y, Ge J, Wang D, Li C, Chen B, et al. Effectiveness of virtual reality in nursing education: meta-analysis. J Med Internet Res 2020 Sep 15;22(9):e18290. [doi: 10.2196/18290] [Medline: 32930664]
36. Jallad ST, Işık B. The effectiveness of virtual reality simulation as learning strategy in the acquisition of medical skills in nursing education: a systematic review. Ir J Med Sci 2022 Jun;191(3):1407-1426. [doi: 10.1007/s11845-021-02695-z] [Medline: 34227032]
37. Ryan GV, Callaghan S, Rafferty A, Higgins MF, Mangina E, McAuliffe F. Learning outcomes of immersive technologies in health care student education: systematic review of the literature. J Med Internet Res 2022;24(2):e30082. [doi: 10.2196/30082]
38. Huai P, Li Y, Wang X, Zhang L, Liu N, Yang H. The effectiveness of virtual reality technology in student nurse education: a systematic review and meta-analysis. Nurse Educ Today 2024 Jul;138:106189. [doi: 10.1016/j.nedt.2024.106189] [Medline: 38603830]

39. Lo Y, Yang C, Yeh T, Tu H, Chang Y. Effectiveness of immersive virtual reality training in nasogastric tube feeding education: a randomized controlled trial. Nurse Educ Today 2022 Dec;119:105601. [doi: 10.1016/j.nedt.2022.105601] [Medline: 36244254]
40. Lee H, Han J, Park J, Min S, Park J. Development and evaluation of extracorporeal membrane oxygenation nursing education program for nursing students using virtual reality. BMC Med Educ 2024 Jan 26;24(1):92. [doi: 10.1186/s12909-024-05057-2] [Medline: 38279179]
41. Shujuan L, Mawpin T, Meichan C, Weijun X, Jing W, Biru L. The use of virtual reality to improve disaster preparedness among nursing students: a randomized study. J Nurs Educ 2022 Feb;61(2):93-96. [doi: 10.3928/01484834-20211213-05] [Medline: 35112954]
42. Fontenot J, Hebert M, Lin H, Kulshreshth AK. Examining the perceptions among undergraduate nursing students using virtual reality in a community course: a mixed-methods explanatory study. J Community Health Nurs 2024;41(3):145-155. [doi: 10.1080/07370016.2023.2280617] [Medline: 37966021]
43. Lau ST, Siah CJR, Loh WL, Rusli KDB, Schmidt LT, Lim FP, et al. Enhancing professional competency in clinical procedures using head-mounted display virtual reality - a mixed method study. Med Educ Online 2023 Dec;28(1):2232134. [doi: 10.1080/10872981.2023.2232134] [Medline: 37406175]

Abbreviations

ARCS: Attention, Relevance, Confidence, and Satisfaction
IMMS: Instructional Materials Motivation Survey
PPE: personal protective equipment
VR: virtual reality
VRS: virtual reality simulation

Edited by B Lesselroth; submitted 28.07.24; peer-reviewed by LH Wang, SY Wang, T Loetscher; comments to author 10.09.24; revised version received 17.09.24; accepted 14.01.25; published 11.02.25.

Please cite as:
Chang W, Lin CC, Crilly J, Lee HL, Chen LC, Han CY
Virtual Reality Simulation for Undergraduate Nursing Students for Care of Patients With Infectious Diseases: Mixed Methods Study
JMIR Med Educ 2025;11:e64780
URL: https://mededu.jmir.org/2025/1/e64780
doi:10.2196/64780
PMID:

Original Paper

Effectiveness of a 5G Local Area Network–Based Digital Microscopy Interactive System: Quasi-Experimental Design

Jie Xu[1], MS; Jihong Sha[2], PhD; Song Jia[1], AD; Jiao Li[1], MD; Lei Xu[1], PhD; Zhihua Shao[2], PhD

[1]Teaching Laboratory Center, School of Medicine, Tongji University, Shanghai, China

[2]Department of Cell Biology, School of Medicine, Tongji University, Shanghai, China

Corresponding Author:
Zhihua Shao, PhD
Department of Cell Biology
School of Medicine
Tongji University
Zhennan Road
Shanghai, 200092
China
Phone: 86 17717409485
Email: shaozhihua@tongji.edu.cn

Abstract

Background: Technological innovation is reshaping the landscape of medical education, bringing revolutionary changes to traditional teaching methods. In this context, the upgrade of the teaching model for microscopy, as one of the core skills in medical education, is particularly important. Proficiency in microscope operation not only affects medical students' pathology diagnosis abilities but also directly impacts the precision of surgical procedures and laboratory analysis skills. However, current microscopy pedagogy faces dual challenges: on one hand, traditional teaching lacks real-time image sharing capabilities, severely limiting the effectiveness of immediate instructor guidance; on the other hand, students find it difficult to independently identify technical flaws in their operations, leading to inefficient skill acquisition. Although whole-slide imaging-based microscopy system technology has partially addressed the issue of image visualization, it cannot replicate the tactile feedback and physical interaction experience of the real world. The breakthrough development of 5G communication technology—with its ultrahigh transmission speed and ultralow latency—provides an innovative solution to this teaching challenge. Leveraging this technological advantage, Tongji University's biology laboratory has pioneered the deployment of a 5G local area network (LAN)–supported digital interactive microscopy system, creating a new model for microscopy education.

Objective: This study aims to investigate the efficacy of an innovative 5G LAN-powered interactive digital microscopy system in enhancing microscopy training efficiency, evaluated through medical students' academic performance and learning experience.

Methods: Using a quasi-experimental design, we quantify system effectiveness via academic performance metrics and learning experience dimensions. A total of 39 students enrolled in the biology course were randomly assigned to 2 groups: one using traditional optical microscopes (control) and the other using the digital microscopy interactive system (DMIS). Their academic performance was evaluated through a knowledge test and 3 laboratory reports. A 5-point Likert-scale questionnaire was used to gather feedback on students' learning experiences. In addition, the DMIS group was required to evaluate the specific functions of the system.

Results: In the knowledge test, no statistical difference was found between the 2 groups; however, the DMIS group scored significantly higher in Lecture 2 (P<.05). In the laboratory reports, the DMIS group performed significantly better than the control group (mean 90.33, SD 2.63 vs mean 80.53, SD 3.52, P<.001). Questionnaire results indicated that the DMIS group has a positive evaluation of the system and expressed greater confidence in its future application. For the evaluation of the laboratory lectures, the DMIS group received higher evaluations on the course content and self-efficacy (P<.05), and higher satisfaction with the laboratory lectures (P<.05).

Conclusions: Overall, the digital microscope interactive system enhances students' learning experiences and improves their academic performance. It offers various interactive functions to facilitate the organization of teaching activities and promote immediate feedback in the classroom. Thus, it is a promising tool for microscopy laboratory teaching.

(JMIR Med Educ 2025;11:e70256) doi:10.2196/70256

KEYWORDS

microscopy technique; digital interactive system; 5G; medical education; undergraduate

Introduction

Emerging technologies are rapidly proliferating, elevating the quality of medical training and better preparing students for future clinical practice [1]. For instance, whole-slide imaging-based microscopy system is accepted as a major technique in the teaching of pathology, anatomy, and histology in several countries [2]. A multimedia-supported manikin system capable of simulating a real clinical operating environment was reported to help students reduce barriers to entering the clinic [3]. Furthermore, 3D printing technology, in addition to its application in clinical practice, provides educators with a cost-effective and efficient educational tool [4]. The maturity of multiple technologies has enabled the potential for technical upgrades to microscopy as a fundamental teaching tool.

Microscopy provides a methodological foundation for high-resolution observation and quantitative analysis of microscopic phenomena [5]. As a core competency for medical students, this technique enhances skill acquisition in pathological diagnostics through microstructural visualization, improves intraoperative spatial localization precision via microscope operational adjustments, and supplies essential analytical capabilities for indispensable scientific experimentation. However, current microscopy instruction faces significant challenges: traditional optical training suffers from delayed feedback, as instructors cannot monitor student operations in real time. This leads to undetected technical errors until learners proactively seek guidance [6]. Compounding this issue, beginners lack the expertise to self-identify technical deficiencies [7], creating a dual problem of instructor feedback absence and limited student self-assessment. Such gaps risk entrenching structural misconceptions. While whole-slide imaging-based microscopy system enables image visualization, its inherent limitations remain critical: it fails to replicate optical parameter adjustment processes and lacks tangible tactile feedback [8], thereby undermining practical skill development. Microscopy expertise necessitates not only the physical ability to visualize specimens but also the cognitive capacity to explore and discriminate among potential morphological variations for accurate characterization of target samples. This dual requirement implies that both knowledge internalization and operational proficiency constitute core learning objectives in microscopy instruction, whereas current pedagogical approaches demonstrate limitations in concurrently achieving these competencies.

The digital microscope interactive system, based on optical microscopes, is a multiterminal digital imaging device connected by 5G local area network (LAN). The system retains the tactile sensation of physical operation while enabling the display of microscopic images on software-equipped devices (such as computers, tablets, or mobile phones) through analog-to-digital conversion. Leveraging 5G's high-bandwidth connectivity, the system enables instantaneous image transmission between terminals. Consequently, instructors can simultaneously observe all students' microscope images on the teacher-terminal computer. This real-time monitoring capability enables immediate detection of operational errors, even in the absence of student feedback. Beyond real-time oversight, the multiplatform compatibility supports interactive pedagogical functions such as distributing laboratory instructions, uploading resulting images, demonstrating operations, and so on. The high popularity of smartphones among students currently makes this mode both flexible and economical [9].

Previous reports have documented the application of 5G technology in multiple educational scenarios [10,11]. While a digital microscope system has been implemented in the pathogenic biology laboratory in China [12], no research has yet reported the deployment of 5G LAN-based digital interactive microscopy systems in laboratory teaching. Consequently, investigating 5G-enabled interactive microscopy systems embodies significant innovative potential to enhance instructional efficacy by addressing the aforementioned limitations in microscopy instruction.

Microscopy instruction in medical education faces challenges of delayed instructor feedback and limited student self-correction capabilities, which impact skill acquisition efficiency. To address this challenge, Tongji University innovatively deployed a 5G LAN-based digital microscopy interaction system (DMIS). This system preserves the tactile sensation of physical operation while leveraging digital microscopy for procedural visualization. Capitalizing on 5G's high-speed transmission, it establishes an immediate feedback mechanism, thereby enhancing instructional efficiency in microscopy training. Using a quasi-experimental design—frequently applied in education [13,14]—this study compared the instructional efficacy between this digital system and conventional optical microscopy training. We hypothesize that these technological features will significantly improve pedagogical outcomes. A comparative analysis of academic performance and learning experience questionnaire data between the DMIS group and the traditional training control group provides evidence-based guidance for medical educators reliant on microscopy technology to optimize pedagogical approaches.

Methods

Experimental Procedures and Participants

The biology course for medical undergraduates at Tongji University is a second-year component of the curriculum, comprising 10 theoretical sessions and 3 laboratory sessions. These laboratory sessions are aligned with the theoretical learning schedule and consist of 4 periods for each session, including the use of microscopes and cell cycle analysis, cell counting and cell culture, and mouse chromosome preparation and karyotype analysis (see Table 1). Students are organized into classes of 20, with 2 classes running simultaneously. In each class, the instructor introduces the experimental content and principles, followed by a demonstration of the procedure. Students then work in pairs to conduct the experiments and complete their laboratory reports according to the requirements outlined in the Microsoft PowerPoint slides provided by the

instructor. All activities described in this article have received informed consent from the students.

Table 1. Laboratory experiments.

Experiment name	Experimental activities
The use of a microscope and cell cycle analysis.	• Microscopy training with commercially prepared slides. • Capture clear and accurate images of each stage of mitosis.
Cell counting and cell culture.	• Prepare the cell sample and count viable cells using microscopy. • Cell passaging.
Preparation and observation of mouse chromosomes.	• Prepare and stain mouse chromosomes. • Capture clear images of mouse chromosomes.

Digital Microscopy Interactive System

The digital microscopy interaction system primarily consists of 3 parts: Panthera I series optical microscopes (Motic Corp), Motic Digilab 3.0 software (Motic Corp), and a computer or mobile device, all of which can connect via a 5G LAN (see Figure 1). In class, students need to download and install the software from the app store onto their digital devices (eg, smartphones and tablets), connect to the designated 5G LAN, and enter their microscope number in the software to view live microscopy images. This setup enables real-time sharing of all microscopic images across the network. Simultaneously, the instructor's terminal computer with the installed software can display all microscope images currently in use, allowing both students and the instructor to visualize the live microscopic manipulations at the same time (see Figure 2A).

Figure 1. The framework of a digital microscopy interactive system based on 5G local area network (LAN).

The Motic Digilab 3.0 software supports 4 functions for both instructors and students (see Table 2). On the student side, students can capture live images, record videos, and ask questions within the interface (see Figure 2C). On the instructor side, instructors can monitor the progress of all students and address issues based on their live images. Since both sides are connected to the same 5G LAN, they can transfer files to each other, thereby enhancing the efficiency of teaching activities (see Figure 2B).

Figure 2. The function of digital microscopy interactive system.

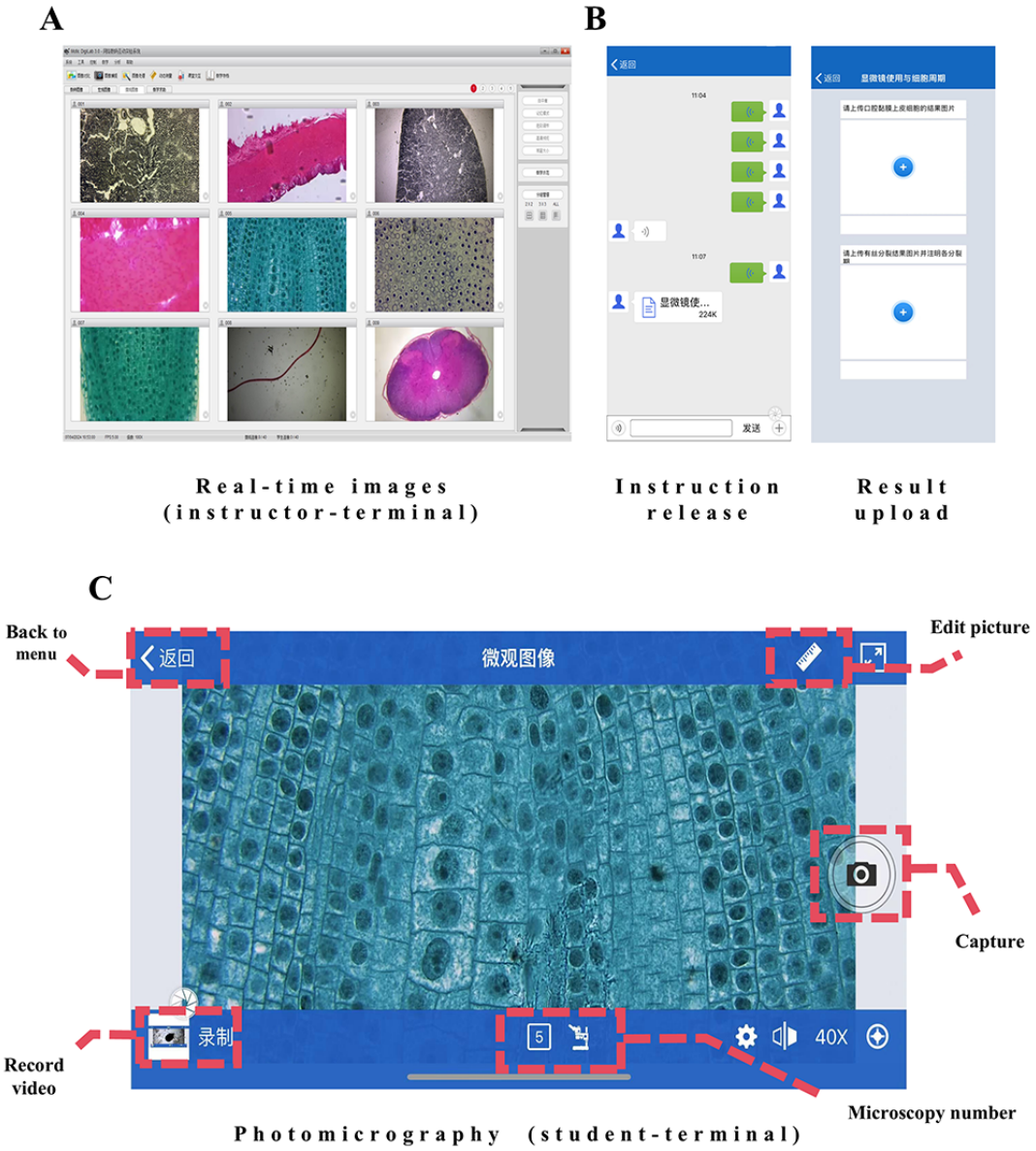


Table 2. The functions of the digital microscopy interactive system.

Function	Instructors' terminal	Students' terminal
Microimage	• View images observed by all student-terminal microscopes in real time.	• View live images observed under the microscope on its mobile app.
Macroimage	• View videos recorded by all student-terminal mobile phones.	• Record the experimental process with a mobile phone camera.
Demonstration	• Use the instructor-terminal microscope to demonstrate proper operation in either forced or unforced mode.	• Continue to use the microscope, while students who have difficulties can refer to the instructor's demonstration.
Interaction	• Deliver the protocol and requirements of this experiment. • Instruct students through their real-time microscope images.	• Upload the result images of this experiment as required. • Ask questions by text or voice in their mobile app.

Grouping

The biology course, conducted in the second semester of the 2023-2024 academic year, involved a total of 39 medical freshmen. In this study, randomization was performed using a computer-generated randomization program integrated within the school education management system. Concealment was ensured by implementing the allocation through the school's system without any manual intervention or researcher influence. Participants were randomly assigned to 2 groups: the DMIS group, which used the digital microscope interactive system, and the control group, which used the traditional optical microscope.

Evaluation

The assessment framework (see Table 3) for the DMIS, designed around the instructional objectives of mastering operation and internalizing knowledge, comprises 4 key components. Specifically, student academic performance was jointly assessed using 1 knowledge test and 3 laboratory reports used to evaluate the internalization of knowledge and proficiency in operation. Furthermore, student learning experience feedback is collected via a 5-point Likert scale questionnaire. In addition, students in the DMIS group complete a supplementary 5-point Likert scale questionnaire specifically designed to evaluate the system's functional features.

Table 3. The evaluation framework of the digital microscopy interactive system.

Dimension	Testing method	Participants	Measure analyzed
Knowledge	Knowledge test	All students	Comprehension of knowledge and mastery of operational protocols.
Skill	Laboratory reports	All students	Microscope operation proficiency and accuracy in result analysis.
Attitude	A 5-point Likert scale questionnaire	All students	Course satisfaction.
Functional evaluation	A 5-point Likert scale questionnaire	DMIS[a] group	System capabilities and future applications.

[a]DMIS: digital microscopy interactive system.

The knowledge test was jointly designed by all instructors participating in the Biology course instruction. To assess comprehension of structural knowledge and mastery of operational protocols, 6 questions were implemented per laboratory lecture. It has been reported that multiple true-false questions are superior in revealing students' misunderstanding of knowledge [15,16]. Therefore, a total of 18 true-false questions were designed. After class, each student anonymously completed the online test through the Sojump platform at the same time.

Their laboratory reports were submitted to the Canvas learning management system at Tongji University, enabling all course instructors to access and evaluate these documents online. The assessment rubric, collaboratively developed by the faculty members, primarily emphasizes microscope operation proficiency and accuracy in result analysis. Each dimension was assessed using a 4-tier performance scale: excellent (45 points), proficient (40 points), satisfactory (35 points), and unsatisfactory (30 points). This contributed to a maximum of 50 points per dimension, thereby yielding a total possible score of 100 points. Student competence in microscopy techniques was evaluated through the scoring of micrographs presented in the reports. High-quality microscopic images reflect proper adjustment of the microscope to achieve optimal resolution, appropriate magnification, and adequate illumination during image capture. Interpretation of results demonstrates students' understanding of target cell biological characteristics and their proficiency in identifying and locating cellular structures through microscopic techniques.

All participants anonymously completed a 5-point Likert scale questionnaire assessing their learning experiences in laboratory lectures. It encompassed four domains: (1) course content, (2) teaching quality, (3) self-efficacy, and (4) teaching effectiveness, with 3-4 items per domain. Specifically, course content assessed the difficulty level, scheduling, interest, and challenge of the laboratory lectures; teaching quality evaluated classroom organization, instructional clarity, and comprehensibility; self-efficacy measured students' confidence in completing tasks and resolving challenges; and teaching effectiveness gauged whether students had a better understanding of knowledge, techniques, and laboratory safety. Finally, an open-ended item invited students to share which technology they have benefited the most from in their classroom learning.

Functional evaluation of the digital microscopy interactive system was conducted by the DMIS group via a 5-point Likert scale questionnaire. This anonymously collected questionnaire assessed user perceptions of system capabilities and future applications.

Analysis

The Sojump platform automatically scored students' responses against predetermined answers and converted results to percentage scores. The Canvas learning management system recorded laboratory report grades based on instructors' online assessments. Open-ended responses regarding "What technology has benefited you the most in classroom learning?" were analyzed through word frequency analysis provided by the Sojump platform. All data were statistically analyzed using IBM SPSS Statistics (version 26). Group comparisons were performed using the independent samples *t* test for normally distributed continuous variables, while the nonparametric test, such as the Mann-Whitney *U* test, was used for categorical variables or continuous variables deviating from normality. The internal consistency of questionnaire items was evaluated using the Cronbach alpha coefficient (Cronbach α), with Cronbach $\alpha \geq 0.70$ considered acceptable for scale reliability. Statistical significance was determined at $P<.05$. Data are presented as mean (SD).

Ethical Considerations

This study was conducted in accordance with established ethical principles, with participation being voluntary in nature. All participants received comprehensive information regarding the study's objectives, procedures, and potential implications. They were assured of their right to withdraw at any stage without penalty. The confidentiality of the data was rigorously maintained through the anonymization of all responses, which were subsequently analyzed and reported in aggregated form.

All procedures involving human subjects were carried out in compliance with the ethical standards set forth by relevant national research committees [17]. The study strictly followed applicable ethical guidelines and data protection laws, thereby safeguarding participants' rights and privacy throughout the research process.

Results

Participants' Demographics

A total of 39 students participated in this course, with 19 in the control group and 20 in the DMIS group. All participants were approximately aged 20 (mean 19.56, SD 0.50) years, with the control group aged 19.74 (SD 0.45) years and the DMIS group aged 19.40 (SD 0.50) years. The demographics of all participants are presented in Table 4. The two groups had a similar composition in terms of gender and GPA (grade point average), which were confirmed to show no statistically significant difference (P>.05). Previous academic performance (mean score of previous courses) also showed no significant difference between the groups (P>.05), with mean scores of 82.09 (SD 5.49) and 80.69 (SD 5.58) for the control and DMIS groups, respectively.

Table 4. The demographics of the participants.

Variable	Total (n=39), n (%)	Control group (n=19), n (%)	DMIS[a] group (n=20), n (%)	*P* value
Students	39 (100)	19 (100)	20 (100)	—[b]
Sex				.66[c]
Male	15 (38.5)	6 (37.5)	7 (35)	
Female	24 (61.5)	10 (62.5)	13 (65)	
GPA[d]				.71[e]
≧4.0	12 (30.8)	6 (37.5)	5 (25)	
3.5-3.9	18 (46.2)	7 (43.8)	10 (50)	
＜3.5	9 (23.1)	3 (18.8)	5 (25)	

[a]DMIS: digital microscopy interactive system.

[b]Not applicable.

[c]Independent sample *t* test.

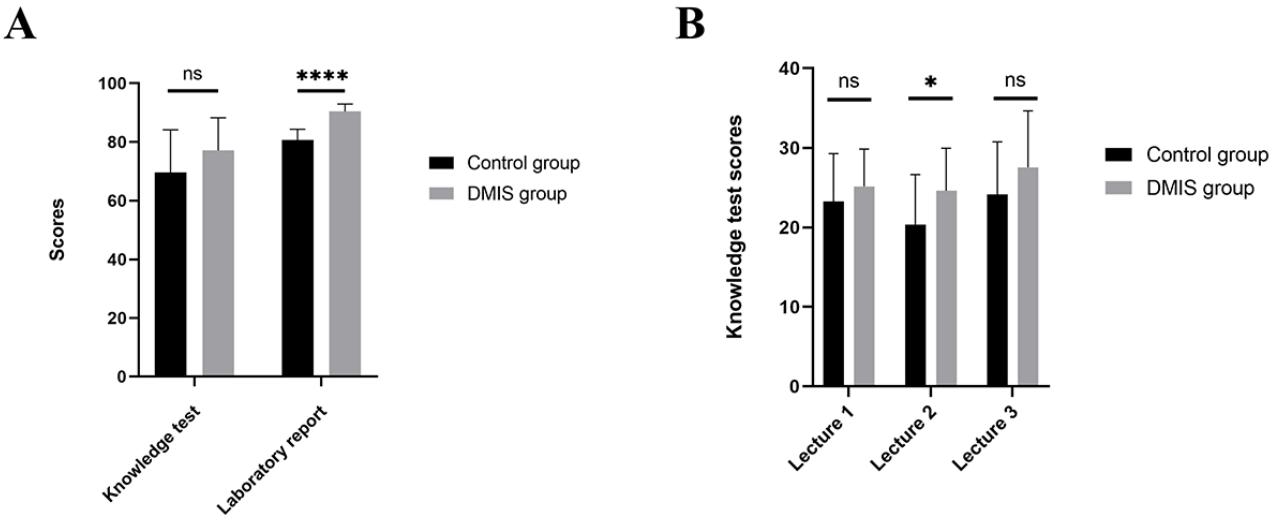
[d]GPA: grade point average.

[e]Chi-square test.

Comparison of the Academic Performance Between the Control Group and the DMIS Group

All 39 students took the knowledge test and submitted their laboratory report. In the knowledge test, the DMIS group had a higher mean score of 77.15 (SD 11.10) compared to 69.58 (SD 14.56) for the control group; however, this difference was not statistically significant (P=.075; see Figure 3A). To evaluate the system's potential impact, we assessed group performance across lectures. The DMIS group demonstrated significantly higher mean scores than controls in Lecture 2 (P<.05; see Figure 3B), which focused on microscopy-based enumeration of viable cells in prepared samples. For the performance of laboratory reports, the DMIS group scored 90.33 (SD 2.63) versus 80.53 (SD 3.52) for the control group (P<.001; see Figure 3A). Overall, the DMIS group achieved better results in both tests and reports.

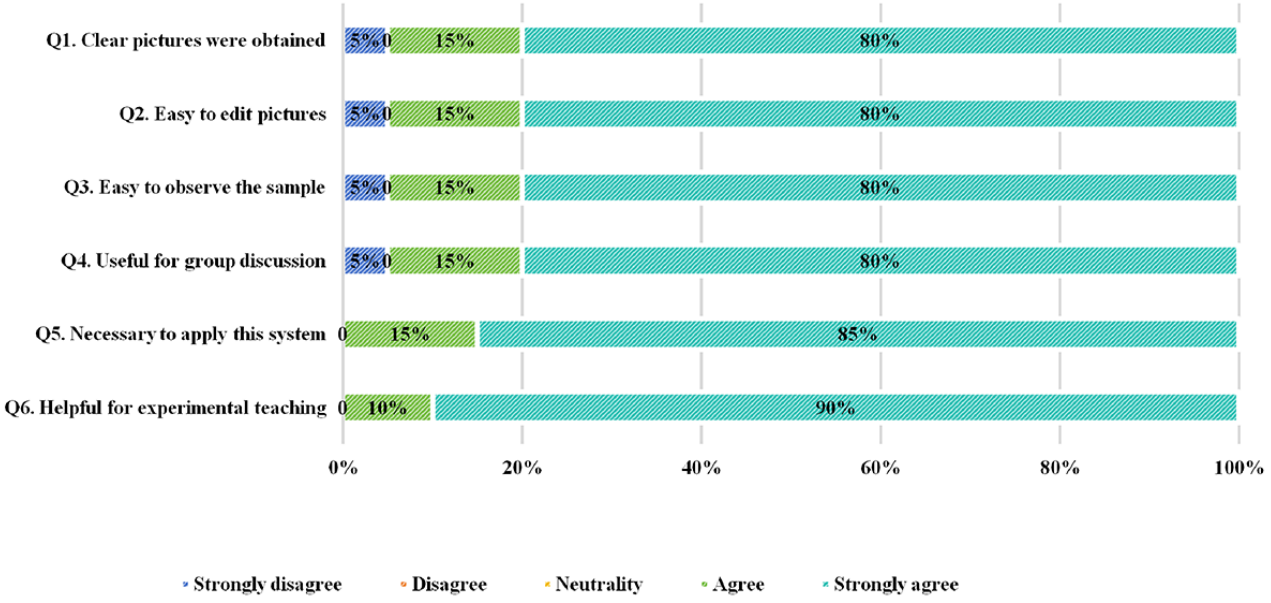
Figure 3. Comparison of scores between the control group and the DMIS group. DMIS: digital microscopy interactive system.



Functional Evaluation of This Digital Microscopy Interactive System

To capture user perspectives regarding the system's capabilities and future application, we distributed a questionnaire survey to the DMIS group. All participants (100%, 20/20) completed the questionnaire, and its Cronbach α was 0.91, indicating good reliability. In their feedback, most students provided positive evaluations and expressed satisfaction with its functions (see Figure 4 Q1-Q4). They indicated that it contributed positively by capturing clear images, providing easy-to-edit images, facilitating sample observation, and promoting group discussions. (95% agree or strongly agree). Furthermore, the students expressed confidence in its future application, and all agreed that this system is necessary and useful for biological experiments (100% agree or strongly agree; see Figure 4 Q5-Q6).

Figure 4. Responses of the digital microscopy interactive system (DMIS) group students regarding the DMIS.



Comparison of Students' Views on the Laboratory Lectures Between the Control Group and the DMIS Group

In addition to evaluating the function of this system, its impact on laboratory lectures was also assessed through the questionnaire. A total of 36 out of 39 (92.3%) students completed the survey, including 20 from the DMIS group (100%, 20/20) and 16 from the control group (84.2%, 16/19). Given its Cronbach α of 0.98, the questionnaire demonstrates high reliability. The results showed a statistically significant difference in two of the four areas addressed in the survey (see Table 5). The first area is course content. Compared to the control group, the DMIS group more strongly supported the appropriateness of the course difficulty, the reasonableness of the schedule, and found it both enjoyable and challenging (P<.05). Similarly, students in the DMIS group reported higher self-efficacy and greater confidence in facing difficulties and challenges after finishing the lectures (P<.05). Notably, there was a statistical difference in overall satisfaction between the

two groups, with the DMIS group scoring higher (mean 4.63, SD 0.50 vs mean 4.95, SD 0.22; $P<.05$). In the open-ended question, when discussing their biggest gains, 30% (6/20) of the DMIS group mentioned microscope technology, compared to 12.5% (2/16) in the control group. From the results, the implementation of this system has significantly improved the effectiveness of microscopy teaching. Furthermore, students have started to recognize the importance and application value of the microscopy technology.

Table 5. Students' perceptions regarding the laboratory lectures.

Domains	Control group (n=16), mean (SD)	DMIS[a] group (n=20), mean (SD)	*P* value
Course content	18.25 (1.88)	19.6 (1.00)	.02[b]
Teaching quality	18.75 (1.92)	19.75 (1.12)	.07
Self-efficiency	18.56 (1.93)	19.8 (0.89)	.03[b]
Teaching effectiveness	14.06 (1.44)	14.85 (0.67)	.07

[a]DMIS: digital microscopy interactive system.

[b]$P \leq .05$ was considered statistically significant.

Discussion

Principal Findings

The core competency in microscopy lies in the precise manipulation of physical components (and potentially software interfaces) to optimize optical system parameters. This process, integrated with internalized knowledge of sample microstructure, enables accurate localization, clear identification, and effective visualization of target structures within the field of view, thereby yielding reliable scientific data or diagnostic evidence. Consequently, it yields reliable scientific data or diagnostic evidence. Therefore, the core objectives of microscopy education lie in knowledge internalization and mastery of operational skills—goals that have remained a persistent challenge for traditional teaching methods. This study implements a 5G-enabled digital interactive microscopy system in biological laboratory instruction, evaluating its effectiveness through students' academic performance, learning experiences, and user evaluations of system functionality.

Our research findings indicate that, regarding academic performance, the knowledge test scores of the DMIS group were generally higher than those of the control group, with a significant difference observed in lecture 2. This disparity likely stems from the greater challenge associated with the microscope operation for counting unstained viable cells covered in that session. The high transparency of unstained viable cells renders them difficult to distinguish from the background, demanding a more robust comprehension of knowledge for identification and proficient operational skills for precise localization. This complex situation notably enhanced the value of the DMIS's technical assistance. Real-time image sharing facilitated group discussion to reach a consensus on accurate identification. Furthermore, the laboratory report scores for the DMIS group were significantly higher than those of the control group, indicating that the DMIS group obtained higher quality microscopic images. This suggests that they effectively chose the objective lens for magnification, captured the images at the correct focal plane, and adjusted the light for the brightness of the images. In addition, they correctly found the cells of interest. Actually, the microscope is a complex tool that requires not only the users' skilled operation but also a solid understanding of morphological concepts for effective image analysis [7]. In the absence of immediate instructor feedback, students are prone to developing persistent misconceptions, often remaining unaware of these errors. More critically, such misunderstandings may lead to systematic misinterpretations, resulting in erroneous interpretations of the information conveyed by images. The real-time imaging capability of the DMIS enhances classroom interactivity, a factor consistently demonstrated by multiple studies [18,19] to improve student learning outcomes. Furthermore, this system retains the tactile feedback and direct manipulation experience that students value, which is missing in whole-slide imaging-based microscopy system [20,21]. These elements contribute to effective learning of microscopy skills and a deeper understanding of image analysis. The observed improved academic performance of the DMIS group supports this conclusion.

In the questionnaire survey, students in the DMIS group provided overwhelmingly positive feedback about the digital microscope interactive system. They generally agreed that it made capturing images clear, facilitated sample observation and picture editing, and supported group discussions. Only a single student reported issues due to his mobile device being incompatible with the app, which negatively impacted his experience. Similar compatibility issues can occur with applications [22] and warrant further optimization. However, this shortcoming did not diminish students' confidence in the system's future application, reflecting their recognition of its benefits. Notably, the DMIS group expressed higher satisfaction with laboratory lectures compared to the control group, especially regarding course content and self-efficacy. This may be attributed to their ability to independently navigate the challenges of microscope technology, enhancing their skills and confidence. As a result, students in the DMIS group identified that the microscope technique was what they gained the most from the class. Given the well-known positive impact of self-efficacy on learning [23], this experience will undoubtedly enable medical students to perform better in high-intensity medical training in the future.

For medical students, proficiency in microscopy is not merely a laboratory technique but a critical component in developing core professional competencies that profoundly influence future clinical practice. First, microscopy skills enhance morphological identification capabilities, serving as a prerequisite for accurate clinical diagnosis. Second, the process of adjusting microscope parameters helps establish efficient and precise visuospatial-motor coordination, thereby improving spatial positioning accuracy in future surgical procedures. In addition, it should be noted that microscopy remains fundamental to experimental research—an essential skill for physicians in China during their clinical practice.

For instructors, this digital microscope interactive system makes the organization of classroom teaching more convenient and efficient. Previously, due to limited laboratory funding, each room was only equipped with one microphotography device, which inevitably led to insufficient usage time for each student group. Now, students can take pictures with at least one available device at any time, greatly reducing the waiting time for equipment. Digital technology can achieve real-time monitoring and feedback, providing students with a personalized learning experience [1,24]. The convenience of 5G networks lies in their ability to transmit multimedia files without delay, enabling rapid feedback between teachers and students. This immediate communication helps resolve issues promptly and significantly enhances the quality of teaching. In addition, the interactive functions provided by the software's main interface are valuable for personalized learning. For example, students can record operation videos using the function of macro image for postclass review and results analysis. If they have questions, these videos can be sent to the instructor for further improvement suggestions. More than that, the demonstration function offers 2 modes: students can freely view the synchronized transmission demonstrated by the teacher, while others can continue to use the system without disruption. While this study primarily centers on investigating how real-time operational visualization enhances microscopy technique acquisition through immediate feedback mechanisms, the system's inherent classroom interaction features—such as video recording—prove instrumental in tracking learning progression and enabling formative assessment. These capabilities offer enhanced potential for developing precision-oriented pedagogical approaches in future microscopy instruction.

Limitations

In this study, although it was observed that the digital microscope interaction system can enhance students' academic performance and learning experience, the sample size might be insufficient. In order to study the impact of this system more specifically, we only investigated the biology course where microscopy is most commonly used. This can lead to some application problems and differential performance that may arise in a large student population being overlooked. In addition, when the number of users increases, whether the system can maintain the stability of the current real-time transmission needs further analysis.

Conclusions

The digital microscopy interaction system presented in this study represents a powerful tool for laboratory instruction. By implementing digital transformation and information technology enhancements to traditional optical microscopy, the system enables real-time visualization of student operations and facilitates immediate instructor feedback. These capabilities provide effective support for achieving the pedagogical objectives of microscopy skill development in medical students. Significant improvements in learning outcomes are evidenced by enhanced academic performance and superior learning experiences. Notably, students exhibited high satisfaction with the system and demonstrated markedly increased engagement during instructional sessions. As a teaching tool, the system's rich interactive functionalities not only assist instructors in organizing teaching activities more efficiently but also support formative assessment of the learning process. This thereby creates favorable conditions for implementing personalized teaching models in the future. Collectively, these findings demonstrate the system's considerable potential as a highly promising instructional aid within medical education.

Acknowledgments

We thank all the participants in this assignment. This work was supported by the 2024-2025 Tongji University Excellent Experimental Project (grant 1500104256). Deepseek V3.1 and ChatGPT-4 were used to improve the clarity of language in the manuscript. All content, design, data analysis, and interpretation are the original work of the authors.

Authors' Contributions

Conceptualization: JX, ZS
Methodology: JX, ZS
Data curation: JS, SJ, JL
Formal analysis: JS, SJ, JL
Writing – original draft: JX
Writing – review & editing: LX, ZS

Conflicts of Interest

None declared.

References

1. Tokuç B, Varol G. Medical education in the era of advancing technology. Balkan Med J 2023;40(6):395-399 [FREE Full text] [doi: 10.4274/balkanmedj.galenos.2023.2023-7-79] [Medline: 37706676]
2. Maity S, Nauhria S, Nayak N, Nauhria S, Coffin T, Wray J, et al. Virtual versus light microscopy usage among students: a systematic review and meta-analytic evidence in medical education. Diagnostics (Basel) 2023;13(3):558 [FREE Full text] [doi: 10.3390/diagnostics13030558] [Medline: 36766660]
3. Yang Y, Cheng G, Xing X, Li Z, Zhang W. Application of a multimedia-supported manikin system for preclinical dental training. BMC Med Educ 2022;22(1):693 [FREE Full text] [doi: 10.1186/s12909-022-03757-1] [Medline: 36167531]
4. Valverde I, Gomez G, Byrne N, Anwar S, Silva Cerpa MA, Martin Talavera M, et al. Criss-cross heart three-dimensional printed models in medical education: A multicenter study on their value as a supporting tool to conventional imaging. Anat Sci Educ 2022;15(4):719-730. [doi: 10.1002/ase.2105] [Medline: 34008341]
5. Hortsch M, Girão-Carmona VCC, de Melo Leite ACR, Nikas I, Koney N, Yohannan D, et al. Teaching cellular architecture: the global status of histology education. Adv Exp Med Biol 2023;1431:177-212. [doi: 10.1007/978-3-031-36727-4_9] [Medline: 37644293]
6. Evans SJM, Moore AR, Olver CS, Avery PR, West AB. Virtual microscopy is more effective than conventional microscopy for teaching cytology to veterinary students: a randomized controlled trial. J Vet Med Educ 2020;47(4):475-481. [doi: 10.3138/jvme.0318-029r1] [Medline: 32105198]
7. Imreh G, Hu J, Le Guyader S. Improving light microscopy training routines with evidence-based education. J Microsc 2024;294(3):295-307. [doi: 10.1111/jmi.13216] [Medline: 37534621]
8. Yang J. Technology-enhanced preclinical medical education (anatomy, histology and occasionally, biochemistry): a practical guide. Adv Exp Med Biol 2023;1431:65-93. [doi: 10.1007/978-3-031-36727-4_4] [Medline: 37644288]
9. Li Z, Zuo T, Wei X, Ding N. ICT Self-efficacy scale: the correlations with the age of first access to the internet, the age at first ownership of a personal computer (PC), and a smartphone. Med Educ Online 2023;28(1):2151068 [FREE Full text] [doi: 10.1080/10872981.2022.2151068] [Medline: 36440825]
10. Barrios-Ulloa A, Cama-Pinto D, Arrabal-Campos FM, Martínez-Lao JA, Monsalvo-Amaris J, Hernández-López A, et al. Overview of mobile communications in Colombia and introduction to 5G. Sensors (Basel) 2023;23(3) [FREE Full text] [doi: 10.3390/s23031126] [Medline: 36772166]
11. Kim M, Son MH, Moon S, Cha WC, Jo IJ, Yoon H. A mixed reality-based telesupervised ultrasound education platform on 5G network compared to direct supervision: prospective randomized pilot trial. JMIR Serious Games 2025;13:e63448 [FREE Full text] [doi: 10.2196/63448] [Medline: 39819654]
12. Xiaofeng Q, Hui C, Jin P, Yujie M, Yuan G. Application of wireless intelligent microscopy interactive system in pathogenic biology experimental teaching [Article in Chinese]. Tongfang Knowledge Network (Beijing) Technology Co, Ltd 2024;26(5):397-400. [doi: 10.13754/j.issn2095-1450.2024.05.10]
13. Hoyt G, Bakshi CS, Basu P. Integration of an audiovisual learning resource in a podiatric medical infectious disease course: multiple cohort pilot study. JMIR Med Educ 2025;11:e55206 [FREE Full text] [doi: 10.2196/55206] [Medline: 39935004]
14. Herman P, M Kibusi S, C Millanzi W. Effectiveness of an interactive web-based clinical practice monitoring system on enhancing motivation in clinical learning among undergraduate nursing students: longitudinal quasi-experimental study in Tanzania. JMIR Med Educ 2025;11:e45912 [FREE Full text] [doi: 10.2196/45912] [Medline: 40267464]
15. Brassil CE, Couch BA. Multiple-true-false questions reveal more thoroughly the complexity of student thinking than multiple-choice questions: a Bayesian item response model comparison. IJ STEM Ed 2019;6(1):16. [doi: 10.1186/s40594-019-0169-0]
16. Hubbard JK, Potts MA, Couch BA. How question types reveal student thinking: an experimental comparison of multiple-true-false and free-response formats. CBE Life Sci Educ 2017;16(2):ar26 [FREE Full text] [doi: 10.1187/cbe.16-12-0339] [Medline: 28450446]
17. Notice on Issuing the Measures for Ethical Review of Life Science and Medical Research Involving Human Subjects (National Health and Family Planning Commission Science and Education Document No. 4 [2023]). Central People's Government of the People's Republic of China. 2023. URL: https://www.gov.cn/zhengce/zhengceku/2023-02/28/content_5743658.htm [accessed 2025-11-06]
18. Liu K, Zhang W, Li W, Wang T, Zheng Y. Effectiveness of virtual reality in nursing education: a systematic review and meta-analysis. BMC Med Educ 2023;23(1):710 [FREE Full text] [doi: 10.1186/s12909-023-04662-x] [Medline: 37770884]
19. Ba H, Zhang L, Yi Z. Enhancing clinical skills in pediatric trainees: a comparative study of ChatGPT-assisted and traditional teaching methods. BMC Med Educ 2024;24(1):558 [FREE Full text] [doi: 10.1186/s12909-024-05565-1] [Medline: 38778332]
20. Ishak A, AlRawashdeh MM, Meletiou-Mavrotheris M, Nikas IP. Virtual pathology education in medical schools worldwide during the covid-19 pandemic: advantages, challenges faced, and perspectives. Diagnostics (Basel) 2022;12(7):1578 [FREE Full text] [doi: 10.3390/diagnostics12071578] [Medline: 35885484]
21. Başer A, Büyük B. Bridging the gap in medical education: comparing analysis of light microscopy and virtual microscopy in histology. PeerJ 2024;12:e17695 [FREE Full text] [doi: 10.7717/peerj.17695] [Medline: 39026537]

22. Iwata Y, Iwata Y, Iida H, Inamori M, Maeda S. Using a smartphone application as a tool for english learning among medical staff and students in Japan. Adv Med Educ Pract 2023;14:167-182 [FREE Full text] [doi: 10.2147/AMEP.S394625] [Medline: 36880091]
23. Hayat AA, Shateri K, Amini M, Shokrpour N. Relationships between academic self-efficacy, learning-related emotions, and metacognitive learning strategies with academic performance in medical students: a structural equation model. BMC Med Educ 2020;20(1) [FREE Full text] [doi: 10.1186/s12909-020-01995-9] [Medline: 7079530]
24. Ma H, Niu A, Tan J, Wang J, Luo Y. Nursing students' perception of digital technology in clinical education among undergraduate programs: A qualitative systematic review. J Prof Nurs 2024;53:49-56. [doi: 10.1016/j.profnurs.2024.04.008] [Medline: 38997198]

Abbreviations

DMIS: digital microscopy interactive system
LAN: local area network

Edited by B Lesselroth; submitted 19.12.24; peer-reviewed by GS Ching, NH Mohamad Zainal, L Kayser; comments to author 13.05.25; accepted 30.09.25; published 24.12.25.

Please cite as:
Xu J, Sha J, Jia S, Li J, Xu L, Shao Z
Effectiveness of a 5G Local Area Network–Based Digital Microscopy Interactive System: Quasi-Experimental Design
JMIR Med Educ 2025;11:e70256
URL: https://mededu.jmir.org/2025/1/e70256
doi:10.2196/70256
PMID:

Pass/Fail Versus Tiered Grades and Academic Performance in Undergraduate Medical Education: Crossover Study

Boris Modrau[1,2], MD, PhD; Karina Frahm Kirk[1], MD, PhD; Sinan Mouaayad Abdulaimma Said[1], MD; Carsten Reidies Bjarkam[1], MD, PhD, Prof Dr; Lone Sunde[1], MD, PhD, Prof Dr; Jacob Bodilsen[1], MD, PhD; Jakob Dal[1], MD, PhD; Jette Kolding Kristensen[1], MD, PhD, Prof Dr; Jeppe Emmersen[3], PhD, PGCME; Mike Bundgaard Astorp[1], MD; Stig Andersen[1], MD, PhD, Prof Dr, PGCME

[1]Department of Clinical Medicine, Faculty of Health, Aalborg University, Selma Lagerløfs Vej 249, Gistrup, Denmark

[2]Stroke Unit, Department of Neurology, Aalborg University Hospital, Aalborg, Denmark

[3]Faculty of Health, Aalborg University, Gistrup, Denmark

Corresponding Author:
Boris Modrau, MD, PhD
Department of Clinical Medicine, Faculty of Health, Aalborg University, Selma Lagerløfs Vej 249, Gistrup, Denmark

Abstract

Background: The impact of Pass/Fail or Tiered grade assessment for exams in undergraduate medical education has caused much debate, but there is little data to inform decision-making. The increasing number of medical schools transitioned to a Pass/Fail assessment has raised concerns about medical students' academic performance. In 2018, during the undergraduate medical curriculum reform at the Faculty of Medicine, Aalborg University changed some exams from Pass/Fail to Tiered grade and vice versa for other exams. These changes provide an opportunity to evaluate the different assessment forms.

Objective: This study aimed to evaluate medical students' academic performance at the final licensing exam in relation to the exam grading principle.

Methods: This single-center cohort study at Aalborg University Medical School, North Denmark Region, assesses the change from 2-digit Tiered grade to Pass/Fail evaluation and vice versa of undergraduate medical students' exams after the 4th and 5th year clinical training modules from Autumn 2015 through Spring 2023. The primary outcome was (1) the average grades at the final licensing exam and (2) the number of students failing exams during the previous two years.

Results: Among the total of 7634 exams, 7164 4th and 5th year clinical training exams were included in the comparisons, of which 3047 (42.5%) were Pass/Fail exams and 4117 (57.5%) were Tiered grade exams. The frequency of students failing exams was 3.3% (n=101/3047) at Pass/Fail and 1.97% (81/4117) with Tiered grade exams (P<.001). This difference was leveled out when counting the near-failure tiered grade as Fail. Tiered grade exams did not differ between semesters (P=.99) nor show a time trend at the 4th year (P=.66). The final licensing exam grades were unaltered (P=.47).

Conclusions: Contrary to our expectation, Pass/Fail exams exhibited a higher fail rate compared to Tiered grade exams without lowering the final academic performance. These results suggest that a shift from Tiered grades to Pass/Fail assessment redirects the focus from rewarding high performance to ensuring standards are maintained among underperforming students.

(JMIR Med Educ 2025;11:e74975) doi:10.2196/74975

KEYWORDS

exam assessment methods; pass/fail; tiered grade; evaluation methods; undergraduate; medical education; exam; academic achievement

Introduction

The importance of exams for undergraduate medical education cannot be overestimated [1], and scores in assessments is a delicate matter gaining much attention [2]. Pass/Fail grading has the benefits of reduced stress, enhanced well-being, supporting a less competitive learning environment with a greater focus on learning [3,4]. This is of interest as a narrative review reported a high number of medical students experienced test anxiety [5]. Furthermore, Tiered grading systems do not consistently correlate with future academic performance [6], which may support a transition towards Pass/Fail grading. Accordingly, a number of medical schools have transitioned to Pass/Fail and the National Board of Medical Examiners in the United States has changed reporting the Step 1 United States Medical Licensing Examination (USMLE) from a scored test to a Pass/Fail assessment.

While these changes support a learning environment in which students learn to become excellent physicians rather than exam-experts, the impact on academic performance has been a major concern. Even though the shift to Pass/Fail grading in the USMLE Step 1 resulted from good intentions, it might have had unintended consequences. Exam-related anxiety might even increase, as medical students now only have one chance to obtain a top score on Step 2 of the USMLE [7]. This might have consequences in the selection for residency programs as top scores have been used as a predictor for a successful residency [8] and as an important selection parameter [9]. However, presumptions and preconceptions may blur a clearer view on grading at exams, and data are needed to support decision-making.

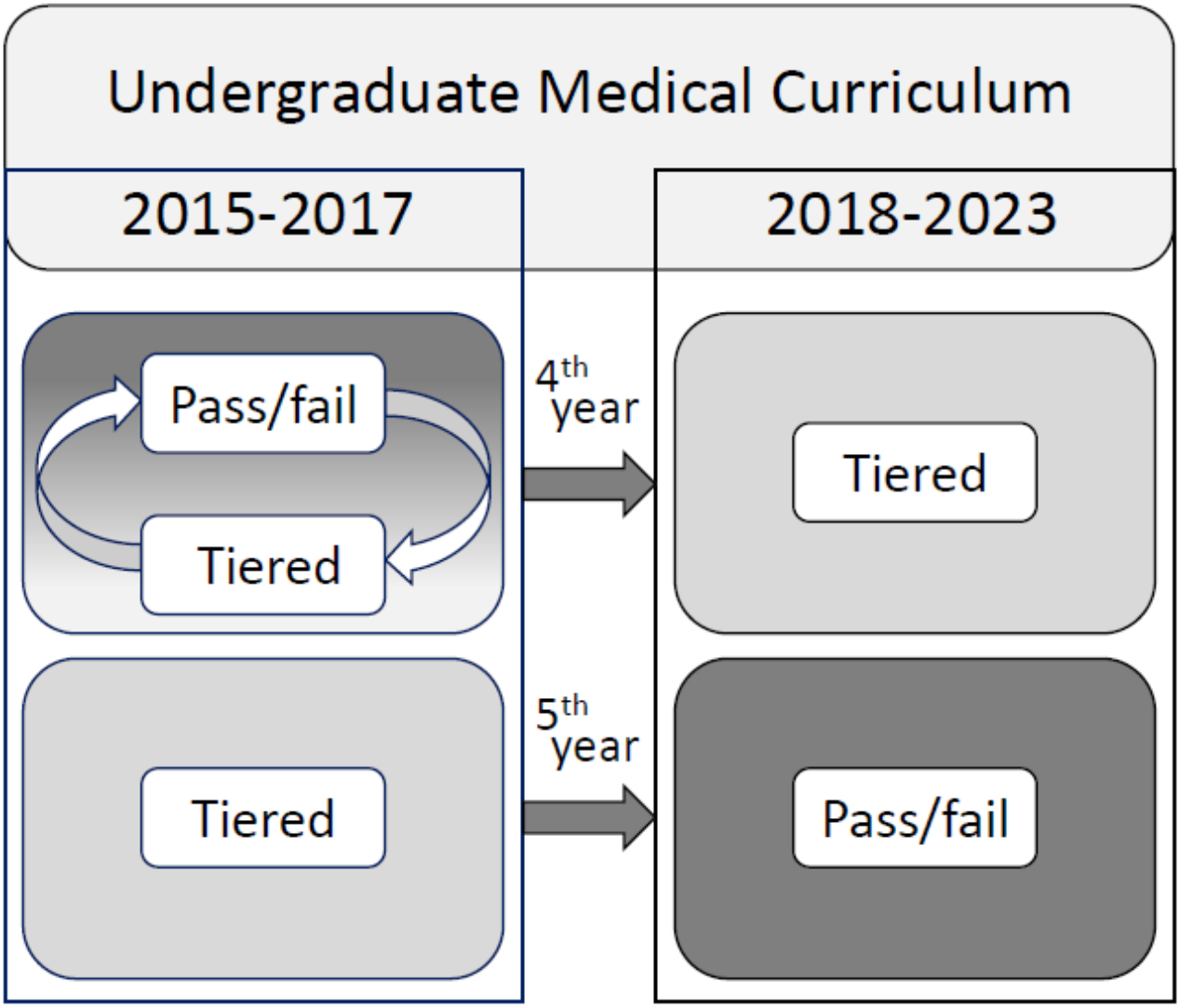
Exam outcomes are readily available as all assessments can be retrieved from Digital Exam, thus supporting quality control by regularly evaluating study outcomes. The curriculum reform of 2018 at Aalborg University Medical School changed assessment at exams between Pass/Fail and Tiered grades while retaining learning objectives, learning principles, and module organization for the 4th and 5th year. Study Board discussions and reflections among a study lead group, leading up to the decision of revising grading principles, inspired the present evaluation, as the changed grading principles provided a unique opportunity for empirically grounded evaluation of the shift from Tiered grades to Pass/Fail assessment in undergraduate medical education.

The aim of this study was to evaluate the shift from Tiered grades to Pass/Fail assessment in undergraduate medical education by comparing the outcome of Tiered grades and Pass/Fail grading at exams at the 4th and 5th year from a large, longitudinal dataset with a crossover design, and to evaluate medical students' academic performance at the 6th year final licensing exam.

Methods

Medical undergraduate education in Denmark is a six-year program split into a Bachelor's program of three years focusing on basic sciences, and a Master's program of three years with a clinical focus. The transformation from basic sciences into clinically based learning included the medical students entering the clinical environment from day 1 in their 4th year. The first clinical year, being the 4th year in the undergraduate medical education, alternated between Pass/Fail at winter exams and Tiered grades at summer exams until the 2018 curriculum reform (Figure 1).

Figure 1. Change of grading in the exams with the 2018 reform of the undergraduate medical curriculum at Aalborg University.



Year-4 exam assessments (upper, left) changed from alternating Pass/Fail and Tiered grades to Tiered grades for all exams (upper, right). Year-5 exam assessments changed from Tiered Grades (lower, left) into Pass/Fail (lower, right).

The 2018 reform [6] changed all 4th year exam assessments to use Tiered grades with the aim of supporting students' awareness of their academic level when entering the clinical years. Year-4 exams were oral exams with patients for Internal Medicine and Surgery, and a multiple choice questionnaire for Pathology (Table 1).

Table . Overview of exam type, scoring, and number of examinations split by the two time periods before and after the 2018 curriculum reform.

Study year and specialty	Examination type	Scoring				Number of examinations		Total number
		2015 - 2017		2018 - 2023		2015 - 2017	2018 - 2023	2015 - 2023
		Autumn	Spring	Autumn	Spring			
4th								
Surgery	Oral patient based[a]	Pass/ Fail	Tiered grades	Tiered grades	Tiered grades	327	593	920
Internal Medicine	Oral patient based[a]	Pass/ Fail	Tiered grades	Tiered grades	Tiered grades	327	602	929
Pathology	Multiple-choice questionnaire[b]	Pass/ Fail	Tiered grades	Tiered grades	Tiered grades	334	1204	1538
5th								
Gynecology/ Obstetrics	Oral case based	Tiered grades	Tiered grades	Pass/ Fail	Pass/ Fail	209	425	634
Neurology/ Neurosurgery	Oral case based	Tiered grades	Tiered grades	Pass/ Fail	Pass/ Fail	201	421	622
Ophthalmology	Oral case based	Tiered grades	Tiered grades	Pass/ Fail	Pass/ Fail	202	429	631
Otorhinolaryngology	Oral case based	Tiered grades	Tiered grades	Pass/ Fail	Pass/ Fail	203	422	625
Pediatrics	Oral patient based[c]	Tiered grades	Tiered grades	Pass/ Fail	Pass/ Fail	209	426	635
Psychiatry	Written case based	Tiered grades	Tiered grades	Pass/ Fail	Pass/ Fail	203	427	630
6th								7164
Final licensing exam	Oral patient based	Tiered grades	Tiered grades	Tiered grades	Tiered grades	—[d]	470	470
								7634

[a]Students handed in a portfolio of 12 admission papers to be granted access to the oral exam.

[b]MCQ, Multiple Choice Questionnaire of a case and photo with 4 - 5 options for an answer.

[c]The patient could be swapped for a case.

[d]Not applicable.

In the succeeding 5th year, students were expected to have adjusted to clinically based learning, and the 2018 reform changed all exam assessments from Tiered Grades to Pass/Fail for all modules (Figure 1). The aim of switching to Pass/Fail evaluation for all exams during the 5th year was to support the students' focus on learning to be a medical doctor rather than studying for exams. The year-5 exams were case-based oral exams for Gynecology/Obstetrics, Neurology, Neurosurgery, Ophthalmology, and Otorhinolaryngology, oral exam with patients for Pediatrics, and a written stepwise case-based exam for Psychiatry (Table 1).

The final 6th year used Tiered Grades throughout the observation period (Table 1) for this study to comply with legislation for final medical exams in Denmark.

Admission criteria in Denmark are set by legislation, and they rely heavily on average grade points from high school. Thus, the admission principle was unchanged throughout the study period, making the students entering medical school on similar terms during the years of the present study.

At Aalborg University Medical School, problem-based learning (PBL) is the learning principle practiced during the three years of the Bachelor's program, which feeds into the Master's program's patient-based learning [10]. The first two years of the Master's program comprises a series of clinical placements in the morning, supported by patient-based, theory-led small-group PBL-tutorials engaging in discussions with their tutors in the afternoon. Medical students are referred to as apprentice-doctors to support a learning focus on what is needed to be a doctor at the basic level. The final year is theory based aiming to foster deep learning when revisiting all aspects of medicine from the previous five years, thus supporting a spiral curriculum.

The learning objectives, module organization, and teaching principles were unaltered throughout the study period. The

academic staff did not change during the years of the study, and examiners were unaware of the present evaluation.

The 4th year consisted of internal medicine, surgery, and pathology. The 5th year comprised Pediatrics, Gynecology/Obstetrics, Neurology, Neurosurgery, Otolaryngology, Ophthalmology, and Psychiatry. All exams for all modules were included in the present evaluation, and the modules were unaltered except for the grading principles. The subgroup with alternating evaluation and the subgroup with permanent change from Tiered Grades to Pass/Fail grading were analysed. Licensing exams in the 6th year were included in the analysis for evaluation of final academic performance.

The Danish grading scale has seven steps, and it is almost comparable to the European Credit Transfer and Accumulation System (ECTS) scale A-F: 12; 10; 7; 4; 02; 00; and −03. While 02 is a Pass, it is the lowest of five pass grades and represents near-fail. Conversely, 00 and −03 represent a clear Fail without modification, and the delineation of failing exams is quite distinct. Thus, Pass/Fail was computed from the 2-digit Tiered grade exam assessments.

Statitical analysis

Results are reported in percent and in the crude numbers, and comparisons were performed using the χ^2 test for comparison of proportions, the Kruskal-Wallis test for comparing numerical grades between semesters, and the Kendall test to test for trend in Tiered grade exams over time. Internal validity was explored by monitoring grade point averages at the 4th year exams using tiered grades. As student cohorts, examiners, and formats remained stable across the years of the study, no additional adjustment for confounders was performed. Student-level variables were not available for the analysis, but in accordance with unaltered entry criteria, age and gender distributions were similar across the study years. Statistical analyses were performed using SPSS Statistics for Windows (version 13.0; IBM Corp).

Ethical Considerations

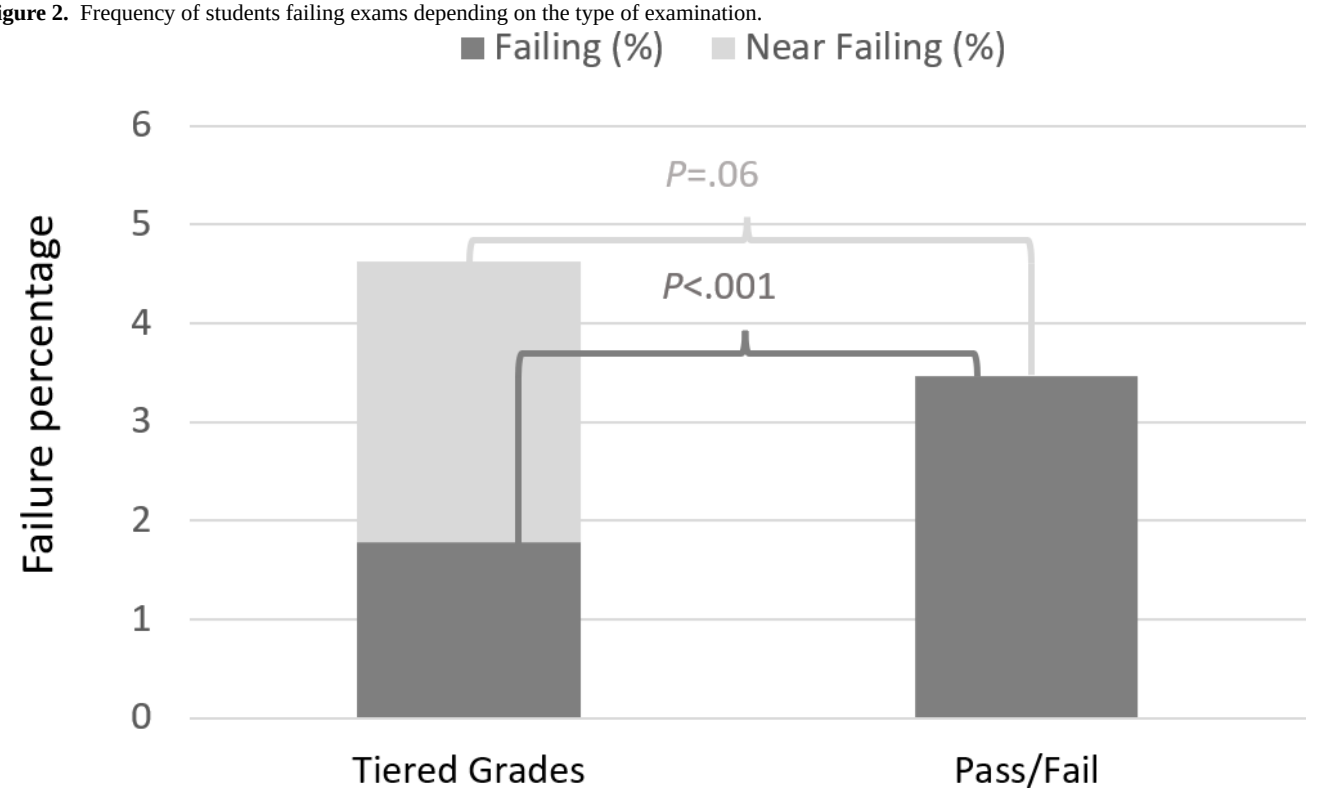
All data on grades was retrieved and reviewed for quality assurance in medical education, and grades were reported anonymously to Aalborg University using a digital solution (Digital Eksamen, Arcanic A/S, Copenhagen). The need for ethical approval or informed consent was waivered for anonymous data retrieved for quality assurance. Still, the presented research meets the recommendations by The Danish Code of Conduct for Research Integrity, with the primary material being digital raw data provided as an anonymous list of grades. Participants or the public were not involved in the design, conduct, reporting or dissemination plans for this study.

Results

Our evaluation comprises 7634 exams for 16 semesters from the Autumn of 2015 through Spring of 2023. In total, 7164 exams with 66 exam sessions prior to, and 78 exam sessions following the 2018 curriculum reform, were included in the comparisons during the 4th and 5th year. Of these 3047 (42.5%) were Pass/Fail exams and 4117 (57.5%) were Tiered grade exams. Additionally, 470-Tiered grade final licensing exams at the 6th year were included for evaluation of the final academic performance.

The overall frequency of students failing exams is illustrated in Figure 2.

Figure 2. Frequency of students failing exams depending on the type of examination.



The frequency of students failing exams (dark gray bars) was 3.3% (n=101/3047) when using Pass/Fail and 2.0% (81/4117) with Tiered grade evaluations (χ^2 test, P<.001). This difference was leveled out (χ^2 test, P=.055) when including the near-failure Tiered grade as fail (light gray bar).

In the subgroup of 988 examinations alternating between Tiered grade and Pass/Fail (upper left box in Figure 1), the frequency of students failing exams was 1.8% (9/497) with Pass/Fail and 0.8% (4/491) with Tiered grade exams (χ^2 test, P=.20). In the subgroup of 5th year students switching permanently from Tiered grade to Pass/Fail (lower boxes in Figure 1), the frequency of students failing exams was 3.6% (92/2550) with Pass/Fail and 1.1% (13/1227) with Tiered grade assessments (χ^2 test, P<.001).

The level of grades at the Tiered grade exams did not differ between semesters (Kruskal-Wallis test, P=.99) nor show a time trend at the 4th year (Kendall test, P=.66). The average grades of the 6th year final licensing exam (Kendall test, P=.47) and SD of about 3 were unchanged over the entire study period (Table 2).

Table . Average grades at year-4 before and after the 2018 reform and the final licence exams at year 6.

Years	Year-4 grades, mean (SD)	Year-6 grades, mean (SD)
2016	8.8 (2.7)	—
2017	8.5 (2.8)	—
2018	8.5 (2.9)	—
2019	8.4 (2.9)	9.3 (2.8)
2020	8.5 (2.9)	8.5 (3.0)
2021	8.3 (3.0)	9.1 (2.8)
2022	8.3 (3.2)	8.4 (3.4)
2023	8.6 (3.0)	8.7 (3.1)

The Danish Tiered Grade scale is numerical, and categories are comparable to the ECTS scale A-F: 12; 10; 7; 4; 02; and 00, −03 are the equivalent of A; B; C; D; E; and F.

Discussion

Principal Findings

Changing from Tiered grades to Pass/Fail assessment for exams in undergraduate medical education during the 4th and 5th years increased the number of students failing exams. This difference leveled out when including the near-fail Tiered grade as Fail. The final academic performance did not change.

A recent meta-analysis reported limited data from the previous millennium and showed no difference in performance when comparing tiered and Pass/Fail examinations [11]. Similar results were found in more recent data reported by Ange et al [12]. Alternating assessment was recently reported to cause higher average grades when using Tiered grades, but data on the number of students passing exams were lacking [13]. Despite these uncertainties [14], a number of medical schools are transitioning to Pass/Fail grading to improve psychological well-being and satisfaction among medical students [15]. This change raises concerns about academic performance and the risk of overlooking underperforming students [16]. Our study addresses these points and may inform decision-makers on the transition from Tiered grades to Pass/Fail in medical education.

Contrary to our expectation, the number of students not passing exams was higher with Pass/Fail compared to Tiered grades. A similar trend was found in the subgroups when Pass/Fail and Tiered grade exams were alternating over the same academic year, and when looking over time. In both comparisons, more students passed the Tiered graded exams compared to the Pass/Fail exams. One possible explanation is the lack of opportunity to address underperformance with a low grade, prompting examiners to shift their focus from rewarding high achievers to ensuring standards are met by underperforming medical students. Another reason might be related to learning differences among students, in that Tiered grades might support students' self-evaluation of performance linked with higher extrinsic motivation. However, we saw no change in the average final licensing exam grades, which suggests an absence of influence on the students' final academic performance.

The choice of grading principles is made by the study board. The curriculum reform in 2018 changed the grading principles based on discussions on study board meetings in the year leading up to the decision, with evaluation of and input to these discussions by members of a study lead group (BM, JE, SA). Only two members of the study board were academic staff representatives of the clinical study years (2013 - 2025, SA). Some clinical academic staff expressed reservations on the switch from Tiered grades to Pass/Fail, but accepted this decision following information on the discussions in the study board, providing a background for the decision. Hence, the evaluators at the 7634 exams implemented the decision on the grading principle without opposing this. Moreover, the final exams all included an external evaluator to oversee if the examination covered the learning objectives and if the grading met the standards set by the curriculum. All external evaluators were academic staff from other universities knowledgeable about the topic being tested. None of these were involved in the change in grading principles at the 4th and 5th years, and the external observation contributed to solid evaluations at the final exams, illustrating that grading principles did not influence the final academic performance by medical students.

Strength and limitations

The study's credibility was strengthened by the large, longitudinal dataset, crossover design, and the broad range of medical specialties included. The retrospective collection of assessment data has likely prevented observer bias, as the assessors were unaware of the present evaluation. Grading is a definite variable, and it strengthens the credibility that all assessments over the years under study were included. The internal validity was strengthened by unaltered admission criteria, teaching principles, learning objectives, and module organization throughout the study period. Still, we cannot rule out confounding factors not accounted for. The overall low failure rate in the Master's program stands out, and it may be attributed to the fact that only medical students passing the Bachelor's program progress into the Master's program. The single-center design may limit the generalizability of the findings. However, medical students are high achievers across settings, and we consider the core results applicable in other settings. We used digital reporting of assessment, which may have reduced observer bias. Many medical schools and educators have transitioned to digital exams. Most of our exams included real patients with a focus on clinical reasoning and demonstration of skills, which restricted us to the use of digital reporting of grades. Extending this thought, the use of the overview of grades provided by the digital reporting of grades for proxy assessment of quality of teaching and learning inspired the present evaluation by making grades easily accessible, a distinct benefit of digital reporting of assessments.

Conclusion

This empirically grounded evaluation of the shift between Tiered grades and Pass/Fail assessment in undergraduate medical education, utilizing a natural experiment following curriculum reform, was strengthened by the large, longitudinal dataset and crossover design.

Pass/Fail grading at exams in undergraduate medical education was related to a higher risk of failing than tiered grad exams, a difference that leveled out when including the near-fail Tiered grade as fail. The data suggested no influence on the students' final academic performance. Thus, we argue that a shift from Tiered grades to Pass/Fail assessment causes a shift in focus from rewarding high performance to upholding standards among underperforming medical students. As Pass/Fail evaluations correlate with enhanced well-being and reduced stress compared to Tiered grade, Pass/Fail may be preferred over Tiered grade assessments in high-stakes exams.

Acknowledgments

Cecilie Mörck Bachmann is acknowledged for data extraction.

Data Availability

Data may be available for a third party upon reasonable request, but they are not publicly available. This retrospective analysis was considered quality control and improvement, and no predefined study protocol was required.

Authors' Contributions

Conceptualization: BM (lead), SA (equal), JE (supporting)
Data curation: SA (lead), BM (supporting)
Formal analysis: SA (lead), BM (equal)
Funding acquisition: BM (lead), SA (equal)
Investigation: BM (lead), SA (equal), KFK (supporting), SS (supporting), CB (supporting), LS (supporting), JB (supporting), JD (supporting), JKK (supporting), JE (supporting), MA (supporting)
Methodology: BM (lead), SA (equal)
Project administration: BM (lead), SA (equal), JE (supporting)
Resources: SA (lead), BM (equal)
Supervision: SA (lead), BM (equal)
Validation: SA (lead), BM (equal)
Visualization: SA (lead), BM (equal)
Writing – original draft: BM (lead) SA (equal)
Writing – review & editing: BM (lead), SA (equal), KFK (supporting), SS (supporting), CB (supporting), LS (supporting), JB (supporting), JD (supporting), JKK (supporting), JE (supporting), MA (supporting)

Conflicts of Interest

JD receives lecture fees and unrestricted research grants from Ipsen, Pfizer, and Recordati.

References

1. Crane MA, Chang HA, Azamfirei R. Medical education takes a step in the right direction: where does that leave students? JAMA 2020 May 26;323(20):2013-2014. [doi: 10.1001/jama.2020.2950] [Medline: 32142102]

2. Schwartzstein RM, Roberts DH. Saying goodbye to lectures in medical school - paradigm shift or passing fad? N Engl J Med 2017 Aug 17;377(7):605-607. [doi: 10.1056/NEJMp1706474] [Medline: 28813217]
3. Astorp MS, Sørensen GVB, Rasmussen S, Emmersen J, Erbs AW, Andersen S. Support for mobilising medical students to join the COVID-19 pandemic emergency healthcare workforce: a cross-sectional questionnaire survey. BMJ Open 2020 Sep 16;10(9):e039082. [doi: 10.1136/bmjopen-2020-039082] [Medline: 32938602]
4. Wood DF. Problem based learning. BMJ 2003 Feb 8;326(7384):328-330. [doi: 10.1136/bmj.326.7384.328] [Medline: 12574050]
5. Alshareef N, Fletcher I, Giga S. The role of emotions in academic performance of undergraduate medical students: a narrative review. BMC Med Educ 2024 Aug 23;24(1):907. [doi: 10.1186/s12909-024-05894-1] [Medline: 39180051]
6. Andersen S, Stentoft D, Emmersen J, Rasmussen S, Birkelund S, Nøhr S. Contention over undergraduate medical curriculum content. Int J Med Educ 2019 Dec 16;10:230-231. [doi: 10.5116/ijme.5de7.7516] [Medline: 31859263]
7. Ozair A, Bhat V, Detchou DKE. The US Residency Selection Process After the United States Medical Licensing Examination Step 1 Pass/Fail Change: overview for applicants and educators. JMIR Med Educ 2023 Jan 6;9(1–13):e37069. [doi: 10.2196/37069] [Medline: 36607718]
8. Bhandarkar AR, Graffeo CS, Johnson J. Stepping Up: How U.S. Neurosurgery training programs can innovatively assess resident applicants in a post-step 1 world. World Neurosurg 2020 Oct;142:291-293. [doi: 10.1016/j.wneu.2020.07.078] [Medline: 32683001]
9. Huq S, Khalafallah AM, Botros D, et al. Perceived impact of USMLE Step 1 pass/fail scoring change on neurosurgery: program director survey. J Neurosurg 2020 Sep 1;133(3):928-935. [doi: 10.3171/2020.4.JNS20748] [Medline: 32559749]
10. Stentoft D. Problem-based projects in medical education: extending PBL practices and broadening learning perspectives. Adv Health Sci Educ Theory Pract 2019 Dec;24(5):959-969. [doi: 10.1007/s10459-019-09917-1] [Medline: 31641941]
11. Wang A, Karunungan KL, Story JD, et al. Reimagining a pass/fail clinical core clerkship: a US residency program director survey and meta-analysis. BMC Med Educ 2023 Oct 24;23(1):788. [doi: 10.1186/s12909-023-04770-8] [Medline: 37875929]
12. Ange B, Wood EA, Thomas A, Wallach PM. Differences in Medical Students' Academic Performance between a Pass/Fail and Tiered Grading System. South Med J 2018 Nov;111(11):683-687. [doi: 10.14423/SMJ.0000000000000884] [Medline: 30392003]
13. Ba-Ali S, Jemec GBE, Sander B, Toft PB, Homøe P, Lund-Andersen H. The effect of two grading systems on the performance of medical students during oral examinations. Dan Med J 2017 Mar;64(3):A5328. [Medline: 28260593]
14. Gonnella JS, Erdmann JB, Hojat M. An empirical study of the predictive validity of number grades in medical school using 3 decades of longitudinal data: implications for a grading system. Med Educ 2004 Apr;38(4):425-434. [doi: 10.1111/j.1365-2923.2004.01774.x] [Medline: 15025644]
15. Spring L, Robillard D, Gehlbach L, Simas TAM. Impact of pass/fail grading on medical students' well-being and academic outcomes. Med Educ 2011 Sep;45(9):867-877. [doi: 10.1111/j.1365-2923.2011.03989.x] [Medline: 21848714]
16. Bloodgood RA, Short JG, Jackson JM, Martindale JR. A change to pass/fail grading in the first two years at one medical school results in improved psychological well-being. Acad Med 2009 May;84(5):655-662. [doi: 10.1097/ACM.0b013e31819f6d78] [Medline: 19704204]

Abbreviations

ECTS: European Credit Transfer and Accumulation System
PBL: Problem-based learning
USMLE : United States Medical Licensing Examination

Edited by P Kanzow; submitted 15.04.25; peer-reviewed by I Drazic, N Mehta; revised version received 14.10.25; accepted 14.10.25; published 02.12.25.

Please cite as:
Modrau B, Kirk KF, Said SMA, Bjarkam CR, Sunde L, Bodilsen J, Dal J, Kristensen JK, Emmersen J, Astorp MB, Andersen S
Pass/Fail Versus Tiered Grades and Academic Performance in Undergraduate Medical Education: Crossover Study
JMIR Med Educ 2025;11:e74975
URL: https://mededu.jmir.org/2025/1/e74975
doi:10.2196/74975

Insights Into History and Trends of Teaching and Learning in Stomatology Education: Bibliometric Analysis

Ziang Zou[1], PhD, MD; Linna Guo[2], PhD, MD

[1]Department of Gynecology and Obstetrics, The Third Xiangya Hopital, Central South University, Changsha, Hunan, China

[2]Department of Stomatology, The Second Xiangya Hopital, Central South University, 139 Renmin Middle Road, Changsha, Hunan, China

Corresponding Author:
Linna Guo, PhD, MD
Department of Stomatology, The Second Xiangya Hopital, Central South University, 139 Renmin Middle Road, Changsha, Hunan, China

Abstract

Background: Stomatology education has experienced substantial transformations over recent decades. Nevertheless, a comprehensive summary encompassing the entirety of this field remains absent in the literature.

Objective: This study aimed to perform a bibliometric analysis to evaluate the research status, current focus, and emerging trends in this field over the last two decades.

Methods: We retrieved publications concerning teaching and learning in stomatology education from the Web of Science core collection covering the period from 2003 to 2023. Subsequently, we conducted a bibliometric analysis and visualization using R-Bibliometrix and CiteSpace.

Results: In total, 5528 publications focusing on teaching and learning in stomatology education were identified. The annual number of publications in this field has shown a consistent upward trend. The United States and the United Kingdom emerged as the leading contributors to research. Among academic institutions, the University of Iowa produced the highest number of publications. The *Journal of Dental Education* was identified as the journal with the highest citation. Wanchek T authored the most highly cited articles in the field. Emerging research hotspots were characterized by keywords such as “deep learning,” “machine learning,” “online learning,” “virtual reality,” and “convolutional neural network.” The thematic map analysis further revealed that “surgery” and “accuracy” were categorized as emerging themes.

Conclusions: The visualization bibliometric analysis of the literature clearly depicts the current hotspots and emerging topics in stomatology education concerning teaching and learning. The findings are intended to serve as a reference to advance the development of stomatology education research globally.

(JMIR Med Educ 2025;11:e66322) doi:10.2196/66322

KEYWORDS

teaching; learning; stomatology; education; bibliometric analysis; trend; stomatology education; visualization; education; R-Bibliometrix; CiteSpace; university; innovation; teaching modality; web of science; WOS

Introduction

In recent years, innovations and restructuring in stomatology education have triggered global reforms in the field. In stomatology education, teaching and learning are both interdependent and mutually reinforcing. Effective education demands not only that instructors possess robust professional knowledge and use sound pedagogical strategies but also that students actively participate, provide feedback (FB), and engage in self-directed learning. The dynamic interaction between teaching and learning facilitates the effective transfer and application of knowledge, ultimately resulting in the cultivation of stomatology professionals who are equipped with both advanced technical skills and strong clinical reasoning abilities.

Teaching and learning in stomatology education have undergone profound and meaningful changes in the past two decades. For instance, significant research efforts have been dedicated to using digital and artificial intelligence (AI) tools to enhance the quality of education and exploring the effectiveness and response of online and distance teaching modalities. Hence, it is crucial to understand the latest trends and developments in teaching and learning in stomatology education. However, keeping up with the rapid changes in stomatology education research is challenging. Scientific publications are important for academic communication and clinical practice guidelines. Evaluative bibliometrics, a quantitative science, assesses research performance to identify influential works in medical practice and research. Unlike literature reviews, which offer qualitative insights, bibliometric analysis provides quantitative data on citations, authorship, and research collaboration [1,2].

Although various fields have been studied, teaching and learning in stomatology education research are limited, with few quantitative analyses. Therefore, bibliometric analysis is essential to fill this gap, offering critical data to identify trends and guide future research.

In this study, we conducted a comprehensive bibliometric analysis of teaching and learning in stomatology education using Citespace software (developed by Chaomei Chen, College of Computing and Informatics, Drexel University, Philadelphia, PA, USA) to generate visual representations. The aim was to systematically evaluate the research landscape, current focal areas, and emerging trends in teaching and learning in stomatology education over the past two decades, emphasizing key achievements and outlining potential future research directions in this field.

Methods

Search Strategy and Eligibility Criteria

Scholarly articles and literature on teaching and learning in stomatology education were retrieved using the Web of Science Core Collection (WoSCC), encompassing the Science Citation Index Expanded and other relevant citation indices. To ensure the validity and reproducibility of our search strategy, we implemented a multistep validation process. A pilot search was conducted in 2023 using a subset of keywords, yielding 200 randomly selected articles that were manually screened by the authors. This process resulted in a precision rate of 92% based on which minor refinements were made to keyword synonyms. Second, the search strategy was developed in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist to ensure methodological rigor and transparency. Third, to assess the comprehensiveness of our chosen database and minimize selection bias, we cross-validated the WoSCC coverage with PubMed and Scopus using a random sample of 100 articles, confirming a 95% overlap in relevant literature. Finally, both controlled vocabulary (MeSH terms) and free-text keywords were used to ensure the search captured a wide range of relevant studies across varying indexing practices. To ensure comprehensive literature retrieval, we designed the search strategy by mirroring MeSH hierarchies and incorporating their synonymous expressions ('Oral medicine' [Mesh] → dentistry/stomatology); ('Education' [Mesh] → teaching/training/learning).

The final search string combined controlled vocabulary (MeSH terms) and free-text terms, as follows:

1# :(((TS=(teaching)) OR TS=(training)) OR TS=(learning)) OR TS=(education)

2# :(((TS=(oral medicine)) OR TS=(stomatology)) OR TS=(dentistry))

3# :1# and 2#

The duration of this span is from January 2003 to December 2023. Articles lacking unique content and those published in languages other than English were excluded.

Ethical Considerations

As the data were directly obtained from the database, ethical approval was not required.

Bibliometric Analysis

The data from these articles were imported and integrated using the bibliometrix package (version 4.4.0) in the R programming language. Our research applied several bibliometric analysis techniques, including keyword co-occurrence and historical direct citation analyses. For this study, we used CiteSpace (version 6.2R6) to perform social network analysis and to identify developmental dynamics, hotspots, future trends, and key aspects within the scientific literature on a specific topic. Burst detection methodology was applied to pinpoint significant keywords or references that experienced a notable surge in frequency within a defined period. In addition, we conducted a clustering (co-citation) analysis of references using a metric based on co-authored publications, allowing for the grouping of institutions or keywords that showed higher levels of collaboration within the same cluster. Clusters of related publications were generated using CiteSpace's fully automated cocitation clustering algorithm with the following parameters:

- Clustering method: Louvain modularity optimization (to maximize intracluster connectivity and intercluster distinction)
- Similarity threshold: default cosine similarity (0.5) to balance sensitivity and specificity
- Time slicing: 2-year intervals (2003–2023) to capture temporal trends
- Node type: author keywords+cited references to integrate semantic and bibliometric signals
- Labeling: algorithm-generated labels via Term Frequency-Inverse Document Frequency (TF-IDF)weighting of high-frequency keywords within each cluster.

No manual adjustments were made to cluster labels to avoid subjective bias. Network maps were used to illustrate variations in the quantity or frequency of published records across clusters of similar research topics, with node size and color indicating these differences. The strength of collaborations between nodes was depicted by connecting lines, where thicker lines indicated stronger collaborations. To further enhance data visualization, we used an overlay visualization technique in which the color of each node represented the average year of occurrence for each institution or keyword. The data were saved in plain text format with full record and cited references from WoSCC and then imported into CiteSpace for further analysis.

The bibliometric data for this study were retrieved from Web of Science, and no additional filtering was applied to remove self-citations. However, as none of the authors have previously published related articles in this field, self-citations do not impact the findings of this analysis.

Results

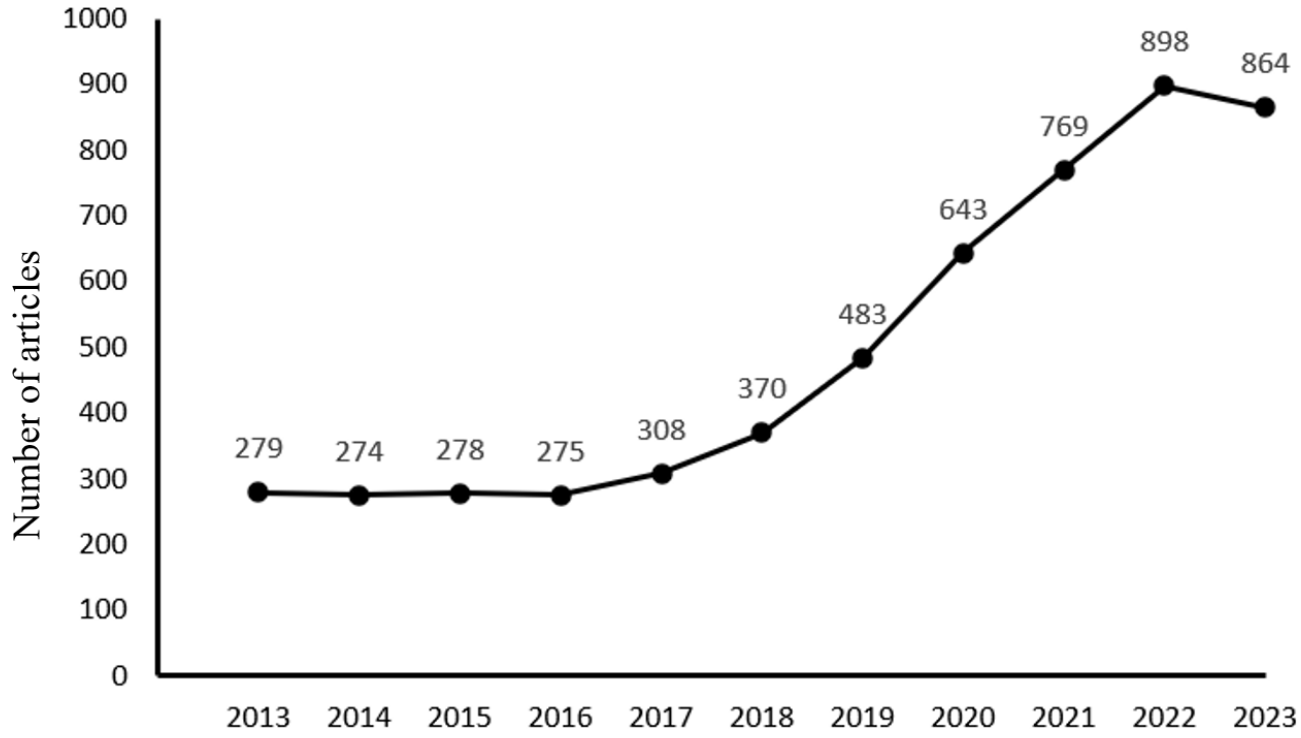
Brief Description of Teaching and Learning in Stomatology Education Literature

As of December 31, 2023, a total of 8831 articles were retrieved from the WoSCC database. Of these, 6532 were published between 2003 and 2023, with 6381 articles written in English. The dataset includes 5528 research papers, representing 86.6% of the total, and 593 review articles, constituting 9.3%. The 5 leading WOS disciplines identified are dentistry oral surgery medicine (n=3375), education scientific disciplines (n=844), medicine general internal (n=510), public environmental occupational health (n=500), and health care sciences services (n=413).

Productivity Analysis

As demonstrated in 1, the annual publication volume from 2013 to 2023 exhibited distinct temporal patterns. The platform period (2012-2017) maintained stable output with minor fluctuations (270-308 articles per year), followed by an exponential growth phase (2018 - 2022) that saw publications surge from 370 to 898 articles per year. Although there was a slight decline in the number of publications from 2022 to 2023, the overall trend remains upward, with the growth rate generally following an exponential pattern. This indicates that international attention to research in stomatology education is steadily increasing and holds promising prospects (Figure 1).

Figure 1. Annual scientific production in teaching and learning of stomatology education.



Countries and Cooperation Networks Analysis

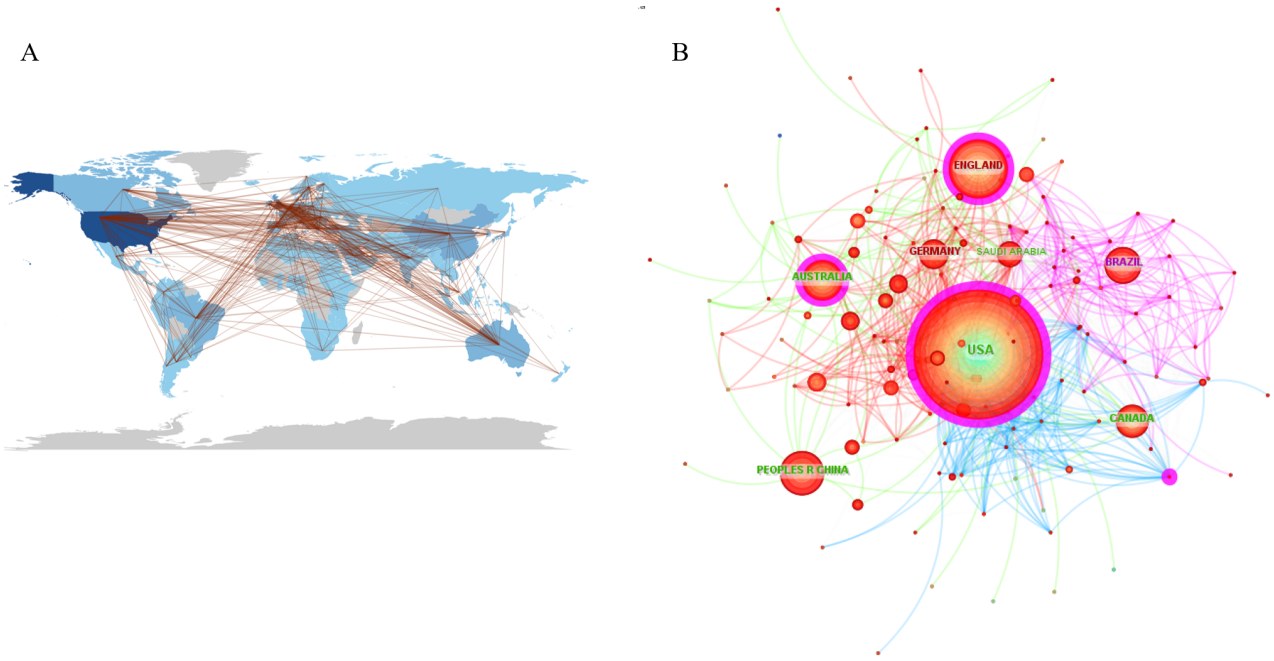
Since 2003, the literature on the research of teaching and learning in stomatology education has been published in 145 countries, mainly led by the following few countries: the United States, the United Kingdom, and China. The number of publications and the frequency of citations are 2 dimensions for analyzing the strength of scientific research, reflecting the attention and impact of a country or institution in a certain field. From the perspective of the number of publications, the top 5 countries are the United States (n=2085), the United Kingdom (n=400), China (n=443), Australia (n=435), and Brazil (n=361). From the perspective of literature citation frequency, the top 5 countries are the United States, the United Kingdom, China, Germany, and Canada. It is worth noting that the United States, the United Kingdom, and China all rank among the top 3 in terms of the number of publications and the frequency of citations, which proves that these 3 countries have strong scientific research capabilities in this field (Table 1).

Table . Top 10 productive countries concerning teaching and learning in stomatology education.

Country	Documents, n (%)	Citations	Average citations	Centrality
The United States	2085 (37.7)	25,599	16.00	0.24
The United Kingdom	800 (14.5)	7942	14.10	0.13
China	443 (8.0)	6125	15.30	0.04
Germany	310 (5.6)	4405	20.10	0.02
Canada	348 (6.3)	3053	14.30	0.02
Australia	435 (7.9)	3023	12.80	0.11
Brazil	361 (6.5)	2756	11.10	0.03
Turkey	161 (2.9)	2387	16.80	0.00

The data were imported into CiteSpace for a cooperation network analysis, where the centrality of each country was calculated. As illustrated in Figure 2B and detailed in Table 1, each circle represents a country; a darker center color indicates earlier publication dates. Higher centrality values reflect stronger collaboration with other nations. Circles with purple outer rings signify particularly high centrality. Research on teaching and learning in stomatology education began earlier in the United States, the United Kingdom, and Australia. The United States exhibits the highest centrality (0.20), followed by the United Kingdom (0.13) and Australia (0.11), suggesting these countries prioritize international collaboration in this field. Notably, while China (0.04), Germany (0.02), Canada (0.02), Brazil (0.03), and Turkey (0.00) rank among the top 10 in terms of publication and citation counts, their centrality is significantly lower than that of the United States, the United Kingdom, and Australia (Table 1). “Centrality” reflects a country’s role as a hub in global research collaboration. A higher centrality value (eg, the United States with 0.20) indicates stronger connectivity and influence in bridging diverse research communities. Therefore, the United States served as a central node connecting European and Asian institutions in stomatology education research. The national cooperation network analysis reveals that the connections between these countries and others are relatively robust, indicating extensive international cooperation (Figure 2).

Figure 2. Leading countries in teaching and learning of stomatology education. (A) Countries’ collaboration world map. (B) Countries’ cooperation network.



Institutions and Cooperation Analysis

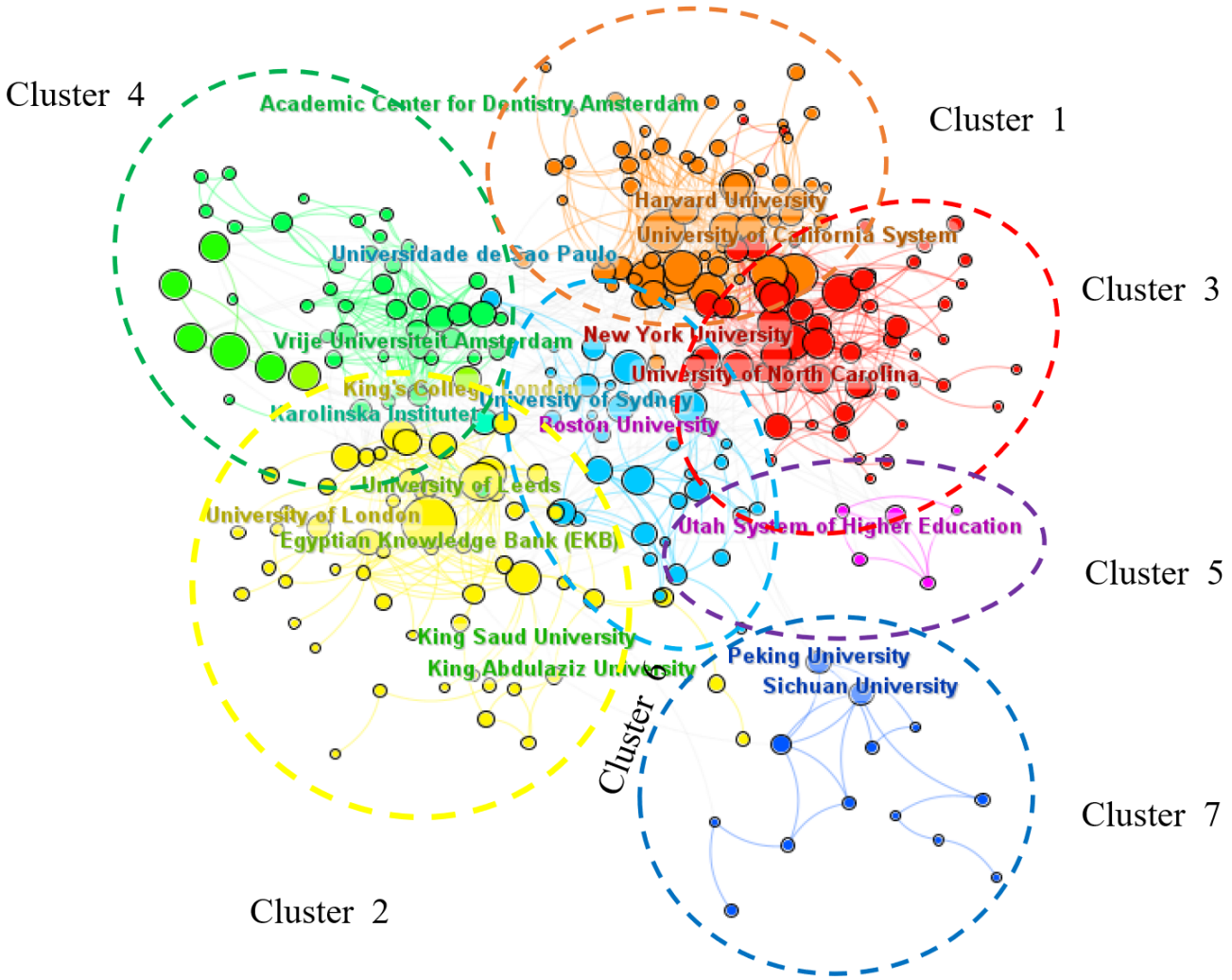
As presented in Table 2, the top 10 institutions in teaching and learning in stomatology education–related research are listed. The top 3 institutions are all from the United States, namely UNIV IOWA (n=305), UNIV MICHIGAN (n=288), UNIV N CAROLINA (n=257, Table 2), while KINGS COLL LONDON (n=236) from the United Kingdom, UNIV SYDNEY from Australia (n=155), and KING SAUD UNIV (n=160) and UNIV HONG KONG (n=153) from Asia have also made outstanding contributions. A comprehensive analysis of cooperative networks among institutions or universities was conducted. In the cooperative network analysis chart, a larger font size signifies that the institution not only began relevant research earlier but has also made more significant contributions to the field (Figure 3). As shown in Figure 3, the University of London,

the University of California, King's College London, Boston University, Harvard University, and Peking University have played an important role in the development of this field. In the network map, cluster analysis of cooperation among institutes and overlay visualization of the largest 7 clusters was shown in Figure 3. The top-ranked item by centrality is the National and Kapodistrian University of Athens (0.14, cluster 2). The second one is the Cleveland Clinic (0.09, cluster 6). The third is the Case Western Reserve University (0.08, cluster 6). The fourth is the University of Sheffield (0.07, cluster 4). The fifth is the Pennsylvania Commonwealth System of Higher Education (0.07, cluster 1).

Table . Top 10 productive institutions concerning teaching and learning in stomatology education.

Affiliation	Country	Publication counts
UNIV IOWA	USA	305
UNIV MICHIGAN	USA	288
UNIV N CAROLINA	USA	257
KINGS COLL LONDON	The United Kingdom	236
UNIV ILLINOIS	USA	185
KING SAUD UNIV	Saudi Arabia	160
UNIV SYDNEY	Australia	155
UNIV HONG KONG	China	153
UNIV WASHINGTON	USA	153

Figure 3. Visualization of active institutes and cluster analysis of cooperation among institutes.



Authors' Analysis

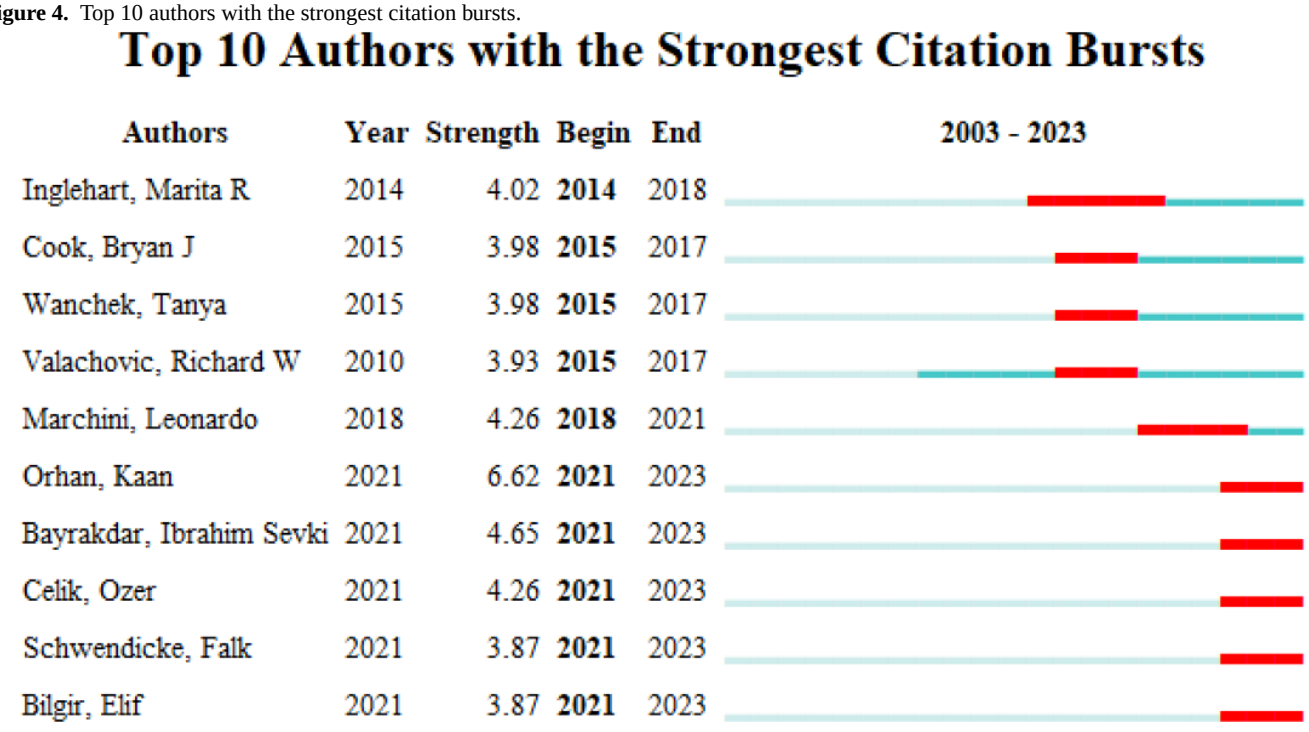
We conducted an analysis of the authors of the published articles and identified the top 5 authors by citation count: Wanchek T (n=132), Cook BJ (n=123), Valachovic RW (n=114), Mattheos N (n=91), and Lynch CD (n=91). The affiliations of these authors, including their respective countries and institutions, are detailed in Table 3. Notably, 3 of these authors are affiliated with the University of Virginia, the United States (Table 3).

Table . Top 10 most productive authors in teaching and learning in stomatology education.

Author	Country	Institution	Citations
Wanchek T	USA	University of Virginia	132
Cook BJ	USA	University of Virginia	123
Valachovic RW	USA	University of Virginia	114
Mattheos N	Sweden	Karolinska Institute	91
Lynch CD	Ireland	University Dental School and Hospital Wilton	88
Kalenderian E	USA	Harvard School of Dental Medicine	72
Anderson EL	USA	American Dental Education Association	71
Lee JY	South Korea	Wonkwang University College of Dentistry	70
Divaris K	USA	University of North Carolina	63
Wilson NHF	The United Kingdom	King's College London Dental Institute	63

Using the "urst" function, we identified the most active authors in the field, as illustrated in Figure 4. Analysis of the burst detection results indicates that teaching and learning in stomatology gained prominence starting in 2014, with significant contributions from Inglehart MR (USA). During 2021 to 2023, Orhan K emerged as the most active author in the field, achieving the highest burst strength of 6.62. Bayrakdar IS, with a burst strength of 4.65, ranked second and had an active period from 2021 to 2023. Consequently, Orhan K and Bayrakdar IS not only demonstrated strongest citation bursts but also positioned themselves as a leading researcher in recent years. In addition, Celik O, Schwendicke F, and Bilgir E have recently held prominent positions in this research frontier (Figure 4).

Figure 4. Top 10 authors with the strongest citation bursts.



Journal Analysis

The top 10 journals with the highest citation counts are presented in the table below. The 5 most cited journals include the *Journal of Dental Education* (n=9120), *European Journal of Dental Education* (n=4963), *Journal of Dental Research* (n=2679), *British Dental Journal* (n=2618), and *Journal of the American Dental Association* (n=2238). Impact factors (averaged over the past 5 years) and Journal Citation Reports (JCR) are detailed in Table 4. When examining the H-index of the cited articles, the leading 5 journals are *Journal of Dental Education* (H=36), *European Journal of Dental Education* (H=31), *British Dental Journal* (H=25), *Journal of the American Dental Association* (H=25), and *BMC Oral Health* (H=23), as presented in Table 4. Consequently, the *Journal of Dental Education* and the *European Journal of Dental Education* have been particularly influential, with a strong focus on research related to teaching and learning in dental education.

Table . Top 10 core journals on teaching and learning in stomatology education.

Journal	H_index	Counts	Citations	IF[a]	JCR[b]
Journal of Dental Education	36	841	9120	1.7	Q3
European Journal of Dental Education	31	444	4963	2.2	Q2
British Dental Journal	25	267	2618	2.1	Q2
Journal of the American Dental Association	25	87	2238	3.4	Q1
BMC Oral Health	23	129	1613	3.2	Q1
Journal of Dental Research	21	50	2679	5.7	Q1
Journal of Dentistry	20	55	1511	5	Q1
Academic Medicine	17	29	699	6.7	Q1
Community Dentistry and Oral Epidemiology	17	50	822	2.7	Q2
International Journal of Paediatric Dentistry	17	38	770	2.9	Q2

[a]IF: impact factors.

[b]JCR: Journal Citation Reports.

Research Focus and Frontiers Analyses on the Nonthermal Plasma Medicine

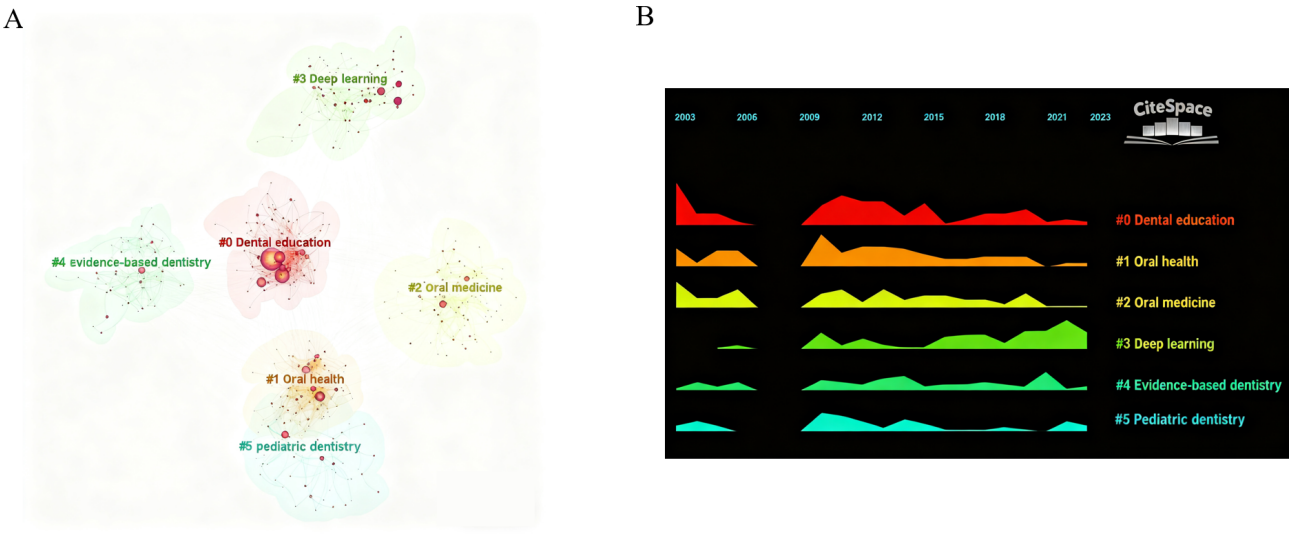
Research Focus Analysis

The word frequency and centrality of keywords reflect the hot topics that researchers have focused on and studied over a period. Keywords mark the core meaning of an article, and conducting keyword analysis can help deeply grasp the core content of this field. In the topic and keyword co-occurrence analysis, nodes reflect the frequency of keywords. Centrality is a basic indicator to measure the weight of nodes, reflecting the importance of nodes in the network. The more prominent the co-occurrence frequency and centrality of keywords, the stronger the importance of nodes. The keyword co-occurrence analysis is shown in Figure 5A. Node size reflected the frequency of keywords. In Figure 5A, the nodes with purple outer circles were those with higher centrality, among which “disease” (0.10), “access” (0.07), “performance” (0.06), “diagnosis” (0.06), and “model” (0.06) possessed higher centrality and frequency, indicating these key contents that have been paid attention to in the promotion and development process of the field (Table 5). Meanwhile, through further sorting and summarizing high-frequency keywords, it was found that the research hotspots in the field of nonthermal plasma medicine from 2003 to 2023 were mainly concentrated in 4 categories. The keyword with the highest level of attention in cluster 0 is “dental education,” the keyword with the highest level of attention in cluster 1 is “oral health,” the keyword with the highest level of attention in cluster 2 is “prevalence,” the keyword with the highest level of attention in cluster 3 is “artificial intelligence,” the keyword with the highest level of attention in cluster 4 is “knowledge,” and the keyword with the highest level of attention in cluster 5 is “impact” and (Figure 5A). Through a timeline analysis of keywords, a chronological map of dental education research was generated. Keywords belonging to the same cluster are aligned horizontally, with their corresponding periods displayed at the top. In this timeline, the density of keywords indicates the significance of the respective clustering domain. Over the past 3 years, “deep learning” has emerged as a prominent research focus in the field of dental education (Figure 5B).

Table . Top 10 highest centrality keywords.

Keywords	Centrality	Counts
Disease	0.10	74
Access	0.07	58
Performance	0.06	90
Diagnosis	0.06	104
Model	0.06	75
Management	0.05	186
Dentists	0.05	126
Quality	0.05	83
Teeth	0.04	67
Adolescents	0.05	59

Figure 5. Visualization of keywords analysis. (A) Keywords co-occurrence and centrality. (B) Timeline of keywords co-occurrence.

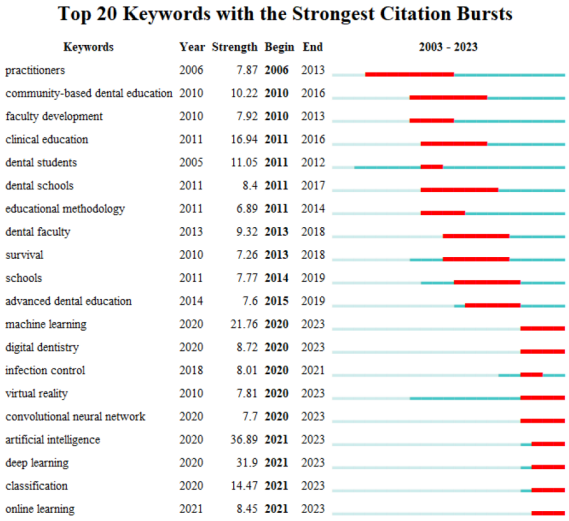


Research Frontiers Analysis

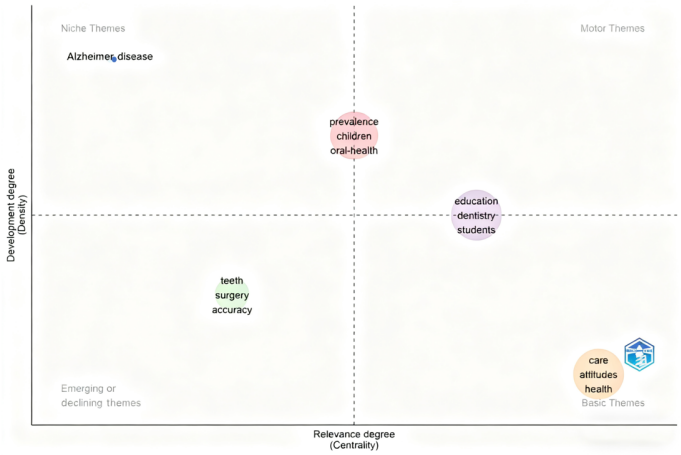
Burst detection identifies keywords or authors that experienced sudden surges in citations over a short period. Keywords with the strongest citation bursts mean that the keywords appear more frequently or are cited more frequently in a relatively short period. The detection indicators of keyword citation bursts generally include intensity and age distribution, which reflect the forefront and development trend of research over a period. Figure 6A lists the top 20 keywords with the strongest citation bursts. “Practitioners” appeared the earliest and lasted the longest (2006 - 2013), with a strength of 7.87. During this period, “faculty development” and “dental students” are also keywords with high attention. Subsequently, “community-based dental education,” “clinical education,” “dental school,” “educational methodology,” and “advanced dental education” began to become a research hotspot for many scholars after 2011 and lasted until 2019. During this period, “clinical education” (2011 - 2016, 16.94), “dental students” (2011 - 2012, 11.05), and "community-based dental education" (2010 - 2016, 10.22) all received more attention. In the past 4 years, “artificial intelligence” showed the highest burst strength (36.89) from 2020 to 2023, signaling its rapid rise as a cutting-edge topic in teaching and learning of stomatology education. In addition, “deep learning,” “machine learning,” “classification,” “digital dentistry,” “online learning,” “infection control,” “virtual reality,” and “convolutional neural network” have also received much attention from scholars in the past 4 years.

Figure 6. (A) Keywords with the strongest citation burst. (B) Thematic map of keywords in teaching and learning of stomatology education.

A



B



This article uses the Bibliometrix package in the R language environment to describe the history of topic evolution based on keywords (Thematic Map). The results show that “care,” “attitudes,” and “knowledge” were in the “basic themes” area, indicating that these are basic concepts for research in this field. In Figure 6B, the analysis revealed that “teeth,” “surgery,” and “accuracy” were classified as “emerging themes,” indicating that researchers are relatively active in these areas, which may have good prospects in the field of teaching and learning in stomatology education.

Co-Cited References Analysis

The historiographic map, proposed by E. Garfield in 2004, graphically represents a chronological network of key direct citations from a bibliographic collection. It illustrates the diachronic evolution of highly cited documents, highlighting shifts in research hotspots. Bibliometrix’s historical citation analysis offers 2 metrics: the local citation score (LCS) for citations within the current database and the global citation score (GCS) for total citations in the Web of Science database. LCS counts how often a paper is cited by other studies in the same field, whereas GCS tracks total citations across all disciplines. These metrics assess the impact of documents and reveal connections across research fields. Use the following command to generate the historical citation network: [options (width=130)] [histResults<- histNetwork (M, min. citations=2, sep=“;")] [net<- histPlot (histResults, n=10, size=5, labelsize=3)].

The analysis results of the historical direct citation network indicated that there were 11 landmark documents in this field, and Table 6 and Figure 7 summarize the relevant information of important documents in the historical cited network atlas. According to the statistics of the 4 articles labeled 4, 5, 8, and 9, the LCS and GCS indices are relatively high, indicating that these 4 articles have a significant impact in the corresponding years, and the research content is highly correlated with the relevant research about teaching and learning in stomatology education.

Table . Eleven landmark documents in teaching and learning of stomatology education.

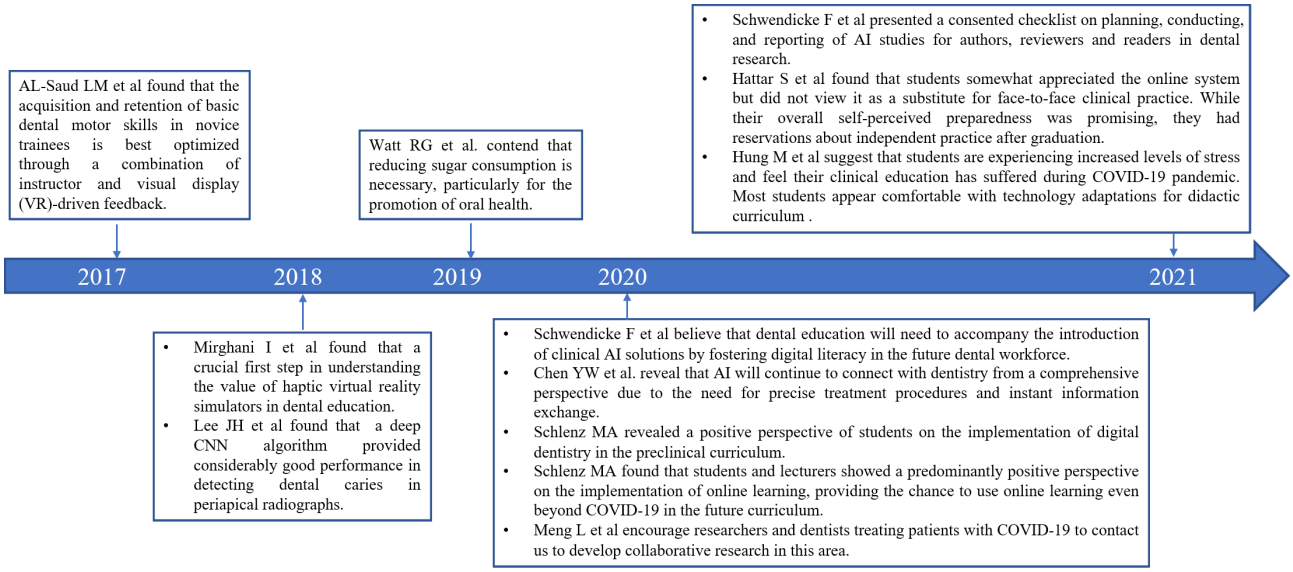
Code	Title	Author	Journal	Country	Year	LCS[a]	GCS[b]
1	Feedback and motor skill acquisition using a haptic dental simulator	Al-Saud et al [3]	Eur J Dent Educ	Denmark	2017	24	67
2	Capturing differences in dental training using a virtual reality simulator	Mirghani et al [4]	Eur J Dent Educ	Denmark	2018	19	52
3	Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm	Lee et al [5]	J Dent	The United Kingdom	2018	51	367
4	Ending the neglect of global oral health: time for radical action	Watt et al [6]	Lancet	The United Kingdom	2019	20	427
5	Artificial intelligence in dentistry: chances and challenges	Schwendicke et al [7]	J Dent Res	USA	2020	40	285
6	Artificial intelligence in dentistry: current applications and future perspectives	Chen et al [8]	Quintessence Int	USA	2020	23	99
7	Artificial intelligence in dental research: checklist for authors, reviewers, readers	Schwendicke et al [9]	J Dent	The United Kingdom	2021	16	136
8	Students' and lecturers on the implementation of online learning in dental education due to SARS-COV-2 (COVID-19): a cross-sectional study	Schlenz et al [10]	BMC Med Educ	The United Kingdom	2020	39	128
9	Coronavirus disease 2019 (COVID-19): emerging and future challenges for dental and oral medicine	Meng et al [11]	J Dent Res	USA	2020	67	914

Code	Title	Author	Journal	Country	Year	LCS[a]	GCS[b]
10	Impact of COVID-19 pandemic on dental education: online experience and practice expectations among dental students at the University of Jordan	Hattar et al [12]	BMC Med Educ	The United Kingdom	2021	25	68
11	In an era of uncertainty: impact of COVID-19 on dental education	Hung et al [13]	J Dent Educ	Denmark	2021	35	101

[a]LCS: local citation score.

[b]GCS: global citation score.

Figure 7. Timeline of part of landmark achievements in teaching and learning of stomatology education. AI: artificial intelligence [5-15].



Discussion

General Information

This study uses bibliometric analysis to examine the key areas of knowledge and emerging trends in stomatology education. Notably, several breakthrough achievements have been identified through these bibliometric methods. The findings reveal a general upward trend in annual publications related to teaching and learning in stomatology education. Furthermore, the geographic distribution of these publications was analyzed based on the country and institutional affiliations of the authors. The results indicate that institutions from the United States, the United Kingdom, China, and Australia are dominant in this field. Specifically, the University of Iowa, the University of Michigan, and the University of North Carolina in the United States; King's College London in Europe; the University of Sydney in Australia; and King Saud University and the University of Hong Kong in Asia have played pivotal roles in the development of stomatology education. In terms of productivity and influence, Wanchek T from the United States has been at the forefront over the past two decades. His focus on how education debt affects career choices among dental students and professionals has earned him numerous citations [14,15]. The primary objective of this study is not to provide exhaustive precision data on teaching and learning in stomatology education but to stimulate interest and encourage further research in this domain. This article also serves as a valuable resource for researchers new to the field, offering a concise overview of current trends and developments in stomatology education.

The observed increase in publications within this field can likely be attributed to the emergence of specialized journals in stomatology education and the growing interest among researchers in the professional and academic evolution of teaching methodologies. Among the top 10 journals, the Journal of Dental Education and the European Journal of Dental Education stand out, having published the highest number of articles and achieving the highest H-index. These journals have become highly esteemed among researchers worldwide, as they highlight the latest advancements and significant discoveries

in stomatology education. In recent years, these journals have particularly focused on the application of AI and its potential impact on stomatology, as well as the role of online learning during the COVID-19 pandemic. These topics align with the findings from keyword bursts, where terms such as "artificial intelligence," "deep learning," "machine learning," "classification," "digital dentistry," "online learning," and "infection control" have emerged as active themes in the field. It is worth mentioning that "artificial intelligence" showed the highest burst strength (36.89) from 2020 to 2023, signaling its explosive growth as a game-changing tool in teaching and learning dental educational skills.

Current Research Focus

In the current research related to stomatology education, the relatively advanced landmark literature was the article published by Al-Saud et al [3] in 2017, which investigated the effect of qualitatively different types of pedagogical FB on the training, transfer, and retention of basic manual dexterity dental skills using a virtual reality (VR) haptic dental simulator. It is reported that the acquisition and retention of basic dental motor skills in novice trainees is best optimized through a combination of instructor and VR-driven FB. Subsequently, Mirghani et al [4] further emphasized the value of haptic VR simulators in dental education. From here on, virtual teaching was paid attention to dental education, and its FB in teaching and learning was investigated by dental educators.

In the realm of teaching and learning in stomatology education, virtual and AI-based instructional methods serve as complementary approaches. Their integration has the potential to substantially enhance educational outcomes. In 2018, Lee et al [5] demonstrated the potential of a deep learning–based convolutional neural network algorithm for the detection and diagnosis of dental caries. This study marked a significant advancement in the application of deep learning technologies within the domain of dental education, garnering substantial attention. Following this, Watt et al [6] highlighted the potential benefits of integrating deep learning and other intelligent systems into dental education, such as personalized learning experiences and predictive analytics. Nonetheless, they cautioned that the adoption of these technologies must be accompanied by ethical considerations to ensure that they augment rather than replace human judgment in educational contexts. Consequently, deep learning, as a facet of AI, is progressively transforming traditional oral medicine education through enhanced image analysis, virtual simulations, personalized learning, and intelligent systems. The incorporation of AI-driven coaching and doctor–patient communication training has notably improved both the quality and efficiency of dental education. In 2020, Schwendicke et al [7] and Chen et al [8] further explored the integration of AI within dentistry. They asserted that future dental education needs to advance alongside the introduction of clinical AI solutions by cultivating digital literacy among emerging dental professionals. They also emphasized that AI will increasingly interface with dentistry from a holistic perspective, driven by the demand for precise treatment procedures and immediate information exchange.

The COVID-19 pandemic has profoundly impacted dental education by accelerating the adoption of digital and remote learning modalities. With physical distancing measures and health concerns limiting in-person interactions, institutions have increasingly turned to virtual platforms and online resources to ensure continuity of education. This shift has highlighted the necessity for adaptable and resilient teaching methods in oral medicine, fostering the development of new pedagogical strategies and technologies. This has also stimulated a large number of studies, including landmark research by Schlenz et al [9,10], who found that students and lecturers have positive attitudes toward the implementation of online learning. Meng et al [11] also encouraged researchers and dentists treating COVID-19 patients to connect with each other online to carry out collaborative research in this field. The pandemic has underscored the significance of integrating digital tools into the curriculum, which not only addresses immediate challenges but also offers long-term benefits in creating more flexible and accessible educational environments. However, experts believe that there are also some potential problems with online teaching or distance learning. Hung et al [13] suggested that students are experiencing increased levels of stress and feel their clinical education has suffered during the COVID-19 pandemic. Hattar et al [12] found that students somewhat appreciated the online system but did not view it as a substitute for face-to-face clinical practice. While their overall self-perceived preparedness was promising, they had reservations about independent practice after graduation. Lee et al [16] found that in the realm of continuing education in oral and maxillofacial radiology for dentists, the online interactive education emerges as an effective tool, fostering positive academic achievements through active learning.

Future Frontiers

The future of teaching and learning in stomatology education will be shaped by advancements in AI, digital simulations, and machine learning strategies. Our bibliometric analysis highlights key emerging trends, including deep learning, convolutional neural networks, VR, and surgical accuracy, which are driving innovation in dental education. These developments indicate a growing emphasis on improving diagnostic accuracy, enhancing clinical training, and integrating AI-driven technologies into educational curricula.

One notable trend is the increasing application of VR and AI-assisted simulations in surgical education. Virtual surgical planning and real-time FB systems allow students to rehearse complex procedures in a risk-free environment, thereby improving both skill acquisition and patient outcomes [17]. Similarly, robotic surgery, which offers enhanced precision and control, is gaining traction in advanced dental training [18]. These findings align with our bibliometric results, which identified "surgery" and "accuracy" as emerging themes. To translate these insights into practice, educators should prioritize competency-based training models that integrate VR-based surgical practice and AI-assisted procedural assessments [19]. For example, Huang et al [20] demonstrated that repeated VR training significantly enhanced students' implantology skills, with positive FB from participants highlighting the technology's effectiveness. Al-Saud et al [3] demonstrated that haptic dental

simulators significantly enhance motor skill acquisition, whereas Mirghani et al highlighted their role in replicating real-world dental procedures, improving student performance and confidence [4]. A study on an inferior alveolar nerve block simulator also found that students who practiced with the simulator were more confident during their first clinical injections, required fewer syringe adjustments, and achieved greater success in numbing patients [21]. These findings highlight the effectiveness of technology-enhanced learning, supporting our bibliometric analysis that identified "VR" and "AI" as emerging trends in stomatology education. Moving forward, further research should explore the long-term impact of these technologies on clinical proficiency and patient care outcomes.

Another key area of transformation is AI-driven diagnostic training. The use of machine learning models to analyze dental imaging data has significantly improved diagnostic precision in fields such as pediatric dentistry and prosthodontics [22]. AI-powered convolutional neural networks have demonstrated high accuracy in detecting caries and occlusal abnormalities [23]. Moreover, the use of AI in dental education is advancing, with studies such as Chen et al discussing its applications in clinical decision-making and personalized training programs [8]. Given this trend, educators should integrate AI-based analysis into preclinical coursework, allowing students to develop advanced diagnostic skills before entering clinical practice.

While AI-driven learning systems enhance efficiency and accessibility, their implementation also raises ethical considerations. Existing research on AI in dentistry exhibits notable deficiencies, impeding its reproducibility and practical implementation. AI should complement, rather than replace, traditional instructor–student interactions [9]. Future research should focus on ethical frameworks for AI integration in education, addressing issues such as student autonomy, decision-making processes, and the potential bias in AI-generated FB. In addition, the interdisciplinary nature of AI-driven education requires stronger collaboration between dental educators, AI developers, and clinical practitioners. Research efforts should focus on optimizing AI algorithms for educational use, ensuring that these technologies align with the pedagogical needs of dental students. By fostering such collaborations, the field can move toward more effective and human-centered educational frameworks.

By aligning bibliometric insights with actionable strategies, educators and researchers can better navigate the evolving landscape of stomatology education. Bibliometric analysis provides a data-driven foundation for curriculum innovation and policy development by identifying emerging trends, evaluating influential studies, and uncovering knowledge gaps. For educators, it helps optimize curriculum design by integrating AI-driven diagnostics, competency-based learning, and virtual simulation training. For researchers, it serves as a strategic tool for identifying high-impact studies, fostering interdisciplinary collaborations, and addressing underexplored areas such as AI ethics and hybrid education models. By leveraging bibliometric methods, stakeholders can make informed decisions that enhance both teaching quality and learning outcomes, ensuring that technological advancements are effectively integrated into stomatology education.

Limitations and Strengths

This study possesses several notable strengths. Primarily, it represents the first bibliometric analysis of publications on teaching and learning in stomatology education using both the Bibliometrix package and CiteSpace. This analysis offers clinicians and scholars a comprehensive overview of the current landscape in stomatology education. In addition, the simultaneous use of 2 bibliometric tools, CiteSpace and Bibliometrix, enhances the robustness of our findings. CiteSpace, in particular, is a widely recognized tool that provides valuable insights into evolving research priorities and trends.

However, this study also has certain limitations. First, the search strategy was specifically focused on teaching and learning in stomatology education, which, despite efforts to expand the search terms, may have resulted in the exclusion of some relevant studies. Second, our study primarily uses bibliometric methods, which focus on citation patterns, co-occurrence networks, and research trends but do not directly assess the pedagogical effectiveness of emerging technologies. While our findings highlight the increasing focus on machine learning, VR, and AI-driven education, further empirical research is needed to evaluate how these innovations impact student learning outcomes, clinical decision-making, and practical skill acquisition. Conducting comparative studies between traditional and technology-enhanced teaching methods could provide deeper insights into their educational value. In addition, our analysis relies on data from WoSCC, which may exclude relevant publications indexed in other databases such as Scopus or PubMed. This could result in a partial representation of the research landscape. Future studies could incorporate multiple databases to ensure a more comprehensive assessment of publication trends. Finally, emerging research areas related to teaching and learning in stomatology education may have been inadvertently overlooked due to the constraints of the used algorithms.

Conclusions

In conclusion, the exploration of innovative teaching and learning methodologies holds significant potential for advancing stomatology education. For example, the integration of deep learning models can enhance diagnostic precision and clinical practice, whereas intelligent teaching systems can facilitate more efficient acquisition of oral medicine knowledge and improve educational outcomes. These technological advancements present new challenges for educators in the field. It is imperative that educators remain informed about the latest research developments and evolving trends in stomatology education. This awareness will not only aid in the dissemination of contemporary concepts but also enable the timely identification of emerging issues, thereby fostering the ongoing development and advancement of stomatology education.

Acknowledgments

This work was supported by the Hunan Provincial Natural Science Foundation of China (2023JJ40817) and the Scientific Research Launch Project for new employees of the Second Xiangya Hospital of Central South University.

Data Availability

All data generated or analyzed during this study are included in this published article.

Conflicts of Interest

None declared.

References

1. Agarwal A, Durairajanayagam D, Tatagari S, et al. Bibliometrics: tracking research impact by selecting the appropriate metrics. Asian J Androl 2016;18(2):296-309. [doi: 10.4103/1008-682X.171582] [Medline: 26806079]
2. Keathley-Herring H, Van Aken E, Gonzalez-Aleu F, Deschamps F, Letens G, Orlandini PC. Assessing the maturity of a research area: bibliometric review and proposed framework. SCIENTOMETRICS 2016 Nov;109(2):927-951. [doi: 10.1007/s11192-016-2096-x]
3. Al-Saud LM, Mushtaq F, Allsop MJ, et al. Feedback and motor skill acquisition using a haptic dental simulator. Eur J Dent Educ 2017 Nov;21(4):240-247. [doi: 10.1111/eje.12214] [Medline: 27324833]
4. Mirghani I, Mushtaq F, Allsop MJ, et al. Capturing differences in dental training using a virtual reality simulator. Eur J Dent Educ 2018 Feb;22(1):67-71. [doi: 10.1111/eje.12245] [Medline: 27864856]
5. Lee JH, Kim DH, Jeong SN, Choi SH. Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm. J Dent 2018 Oct;77:106-111. [doi: 10.1016/j.jdent.2018.07.015] [Medline: 30056118]
6. Watt RG, Daly B, Allison P, et al. Ending the neglect of global oral health: time for radical action. Lancet 2019 Jul 20;394(10194):261-272. [doi: 10.1016/S0140-6736(19)31133-X] [Medline: 31327370]
7. Schwendicke F, Samek W, Krois J. Artificial intelligence in dentistry: chances and challenges. J Dent Res 2020 Jul;99(7):769-774. [doi: 10.1177/0022034520915714] [Medline: 32315260]
8. Chen YW, Stanley K, Att W. Artificial intelligence in dentistry: current applications and future perspectives. Quintessence Int 2020;51(3):248-257. [doi: 10.3290/j.qi.a43952] [Medline: 32020135]
9. Schwendicke F, Singh T, Lee JH, et al. Artificial intelligence in dental research: checklist for authors, reviewers, readers. J Dent (Shiraz) 2021 Apr;107:103610. [doi: 10.1016/j.jdent.2021.103610]
10. Schlenz MA, Schmidt A, Wöstmann B, Krämer N, Schulz-Weidner N. Students' and lecturers' perspective on the implementation of online learning in dental education due to SARS-CoV-2 (COVID-19): a cross-sectional study. BMC Med Educ 2020 Oct 9;20(1):354. [doi: 10.1186/s12909-020-02266-3] [Medline: 33036592]
11. Meng L, Hua F, Bian Z. Coronavirus disease 2019 (COVID-19): emerging and future challenges for dental and oral medicine. J Dent Res 2020 May;99(5):481-487. [doi: 10.1177/0022034520914246] [Medline: 32162995]
12. Hattar S, AlHadidi A, Sawair FA, Alraheam IA, El-Ma'aita A, Wahab FK. Impact of COVID-19 pandemic on dental education: online experience and practice expectations among dental students at the University of Jordan. BMC Med Educ 2021 Mar 8;21(1):151. [doi: 10.1186/s12909-021-02584-0] [Medline: 33685451]
13. Hung M, Licari FW, Hon ES, et al. In an era of uncertainty: impact of COVID-19 on dental education. J Dent Educ 2021 Feb;85(2):148-156. [doi: 10.1002/jdd.12404] [Medline: 32920890]
14. Wanchek T, Nicholson S, Vujicic M, Menezes A, Ziebert A. Educational debt and intended employment choice among dental school seniors. J Am Dent Assoc 2014 May;145(5):428-434. [doi: 10.14219/jada.2014.12] [Medline: 24789235]
15. Nicholson S, Vujicic M, Wanchek T, Ziebert A, Menezes A. The effect of education debt on dentists' career decisions. J Am Dent Assoc 2015 Nov;146(11):800-807. [doi: 10.1016/j.adaj.2015.05.015] [Medline: 26514885]
16. Lee N, Huh J, Jeong H, Park W. Effectiveness of Online Interactive Education in Dental Radiology. Int Dent J 2024 Oct;74(5):1024-1032. [doi: 10.1016/j.identj.2024.03.016] [Medline: 38644105]
17. Deng H, Bian H, Li C, Li Y. Autonomous dental robotic surgery for zygomatic implants: a two-stage technique. J Prosthet Dent 2025 May;133(5):1132-1138. [doi: 10.1016/j.prosdent.2023.05.033] [Medline: 37567843]
18. Kalinov T, Georgiev T, Bliznakova K, Zlatarov A, Kolev N. Assessment of students' satisfaction with virtual robotic surgery training. Heliyon 2023 Jan;9(1):e12839. [doi: 10.1016/j.heliyon.2023.e12839] [Medline: 36699266]
19. Revilla-León M, Gómez-Polo M, Vyas S, et al. Artificial intelligence models for tooth-supported fixed and removable prosthodontics: a systematic review. J Prosthet Dent 2023 Feb;129(2):276-292. [doi: 10.1016/j.prosdent.2021.06.001]
20. Huang Y, Hu Y, Chan U, et al. Student perceptions toward virtual reality training in dental implant education. PeerJ 2023;11:e14857. [doi: 10.7717/peerj.14857]
21. Bevizova K, Falougy HE, Thurzo A, Harsanyi S. Is virtual reality enhancing dental anatomy education? A systematic review and meta-analysis. BMC Med Educ 2024 Nov 29;24(1):1395. [doi: 10.1186/s12909-024-06233-0] [Medline: 39614238]

22. Vishwanathaiah S, Fageeh HN, Khanagar SB, Maganur PC. Artificial intelligence its uses and application in pediatric dentistry: a review. Biomedicines 2023 Mar 5;11(3):788. [doi: 10.3390/biomedicines11030788] [Medline: 36979767]
23. Kaya E, Güneç HG, Ürkmez E, Aydın KC, Fehmi H. Deep learning for diagnostic charting on pediatric panoramic radiographs. Int J Comput Dent 2024 Oct 15;27(3):225-233. [doi: 10.3290/j.ijcd.b4200863] [Medline: 37417445]

Abbreviations

AI: artificial intelligence
FB: feedback
GCS: global citation score
LCS: local citation score
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
VR: virtual reality
WoSCC: Web of Science Core Collection

Edited by B Lesselroth; submitted 10.09.24; peer-reviewed by E Feofanova, Z Ehtesham; revised version received 07.04.25; accepted 23.09.25; published 20.10.25.

Please cite as:
Zou Z, Guo L
Insights Into History and Trends of Teaching and Learning in Stomatology Education: Bibliometric Analysis
JMIR Med Educ 2025;11:e66322
URL: https://mededu.jmir.org/2025/1/e66322
doi:10.2196/66322

Digital Literacy Training for Digitalization Officers ("Digi-Managers") in Outpatient Medical and Psychotherapeutic Care: Conceptualization and Longitudinal Evaluation of a Certificate Course

Anne Mainz[1], MSc; Timo Neunaber[1], MA; Paula Cara D'Agnese[2], MA; Alexander Eid[2], BA; Tanja Galla[2], BSc; Christoph Ellers[3], BA; Sven Meister[1,4], Prof Dr

[1]Health Informatics, Faculty of Health, Witten/Herdecke University, Pferdebachstr. 11, Witten, Germany

[2]IT & Digital Health Division, Kassenärztliche Vereinigung Westfalen-Lippe, Dortmund, Germany

[3]Department of Education, Ärztekammer Westfalen-Lippe, Münster, Germany

[4]Department of Healthcare, Fraunhofer Institute for Software and Systems Engineering, Dortmund, Germany

Corresponding Author:
Anne Mainz, MSc
Health Informatics, Faculty of Health, Witten/Herdecke University, Pferdebachstr. 11, Witten, Germany

Abstract

Background: Digital tools, services, and information in patient care demand new competencies in outpatient care, and the workforce is faced with the need to deal with digitalization.

Objective: In a targeted certificate course (Certification of Digitalization Officers in Medical Practices and Psychotherapeutic Practices, Digi-Manager), medical assistants are trained to serve as digitalization officers, enabling them to implement the requirements of digitalized health care within their practices.

Methods: As part of an accompanying study, the course is evaluated by the participants, and the change in their digital literacy is recorded. We measured different knowledge, skills, and attitude dimensions at 3 different times—before, during, and after the course—and used ANOVA to examine significant changes.

Results: Digi-Managers started the course with an already high self-assessment of their digital literacy. Skills and knowledge increased significantly in all categories (cognitive, technical, ethical, and health information) from the initial to the final measurement, as did self-confidence in the use of general software and hardware. Positive attitude remained stable over the training period, and the course was rated very positively by participants across all areas.

Conclusions: Training programs on digital topics for health care professionals are necessary, and this certification course is a role model for successful further education through a mixture of theoretical knowledge transfer and practical application. Especially, the use of a digital maturity model and a digital laboratory was a unique and useful feature. Further research needs to go into alternative assessment methods of digital literacy, as the results suggest that self-assessment measures self-efficacy and confidence, rather than pure competence. Nevertheless, the increase in self-assessed competence suggests that the training was successful.

(JMIR Med Educ 2025;11:e70843) doi:10.2196/70843

KEYWORDS

digital competence; digital literacy; digital maturity; certificate course; training course; longitudinal study; medical assistant; health care professional

Introduction

Background

The use of digital innovations with different tools, services, and information in patient care is associated with new requirements for day-to-day work in outpatient medical care, and the spectrum of them is expanding more and more. Not only physicians but also other medical professionals are faced with the need to deal with digitalization [1]. Telemedicine, eHealth, and mobile health (mHealth) enable access to health information, improve communication, allow personalized care, make remote monitoring possible, and support self-management of patients [2]. Administrative processes are also increasingly supported digitally [1].

The digital transformation in Germany's health care system has led to a few important digital tools in recent years: the electronic patient record (ePA), electronic sick leave certificates (eAU), electronic prescription (eRezept), the electronic medication plan

(eMP), electronic communication systems (KIM and TIM), telehealth applications, and digital health applications (DiGA). Additionally, various legal initiatives to establish more technologies in health care are currently coming into force in Germany [3]. On the side of the outpatient medical practices in Germany, the level of digitization or digital maturity depends heavily on the individual personality, motivation, and competence of the people involved [4-6]. You will find mostly outdated IT systems within the practices [3], and many things still happen in paper form, such as the storage of data or communication with other health care providers or patients [4].

Besides the availability of digital tools, services, and information, individuals need more motivation to use these technologies. Following the self-determination theory [7], the feeling of competence is one of the three pillars to develop motivation. Otherwise, feelings of incompetence regarding the use of health informatics technology tools lead to reluctance in using these tools and are one of the main reasons to avoid them. Digital competence or digital literacy in health care could be defined as the ability to integrate and apply context-appropriate knowledge, skills, and psychosocial factors—such as attitudes, beliefs, values, and motivation—to perform within the health care domain [8]. Especially when health care workers were not given enough time to learn on-the-job or did not have enough support from peers, they were not willing to use new technology [9]. To enable individuals to gain digital literacy in health care, education is needed.

In their focus group study, Mannevaara et al [10] identified knowledge and skill issues regarding IT-related management and IT background knowledge as the main challenges faced in health care. Competencies related to direct patient care, communication, ethics in health IT, project- and change management, digital literacy, information and knowledge management, teaching, and education were essential in today's health care practice. Especially, competencies in decision-making, information and knowledge management, teaching, training, and data security were highlighted as important by German participants in this study [10]. Hübner et al [11] identified the top 3 core competency areas that need to be addressed for employees in the health care sector that work in direct patient care: (1) communication, (2) documentation, and (3) information and knowledge management in patient care, which coincides with the focus of the Digi-Manager training course.

Goal of This Study

With this in mind, the project "Certification of Digitalization Officers in Medical Practices and Psychotherapeutic Practices (Digi-Manager)," funded by the Federal Ministry of Health in Germany, launched an educational program for medical assistants to train them as digitalization officers. In Germany, medical assistants make appointments for patients, document treatment procedures for patient files, take care of billing for services rendered, and organize practice procedures. They apply bandages, prepare syringes, or draw blood for laboratory tests. They also inform patients about pre- and posttreatment options, maintain medical instruments, and carry out laboratory work [12].

Digitalization officers, or how they are called in this certification program, "Digi-Managers," create digitalization strategies for their own practices, can get digitalization projects off the ground, and act as a point of contact for the digitalization of patient care.

To evaluate if the digital literacy of the participants could be increased through participation in the Digi-Manager course, their digital literacy before (t0), during (t1), and after (t2) the course was measured and compared. To measure the whole concept of digital literacy, participants are asked about their knowledge, skills, and attitudes regarding digital health technologies.

Hypotheses

The transfer of skills and abilities should be verifiable in a successful training program, which is why hypothesis H1 is formulated. A scoping review [13] suggests that discussions, group workshops, self-directed learning materials, and providing practice opportunities, among other things, help to grow digital confidence, which are provided within the Digi-Manager training. This leads to our hypothesis H2. Other studies show that increased usage causes a rise in computer confidence, which also increases positive attitudes toward computers [14], which supports the assumption of hypothesis H3.

- H1: Attending the Digi-Manager training course significantly improves the specific knowledge and skills imparted to participants.
- H2: Attending the Digi-Manager training course significantly improves participants' general confidence level for technology usage.
- H3: Attending the Digi-Manager training course significantly improves participants' attitude toward digital (health) applications.

Methods

Ethical Considerations

The ethics application was reviewed and approved by the ethics committee of Witten/Herdecke University on April 20, 2023, and no ethical or legal concerns were raised (application/approval no. S-93/2023). The participants received no further compensation for participating. Before each survey, the participants received information about the duration, procedure, and content of the study and had to provide consent. For each question, there was the option to refuse to answer. The responses of the participants were pseudonymized using an identification code, and no identifying information was queried.

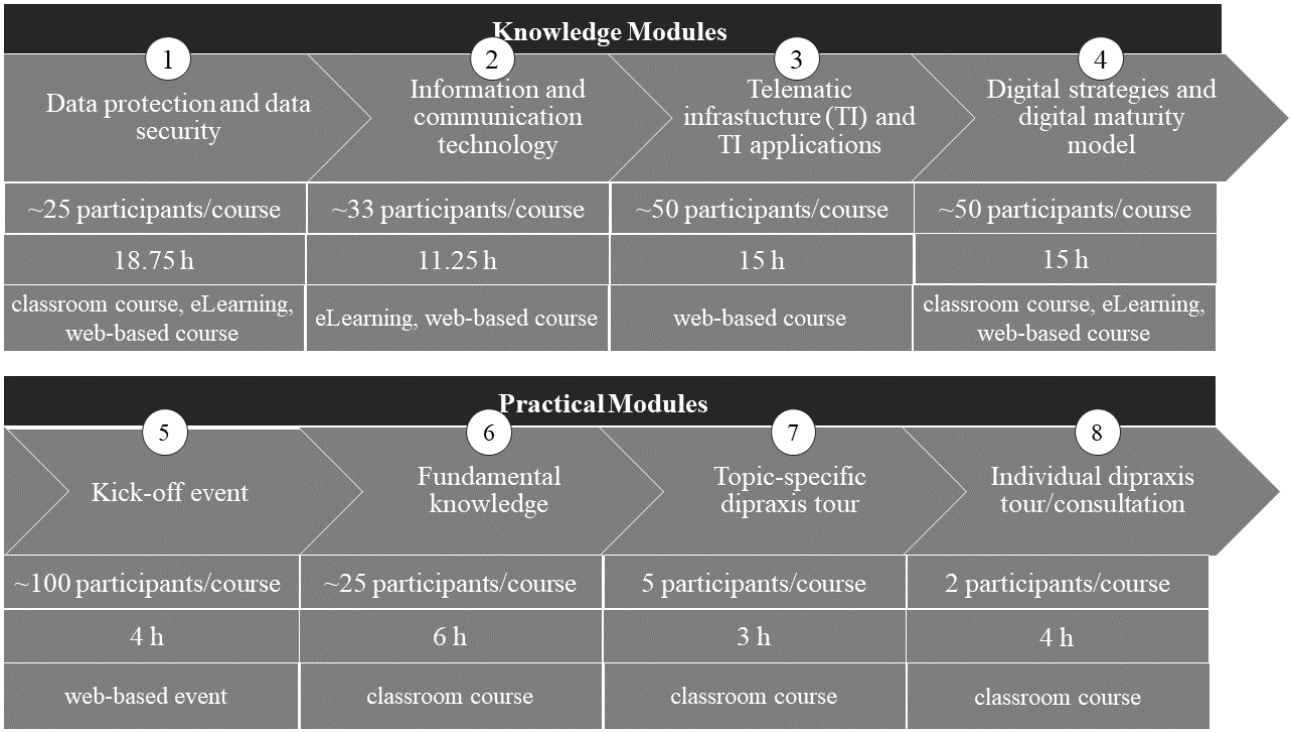
Course Concept

The aim of the Digi-Manager certificate course was to enable medical assistants in outpatient medical practices to derive and implement digitalization strategies and projects for their own practices and become contact persons for the digitalization of patient care. Therefore, they had to acquire skills in the areas of technology, data use, IT security, and data protection.

The participants were released from work for a total of 205 hours during the certificate course and received no further incentives for participation. The participants spend 60 hours on the knowledge modules and 17 hours on the practical modules

in web-based and on-site classroom courses. Approximately 128 hours were planned additionally for orientation, organization, self-study, exams, and creating a digitalization strategy (Figure 1). Different time slots (2 - 4 slots per module) were offered for each course, and participants were allowed to choose them freely. Participation in each module was mandatory to pass the certificate course, and attendance was assessed through an attendance list. The content of the course was developed based on an existing certified course of ÄKWL (Ärztekammer Westfalen-Lippe, a medical association representing 40,000 doctors of the Westphalia-Lippe region) that has been running for many years and was adapted and supplemented by a panel of experts with regard to the requirements of digitalization officers (Multimedia Appendix 1).

Figure 1. Schedule of the certificate course.



The course consisted of so-called knowledge and practical modules. The knowledge modules lasted from May until September 2023 and were held as blended learning courses with e-learning materials and web-based and on-site courses. The e-learning materials and web-based classes were distributed by the learning management platform ILIAS.

The practical modules lasted from October 2023 until May 2024 and consisted of a kick-off event, a fundamental knowledge course, and consultations in smaller groups. The courses took place at the dipraxis, a digital laboratory for testing digital tools and analyzing processes in outpatient care. A major component of the practical modules was the digital maturity model of the KVWL (Kassenärztliche Vereinigung Westfalen-Lippe, an association of statutory health insurance physicians of the Westphalia-Lippe region). The model was developed on the findings of Neunaber and Meister [15] for this project. It measures the digital maturity of the practices on 5 assessment categories: corporate management, infrastructure (IT security, data protection, interoperability, data processing, telematic infrastructure, and data collection), treatment and therapy, patient management, and administration. Course participants classified their own practice by answering 25 items. A digital tool visualized the digitalization status of the practice based on the answers using a radar chart. The model could be used by Digi-Managers to assess the current level of digitalization in the practice and identify potential for improvement of the digital situation. With the results of the digital maturity model and the consulting in the dipraxis, the digital managers developed practice-specific digitalization strategies.

Recruitment

The respondents for the survey were the participants of the Digi-Manager training course. Of the total of 100 participants, all were asked to take part in the surveys.

Guideline

The GREET (Guideline for Reporting Evidence-Based Practice Educational Interventions and Teaching) checklist was used to describe the educational intervention in detail [16]. It comprises 17 items to describe why, what, who, how, where, when, how much, and how well an educational intervention took place and what planned and unplanned changes occurred.

Questionnaires

Within the initial survey (Multimedia Appendix 2), participants were asked about their basic demographic information to describe the study population and identify possible biases: their age, gender identity, duration of employment, and the medical field of the doctor's office they work at. Various studies indicate that age and gender are associated with different usage behavior and perception of digital tools [17,18], but other studies show that the influence of these variables is often overestimated

[19,20]. In order to control these possible effects, demographic information was collected and analyzed.

Digital literacy of the participants was measured using 2 existing questionnaires, which recorded both the skills and knowledge relating to digital (health) systems as well as attitudes toward digital (health) systems. Since digital literacy is understood as the competence to deal with digital tools in the everyday professional life of medical assistants in this study, an instrument was needed that operationalizes different dimensions of competence and refers to the wide range of digital tools in medical practices. Fitting tools in terms of content and quality criteria were identified using the results of a preceding scoping review [21]. The first questionnaire used was the Public Health Informatics Competencies for Primary Health Care (PHIC4PHC) questionnaire [22], with a total of 42 items on a 5-point Likert scale. The items are divided into the following dimensions: *cognitive proficiency* (digital health system knowledge and digital health system skills), *technical proficiency* (general computer skills, office application skills, and network skills), *ethical proficiency* (privacy, security, and legal knowledge), and *health information literacy* (health information access, management, integration, and evaluation).

In addition to knowledge and skills, the questionnaire by Kuek and Hakkennes [23] was used to record the attitude toward the use of digital health care systems via items on (1) confidence in usage behavior, (2) technology acceptance, and (3) acceptance and use of technology.

The first part of the questionnaire required participants to indicate their confidence level for different commonly used hardware and software devices on a 5-point scale from not at all confident to completely confident. The devices and applications in question were computers, office applications, smartphones, tablets, email, and social media. The second and third parts of the questionnaire were based on the TAM (technology acceptance model) and UTAUT (unified theory of acceptance and use of technology) questionnaires, which have been used and validated in various studies. Technology acceptance of health information systems was measured using the dimensions *perceived usefulness* and *perceived ease of use* of the TAM, with 12 items, and supplemented with the dimensions *attitude toward technology*, *social influence*, *facilitating conditions*, and *anxiety* of the UTAUT, with 15 items, all on a 7-point Likert scale.

The system usability scale (SUS) was used as an established, standardized, and quick questionnaire for the assessment of perceived usability of the digital maturity level tool (Brooke, 1996, quoted from [24]). The 10 items were measured on a 5-point Likert scale.

Procedure

Digital literacy was assessed in the form of a longitudinal study parallel to the certificate course. The measurement points were based on the participants' progress in the course: before completing the first knowledge module (t0), after completing all 4 knowledge modules (t1), and after completing the practical modules (t2). The links to the questionnaires were distributed via the learning management platform ILIAS, which was used to provide communication and content for the Digi-Manager training. This made it easier for participants to access the surveys, and the integrated reminder and activation functions ensured that participation in the surveys was not forgotten. In addition, the completion of the evaluation surveys could be used as a condition for further progress within the learning material in ILIAS. As no further learning content was planned after the final survey, reminder emails were sent to the participants for this purpose.

Before starting every survey, the participants were informed about the content of the study, data processing, and data protection and were asked for their written consent to participate. To statistically evaluate the change in digital literacy without violating anonymity, participants were asked to provide each survey with their individual identification code. This code was a 6-digit character string made up of time-stable personal characteristics.

The first questionnaire for t0 collected basic demographic information and measured the participants' existing digital skills. The second questionnaire for t1 was intended to assess the digital literacy of the participants again and additionally asked about the evaluation of the first part of the training. At this second measurement point, the Digi-Managers also had the opportunity to evaluate the certificate course in terms of support and design (participant support, technical moderation, quality of scripts, and atmosphere), content (topicality of content, content structure, selection of speakers and authors, discussion/interaction, practical relevance, and personal goal achievement), planning and organization (program announcement, selection of dates, and time frame), and the web conferencing system (technical functionality, user friendliness of the screen, sound quality, and image quality). Participants were able to rate the various aspects of these dimensions on a scale of 1 to 6, with 1 as the best and 6 as the worst rating option. In addition, there were free-text questions in which the participants were asked what they perceived as particularly positive or particularly negative about the training. In the final questionnaire for t2, digital literacy was assessed one last time, the second part of the training was evaluated, and the applicability and comprehensibility of the digital maturity level tool were recorded.

Data Analysis

The datasets of the different surveys were matched through the 6-digit individual identification code. Since some data records did not have fully matching codes, data records were also assigned that were at most 2 characters different or in a reversed order.

The collected data were analyzed with the IBM SPSS Statistics 28 analysis software. To analyze the data descriptively, the frequency, mean values, and standard deviations were reported. Normal distribution as a prerequisite was tested by the Shapiro-Wilk test because it is more powerful than the Kolmogorov-Smirnov test [25], and sphericity was tested with the Mauchly test. The Levene test was used to test the homogeneity of variances.

Correlations were tested through Pearson *r* and rated with the classification from Cohen [26], with $|r|=0.1$ as weak correlation, $|r|=0.3$ as moderate correlation, and $|r|=0.5$ as strong correlation. ANOVA with repeated measures was used to test if the mean values of digital literacy at different measurement points differ significantly. Post hoc tests were used to determine between which measurement times significant differences exist. For evaluation of effect sizes, the classification according to Cohen [26] was chosen with 0.01 as a weak effect, 0.06 as a moderate effect, and 0.14 as a large/strong effect. As η^2 systematically overestimates the effect size, ω^2 and ε^2 are also calculated, as these have lower bias [27]. Between-subject effects such as age, gender, and education level were also tested to gain insights into their additional effect.

The SUS score was calculated as the sum of item scores with the negative worded items inverted and multiplied by 2.5, resulting in a value between 0 and 100 [24].

The free-text answers were formed into inductive categories, and only the top 3 of particularly positive or particularly negative aspects were reported. All other named categories could be seen in Multimedia Appendix 3.

Results

Digi-Manager Characteristics

A total of 100 participants started the Digi-Manager training course. The participants were aged between 20 and 61 years, with an average age of 37.4 (SD 11.3) years, and the vast majority identified as female (95/100, 95%). Only 4 (4%) identified as male, and 1 (1%) did not provide any information on their gender identity. The participants worked at doctors' offices with different primary medical fields, as shown in Table 1.

Table . Participants' demographic data (N=100).

Characteristics	Participants (N=100)
Age (years), mean (SD)	37.4 (11.3)
Gender, n (%)	
Woman	95 (95)
Man	4 (4)
Prefer not to say	1 (1)
Primary medical field of doctor's office, n (%)	
General practice	50 (50)
Gynecology	9 (9)
Dermatology	5 (5)
Orthopedics	5 (5)
Psychiatry or psychotherapeutic practice	4 (4)
Pediatrics	3 (3)
Internal medicine	3 (3)
Neurology	3 (3)
Urology	2 (2)
Ophthalmology	2 (2)
Gastroenterology	2 (2)
Child and adolescent psychiatry	2 (2)
Oral and maxillofacial surgery	2 (2)
Otorhinolaryngology	1 (1)
Pneumology	1 (1)
Nuclear medicine	1 (1)
Radiology	1 (1)
Reproductive medicine	1 (1)
Not specified	3 (3)

For the first survey, the participation rate was 100%. In the second survey, 97 out of the 100 Digi-Managers participated. For the third and last survey, only 64 of the 100 answered the entire questionnaire, despite repeated reminders via email and ILIAS.

Gender as a between-subject effect was not further monitored because of the vast majority (95/100, 95%) of female-identifying participants. The variables that showed a significant correlation with age were all 3 measurement points for confidence in usage behavior with a medium negative effect (r_{t0}=–0.422, P<.001; r_{t1}=–0.423, P<.001; r_{t2}=–0.381, P=.005). The other correlation values can be seen in Multimedia Appendix 4.

Evaluation of the Certificate Course

The course was evaluated separately for the knowledge and the practical modules, but for both, the review was very positive. Most ratings were equal to or above a mean value of 2; only the *selection of dates* for the knowledge modules and the *time frame* for both the knowledge and practical modules were just below this value. All other categories for support and design, content, and web conferencing systems were ranked 1.4 - 2 (Table 2).

Table . Evaluation of the certificate course aspects by the Digi-Managers (scale from 1 (very good) to 6 (very bad)).

Evaluation categories	Knowledge modules, mean (SD)	Practical modules, mean (SD)
Support and design		
Participant support	1.7 (0.9)	1.4 (0.7)
Technical moderation	1.6 (0.7)	—[a]
Quality of scripts	1.9 (0.9)	—
Atmosphere	1.7 (0.8)	1.4 (0.7)
Content		
Topicality of content	1.6 (0.6)	1.4 (0.7)
Content structure	1.8 (0.8)	1.7 (0.8)
Selection of speakers/authors	1.9 (0.8)	1.6 (0.8)
Discussion/interaction	2 (1)	1.8 (0.9)
Practical relevance	2 (1)	1.7 (1)
Personal goal achievement	1.9 (0.9)	1.7 (0.8)
Planning and organization		
Program announcement	1.9 (1.1)	1.5 (0.7)
Selection of dates	2.2 (1.1)	1.7 (0.8)
Time frame	2.1 (1)	2.1 (1.1)
Web conferencing system		
Technical functionality	1.8 (0.7)	—
User friendliness of the screen	1.9 (0.9)	—
Sound quality	1.8 (0.9)	—
Image quality	1.8 (0.8)	—

[a]Not applicable.

The most named positive aspects of the knowledge modules were the content (n=24), with comprehensive scripts, refreshment of knowledge, and new impulses in a comprehensible manner, and the possibility to have access before and after the materials. The second most named was the good support (n=24), which was fast, friendly, and competent, gave individual help, and was very helpful with further questions. The practical relevance and application orientation were named as a positive aspect, the third most (n=14). In turn, most participants said that they could not name any negative aspects (n=25). Some mentioned that it was too much input for the short period of time (n=7), and others said that the modules had too much frontal teaching (n=6), which made the courses dry, difficult to follow, boring, and with too little interaction.

For the practical modules, the exchange with other participants in small groups was particularly positively highlighted by the Digi-Managers (n=26) for offering new perspectives or solutions in conversations. Furthermore, positively perceived was the very informative and instructive nature of the practice modules (n=10) and direct transfer to everyday practice (n=9). When asked about negative aspects, the most common response was that the participants could not name any (n=12). Some participants said that they wished for better day and time selection options (n=4), because the option selections were confusing, could not always be set up, and the always-changing slots were problematic. Some participants wished for more feedback on the tasks they had completed (n=4).

Evaluation of the Digital Maturity Tool

The usability of the digital maturity tool was ranked via SUS score with a mean value of 85 (SD 12.9), which could be classified as an A+ grade after the Sauro-Lewis grading scale [28]. The content of the tool was evaluated very positively in the different aspects with values between 1.39 and 1.79 (Table 3).

Table . Evaluation of the digital maturity tool content aspects by the Digi-Managers (scale from 1 (very good) to 6 (very bad)).

Content evaluation categories	Content rating, mean (SD)
Topicality of content	1.4 (0.6)
Content structure	1.6 (0.7)
Practical relevance	1.6 (0.7)
Personal goal achievement	1.8 (1)

Regarding the digital maturity tool, the participants especially liked the visual representation as a radar chart (n=23) because it was appealing, clear, and colorful; showed all relevant information at a glance; and made the digitalization tangible. A lot of participants positively highlighted the possibility to determine the "status quo" of their practice (n=15) and the potential to uncover gaps and deficits (n=11). Most had no negative aspects (n=16), but some mentioned that the answer options to the items were not always clearly distinguishable, or there were no answer options that fitted their practice perfectly, but it was possible to clarify answers through free-text fields (n=8). Some participants had problems with lots of technical jargon (n=5).

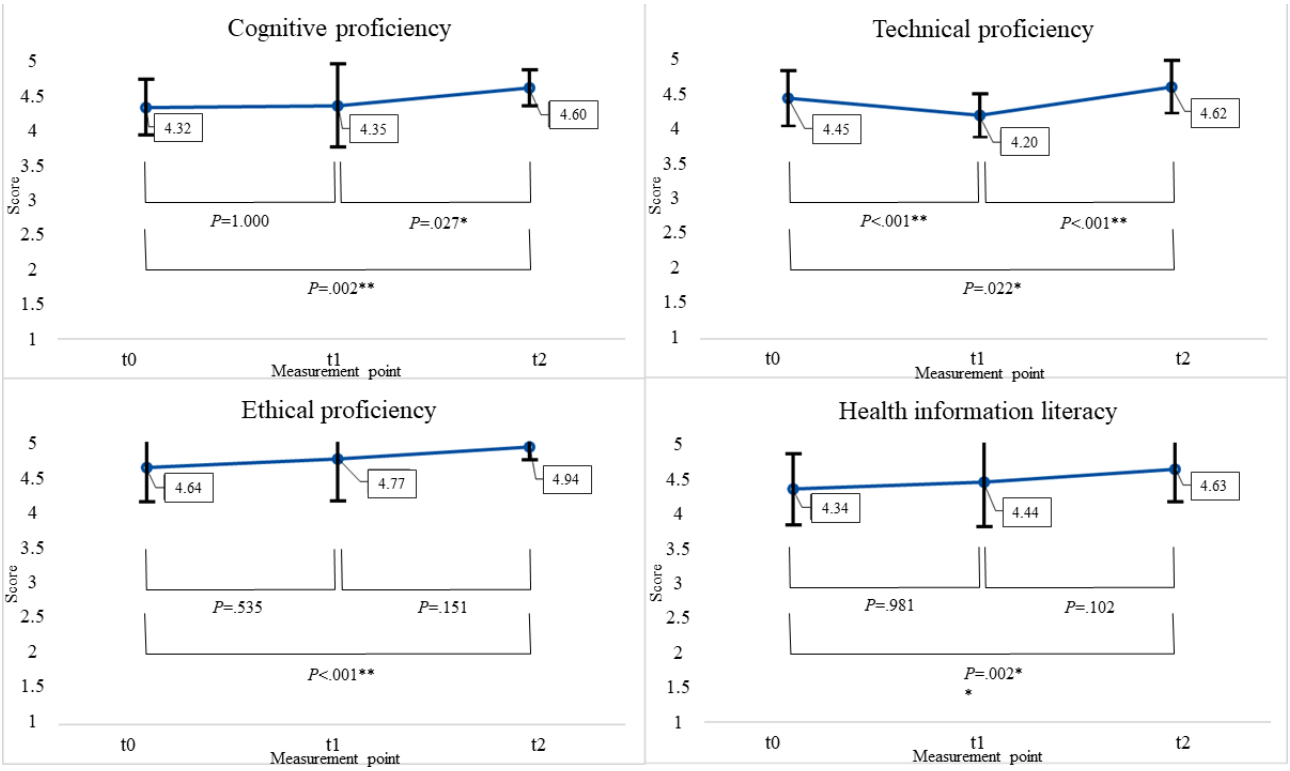
Progression of Digital Literacy

An ANOVA with repeated measures showed that *cognitive proficiency* changed significantly ($F_{1.67,63.55}$=5.5, P<.009; partial η^2=0.13). ω^2=0.1 and ε^2=0.1 could be classified as a medium effect. Because of violations of sphericity (P=.018), the Greenhouse-Geisser adjustment was used. Bonferroni-adjusted post hoc analysis revealed a significant increase (P=.027) in cognitive proficiency from the second to last measurement point ($mean_{Diff}$ –0.25, 95% CI –0.48 to –0.22) and a significant increase (P=.002) between the first and last survey ($mean_{Diff}$ –0.28, 95% CI –0.46 to –0.09) ($mean_{t0}$ 4.3, SD_{t0} 0.4; $mean_{t1}$ 4.4, SD_{t1} 0.6; $mean_{t2}$ 4.6, SD_{t2} 0.3) (Figure 2).

Technical proficiency initially decreased before rising again at the last measurement point ($mean_{t0}$ 4.5, SD_{t0} 0.4; $mean_{t1}$ 4.2, SD_{t1} 0.3; $mean_{t2}$ 4.6, SD_{t2} 0.4). None of the variables were normally distributed. Because of violations of sphericity, the Greenhouse-Geisser adjustment was used to correct the repeated measures ANOVA, and it showed significant differences ($F_{1.83,75.05}$=31.7, P<.001; partial η^2=0.44). ω^2=0.42 and ε^2=0.42 showed a large effect. Bonferroni-adjusted post hoc analysis revealed significant differences between all measurement points: a significant decrease (P<.001) between the first and second survey ($mean_{Diff}$ 0.25, 95% CI 0.14 to 0.36), as well as a significant increase (P<.001) between the second and third survey ($mean_{Diff}$ –0.41, 95% CI –0.54 to –0.28), and an overall significant increase (P=.022) between the first and third survey ($mean_{Diff}$ –0.17, 95% CI –0.31 to –0.02).

Ethical proficiency increased over the entire training period ($mean_{t0}$ 4.6, SD_{t0} 0.5; $mean_{t1}$ 4.8, SD_{t1} 0.6; $mean_{t2}$ 4.9, SD_{t2} 0.2), with a statistically significant difference between measurements, assessed by repeated measures ANOVA with the Greenhouse-Geisser correction (Mauchly test P=.003; $F_{1.68,94.3}$=6.54, P=.004; partial η^2=0.11). ω^2=0.09 and ε^2=0.09 could be classified as a medium effect. Bonferroni-adjusted post hoc analysis showed a significant increase (P<.001) from the first to last measurement point ($mean_{Diff}$ –0.30, 95% CI –0.46 to –0.14).

Figure 2. Means, SDs, and post hoc results for the 4 PHIC4PHC (Public Health Informatics Competencies for Primary Health Care) domains. *P<.05, **P<.01.



The change in *health information literacy* was assessed as significant through repeated measures ANOVA ($F_{2,102}$=5.71, P=.004; partial η^2=0.1). ω^2=0.08 and ε^2=0.08 show a medium effect. Sphericity could be assumed (P=.361). Post hoc analysis with the Bonferroni correction revealed a significant increase (P=.002) from the first ($mean_{t0}$ 4.3, SD_{t0} 0.5) to last ($mean_{t2}$ 4.6, SD_{t2} 0.5) measurement point ($mean_{Diff}$ –0.29, 95% CI –0.49 to –0.10).

The confidence level for technology usage increased from one measuring time to the next ($mean_{t0}$ 3.6, SD_{t0} 0.8; $mean_{t1}$ 3.7, SD_{t1} 0.6; $mean_{t2}$ 4.1, SD_{t2} 0.6). None of the variables were normally distributed, as assessed by the Shapiro-Wilk test, but since the sample size was sufficiently large (n>30) and due to the robustness of the ANOVA with repeated measures, no further actions need to be taken [29]. The Greenhouse-Geisser adjustment was used to correct the violations of sphericity due to a significant result of the Mauchly test (P=.002) [30]. A repeated measures ANOVA with a Greenhouse-Geisser correction determined that the confidence level showed a statistically significant difference between measurements ($F_{1.6,71.72}$=22.59, P<.001; partial η^2=0.33). ω^2=0.32 and ε^2=0.32 could be classified as a large effect.

Bonferroni-adjusted post hoc analysis revealed significantly (P<.001) higher confidence scores in the comparison between the first and last measurement points ($mean_{Diff}$ –0.48, 95% CI –0.32 to 0.03), and significantly (P<.001) higher between the second and last measurement points ($mean_{Diff}$ –0.34, 95% CI –0.48 to –0.20).

The attitude toward (health) technology was assessed by TAM and UTAUT items. There was no statistically significant difference for the different measurement points of the TAM ($F_{1.65,70.91}$=1.89, P=.166) as assessed by repeated measures ANOVA with the Greenhouse-Geisser correction (Mauchly test P=.007). The mean TAM score was already high in the first survey ($mean_{t0}$ 6, SD_{t0} 0.7) and remained at a similar height for the second ($mean_{t1}$ 6.2, SD_{t1} 0.7) and third ($mean_{t2}$ 6.3, SD_{t2} 0.8) surveys.

For the UTAUT scores, normal distribution (P_{t0}=.174; P_{t1}=.200; P_{t2}=.200) and sphericity (P=.080) could be assumed. The repeated measures ANOVA showed no significant differences ($F_{1.65,70.91}$=1.89, P=.166). The UTAUT score remained—similar to the TAM score—high over the entire data collection period ($mean_{t0}$ 6, SD_{t0} 0.3; $mean_{t1}$ 6.1, SD_{t1} 0.7; $mean_{t2}$ 6.2, SD_{t2} 0.4).

Discussion

Principal Results

The study results showed that the training was both successful and satisfactory for the participants of the certificate course “Certification of Digitalization Officers in Medical Practices and Psychotherapeutic Practices (Digi-Manager).” Although the Digi-Managers already started with a high level of self-evaluated digital literacy, this increased significantly from the beginning to the end of the training program. Hypothesis H1 was confirmed: after the training course, the Digi-Manager had significantly higher values in the cognitive proficiency, ethical proficiency, and health information literacy with a medium effect. The significant increase in technical proficiency even had a large effect. Hypothesis H2 was also confirmed, and attending the Digi-Manager training course significantly improved participants’ general confidence level for technology

usage with a large effect. The attitude toward digital (health) applications remained stable at a high positive level over the entire course duration. Therefore, hypothesis H3 was rejected because the attitude toward digital (health) applications did not improve during and after the Digi-Manager training course. We assumed that this was because the practices had to actively apply for the course. Although the participants were finally drawn by lot, it is likely that practices that were interested in digital topics anyway applied more often and sent their most digitally savvy colleague. The very positive perception of the training—both for the knowledge and practical modules—also remained stable throughout the course. Age had no effect on most of the variables. The only variable that was negatively influenced by older age was self-confidence in use. This effect was stable for all 3 measurement times. Similar results of lower self-confidence when dealing with digital topics in older adults were found in other studies [20]. Our survey therefore reflects the state of research that there are hardly any age effects in the use of digital tools. Gender effects could not be examined because of the vast majority of female-identifying participants. However, the low participation rate of men reflects the real gender proportion of men in this occupational field, which was only 2% in the year 2023 in Germany, as reported by the federal employment agency (Bundesagentur für Arbeit).

To our knowledge, this is the first training program for medical assistants for general digital literacy that was scientifically evaluated. Multiple review papers show that training courses for health care staff teach mostly about specific technologies such as electronic medical records or telehealth [31,32], take place for the most part in an academic context [33], or are intended for physicians [34]. Many authors criticize the lack of sufficient training [34-36].

A competence measure study in the more digital-savvy country of Finland [37] showed that further education for health professionals is needed not only in Germany but all over the world. The level of digital competence among health care professionals also varies in other countries. Especially, human-centered remote counseling competence was identified as the category with the weakest score. Health care professionals' knowledge of ethical, legal, and regulatory requirements, as well as privacy and security issues regarding digital tools, was named as a mandatory subject matter in training. In the study of Jarva et al [37], higher age was associated with lower evaluation of digital solutions as part of work and a decrease in self-evaluated competence. This was not further confirmed in our study.

Limitations

One limitation of the results is that, despite repeated reminders both digitally and in person, there was a high level of nonparticipation at the time of the last survey. Only 64% (64/100) completed the last survey. It cannot be ruled out that this might slightly distort the results as it is more likely that the committed participants, who got a lot out of the training anyway, participated until the end than those who perceived the training as less enriching. In future training courses, an attempt will be made to combine the evaluation with the last content-related work in the course in order to increase participation in the final evaluation.

As a further limitation on the part of the participants, it should be considered that, as already briefly mentioned above in the discussion of H3, a self-selection bias could exist through the application process and registration through the practice owner. Participants were either self-motivated to take part or were perceived by their practice owner as the "most suitable" and therefore probably the most interested person of the practice in digital topics.

Furthermore, it must be questioned whether the self-assessment questionnaires were really suitable for measuring competence. According to the results, competence in the domain of technical proficiency decreased when participating in this training course, which appears illogical. Since the questionnaire measures how capable the participants see themselves, this value can decrease if they learn what they do not yet know. Self-perception does not always match actual performance [38]. This is further backed up by another study that questions the suitability of self-assessment scales for measuring competence: "Perceived skills [...] do not predict actual performance," as van der Vaart et al [39] stated in their paper, comparing the results of self-assessment scales that aim to measure the eHealth literacy of participants, with their actual performance in different skill tests [39]. In this study, correlations between the used eHEALS (eHealth Literacy Scale) measure and successfully completed tasks were nonsignificant and weak, and no group differences between participants who scored below and above the median in the performance tests for the eHEALS scores were found. Jarva et al [37] supported this thesis by using the specified term "self-evaluated competence" in their study. The reliability and validity of estimating one's own competence have already been questioned by many authors [40-42]. Despite this, self-assessment of specific skills and knowledge is, to date, the most commonly used form of measuring digital competence [21]. The question arises as to whether the questionnaires—tested for quality characteristics such as reliability—are in fact simply measuring a different concept than competence. Ulfert-Blank and Schmidt [43] suggested *digital self-efficacy* as a possible characteristic that could be measured through these instruments, defined as "an individual's perception of efficacy in performing tasks related to the use of digital system." Bancroft et al [13] proposed that self-assessment of competence measures a mixture of competence and confidence and that both concepts are closely related and sometimes conflated, but also could be out of alignment, and a lack of confidence could hold back people who are per se competent. New paths must be found to measure the actual digital competence of health care professionals.

Future Research Directions

As mentioned above, new alternative ways to measure digital literacy/competence—besides self-assessment scales—must be found. One promising approach could be the use of performance measures, which were already used for the measurement of other concepts, like eHealth literacy [39] or data literacy [44,45].

It is noticeable among the participants of the training course that they all started with a very positive attitude. This is probably

because there was an active application process for the training, and people who were already digitally interested were more likely to want to be trained as Digi-Managers. It would be interesting to see whether further training would lead to an improvement in the attitudes of people who are not yet so positively disposed.

Learnings for Future Courses

The success in positive learning outcomes and satisfaction of the participants shows the relevance of the continuation of the course. The training program will be carried on with slight changes, following the feedback of participants, instructors, and organizers. Future courses will be shorter in time to enable smaller practices to participate with fewer lost hours of their employees and more e-learning and web-based courses in order to travel less. That should also improve the day and time selection options. Course sizes are to be reduced to enable more active involvement of participants and, above all, to further support practical networking and the exchange of experience. A streamlined concept with consistent quality should focus on the unique selling points of the Digi-Manager training: the digital laboratory and maturity model. The additional focus on soft skills (project management, communication, and conflict management) is helpful for the effective transfer and realization of digital projects in practice.

Conclusions

The Digi-Manager program was a successful and long-needed training program for health care professionals in the German region of Westfalen-Lippe. More training programs and courses for health care professionals are needed not only in Germany but all over the world. The mixture of transferring theoretical knowledge and practical applications with reference to one's own everyday work through soft skill training, the maturity model, and digitalization strategy results in a unique and effective further education concept.

Acknowledgments

This research was funded by the German Federal Ministry of Health (BMG, Bundesministerium für Gesundheit), grant ZMI5-2523FEP30B.

Authors' Contributions

All authors contributed to the conceptualization, funding acquisition, manuscript writing, and revision. SM provided overall supervision, and TG oversaw project administration. AM and SM were responsible for designing the methodology and conducting data analysis.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Equivalent of the ÄKWL certificate course and content of the different modules. ÄKWL: Ärztekammer Westfalen-Lippe (medical association of the Westphalia-Lippe region).
[PDF File, 148 KB - mededu_v11i1e70843_app1.pdf]

Multimedia Appendix 2
Questionnaires of the web-based surveys.
[PDF File, 581 KB - mededu_v11i1e70843_app2.pdf]

Multimedia Appendix 3
Inductively formed categories from free-text answers regarding the course modules.
[PDF File, 176 KB - mededu_v11i1e70843_app3.pdf]

Multimedia Appendix 4
Correlation coefficients and significance values.
[PDF File, 221 KB - mededu_v11i1e70843_app4.pdf]

Checklist 1
Checklist items of GREET (Guideline for Reporting Evidence-Based Practice Educational Interventions and Teaching).
[PDF File, 309 KB - mededu_v11i1e70843_app5.pdf]

References

1. Nazeha N, Pavagadhi D, Kyaw BM, Car J, Jimenez G, Tudor Car L. A digitally competent health workforce: scoping review of educational frameworks. J Med Internet Res 2020 Nov 5;22(11):e22706. [doi: 10.2196/22706] [Medline: 33151152]

2. Yeung AWK, Torkamani A, Butte AJ, et al. The promise of digital healthcare technologies. Front Public Health 2023;11:1196596. [doi: 10.3389/fpubh.2023.1196596] [Medline: 37822534]
3. Stachwitz P, Debatin JF. Digitalisierung im Gesundheitswesen: heute und in Zukunft. Bundesgesundheitsbl 2023 Feb;66(2):105-113. [doi: 10.1007/s00103-022-03642-8]
4. Digitalisierung und Datennutzung für Gesundheitsforschung und Versorgung – Positionen und Empfehlungen [Article in German]. German Science and Humanities Council. 2022 Jul. [doi: 10.57674/bxkz-8407]
5. Weik L, Fehring L, Mortsiefer A, Meister S. Big 5 personality traits and individual- and practice-related characteristics as influencing factors of digital maturity in general practices: quantitative web-based survey study. J Med Internet Res 2024 Jan 22;26:e52085. [doi: 10.2196/52085] [Medline: 38252468]
6. Weik L, Fehring L, Mortsiefer A, Meister S. Understanding inherent influencing factors to digital health adoption in general practices through a mixed-methods analysis. NPJ Digit Med 2024 Feb 27;7(1):47. [doi: 10.1038/s41746-024-01049-0] [Medline: 38413767]
7. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 2000 Jan;55(1):68-78. [doi: 10.1037//0003-066x.55.1.68] [Medline: 11392867]
8. Vitello S, Greatorex J, Shaw S. What is competence? A shared interpretation of competence to support teaching, learning and assessment. Cambridge University Press & Assessment. 2021 Dec 20 URL: https://www.cambridgeassessment.org.uk/Images/645254-what-is-competence-a-shared-interpretation-of-competence-to-support-teaching-learning-and-assessment.pdf [accessed 2025-07-30]
9. De Leeuw JA, Woltjer H, Kool RB. Identification of factors influencing the adoption of health information technology by nurses who are digitally lagging: in-depth interview study. J Med Internet Res 2020 Aug 14;22(8):e15630. [doi: 10.2196/15630] [Medline: 32663142]
10. Mannevaara P, Kinnunen UM, Egbert N, et al. Discovering the importance of health informatics education competencies in healthcare practice: a focus group interview. Int J Med Inform 2024 Jul;187:105463. [doi: 10.1016/j.ijmedinf.2024.105463] [Medline: 38643700]
11. Hübner U, Thyea J, Shaw T, et al. Towards the TIGER international framework for recommendations of core competencies in health informatics 2.0: extending the scope and the roles. In: Studies in Health Technology and Informatics, Volume 264: MEDINFO 2019: Health and Wellbeing e-Networks for All: IOS Press; 2019:1218-1222. [doi: 10.3233/SHTI190420]
12. Medizinische/r Fachangestellte/r – Ausbildungsberuf [Article in German]. Bundesagentur für Arbeit. 2024. URL: https://web.arbeitsagentur.de/berufenet/beruf/33212 [accessed 2024-10-19]
13. Bancroft R, Challen R, Pearce R. Searching for a shared understanding of digital confidence in a tertiary context: a scoping review. J Learn Dev High Educ 2024(30). [doi: 10.47408/jldhe.vi30.1061]
14. Levine T, Donitsa-Schmidt S. Computer use, confidence, attitudes, and knowledge: a causal analysis. Comput Human Behav 1998 Jan;14(1):125-146. [doi: 10.1016/S0747-5632(97)00036-8]
15. Neunaber T, Meister S. Digital maturity and its measurement of general practitioners: a scoping review. Int J Environ Res Public Health 2023 Feb 28;20(5):4377. [doi: 10.3390/ijerph20054377] [Medline: 36901387]
16. Phillips AC, Lewis LK, McEvoy MP, et al. Development and validation of the guideline for reporting evidence-based practice educational interventions and teaching (GREET). BMC Med Educ 2016 Sep 6;16:237. [doi: 10.1186/s12909-016-0759-1] [Medline: 27599967]
17. van Volkom M, Stapley JC, Amaturo V. Revisiting the digital divide: generational differences in technology use in everyday life. N Am J Psychol 2014;16(3):557-574 [FREE Full text]
18. Acilar A, Sæbø Ø. Towards understanding the gender digital divide: a systematic literature review. Glob Knowl Mem Commun 2023;72(3):233-249. [doi: 10.1108/GKMC-09-2021-0147]
19. Siddiq F, Scherer R. Is there a gender gap? A meta-analysis of the gender differences in students' ICT literacy. Educ Res Rev 2019 Jun;27:205-217. [doi: 10.1016/j.edurev.2019.03.007]
20. Helsper EJ, Eynon R. Digital natives: where is the evidence? Br Educ Res J 2010 Jun;36(3):503-520. [doi: 10.1080/01411920902989227]
21. Mainz A, Nitsche J, Weirauch V, Meister S. Measuring the digital competence of health professionals: scoping review. JMIR Med Educ 2024 Mar 29;10:e55737. [doi: 10.2196/55737] [Medline: 38551628]
22. Rachmani E, Hsu CY, Chang PW, et al. Development and validation of an instrument for measuring competencies on public health informatics of primary health care worker (PHIC4PHC) in Indonesia. Prim Health Care Res Dev 2020 Jul 6;21:e22. [doi: 10.1017/S1463423620000018] [Medline: 32624060]
23. Kuek A, Hakkennes S. Healthcare staff digital literacy levels and their attitudes towards information systems. Health Inform J 2020 Mar;26(1):592-612. [doi: 10.1177/1460458219839613] [Medline: 30983476]
24. Lewis JR. The system usability scale: past, present, and future. Int J Hum Comput Interact 2018 Jul 3;34(7):577-590. [doi: 10.1080/10447318.2018.1455307]
25. Razali NM, Wah YB. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. J Stat Model Analyt 2011;2(1):21-33 [FREE Full text]
26. Cohen J. Statistical Power Analysis for the Behavioral Sciences: Routledge; 1988. [doi: 10.4324/9780203771587]

27. Okada K. Is omega squared less biased? A comparison of three major effect size indices in one-way ANOVA. Behaviormetrika 2013 Jul;40(2):129-147. [doi: 10.2333/bhmk.40.129]
28. Sauro J, Lewis JR. Quantifying the User Experience: Practical Statistics for User Research: Morgan Kaufmann; 2016.
29. Blanca MJ, Alarcón R, Arnau J, Bono R, Bendayan R. Non-normal data: is ANOVA still a valid option? Psicothema 2017 Nov;29(4):552-557. [doi: 10.7334/psicothema2016.383] [Medline: 29048317]
30. Girden ER. ANOVA: Repeated Measures: SAGE Publications; 1992.
31. Samadbeik M, Fatehi F, Braunstein M, et al. Education and training on electronic medical records (EMRs) for health care professionals and students: a scoping review. Int J Med Inform 2020 Oct;142:104238. [doi: 10.1016/j.ijmedinf.2020.104238] [Medline: 32828034]
32. Edirippulige S, Armfield NR. Education and training to support the use of clinical telehealth: a review of the literature. J Telemed Telecare 2017 Feb;23(2):273-282. [doi: 10.1177/1357633X16632968] [Medline: 26892005]
33. Fernández-Luque AM, Ramírez-Montoya MS, Cordón-García JA. Training in digital competencies for health professionals: systematic mapping (2015-2019). Inf Prof 2021;30(2). [doi: 10.3145/epi.2021.mar.13]
34. Jimenez G, Spinazze P, Matchar D, et al. Digital health competencies for primary healthcare professionals: a scoping review. Int J Med Inform 2020 Nov;143:104260. [doi: 10.1016/j.ijmedinf.2020.104260] [Medline: 32919345]
35. Konttila J, Siira H, Kyngäs H, et al. Healthcare professionals' competence in digitalisation: a systematic review. J Clin Nurs 2019 Mar;28(5-6):745-761. [doi: 10.1111/jocn.14710] [Medline: 30376199]
36. Kinnunen UM, Heponiemi T, Rajalahti E, Ahonen O, Korhonen T, Hyppönen H. Factors related to health informatics competencies for nurses: results of a national electronic health record survey. Comput Inform Nurs 2019 Aug;37(8):420-429. [doi: 10.1097/CIN.0000000000000511] [Medline: 30741730]
37. Jarva E, Oikarinen A, Andersson J, Pramila-Savukoski S, Hammarén M, Mikkonen K. Healthcare professionals' digital health competence profiles and associated factors: a cross-sectional study. J Adv Nurs 2024 Aug;80(8):3236-3252. [doi: 10.1111/jan.16096] [Medline: 38323687]
38. Zell E, Krizan Z. Do people have insight into their abilities? A meta-synthesis. Perspect Psychol Sci 2014 Mar;9(2):111-125. [doi: 10.1177/1745691613518075] [Medline: 26173249]
39. van der Vaart R, van Deursen AJ, Drossaert CH, Taal E, van Dijk JA, van de Laar MA. Does the eHealth Literacy Scale (eHEALS) measure what it intends to measure? Validation of a Dutch version of the eHEALS in two adult populations. J Med Internet Res 2011 Nov 9;13(4):e86. [doi: 10.2196/jmir.1840] [Medline: 22071338]
40. van Deursen AJAM, van Dijk JAGM. Measuring internet skills. Int J Hum Comput Interact 2010 Sep 17;26(10):891-916. [doi: 10.1080/10447318.2010.496338]
41. Merritt K, Smith KD, Di Renzo JC. An investigation of self-reported computer literacy: is it reliable? Issues Inf Syst 2005;6(1):289-295. [doi: 10.48009/1_iis_2005_289-295]
42. Hargittai E. Survey measures of web-oriented digital literacy. Soc Sci Comput Rev 2005 Aug;23(3):371-379. [doi: 10.1177/0894439305275911]
43. Ulfert-Blank AS, Schmidt I. Assessing digital self-efficacy: review and scale development. Comput Educ 2022 Dec;191:104626. [doi: 10.1016/j.compedu.2022.104626]
44. Larasati PE, Yunanta DRA. Validity and reliability estimation of assessment ability instrument for data literacy on high school physics material. J Phys Conf Ser 2020 Jan 1;1440:012020. [doi: 10.1088/1742-6596/1440/1/012020]
45. Pratama MA, Lestari DP, Sari WK, Putri TSY, Adiatmah VAK. Data literacy assessment instrument for preparing 21 Cs literacy: preliminary study. J Phys Conf Ser 2020 Jan 1;1440:012085. [doi: 10.1088/1742-6596/1440/1/012085]

Abbreviations

ÄKWL: Ärztekammer Westfalen-Lippe (medical association of the Westphalia-Lippe region)
eHEALS: eHealth Literacy Scale
GREET: Guideline for Reporting Evidence-Based Practice Educational Interventions and Teaching
KVWL: Kassenärztliche Vereinigung Westfalen-Lippe (association of statutory health insurance physicians of the Westphalia-Lippe region)
mHealth: mobile health
PHIC4PHC: Public Health Informatics Competencies for Primary Health Care
SUS: system usability scale
TAM: technology acceptance model
UTAUT: unified theory of acceptance and use of technology

Edited by D Chartash; submitted 03.01.25; peer-reviewed by A Azizi, J Busch-Casler, M Chomali; revised version received 30.06.25; accepted 02.07.25; published 29.08.25.

Please cite as:
Mainz A, Neunaber T, D'Agnese PC, Eid A, Galla T, Ellers C, Meister S
Digital Literacy Training for Digitalization Officers ("Digi-Managers") in Outpatient Medical and Psychotherapeutic Care: Conceptualization and Longitudinal Evaluation of a Certificate Course
JMIR Med Educ 2025;11:e70843
URL: https://mededu.jmir.org/2025/1/e70843
doi:10.2196/70843

Mapping the Evolution of China's Traditional Chinese Medicine Education Policies: Insights From a BERTopic-Based Descriptive Study

Tao Yang[1], PhD; Fan Yang[2,3], PhD; Yong Li[2], PhD

[1]School of Basic Medical Sciences, Chengdu University of Traditional Chinese Medicine, Chengdu, China

[2]School of Management, Chengdu University of Traditional Chinese Medicine, Sichuan Province, Chengdu, China

[3]School of Teacher Education, East China Normal University, Shanghai, China

Corresponding Author:
Fan Yang, PhD
School of Management, Chengdu University of Traditional Chinese Medicine, Sichuan Province, Chengdu, China

Abstract

Background: Traditional Chinese medicine (TCM) education in China has evolved significantly, shaped by both national policy and social needs. Despite this, the academic community has yet to fully explore the long-term trends and core issues in TCM education policies. As the global interest in TCM continues to grow, understanding these trends becomes crucial for guiding future policy and educational reforms. This study used cutting-edge deep learning techniques to fill this gap, offering a novel, data-driven perspective on the evolution of TCM education policies.

Objective: This study aimed to systematically analyze the research topics and evolutionary trends in TCM education policies in China using a deep learning–based topic modeling approach, providing valuable insights to guide future policy development and educational practices.

Methods: TCM policy–related documents were collected from major sources, including the Ministry of Education, the National Administration of Traditional Chinese Medicine, PKU Lawinfo, and archives of TCM colleges. The text was preprocessed and analyzed using the BERTopic model, a state-of-the-art tool for topic modeling, to extract key themes and examine the policy development trajectory.

Results: The analysis revealed 27 core topics in TCM education policies, including medical education, curriculum reform, rural health care, internationalization, and the integration of TCM with modern education systems. These topics were clustered into 5 stages of policy evolution: marginalization, standardization, specialization, systematization, and restandardization. These stages reflect the ongoing balancing act between modernizing TCM education and preserving its traditional values, while adapting to national political, social, and economic strategies.

Conclusions: This study offers groundbreaking insights into the dynamic and multifaceted evolution of TCM education policies in China. By leveraging the BERTopic model, it provides a comprehensive framework for understanding the forces shaping TCM education and offers actionable recommendations for future policy making. The findings are essential for educators, policymakers, and researchers aiming to refine and innovate TCM education in an increasingly globalized world.

(JMIR Med Educ 2025;11:e72660) doi:10.2196/72660

KEYWORDS

traditional Chinese medicine education policy; BERTopic; policy evolution; descriptive study; topic analysis

Introduction

Traditional Chinese medicine (TCM) is an important part of China's 5000-year-old culture. Its education policy not only concerns the inheritance and development of TCM but also has a profound impact on the global health system [1]. In the modernization process of China, the evolution of TCM education policies reflects the country's emphasis on and adjustment of traditional medical education, as well as the complex interaction between social demands, political environment, and cultural confidence. With the enhancement of China's comprehensive national power and its rising international status, TCM education has gradually become an important part of global health governance. The direction and effectiveness of its policies directly affect the influence and adaptability of TCM globally [1]. Against this backdrop, in-depth research and analysis of the evolution and development path of TCM education policies are of great academic and practical significance. By tracing the evolution of TCM education policies, we can reveal the driving forces and logical relationships behind the policies, providing a solid basis for decision makers. Moreover, analyzing the

evolutionary trends of TCM education policies can also help educational institutions and researchers better understand the future direction of TCM education, thereby promoting the reform and innovation of TCM education to meet the diverse needs of modern society and international challenges.

Text mining technology, as an advanced data analysis method, can extract potential topics and patterns from a large number of policy texts, providing a new perspective for understanding the evolution of TCM education policies. Through systematic analysis of policy texts, we can identify the changing trends, weight distribution, and logical structure of policy topics, thereby revealing the focus and orientation of TCM education policies in different historical stages. This analysis not only provides empirical support for the theoretical research of TCM education but also offers scientific references for policymakers when formulating future TCM education policies [2]. This study uses Bert and dynamic topic modeling (DTM) and other text mining methods to conduct topic mining [3,4], identification, and analysis of TCM education policy texts to study the development direction and evolutionary path of TCM education policies, providing guidance and references for scientific research and scientific management practices in relevant fields.

Methods

Data Sources

This study obtained more than 200 national-level policy texts related to TCM education from 1902 to 2024 from the official websites of the Central People's Government of the People's Republic of China, the Ministry of Education, the National Health Commission, and the National Administration of Traditional Chinese Medicine; China National Knowledge Infrastructure (including yearbooks, mainly the China TCM Yearbook); publicly published books, journals, and newspapers; and internal material compilations and archive files of relevant departments, supplemented by important policies mentioned in literature studies. These data cover key stages and important nodes of TCM education policies, ensuring the comprehensiveness and systematic nature of the study.

Inclusion Criteria

Only national-level policy documents directly related to TCM education were included. This means policies that explicitly addressed TCM educational objectives, curriculum design, teaching methods, teacher qualifications, student enrollment, or educational resource allocation in the context of TCM education were selected. Policies that were tangentially related, such as general health policies with only a passing mention of TCM without specific educational implications, were excluded.

Exclusion Criteria

Documents that were not official policies, such as opinion pieces, academic commentaries, and nongovernmental reports on TCM education, were excluded. Additionally, policies that were duplicates, drafts, or had been superseded by subsequent policies were not included in the final dataset.

The data collection focuses on national-level policy documents, combined with TCM education reform plans and guiding opinions issued by the Ministry of Education and health departments, as well as relevant historical records and academic research reports, to build a complete policy evolution database.

Overall Research Framework

The overall research framework of this paper is divided into data collection, model improvement, data cleaning, BERTopic data processing, result output, and result interpretation. BERTopic is a topic modeling technique based on BERT word vectors [5], which creates dense clusters using Transformers and class-based Term Frequency–Inverse Document Frequency (c-TF-IDF) while retaining important words in topic descriptions. The topic modeling analysis using this model can be carried out in the following 5 steps: embeddings, dimensionality reduction, clustering, tokenizer, and weighting scheme [6]. The model has modular features, and processing methods can be chosen autonomously at each step according to research needs [7].

Compared with traditional topic models such as Latent Dirichlet Allocation (LDA), BERTopic exhibits distinct advantages in analyzing TCM policies. Traditional models rely on statistical co-occurrence of words, which often fails to capture semantic relationships effectively. For instance, LDA assumes that topics are generated from a Dirichlet distribution over words, struggling to differentiate between polysemous terms [8]. In contrast, BERTopic leverages pretrained BERT embeddings, enabling it to understand the context of words, thus better handling TCM-specific terminologies that are often ambiguous and context-dependent. BERTopic, by integrating c-TF-IDF, emphasizes the importance of words within topics based on semantic similarity, leading to more interpretable and accurate topic representations in the TCM policy domain [5].

When compared with existing TCM policy research using methods such as LDA-based analysis, BERTopic demonstrates significant superiority in semantic understanding. Traditional LDA-based studies often produce topics with overlapping meanings and struggle to distinguish between terms with similar statistical distributions but different semantic interpretations [9]. For example, in TCM policies, terms such as "herbal medicine" may have multiple connotations, such as medicinal plant resources, prescription formulations, or quality control standards. BERTopic, powered by its deep bidirectional Transformer architecture, can analyze the surrounding text to disambiguate these polysemous terms accurately [10]. Moreover, TCM has a rich set of domain-specific jargon, and BERTopic's pretraining on extensive language corpora allows it to recognize and leverage these specialized terminologies more effectively, providing a more nuanced and comprehensive understanding of TCM policy documents [11].

Data Cleaning and BERTopic Model Improvement Methods

After obtaining the TCM education policy-related texts, since the BERTopic model itself has a limit on text length, it will truncate texts exceeding 512 characters. However, most policy texts have more than 1000 characters. Therefore, this study improved the BERTopic model by extracting BERT features

for every 512 characters of long texts and then calculating the mean value as the text feature.

In addition, conventional data cleaning of policy texts is also required. First, the text content of Word and PDF policy documents is extracted, and obvious recognition errors and redundant invalid data are removed or replaced. Then, using the Jieba library and the HIT stop word list, common Chinese stop words are removed. Furthermore, a specialized word dictionary for the TCM policy field is used to segment the text, obtaining preprocessed text data.

Data Processing Methods

Text Vectorization

The first step in processing the text input into the model is text vectorization, that is, embedding. Before conducting topic modeling and clustering, it is necessary to convert text data into numerical vector representations, which is the basis of topic modeling [12]. In the BERTopic constructed in this study, the preprocessed documents are converted into dense vector representations based on the Sentence-BERT framework. By adding pooling operations in the output of BERT, fixed-size sentence embeddings are obtained. This study uses the paraphrase-multilingual-MiniLM-L12-V2 sentence vector model to represent the preprocessed text embeddings. This model supports more than 50 languages, including Chinese, and is very effective for most use cases. It not only greatly improves training speed but also maintains superior performance.

Dimensionality Reduction

After text vectorization, dimensionality reduction is required, that is, mapping high-dimensional data to low-dimensional space. The purpose of dimensionality reduction is to reduce the dimension of the embedding vectors, thereby simplifying computation and storage and enhancing clustering effects [13]. Common dimensionality reduction techniques include Uniform Manifold Approximation and Projection (UMAP), Principal Component Analysis, and t-Distributed Stochastic Neighbor Embedding, with UMAP being the default choice for BERTopic. This method effectively reduces dimensions while retaining the local and global structure of the data, ensuring the accuracy of clustering semantically similar documents. Compared with other dimensionality reduction algorithms, UMAP is better at preserving the characteristics of the original data in low-dimensional space, making it the preferred method for text clustering dimensionality reduction.

Clustering

After text dimensionality reduction, the next step is to use clustering algorithms to assign embedding vectors to different topics. In this study, the HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) algorithm is used for topic clustering of the reduced dimension document embeddings [14]. HDBSCAN extends DBSCAN (Density-Based Spatial Clustering of Applications with Noise) into a hierarchical clustering algorithm, which can automatically select the optimal clusters and treat noise as outliers, thereby avoiding the incorrect allocation of unrelated documents. This algorithm can generate dynamic topic representations, allowing the same topic to be expressed in different forms at different times, providing flexibility for studying the temporal changes of policy topics [15].

Topic Word Extraction

After topic clustering is completed, the next step is to extract and represent topic words using the Bag-of-Words model. This study used the c-TF-IDF method, which combines all documents in each cluster into a long document and calculates the frequency of each word [5]. To improve the coherence and diversity of topics, maximal marginal relevance is further applied to optimize the selection of topic words, reducing the repetition of synonyms and ensuring the diversity and accuracy of topic representation.

Dynamic Topic Modeling

DTM takes into account the changes in topics over time and can better capture the evolution of policies than static topic modeling [16]. BERTopic combines temporal factors by calculating topic representations at each time point to achieve DTM. First, the model is fitted globally and then the topic representations at each time point are fine-tuned. This allows for more precise analysis of topic evolution through global and evolutionary adjustments, providing an effective tool for understanding changes in policy topics over different periods.

Results

TCM Education Policy Topic Extraction Results

Using the BERTopic model for topic modeling of policy abstracts, this study adjusted the model parameters based on multiple experimental results. Under the set parameter conditions, a total of 27 main research topics in TCM education policies were identified (from topic 0 to topic 26), covering 183 policy texts in the dataset. Another 27 policy texts were classified as noise data (assigned to the −1 category). The noise data, aside from a few outlier policy texts, mainly consisted of research topics that did not reach the set parameter cluster size. These were mostly secondary or minor topics under the main themes, and their impact on the main research topics of TCM education policies was minimal. The representative keywords generated by the model for each topic are shown in Table 1.

Table . Topic identification results and related document counts based on BERTopic.

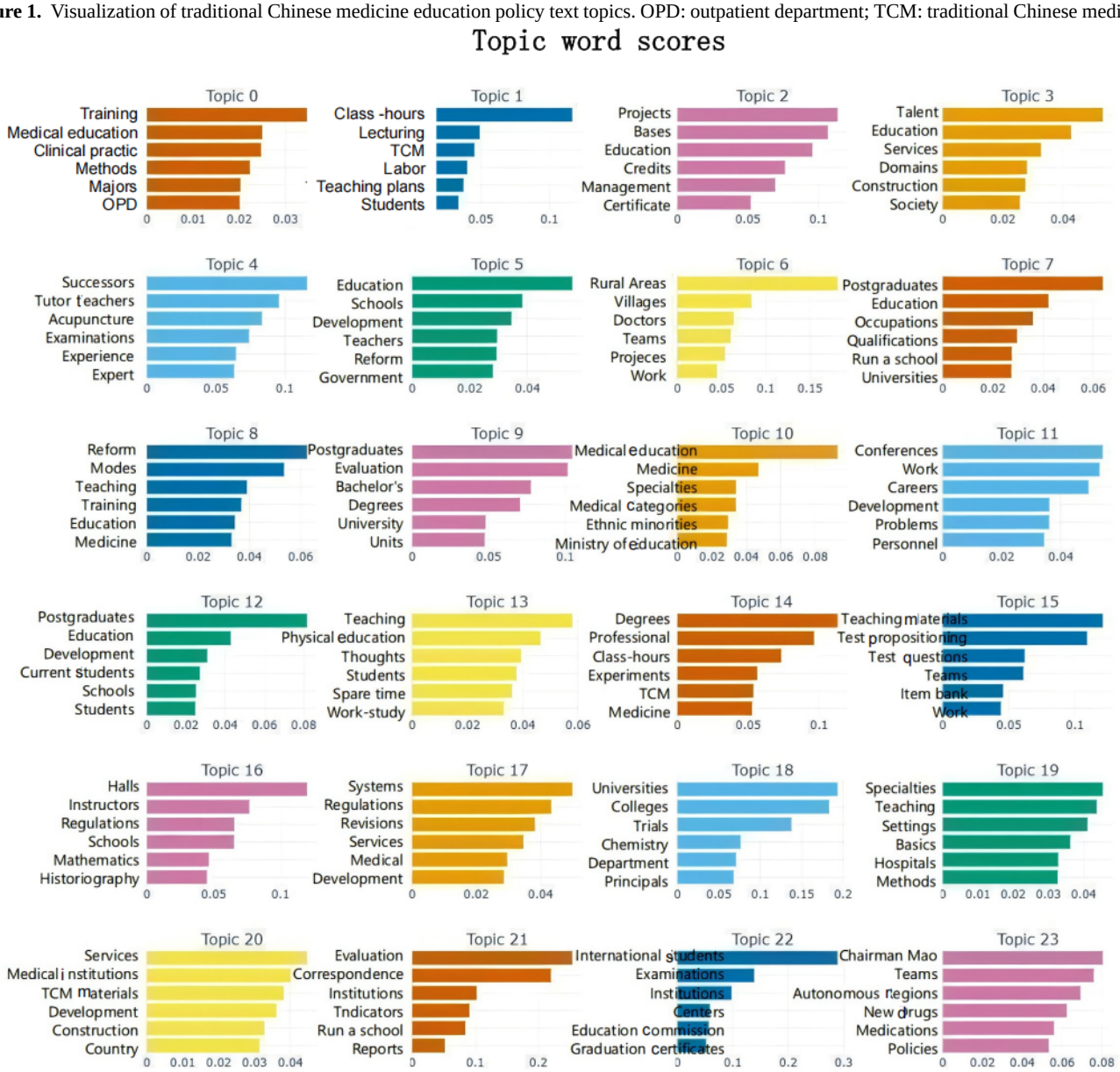
Topic number	Document count	Percentage	Keywords
Topic 0	20	10.9	“Training,” “Medical Education,” “Clinical,” “Method,” “Major,” “Outpatient,” “Technology,” “Principle,” “Content,” and “Trainee”
Topic 1	14	7.7	“Class Hours,” “Lecture,” “‘Chinese Medicine,” “Labor,” “Teaching Plan,” “Student,” “Basic,” “Assessment,” “Teaching,” and “Medic”
Topic 2	10	5.5	“Project,” “Base,” “Education,” “Credit,” “Management,” “Certificate,” “Committee,” “Method,” “Sponsor,” and “Office”
Topic 3	10	5.5	“Talent,” “Education,” “Service,” “Field,” “Construction,” “Society,” “Hospital,” “Teaching,” “Development,” and “Clinical”
Topic 4	10	5.5	“Successor,” “Mentor,” “Acupuncture,” “Exam,” “Expert,” “Experience,” “International,” “Graduation,” “Academic,” and “Work”
Topic 5	9	5.0	“Education,” “School,” “Development,” “Teacher,” “Reform,” “Government,” “Society,” “Education,” “Vocational,” and “System”
Topic 6	9	5.0	“Rural,” “Village,” “Doctor,” “Team,” “Project,” “Work,” “Personnel,” “Rural Health,” “Plan,” and “Training”
Topic 7	9	5.0	“Graduate,” “Education,” “Vocation,” “Academic,” “University,” “Quality,” “Development,” “Reform,” and “Tutor”
Topic 8	9	5.0	“Reform,” “Model,” “Teaching,” “Training,” “Education,” “Medicine,” “Five-Year Program,” “Standardization,” “Construction,” and “General Practice“
Topic 9	9	5.0	“Graduate,” “Assessment,” “Degree,” “Bachelor’s Degree,” “University,” “Unit,” “Work,” “Quality,” and “Discipline”
Topic 10	9	5.0	“Medical Education,” “Medicine,” “Major,” “Ethnic Minority,” “Ministry of Education,” “Construction,” “Medical College,” “Student,” and “Teaching”
Topic 11	7	3.8	“Meeting,” ”Work,” “Enterprise,” “Development,” “Problem,” “Personnel,” “National,” “Reform,” “Western Medicine,” and “Comrade”
Topic 12	6	3.3	“Graduate,” “Education,” “Development,” “Student,” “School,” “Student,” “Vocation,” “Education,” “Guidance,” and “Society“

Topic number	Document count	Percentage	Keywords
Topic 13	5	2.7	“Teaching,” “Physical Education,” “Ideology,” “Student,” “Part-time,” “Half-work,” “Medical Skill,” “Physical Education Class,” “Work,” and “Patient”
Topic 14	5	2.7	“Degree,” “Professional Degree,” “Class Hours,” “Experiment,” “Chinese Medicine,” “Medicine,” “Master’s Degree,” “Set,” “Clinical,” and “Assessment”
Topic 15	5	2.7	“Textbook,” “Question Setting,” “Question,” “Group,” “Question Bank,” “Work,” “Course,” “Database Construction,” “Question Bank,” and “Planning”
Topic 16	5	2.7	“School,” “Teaching,” “Regulation,” “School,” “Mathematics,” “History,” “Physics,” “Graduation,” “Geography,” and “Student”
Topic 17	4	2.2	“Research,” “Construction,” “Chinese Medicine,” “Service,” “Medicine,” “Development,” “AIDS,” “Advantage,” “Medical,” and “Key”
Topic 18	4	2.2	“University,” “College,” “Experiment,” “Chemistry,” “Department,” “President,” “Subject,” “Lecture Notes,” “Dean,” and “Enrollment”
Topic 19	4	2.2	“Major,” “Teaching,” “Setting,” “Basic,” “Hospital,” “Method,” “Full-time Teacher,” “Course,” “Pathogenesis,” and “Clinical”
Topic 20	4	2.2	“Service,” “Medical Institution,” “Chinese Medicinal Materials,” “Development,” “Construction,” “National,” “Medical,” “System,” “Research,” and “Institution”
Topic 21	3	1.6	“Assessment,” “Correspondence Education,” “Institution,” “Indicator System,” “Education,” “Report,” “Level,” “Work,” “Aspect,” and “Experience”
Topic 22	3	1.6	“International Student,” “Exam,” “Institution,” “Center,” “Education Commission,” “Graduation Certificate,” “Regulation,” “Time,” “Academic Certificate,” and “Question Setting”
Topic 23	3	1.6	“Chairman Mao,” “Team,” “Autonomous Region,” “New Medicine,” “Medical Science,” “Policy,” “Medicine,” “Work,” “Lack of Successors,” and “Important Instruction”
Topic 24	3	1.6	“Credit,” “Committee,” “Education,” “Academic Conference,” “Score,” “Chairman,” “Project,” “Secretary-General,” “Assessment,” and “Method”

Topic number	Document count	Percentage	Keywords
Topic 25	2	1.1	"Talent," "Construction," "Service," "Medicine," "Development," "Standardization," "Chinese Medicine," "Ethnic," "Culture," and "Talent Team"
Topic 26	2	1.1	"Professional Degree," "Master," "Standardization," "Graduate," "Base," "Family Planning Commission," "Training," "Office," "Collaboration," and "Medicine and Education"

According to the results shown in Figure 1, the BERTopic model successfully extracted the main topics from the TCM education policy literature and effectively expressed these topics through feature words and weights. The following are specific descriptions of several topics in the figure: topic 3, with feature words such as "talent," "education," "service," and "construction," indicates that this topic focuses on the cultivation of TCM talents and the improvement of educational quality, showing the policy's emphasis on enhancing the capabilities of TCM talents. Topic 6, with feature words including "rural," "village," "doctor," and "team," indicates that this topic is related to rural medical services and grassroots hospital construction, reflecting the efforts of TCM policies in promoting grassroots medical services. Topic 10, with feature words such as "medical education," "Ministry of Education," "medical major," and "professional," indicates that this topic mainly involves educational planning and professional construction related to the Ministry of Education, reflecting the focus of TCM education policies on overall planning and disciplinary development. Topic 14, with feature words such as "class hours," "experiment," and "degree," indicates that this topic is mainly related to the degree settings and graduate education in TCM education, reflecting the policy inclination toward specialized cultivation in higher education. Topic 18, with feature words such as "university," "college," "department," and "president," focuses on the academic development and overall educational quality of TCM higher education institutions, showing the role of policies in promoting the systematic development of TCM education. Topic 21, with feature words such as "assessment," "institutions," "indicator system," and "education," indicates that this topic is related to the assessment policies of TCM education, focusing on the standardization of TCM education. Through these feature words and their weight distributions, it can be seen that the BERTopic model has effectively captured the differences between various topics and can accurately express the core content of each topic. These topics reflect the multiple dimensions of concern in TCM education policies, including educational planning, talent cultivation, grassroots medical services, systematization, specialization, and standardization. This provides a clear direction and basis for subsequent policy analysis and improvement.

Figure 1. Visualization of traditional Chinese medicine education policy text topics. OPD: outpatient department; TCM: traditional Chinese medicine.

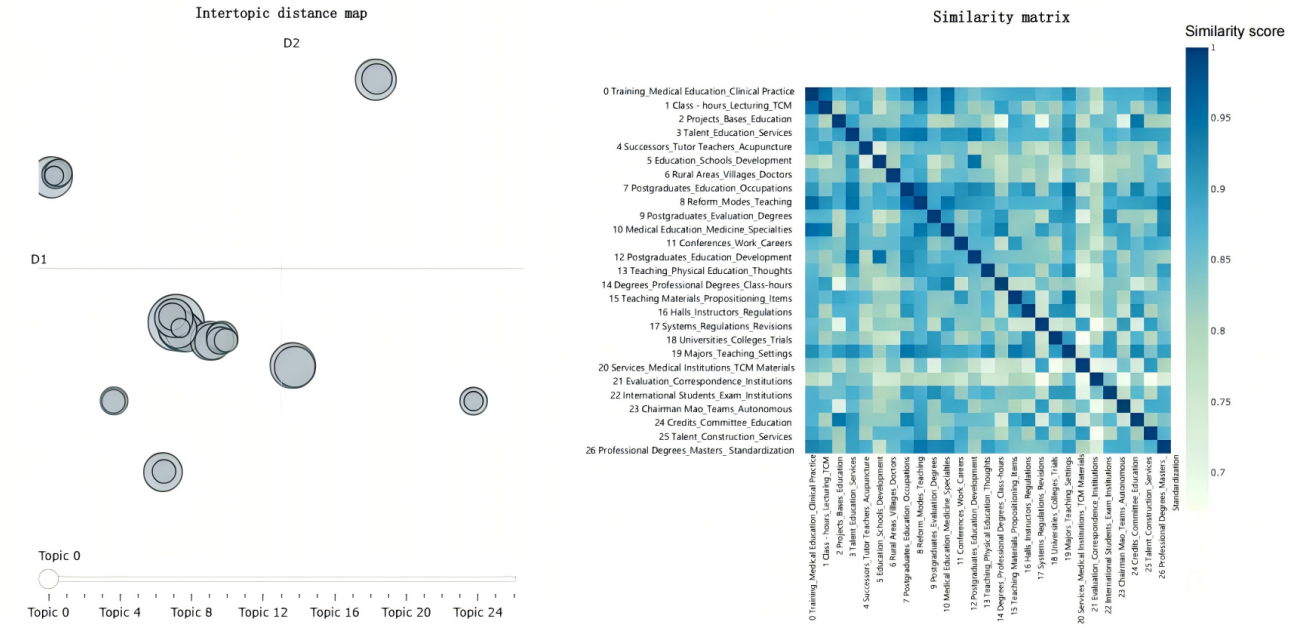


According to the 2D visualization distribution results of the topic-related documents shown in Figure 2, the clustering of different topics in the 2D space can be observed. The documents in the figure are roughly distributed in 4 quadrants, with each quadrant concentrating a category of topic documents, showing a clear trend of classification concentration. The first quadrant mainly focuses on topics related to educational curriculum reform and development. These documents involve the optimization of the educational system, curriculum reform, and the improvement of graduate education, reflecting the attention of TCM education policies on the development of the educational system. The second quadrant gathers topics related to academic research and teaching quality. These documents mainly focus on improving teaching quality and promoting academic research, showing the efforts and achievements of TCM education in the academic and teaching practice aspects. The third quadrant mainly concentrates on topics related to grassroots medical services and talent cultivation. These documents emphasize the role of TCM in grassroots medical services and how to cultivate suitable talents through educational policies to meet the needs of grassroots medical services, reflecting the policy's focus on strengthening rural medical services. The fourth quadrant gathers topics related to the specialization of TCM education and degree settings. These documents discuss the professionalization, institutionalization, and degree settings in TCM education, reflecting the strategies of TCM education in specialized and higher education development. From the figure, it can be seen that the topic documents in each quadrant have a clear trend of classification concentration, reflecting the targeted adjustments and optimizations made by TCM education policies in multiple dimensions. The first quadrant focuses on the development of the educational system and curriculum reform, the second quadrant highlights the improvement of academic research and teaching quality, the third quadrant pays attention to the strengthening of grassroots medical services and talent cultivation, and the fourth quadrant focuses on the specialization and higher education development of TCM education. This classification concentration trend further verifies the diversified promotion of TCM education policies at different levels.

Figure 2. Thematic—term-clustering diagram. TCM: traditional Chinese medicine.

Based on the identified topics, the study used the model.visualize-topics() function to generate an interactive visualization of all topics, as presented in Figure 3. Each circle in the graph represents a topic, with its size indicating the frequency of the topic's appearance in all documents. The distance between circles represents the similarity between topics, with closer distances indicating higher similarity. It can be seen that topic 0 and topic 1 are the most frequently occurring topics, involving key issues such as curriculum settings, clinical teaching, and teaching methods in TCM education. In addition, multiple topics such as topic 5, topic 7, topic 8, and topic 9 are clustered in the lower left corner of the graph, indicating higher similarity among these topics. They cover common aspects of TCM education policies, such as educational reform, improvement of teaching methods, and educational management.

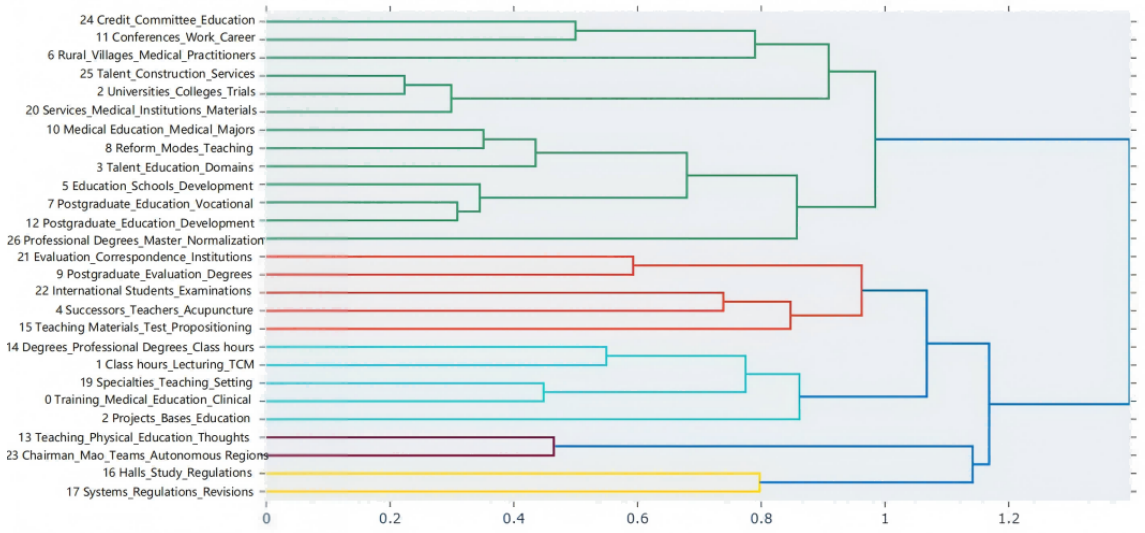
Figure 3. Visualization of traditional Chinese medicine education policy topic distance and cosine similarity. TCM: traditional Chinese medicine.



To understand the potential hierarchical structure of the topics, hierarchical clustering was performed on the relationships between topics based on cosine similarity, and the clustering results were visualized, as shown in Figure 4. The hierarchical relationships between topics can be intuitively understood from the figure. For example, topic 18 and topic 24 are located on the same clustering branch. These topics mainly involve academic development, degree settings, and overall educational development in higher education, which are important components of the systematic construction of TCM education. Topic 21 and topic 9 are clustered together, reflecting the policy's emphasis on standardization processes such as

professional assessment, educational standards, and school management. This ensures consistency in educational content and management methods, thereby improving educational quality and professional standards. Topic 14 and topic 1 form a distinct clustering branch, indicating the policy's focus on maintaining and developing the unique disciplinary advantages of TCM education through specialized education, curriculum design, and professional education content. Topic 13 and topic 23 are clustered together, showing a common focus on academic inheritance and team building. The policy promotes the standardized management of TCM education through these topics. Topic 16 and topic 17 are clustered together, involving topics such as schools, regulations, old medicine, and abolition, indicating the marginalization and stagnation of TCM education.

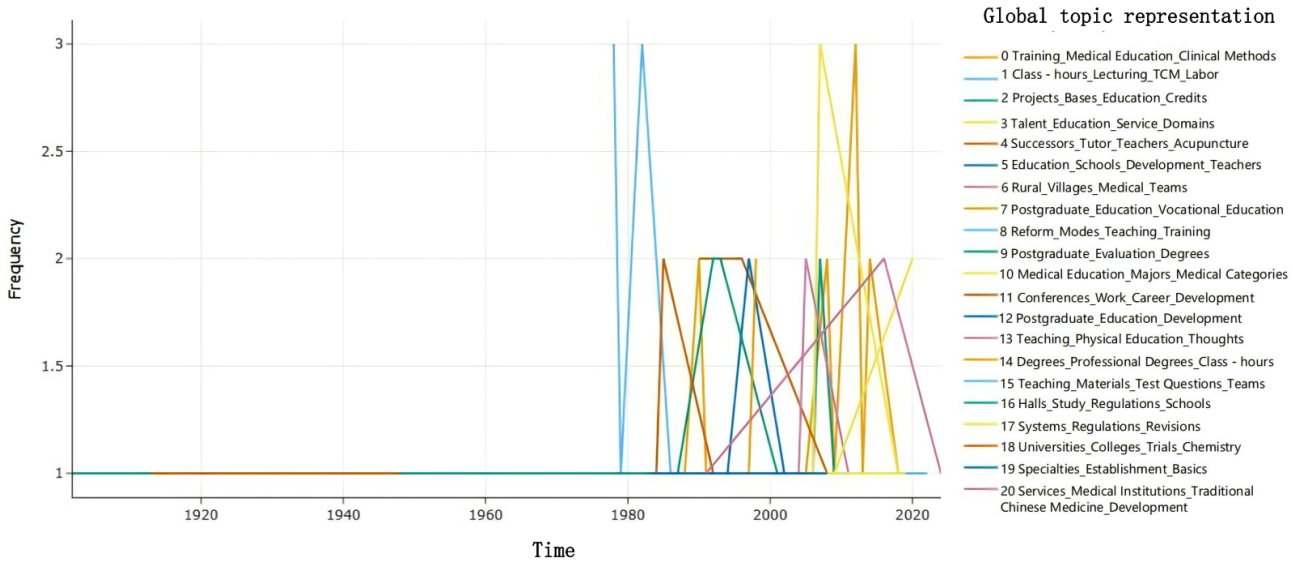
Figure 4. Hierarchical clustering of topics based on cosine similarity. TCM: traditional Chinese medicine.



Evolutionary Analysis of TCM Education Policy Topics

The DTM of BERTopic can intuitively present the research hotspots and scholar attention changes in the field, facilitating the analysis of the evolutionary trends of various research directions in information resource management. Based on the model results outlined in the preceding analysis, further analysis was conducted on the historical evolution and trends of the extracted topics over time, as shown in Figure 5. It can be seen that before the 1970s, the occurrence of topics was relatively sparse, mainly focusing on basic education and training issues. After the reform and opening up, the frequency of topic occurrence increased significantly, indicating high policy activity during this period. The alternating appearance of topics suggests that TCM education policies underwent numerous reforms and adjustments during this period. For example, topic 2 and topic 10 frequently appeared during this period, indicating increasing attention to the specialized development of TCM education, such as the construction of education bases, project management, and professional education. In the 21st century, the frequency of multiple topics increased significantly and alternated frequently, indicating the complexity and diversity of policies during this period. This suggests that with the development of society, the field of TCM education is continuously adapting to new demands, and policies are constantly being adjusted and evolved accordingly. Overall, over time, the themes of TCM education policies have become richer and more complex, reflecting the developmental trajectory and evolving academic focus of TCM education policies.

Figure 5. Evolutionary trends of traditional Chinese medicine education policy topics. TCM: traditional Chinese medicine.



Discussion

Evolutionary Trajectory of TCM Education Policies

Drawing on the policy cycle theory [17] and the multiple streams framework [18], this study systematically analyzes the evolutionary trends of TCM education policies across multiple key topics. These theoretical frameworks offer a structured lens to interpret how policy problems are identified, solutions formulated, and implemented over time, thereby enhancing the interpretability of the 5-stage division of TCM education policy evolution [19].

Policy cycle theory posits that policies undergo sequential stages of agenda setting, formulation, implementation, evaluation, and termination or continuation. In the context of TCM education, this theory helps understand how policy makers have navigated challenges and opportunities at each stage to advance the systematization, standardization, specialization, and normalization of TCM education [20]. The multiple streams framework, on the other hand, emphasizes the confluence of problem streams, policy streams, and political streams in driving policy change. This perspective elucidates how external factors, such as social needs, technological advancements, and political climates, have interacted to shape the evolution of TCM education policies [21]. Through the analysis of national-level policy texts using the BERTopic model, we have identified the following evolutionary stages of TCM education policies.

Through the analysis of national-level policy texts using the BERTopic model, we have identified the following evolutionary stages of TCM education policies:

Marginalization Stage (1840-1949)

During this period, influenced by the social and political upheaval and the influx of Western medicine, TCM education faced marginalization. The lack of a stable political environment and the dominance of Western medical concepts in the policy agenda hindered the development of TCM education, aligning with the problem stream and political stream dynamics described in the multiple streams framework.

During this period, policy themes focused on “abolishing traditional Chinese medicine” and “prohibiting old medical schools” [22]. Specifically, policymakers believed that TCM could not be verified within the framework of modern science and thus did not conform to contemporary educational principles. This perspective was reflected in several important educational policy documents, such as the “Regulations for the Establishment of Universities (Gui Mao School System)” and the “University Regulations (Ren Zi Gui Chou School System),” which explicitly called for the abolition of traditional TCM education and prohibited the establishment of any form of old medical schools [23,24]. During this time, the government’s attitude toward TCM education was extremely negative, considering it unscientific and unsystematic, and attempted to exclude it entirely from the national formal education system through coercive educational reforms. Under government suppression, TCM educational institutions were closed one after another, and the traditional TCM education model fell into a state of stagnation. This situation led to a sharp decline in the overall influence of TCM in Chinese society, and its very existence was facing unprecedented challenges. However, despite the systemic suppression of TCM education at the policy level, the demand for and support of TCM among the public did not completely disappear. Many TCM practitioners and supporters continued to preserve and develop this ancient medical system through private tutoring and family inheritance, trying to sustain TCM in a difficult environment. Although this informal education model was far smaller in scale and influence compared with the government-supported modern education system, it preserved the basic knowledge and skills of TCM to a certain extent, providing an important foundation for the later revival of TCM education.

Normalization Stage (1949-1978)

After the founding of the People’s Republic of China, the government began to recognize the value of TCM, and policies were formulated to normalize TCM education. This stage reflects the agenda-setting and formulation phases of the policy cycle, as policy makers worked to integrate TCM into the national education system.

The policy themes during this period mainly focused on "restoring TCM education," "Western medicine practitioners learning TCM," "master-apprentice TCM education," and "inheriting the academic experience of senior TCM practitioners" [25]. With the establishment of the new China, the national ideology reevaluated the value of TCM and began to protect and develop it as an important part of Chinese culture. Against this backdrop, the government issued a series of policies aimed at restoring and standardizing TCM education. For example, the "Notice on the Regulations for the Organization of TCM Advanced Study Schools and Classes" and the "Instruction on Carrying Out Master-Apprentice TCM Education" clarified the measures for the restoration of TCM education and attempted to combine modern education with traditional master-apprentice inheritance to ensure the effective transmission and development of TCM knowledge and skills. The policy of "Western medicine practitioners learning TCM" was a significant orientation during this period. The government hoped to integrate the strengths of both traditional and modern medicine by encouraging Western medical practitioners to study TCM, thereby addressing the shortage of medical resources at the time. Additionally, the government advocated for the continuation of traditional TCM knowledge through the "master-apprentice" approach, providing a dual safeguard for the revival of TCM education [26]. Furthermore, recognizing that the essence of TCM was often preserved in the experience of senior practitioners, the government issued the "Urgent Notice on Inheriting the Academic Experience of Senior TCM Practitioners" to ensure that this valuable knowledge would not be lost over time. This policy demonstrated the government's emphasis on balancing inheritance and innovation in TCM education and its commitment to preserving and promoting the core aspects of TCM through institutional means [27]. During this stage, the normalization of TCM education was reflected not only in policy making but also in the gradual improvement of the educational system. The government established TCM colleges nationwide and formulated unified teaching syllabuses and textbooks, such as the "Opinions on the Compilation of Teaching Syllabuses and Textbooks for TCM College Courses." These measures ensured the standardization of TCM education, leading to its expansion in scale and enhancement in content and quality.

Specialization Stage (1978-1999)

With the reform and opening-up policy, TCM education entered a stage of specialization. The emphasis on developing specialized TCM education programs and improving the quality of TCM education can be seen as a response to the evolving social needs and technological progress, in line with the implementation and evaluation stages of the policy cycle.

After China implemented the reform and opening-up policy in 1978, TCM education entered a new stage characterized by specialization. The policy themes during this period mainly focused on "promoting and developing TCM education," "seven-year programs," "teaching content," and "academic inheritance." These themes not only reflected the government's attention and investment in TCM education but also demonstrated the deepening and expansion of TCM education within the modern educational system. "Promoting and developing TCM education" was the core policy orientation of this stage [28]. The government recognized the unique value of TCM education in the modern medical system and issued a series of policies aimed at supporting and developing TCM education. For example, the "Report on Seriously Implementing the Party's TCM Policy and Solving the Problem of Succession in the TCM Workforce" emphasized the issue of TCM talent cultivation and proposed strengthening education to address the shortage of successors in the TCM field [29]. The "seven-year program" was a significant innovation in TCM education during this period. To improve the quality of TCM education, the government began to extend the duration of TCM education from the traditional 5-year program to a 7-year program. This measure not only increased the students' learning time but also broadened and deepened the educational content, enabling students to graduate with more solid theoretical knowledge and practical skills in TCM. The "Notice on the Trial Implementation of Seven-Year Higher TCM Education" officially launched this educational reform, marking a step toward higher academic standards in TCM education. In terms of teaching content and academic inheritance, the government ensured the integrity and systematic nature of TCM education through a series of policies. For example, the "Opinions on Strengthening Clinical Teaching in Higher TCM Education" emphasized the importance of clinical teaching, ensuring that students gained sufficient clinical practice experience in addition to theoretical learning. This combination of theory and practice made TCM education more practical and capable of cultivating TCM talents that met societal needs. Moreover, the government focused on the academic inheritance of TCM education, ensuring the intergenerational transmission of traditional TCM knowledge and skills [30]. The "1988 - 2000 TCM Education Development Strategy Plan" set clear academic inheritance goals, using apprenticeship and other inheritance methods to ensure that the core knowledge of TCM was not diluted by the modernization process.

Systematization Stage (1999-2012)

Entering the 21st century, TCM education entered a new stage of systematization. During this stage, policies aimed at systematizing TCM education, integrating various aspects of education, including curriculum design, teaching methods, and teacher training. This comprehensive approach to policy development is consistent with the continuous improvement and refinement process within the policy cycle.

The policy themes during this period mainly focused on "continuing education in TCM," "reform of management systems and mechanisms," and "internationalization of TCM education." These themes reflected the government's intention to improve the overall level and international competitiveness of TCM education through the perfection of the educational system, deepening of management system reforms, and promotion of international cooperation [31]. "Continuing education in TCM" was a key focus of policies during this period. With the development of society and the changing medical demands, the government realized that traditional TCM education alone was insufficient to meet the complex needs of modern medicine. Therefore, continuing education became an important component of TCM education. For example, the

"TCM Continuing Education Fifteenth Five-Year Plan" and the "Regulations on TCM Continuing Education" clarified the goals and implementation methods of TCM continuing education, requiring the enhancement of TCM practitioners' knowledge and skills through continuing education to address new medical challenges. "Reform of management systems and mechanisms" was another important theme in the development of TCM education during this stage [32]. As educational system reforms deepened, the management and operation mechanisms of TCM education also underwent significant adjustments. For example, the "Opinions on Several Issues Concerning TCM Education" and the "TCM Innovation Development Plan Outline (2006 - 2020)" proposed the direction of management system reform in TCM education, aiming to improve the efficiency and quality of TCM education through educational system and mechanism reforms. "Internationalization of TCM education" was also an important policy orientation during this stage. With the enhancement of China's international status and the growing global influence of TCM, the government began to actively promote the internationalization of TCM education [33]. Policy documents such as the "Several Opinions of the State Council on Supporting and Promoting the Development of the TCM Cause" and the "Basic Requirements for the Establishment of TCM Undergraduate Education in Colleges and Universities (Trial)" clarified the goals and strategies for the internationalization of TCM education, encouraging TCM colleges and universities to cooperate with international academic institutions to promote the globalization of TCM education and research.

Standardization Stage (2012 to Present)

In recent years, TCM education policies have focused on standardization, aiming to establish unified standards for TCM education at home and abroad. This stage reflects the efforts to enhance the international competitiveness of TCM education and ensure its quality and consistency, in line with the ongoing evaluation and potential continuation phases of the policy cycle.

Since 2012, TCM education in China has entered a new stage of standardization. The policy themes during this period mainly focused on "standardization," "integration of medical and educational collaboration," "cultivation of general practitioners," and "high-quality development." These themes reflected the government's intention to ensure the high quality and level of TCM education through standardization, further enhancing its international competitiveness. "Standardization" has been a core theme of TCM education policies during this stage. To achieve the normalization and scientification of TCM education, the government has issued a series of policy documents specifying the standardization requirements for TCM education [34]. For example, the "Law of the People's Republic of China on Traditional Chinese Medicine" and the "Several Opinions on Further Promoting the '5+3' Integrated Medical Talent Training Work" detailed the teaching standards, curriculum settings, and talent training models for TCM education. "Integration of medical and educational collaboration" has been an important development model for TCM education during this stage. The government realized that further development of TCM education required a closer combination with medical practice. Therefore, the "Guiding Opinions on Deepening the Reform and Development of TCM Education through Medical and Educational Collaboration" was issued, specifying the requirements for medical and educational collaboration and emphasizing the improvement of TCM students' practical abilities and comprehensive quality through the integration of teaching and clinical practice. "Cultivation of general practitioners" has been another important theme in TCM education policies during this stage. With the increasing demand for general practitioners in society, the government began to introduce a general practitioner training model into TCM education. For example, the "Implementation Measures for the Standardized Training of TCM Resident Physicians (Trial)" and the "Guiding Opinions on Deepening the Succession Education of TCM" proposed a combination of standardized training and master-apprentice education to cultivate general TCM practitioners with comprehensive capabilities. "High-quality development" has been the overall goal of TCM education during this stage. Against the backdrop of the rapid development of the global health sector, the government required TCM education to focus on quality and promote high-level, high-standard development [35]. For example, the "Notice on the Implementation of the 'Double Ten Thousand Plan' for First-Class Undergraduate Major Construction" and the "Several Policy Measures for Accelerating the Characteristic Development of TCM" emphasized the improvement of teaching quality, optimization of curriculum settings, and strengthening of scientific research innovation to achieve high-quality development in TCM education.

Logic Behind the Evolution of TCM Education Policies

Logic of Balancing Modernization and Traditional Inheritance

The evolution of TCM education policies in China centers on the balance between modernization and traditional inheritance [36]. Since the mid-19th century, the invasion of Western powers and the introduction of modern science had a profound impact on traditional Chinese culture. In an effort to adapt to the new external environment, policymakers introduced the Western medical education system and attempted to abolish traditional TCM education to promote the modernization of Chinese medicine. This reflected the then-prevailing belief that modernization was the only path to national strength, and traditional TCM education was seen as an obstacle to this process. However, this logic of excluding tradition did not fully succeed. Despite the policy attempts to suppress TCM education, TCM, as an important part of Chinese culture, did not disappear due to policy pressure. Many TCM practitioners continued to preserve and spread TCM knowledge through private tutoring and apprenticeship, laying the foundation for the survival of TCM [37]. After the founding of the People's Republic of China in 1949, the state reevaluated the value of TCM and began to restore and standardize TCM education. For example, the "Notice on the Regulations for the Organization of TCM Advanced Study Schools and Classes" was a typical example, reflecting the state's emphasis on TCM education and aiming to restore the legitimate status of TCM in the medical system through the reconstruction of the educational system. This logic reflects the state's gradual recognition in the modernization

process that Western medicine alone could not meet China's needs, and a balance between modernization and traditional inheritance had to be found [38].

Logic Driven by Political and Social Demands

The evolution of TCM education policies in China has been deeply influenced by both political environment and social demands, with each stage of policy change reflecting the political background and social needs of the time. After the founding of New China, the government rerecognized the importance of TCM as an important part of Chinese traditional culture, and policies such as the "Instruction on Carrying Out Master-Apprentice TCM Education" were issued in the 1950s to restore the legitimate status of TCM education and address the challenge of medical resource shortages. During the reform and opening-up period, with rapid economic development and changes in social structure, TCM education policies began to emphasize specialized development to meet the diverse medical service needs. The "1988 - 2000 TCM Education Development Strategy Plan" was a representative policy of this period, proposing to extend the academic system, enrich the curriculum content, and strengthen clinical practice to cultivate high-quality talents with TCM characteristics. Entering the 21st century, with the enhancement of China's international status and the acceleration of globalization, the international demand for TCM increased, and the government integrated international considerations into educational policies. For example, the "Several Opinions of the State Council on Supporting and Promoting the Development of the TCM Cause" explicitly proposed to promote the internationalization of TCM education, encouraging TCM colleges and universities to cooperate with international academic institutions to enhance its global influence. These policies not only responded to domestic and international demands for TCM but also reflected the strategic goal of enhancing China's cultural soft power through the internationalization of TCM education.

Logic of Adaptation to Reform and Innovation

Reform and innovation are key logics in the evolution of TCM education policies in China. With social development and technological progress, TCM education faces new challenges and opportunities, and policymakers have gradually placed reform and innovation at the core of policymaking. Since the reform and opening-up, Chinese society has undergone tremendous changes, and TCM education policies have also entered a new stage oriented by innovation. In the mid-1980s, the government realized that merely restoring and standardizing TCM education was insufficient to meet modern demands, and it was necessary to enhance the quality and international competitiveness of education through reform and innovation. For example, the "Report on Seriously Implementing the Party's TCM Policy and Solving the Problem of Succession in the TCM Workforce" emphasized educational model innovation through academic system reform, curriculum enrichment, and teaching method improvement. The introduction of the 7-year program was a specific manifestation of this innovative logic [39], and the "Notice on the Trial Implementation of Seven-Year Higher TCM Education" marked a milestone in the modernization of TCM education. With the upgrading of medical service demands, standardization became a focus of reform in the early 21st century. The government implemented standardization policies such as the "Law of the People's Republic of China on Traditional Chinese Medicine" and the "Several Policy Measures for Accelerating the Characteristic Development of TCM" to ensure the consistency and global competitiveness of educational quality. This standardization not only guaranteed high-quality output but also made TCM education more acceptable and recognizable by the international community.

Logic of Integration of Cultural Confidence and National Strategy

The deep logic of TCM education policies includes the integration of cultural confidence and national strategy. As an important part of Chinese traditional culture, TCM education policies are not only related to medical education but also bear the mission of cultural inheritance and national strategy. In the early days of New China, the restoration of cultural confidence was a key driving force for the revival of TCM education policies. With the enhancement of China's comprehensive national power, the government has gradually strengthened its emphasis on TCM education, consolidating its important position in national culture and social development through institutional means. For example, the "Urgent Notice on Inheriting the Academic Experience of Senior TCM Practitioners" emphasized the protection and inheritance of the valuable experience of senior TCM practitioners, laying the foundation for the long-term development of TCM. At the national strategy level, the evolution of TCM education policies has always been closely linked to the overall national development strategy. After the reform and opening-up, with the enhancement of China's economy and international influence, the government has paid more attention to the globalization and standardization of TCM education. This not only reflects cultural confidence but also demonstrates the important role of TCM education in showcasing China's soft power [40]. The integration of cultural confidence and national strategy runs through all stages of TCM education policy evolution, promoting the modernization and internationalization of TCM education and enhancing the global influence of Chinese culture.

Policy Effectiveness in TCM Education

Assessing policy effectiveness is crucial as it reveals the practical impact of policy initiatives and provides insights for future improvements, aligning with the tenets of policy evaluation research. During the Standardization Stage (2012 to Present), TCM education policies aimed to establish unified standards both domestically and internationally. A notable case is the implementation of the "Undergraduate Medical Education Standards-Traditional Chinese Medicine Major (Provisional) (2012)" in China. This policy stipulated specific requirements for curriculum design, teaching resources, and faculty qualifications, leading to a significant improvement in the quality of TCM education. For instance, Henan University of Chinese Medicine achieved a 99.33% course evaluation coverage rate in 2018 - 2019, while Zhejiang University of Chinese Medicine reported an 89.22% national licensing

examination pass rate in 2020, 24.6 percentage points above the national average [41].

In the Specialization Stage (1978 - 1999), policies promoting the development of specialized TCM education programs also demonstrated remarkable effectiveness. For example, the establishment of TCM clinical specialization courses in major TCM universities across China was a direct result of policy support. These courses focused on in-depth training in areas such as acupuncture, Chinese herbal medicine, and TCM diagnosis. Take Beijing University of Chinese Medicine as a case. Beijing University of Chinese Medicine enrolled its first group of 36 master's students in 1978, and clinical program enrollment increased by 40% from 1985 to 1995 [42]. Moreover, 65.17% of Chengdu University of Traditional Chinese Medicine graduates serve at the grassroots level, with more than 80% relevance to their major [43]. This case study illustrates how specialization-oriented policies successfully cultivated high-quality TCM professionals, meeting the practical needs of the industry.

However, policy effectiveness was not uniformly positive across all stages. During the Marginalization Stage (1840 - 1949), due to social unrest and the influence of Western medicine, TCM education policies struggled to achieve their intended goals. The lack of stable political support and limited resources meant that TCM education institutions faced difficulties in maintaining educational quality and scale. For instance, many traditional TCM private schools were forced to close, and the number of TCM students declined sharply during this period [44]. This historical case highlights the importance of a conducive social and political environment for policy implementation and effectiveness.

In conclusion, by analyzing these case studies, it is evident that TCM education policies have achieved varying degrees of success. Effective policies often require a combination of clear goals, adequate resource allocation, and a supportive social context. Insights from these case studies can inform future policy making, helping to optimize TCM education development and better meet the needs of society and the medical field.

Future Trends of TCM Education Policies

First, the deepening of standardization and internationalization. In the future, TCM education needs to further promote standardization, especially in the context of globalization, where alignment with international standards will become crucial. With China's increasing influence in the international community, TCM, as a representative of Chinese culture, will face more international cooperation demands in its educational system. Accelerating the formulation and promotion of TCM education standards and incorporating TCM courses into the international medical education system will be key to promoting its global dissemination, attracting more international students and researchers, and enhancing the international influence of TCM, thereby promoting the diversification and innovation of education.

Second, the strengthening of innovation and integration with technology. TCM education needs to place greater emphasis on the integration with modern technology, with information technology and biotechnology becoming important drivers. The development of artificial intelligence, big data, blockchain, and other technologies will greatly improve the efficiency and quality of TCM education. Digital platforms and web-based education tools may be more widely applied in TCM education, providing convenient ways to access knowledge and virtual laboratories to enhance students' learning outcomes and practical abilities.

Third, the balanced development of cultural confidence and globalization. With the enhancement of China's international status, TCM education needs to maintain cultural confidence in the context of globalization. This confidence should be reflected not only in the inheritance and innovation of TCM educational content but also in how to play a greater role in global health governance. Future policies will place greater emphasis on the cultural export of TCM, showcasing its unique value through international exchanges, academic cooperation, and cultural activities to enhance the international image of TCM and China's discourse power in international medical education.

Fourth, the expansion of policy support and diversified cooperation. Future TCM education policies should place greater emphasis on policy support and diversified cooperation. Through measures such as financial investment, legal protection, and talent introduction, support should be provided for TCM colleges and universities in their exploration of scientific innovation, international cooperation, and digital transformation. At the same time, cooperation with international organizations, multinational corporations, and nongovernmental organizations should be strengthened to promote the global allocation and sharing of TCM educational resources. This diversified cooperation will attract more global resources for TCM education, promoting its continuous innovation and development.

Conclusions and Outlook

The evolution of TCM education policies is not only the result of historical development and social demands but also a comprehensive reflection of cultural inheritance, innovative reform, and international strategy in China's modernization process. From the early marginalization to the normalization after the founding of New China, from the specialized development during the reform and opening-up period to the standardization and internationalization in the 21st century, each adjustment and evolution of TCM education policies reflects China's rethinking and redefinition of the positioning of TCM education in different historical stages. The core logic of this policy evolution process not only provides an effective framework for understanding the past of TCM education but also offers important clues for predicting its future development direction. In the future, the formulation of TCM education policies should maintain the traditional advantages of TCM while actively addressing new challenges brought by modern technology and globalization, promoting the continuous optimization and improvement of the TCM education system. It is necessary not only to meet the domestic demand for medical services but also to inherit and promote Chinese culture globally, becoming an important bridge connecting Eastern and Western

medicine. Through continuous policy innovation and improvement, TCM education will contribute more to the development of health care in China and around the world.

Acknowledgments

The authors would like to thank their supervisor, Fan Yang, PhD, for the guidance and support throughout this research. They also appreciate the support from the Youth Project of Education under the National Social Science. This study was supported by the Youth Project of Education under the National Social Science Fund, China (project number CGA220304).

Data Availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

TY contributed to writing—original draft, methodology, data curation, and software. FY contributed to writing—original draft, data curation, formal analysis, and project management. YL contributed to writing—review and editing, supervision, conceptualization, formal analysis, and resources.

Conflicts of Interest

None declared.

References

1. Yao L, Jin Z, Mao C, Zhang Y, Luo Y. Traditional Chinese medicine clinical records classification with BERT and domain specific corpora. J Am Med Inform Assoc 2019 Dec 1;26(12):1632-1636. [doi: 10.1093/jamia/ocz164]
2. Zhou X, Peng Y, Liu B. Text mining for traditional Chinese medical knowledge discovery: a survey. J Biomed Inform 2010 Aug;43(4):650-660. [doi: 10.1016/j.jbi.2010.01.002]
3. Vig J, Belinkov Y. Analyzing the structure of attention in a transformer language model. arXiv. Preprint posted online on Jun 7, 2019. [doi: 10.48550/arXiv.1906.04284]
4. Alsentzer E, Murphy JR, Boag W, et al. Publicly available clinical BERT embeddings. arXiv. Preprint posted online on Apr 6, 2019. [doi: 10.48550/arXiv.1904.03323]
5. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv. Preprint posted online on Mar 11, 2022. [doi: 10.48550/arXiv.2203.05794]
6. Wolf T, Sanh V, Chaumond J, et al. Hugging face's transformers: state-of-the-art natural language processing. arXiv. Preprint posted online on Oct 9, 2019. [doi: 10.48550/arXiv.1910.0377]
7. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. Presented at: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2019 p. 119-124. [doi: 10.18655/naacl-hlt.2019.119]
8. Griffiths TL, Steyvers M. Finding scientific topics. Proc Natl Acad Sci USA 2004 Apr 6;101(suppl_1):5228-5235. [doi: 10.1073/pnas.0307752101]
9. Wu G, Ning H, Yuan Y, et al. Topic identification and content analysis of internet medical policies under the background of Healthy China 2030. Health Res Policy Sys 2024;22(1):132. [doi: 10.1186/s12961-024-01226-3]
10. El-Gayar O, Al-Ramahi MA, Wahbeh A, Nasralah T, El Noshokaty A. A comparative analysis of the interpretability of LDA and LLM for topic modeling: the case of healthcare apps. Presented at: AMCIS 2024 Proceedings 22; 2024 URL: https://aisel.aisnet.org/amcis2024/health_it/health_it/22 [accessed 2025-09-10]
11. ChinaXiv. Study on the standardization of terms in TCM diagnosis and treatment knowledge base [article in chinese]. ChinaXiv. Preprint posted online on 2023. [doi: 10.12074/202304.00735V1]
12. Bursztein E, Zhang M, Vallis O, Jia X, Kurakin A. RETVec: resilient and efficient text vectorizer. arXiv. Preprint posted online on Apr 23, 2023. [doi: 10.48550/arXiv.2302.09207]
13. Singh K, Devi SD, Devi HM, Mahanta AK. A novel approach for dimension reduction using word embedding: an enhanced text classification approach. International Journal of Information Management Data Insights 2022 Apr;2(1):100061. [doi: 10.1016/j.jjimei.2022.100061]
14. Campello R, Moulavi D, Sander J. Density-based clustering based on hierarchical density estimates. Presented at: Pacific-Asia Conference on Knowledge Discovery and Data Mining; 2013; Springer p. 160-172. [doi: 10.1007/978-3-642-37456-2_14]
15. McInnes L, Healy J, Astels S. HDBSCAN: Hierarchical density based clustering. JOSS 2017;2(11):205. [doi: 10.21105/joss.00205]
16. Guillén-Pacho I, Badenes-Olmedo C, Corcho O. Dynamic topic modelling for exploring the scientific literature on coronavirus: an unsupervised labelling technique. Int J Data Sci Anal 2024. [doi: 10.1007/s41060-024-00610-0]

17. Sabatier PA. Theories of the Policy Process, 1st edition: Westview Press; 1999. URL: https://ruby.fgcu.edu/courses/twimberley/EVR2861/theorypolprocess.pdf [accessed 2025-09-10]
18. Kingdon JW. Agendas, Alternatives, and Public Policies: Little, Brown; 1984. [doi: 10.1017/S0143814X00003068]
19. Lin HX, Zhao J, Liu B. The policy process of establishing professional degrees in acupuncture and moxibustion disciplines in china: an analysis based on the multiple streams theory. : University and Discipline; 2024 URL: https://kns.cnki.net/KCMS/detail/detail.aspx?DBName=cjfqtotal&dbcode=cjfq&filename=DXXK202401011 [accessed 2025-09-16]
20. Bu D. Common characteristics and developmental insights of Chinese medical education policies in the 21st century. Med Educ Res Pract 2023;31(4):397-401. [doi: 10.13555/j.cnki.c.m.e.2023.04.001] [Medline: 36945697]
21. Sun S, Zhu W. Research on the policy process of collaborative medical education reform based on the multiple streams framework. China High Med Educ 2019;1:17-18. [doi: 10.3969/J.ISSN.1002-1701.2019.01.009]
22. Li Y, Huang Z, Luan Z, et al. Efficient evidence selection for systematic reviews in traditional Chinese medicine. BMC Med Res Methodol 2025 Jan 15;25(1):10. [doi: 10.1186/s12874-024-02430-z] [Medline: 39815209]
23. Xiang W. A review of the “Regulations for Traditional Chinese Medicine” of the Nanjing Nationalist Government. Republican Archive 2004;04:82-86 [FREE Full text]
24. Jiao Y, Ling T. A brief history of traditional Chinese medicine education in Zhejiang in modern times. J Zhejiang Chin Med Univ 2018;42(9):752-755. [doi: 10.16466/j.issn1005-5509.2018.09.019]
25. Zheng G. Historical review of new china’s policies on traditional chinese medicine. Doctoral dissertation.: Party School of the CPC Central Committee; 2011 URL: https://kns.cnki.net/kcms2/article/[illegible] [accessed 2025-09-10]
26. Qiaqiao G. A study on the traditional Chinese medicine apprenticeship movement from 1956 to 1966. Master’s thesis.: Guangzhou University of Chinese Medicine; 2012 URL: https://kns.cnki.net/kcms2/article/[illegible] [accessed 2025-09-02]
27. Hong J. Actively inheriting the academic experience of senior traditional Chinese medicine practitioners. J Tradit Chin Med 1958;02:74-75. [doi: 10.13288/j.11-2166/r.1958.02.002]
28. Zhou ZY. Strengthening the study of traditional Chinese medicine culture and maintaining the advantages and characteristics of TCM. J Nanjing Univ Tradit Chin Med 2007(2):63 [FREE Full text]
29. Tian JF. A pilot study on the seven-year system to improve the structure of higher traditional Chinese medicine education. Chin Med Educ 1991(1):6 [FREE Full text]
30. Ning ZP, Cai TR, Bu XC, Guo ZH, Xiao WM, Liu J. Legislative research on the academic inheritance system of traditional Chinese medicine. Hunan J Tradit Chin Med 2017;33(1):1-6. [doi: 10.16808/j.cnki.issn1003-7705.2017.01.001]
31. Beijing University of Chinese Medicine. The “Tenth Five-Year” plan for continuing education in traditional Chinese medicine. Chin Med Educ 2002(5):1-2 [FREE Full text]
32. Li LL, Jiang FA. Discussion on the transformation of management in traditional Chinese medicine higher education institutions from “Control” to “Management”. J Tradit Chin Med Manag 2011;19(1):41-42. [doi: 10.16690/j.cnki.1007-9203.2011.01.048]
33. Dong W, Zheng L, Xiao Z, Zhou D. A discussion on the current status and development strategies of traditional Chinese medicine education internationalization. J Nanjing Univ Tradit Chin Med 2012;13(1):61-64 [FREE Full text]
34. Li JW, Zhang YZ. A study on the reform of the “5+3” integrated excellent traditional Chinese medicine talent training model under the background of medical-education collaboration. China High Med Educ 2016(2):9-10 [FREE Full text]
35. Wen LJ, Song ZJ, Peng L, Tang K, Zhou YF, Guo J. A discussion on the standardized training model for traditional Chinese medicine general practitioners. Shizhen J Tradit Chin MedPharm 2018;29(3):723-725 [FREE Full text]
36. Wang YY, Tian JZ. Traditional Chinese medicine inheritance and innovation in the new situation. J Beijing Univ Tradit Chin Med 2018;41(7):533-536 [FREE Full text]
37. Feng Z, Huang YR. Innovation in traditional Chinese medicine education and the mentor-apprenticeship model. J Tradit Chin Med 2007(1):90-91. [doi: 10.13288/j.11-2166/r.2007.01.056]
38. Sun D, Guo LM, Tai DM. The development of traditional Chinese medicine international education in the new era: logic, challenges, and pathways. Med Educ Res Pract 2021;29(1):1-3. [doi: 10.13555/j.cnki.c.m.e.2021.01.001]
39. Jiang J, Lin X, Gao YQ, et al. A review and reflection on the reform of practical teaching in the seven-year program of traditional Chinese medicine. J Tradit Chin Med Educ 2007(6):46-49 [FREE Full text]
40. Liu XX, Zhang HL. The influence factors of traditional Chinese medicine cultural soft power under the perspective of “Cultural Power”. J Tradit Chin Med 2020;61(9):762-765. [doi: 10.13288/j.11-2166/r.2020.09.007]
41. Zhejiang Chinese Medical University. 2020-2021 academic year undergraduate teaching quality report [EB/OL]. 2021 URL: https://xxgk.zcmu.edu.cn/2020-2021.pdf [accessed 2025-09-02]
42. Yizhen Z, Jiayi A, Xi T, et al. Research on the characteristics of policy options and optimization paths for postgraduate education in traditional Chinese medicine. Tradit Chin Med Educ 2025;44:84-92. [doi: 10.3969/j.issn.1003-305X.2025.01.238]

43. Chengdu University of Traditional Chinese Medicine. 2023 graduate employment quality annual report [EB/OL]. 2024 URL: http://wg85.jswtbj.com/Upload/main/ContentManage/Article/File/2024/08/22/202408221719216578.pdf [accessed 2025-09-02]
44. Xiaofei Z. Contemporary value of the Beijing Chinese medicine lecture hall relics. Beijing J Tradit Chin Med 2024;43:1221-1226. [doi: 10.16025/j.1674-1307.2024.11.001]

Abbreviations

c-TF-IDF: class-based Term Frequency–Inverse Document Frequency
DBSCAN: Density-Based Spatial Clustering of Applications with Noise
DTM: dynamic topic modeling
HDBSCAN: Hierarchical Density-Based Spatial Clustering of Applications with Noise
LDA: Latent Dirichlet Allocation
TCM: Traditional Chinese medicine
UMAP: Uniform Manifold Approximation and Projection

Edited by J Gentges; submitted 14.02.25; peer-reviewed by D Wang, Z Lin; revised version received 26.05.25; accepted 27.05.25; published 25.09.25.

Please cite as:
Yang T, Yang F, Li Y
Mapping the Evolution of China's Traditional Chinese Medicine Education Policies: Insights From a BERTopic-Based Descriptive Study
JMIR Med Educ 2025;11:e72660
URL: https://mededu.jmir.org/2025/1/e72660
doi:10.2196/72660

Mono-Professional Simulation-Based Obstetric Training in a Low-Resource Setting: Stepped-Wedge Cluster Randomized Trial

Anne A C van Tetering[1,2,3], PhD, MD; Ella L de Vries[1,3], MSc, MD; Peter Ntuyo[4], MD; E R van den Heuvel[5], Prof Dr; Annemarie F Fransen[1,3], PhD, MD; M Beatrijs van der Hout-van der Jagt[1,3,6], MSc, PhD; Imelda Namagembe[4], PhD, MD; Josaphat Byamugisha[7], Prof Dr, MD; S Guid Oei[1,3], Prof Dr, MD

[1]Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands

[2]Department of Obstetrics and Gynaecology, Amphia Ziekenhuis, Breda, The Netherlands

[3]Department of Obstetrics and Gynaecology, Máxima Medical Center, De Run 4600, Veldhoven, The Netherlands

[4]Department of Maternal Fetal Medicine, Mulago Specialised Women and Neonatal Hospital, Kampala, Uganda

[5]Dean of Mathematics & Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands

[6]Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands

[7]Department of Obstetrics and Gynaecology, School of Medicine, Makerere University College of Health Sciences, Kampala, Uganda

Corresponding Author:
Anne A C van Tetering, PhD, MD
Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands

Abstract

Background: Emergency obstetric simulation-based training has increasingly been used to improve emergency obstetric care provision in sub-Saharan Africa. For determining the optimal methodology for effective training sessions in resource-constrained settings, it is crucial to conduct high-quality research.

Objective: We aim to investigate the impact of a train-the-trainer model for providing technology-enhanced, mono-professional, simulation-based training in obstetrics in a resource-constrained setting on maternal and perinatal outcomes.

Methods: A stepped-wedge cluster randomized trial was conducted from October 2014 until March 2016 at the medium- to high-risk ward at Mulago National Referral Hospital, Uganda, with an annual delivery rate of over 23,000. The intervention consisted of a train-the-trainer model in which training was cascaded down from master trainers to local facilitators (obstetric senior staff members) to learners (senior house officers). The training of senior house officers was provided to 7 fixed clusters by a computer-generated random sequential roll-out. The training comprised a 1-day (8 h), mono-professional, simulation-based training in obstetrics, and half-day repetition training sessions targeted at every 7 weeks. Both medical technical skills and teamwork skills were taught. The primary outcome comprised a combined maternal and perinatal mortality rate. Secondary outcomes comprised the maternal mortality rate, the perinatal mortality rate, the percentage of births by vacuum extraction and cesarean section, and the Weighted Adverse Outcome Score.

Results: Overall, there were 17,496 births. The combined mortality rate was 9.05% (95% CI 8.37% - 9.77%) in the intervention group, and 8.73% (95% CI 8.21% - 9.28%) in the control group (odds ratio [OR] 0.98, 95% CI 0.86 - 1.12; *P*=.81). No statistically significant change was found in the maternal mortality rate (OR 0.80, 95% CI 0.27 - 2.32; *P*=.68) or the perinatal mortality rate (OR 0.99, 95% CI 0.87 - 1.13; *P*=.87). This study did not identify any difference in the percentage of vacuum extractions, the percentage of cesarean sections, or Weighted Adverse Outcome Scores.

Conclusions: This train-the-trainer model for providing technology-enhanced, mono-professional, simulation-based training in obstetrics was not able to change maternal and perinatal mortality outcomes. This study, in combination with literature, suggests that future research should consider multiprofessional team training in obstetrics involving all staff within their units.

Trial Registration: ISRCTN Registry ISRCTN98617255; https://www.isrctn.com/ISRCTN98617255

International Registered Report Identifier (IRRID): RR2-10.2196/17277

(JMIR Med Educ 2025;11:e54911) doi:10.2196/54911

KEYWORDS

obstetric; simulation; training; low income; middle income; LMIC; simulation training; medical education; mono-professional; obstetric training; low-resource setting; low-income setting; stepped-wedge; RCT; obstetric care; pregnancy care; Africa; maternal; perinatal; outcome; hospital; train-the-trainer; facilitator; trainer; learner; mortality rate; randomized controlled trial

Introduction

Emergency Obstetric Care in Uganda

Uganda continues to face challenges in providing safe obstetric care. Despite an increase in the rate of institutional births from 59% to 74%, the maternal mortality ratio was still high at 375 per 100,000 live births in 2017, accompanied by a neonatal mortality rate of 21 deaths per 1000 live births [1]. Key barriers for providing safe childbirth include shortcomings in the management of emergency obstetric care, delays in referral practices, and insufficient coordination among health care staff, all of which obstruct the provision of adequate emergency obstetric care [2].

Simulation-Based Obstetric Training

To address these challenges, simulation-based training for emergency obstetric care has evolved as a promising approach in sub-Saharan Africa. Growing evidence suggests that this type of training improves health care providers' knowledge and skills, while also leading to positive changes in their behavior [3-5]. Additionally, evidence from other studies has shown encouraging effects on patient outcomes, including reported reductions in neonatal and perinatal mortality rates, as well as potential decreases in maternal mortality and postpartum hemorrhage [6-9]. Despite these promising findings, assessments of patient outcomes remain infrequent, and the results are often inconsistent [3,4].

Evaluating Simulation-Based Training

One limitation of current evaluations is the reliance on 1-group pretest-posttest designs, which often fail to control for external variables that may influence the results. Furthermore, significant variability exists in training length, content, and design, with programs ranging from mono-professional to multi-professional approaches. This variation makes it difficult to identify which components most effectively contribute to the success of the training. Additional challenges, such as resource constraints, difficulties in sustaining training programs, staff shortages, and high turnover rates, further hinder the implementation and long-term impact of simulation-based training in sub-Saharan Africa. To overcome these challenges, high-quality research is essential to determine the most effective methodologies for emergency obstetric simulation-based training in sub-Saharan Africa.

This study aimed to evaluate the effect of a train-the-trainer program designed to provide technology-enhanced, mono-professional, simulation-based obstetric training on patient outcomes in Uganda [10].

Methods

Setting

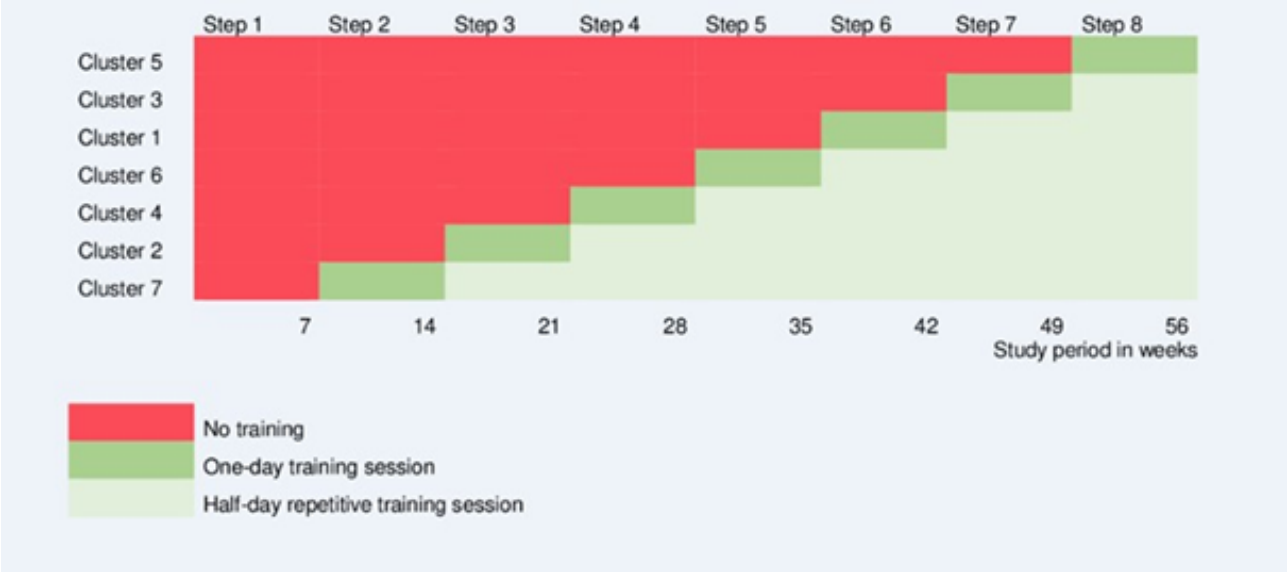
A stepped-wedge cluster randomized trial was conducted from October 2014 until March 2016 at the medium- to high-risk labor ward at Mulago National Referral Hospital in Uganda. This hospital also functions as the main teaching facility for Makerere University College of Medicine and Health Sciences. During this study's period, over 23,000 women gave birth annually at the medium to high-risk labor ward.

Design and Recruitment

The stepped-wedge cluster randomized trial design facilitated the phased implementation of the training program, with different clusters receiving the intervention at different periods to assess its impact on patient outcomes. This approach allowed for the measurement of the intervention's effect both within and between clusters. Additionally, it enabled the intervention to be provided as a standard service to all participants, while being implemented in stages [11]. As training of all obstetric ward staff was not feasible due to financial and logistical challenges, senior house officers (SHOs) were chosen as the target group for the training program due to their coordinating role in providing emergency obstetric care.

In October 2014, a total of 7 fixed clusters of SHOs were recruited to receive the training. All participants provided written informed consent before this study began. A computer-generated random sequential roll-out of the training program was conducted to determine the order in which the different clusters would receive the intervention (Figure 1). Examination and holiday periods were excluded from the schedule, as fixed clusters could not be maintained in the SHOs' work schedules during these times.

Figure 1. Stepped-wedge cluster randomized design with 7 clusters and 8 steps.



Train-the-Trainer Model

The training program was conducted using a train-the-trainer model, in which training was cascaded down from master trainers to local facilitators, and then to the learners, who were the SHOs. In this model, the master trainers, who were obstetricians from a high-resource setting, had been previously certified as simulation-based trainers by institutions such as EuSim or the Center for Medical Simulation. These master trainers provided a 4-day train-the-trainer program to 14 local facilitators. The facilitators, all gynecologists, were selected based on their clinical and teaching experience by the head of the department. The program included both lectures and practical teaching sessions using simulation-based obstetric scenarios. The train-the-trainer course concluded with an assessment session. During this session, the local facilitators trained intern doctors using a draft SHO training program. Afterward, the master trainers provided feedback to the facilitators. Based on this, 12 of the facilitators were certified as simulation trainers. The SHO training program was then adjusted based on feedback from both trainers and trainees. Subsequently, the local facilitators delivered the training to fixed clusters of 6 to 9 SHOs, comprising first-, second-, and third-year SHOs. A 1-day annual refresher training was offered to all local facilitators. The local facilitators were compensated for lost clinical income by being paid for their participation in the training sessions.

Course Content

Course content was developed by Medsim, a medical simulation center in the Netherlands, in cooperation with senior staff members of Mulago National Referral Hospital. The SHO training program included a 1-day (8 h) mono-professional, simulation-based sessions, followed by half-day refresher sessions every 7 weeks. These refresher sessions started after the switch from the control to the intervention group. Each training session was provided by 2 local facilitators. Scenarios were based on the main local causes of maternal and perinatal mortality and tailored to local clinical protocols and availability of medical equipment. This led to the creation of 2 different scenarios for postpartum hemorrhage, a scenario for eclampsia, a scenario involving fetal distress with a ventouse delivery, and a breech delivery scenario. Both medical-technical and teamwork skills were included in the training, with the difficulty level increasing throughout the day. Every SHO participated in at least 2 scenarios during the 1-day training, while having an observer role in the nonparticipating scenarios. During the repetition training sessions, a single clinical scenario was executed and repeated until skills were mastered.

Data Collection and Outcomes

The primary outcome of this study was the combined maternal and perinatal mortality rate, expressed as a percentage of maternal and perinatal deaths per total number of births. Perinatal deaths were defined as stillbirths and deaths occurring within the first week of life in the special care unit. Data about each delivery and maternal and perinatal outcomes were prospectively registered using the maternity register and transcribed without identification of the subjects. Data about maternal deaths in the high dependency unit, and neonatal deaths in the special care unit were obtained from registration books in these units. These data were matched to and merged with data from the maternity register of the medium to high-risk ward into 1 final electronic database.

Secondary outcomes comprised the maternal mortality rate (maternal deaths per 100,000 births), the perinatal mortality rate (perinatal deaths per 1000 births), percentage of births by vacuum extraction, percentage of births by cesarean section, and the Weighted Adverse Outcome Score (WAOS). The WAOS was defined as the total weighted score of each adverse outcome divided by the total number of births [12]. Four out of 10 index measures (maternal death [750 points], intrapartum or perinatal death [400 points], uterine rupture [100 points], Apgar score less than 7 after 5 minutes [25 points]) were available for registration and assessment. Finally, the maternal mortality ratio (maternal mortality per 100,000 live births), and the perinatal mortality ratio (perinatal mortality per 1000 live births) were calculated for the control and intervention group. As data were analyzed on the cluster level, the authors could not identify individual participants' results.

To provide a comprehensive understanding of the training's effectiveness, additional secondary outcomes were included and published separately, such as the evaluation of the instructional design, participants' reactions (corresponding to Kirkpatrick level 1), and the effects on knowledge, teamwork, and medical-technical skills (corresponding to Kirkpatrick level 2) [13].

Sample Size Calculation

The power calculation was conducted following the methods described by Hussey et al [14] and Woertman et al [10,15,16]. Initially, the sample size for a standard randomized clinical trial was calculated. To show a 20% reduction in combined maternal and perinatal mortality with an α of .05 and a power of 80%, a total of 6398 births would be required for a simple randomized clinical trial design. The design effect was then determined, assuming an intracluster correlation of 0.05, a cluster size of 3343 births per year, and 7 clusters in total. Accounting for the design effect, 2367 births per measurement period would be required. To meet this target, each measurement period would need to last at least 5 weeks. However, for logistical feasibility, the duration of each step was set at 7 weeks, resulting in a total study duration of 56 weeks. Statistical significance was defined as a 2-sided *P* value of <.05.

Statistical Analysis

Patient baseline characteristics were summarized with medians and IQRs for continuous variables and with counts (percentages) for categorical variables. A generalized linear mixed-effects model was used for the estimation of an intervention effect. Here, the outcome is the binary event on the individual (whether a birth was or was not complicated by maternal mortality, perinatal mortality, or both), and a logit link function was used to model the probability of the event. In the logit scale, the cluster indicator served as a random effect on the intercept, and the period of the stepped wedge was treated as a fixed effect, as well as the intervention effect. The effect of the intervention was reported as an odds ratio, and the performance per treatment group was reported as a percentage or as an incidence rate ratio with 95% CI.

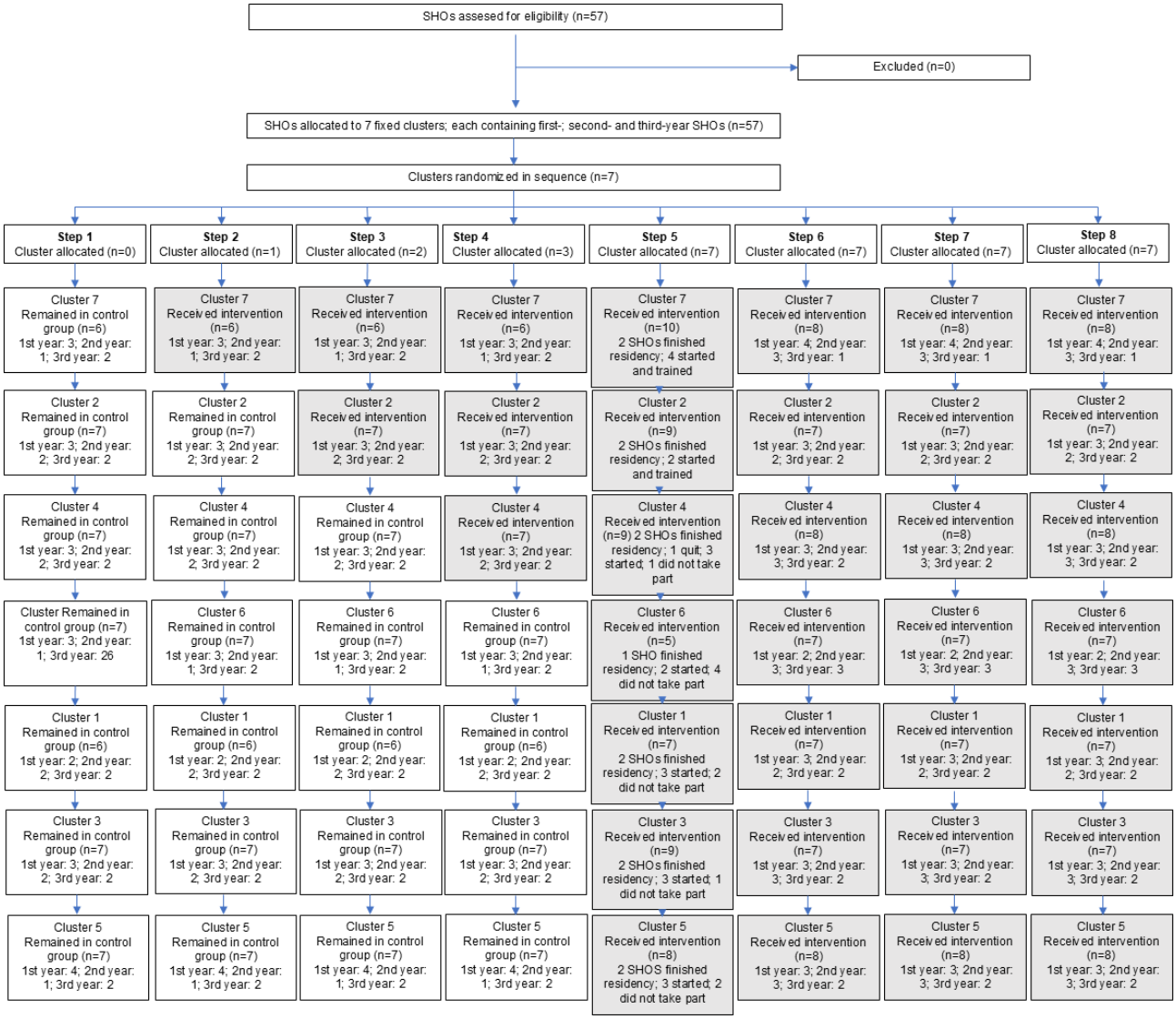
Ethical Considerations

Ethical approval was obtained from the Mulago Research and Ethics Committee (Protocol MREC: 674), and the Uganda National Council for Science and Technology (UNCST, SS 3927). Written informed consent to participants in this study was obtained at the beginning of the first training day, and all participants had the ability to opt out during this study's period. No compensation was provided for the participants. All data were anonymized.

Results

From October 2014 until March 2016, a total of 57 SHOs were randomized into 7 clusters. The first 3 included clusters received the main training on schedule according to protocol. Afterward, when SHOs heard about the training experience from their peers, they requested to expedite the training schedule, rather than waiting for the allocated time slot. Therefore, these 4 remaining clusters were trained within the same week and they simultaneously switched from the control group to the intervention group during this week (Figure 2). After the deviation from the stepped-wedge design in the timing of the intervention, no amendment was needed by the ethical committee, as the intervention itself had not changed.

Figure 2. Flowchart of stepped-wedge cluster randomized design. SHO: senior house officer.



Baseline characteristics are shown in Table 1. Overall, there were 17,496 births. There were fewer female and more male neonates in the intervention group compared to the control group. The results on maternal and perinatal mortality rates, mode of delivery, and the WAOS are shown in Table 2. No differences were found between the intervention and the control group in the combined maternal and perinatal mortality rate, the maternal mortality rate, and the perinatal mortality rate. Results on the percentage of vacuum extractions, the percentage of cesarean sections, and the WAOS did not show any difference between the intervention and the control group. The maternal mortality ratio in the control and intervention group was, respectively, 160.4 (95% CI 94.9 - 266.6) and 116.9 (95% CI 51.2 - 252.4). The perinatal mortality ratio in the control and intervention group was 94.4 (95% CI 88.8 - 100.4) and 98.7 (95% CI 91.3 - 106.6).

Table . Baseline characteristics.

Characteristics		Total	Control group	Intervention group	*P* value
Maternal characteristics					
Age (years), median (IQR)		24 (21-28)	24 (21-28)	24 (21-29)	.11[a]
Parity, n					.71[b]
	Primiparous	6760	4219	2541	
	Multiparous	9594	5960	3634	
Gestation, n (%)					.18[b]
	Preterm (<37 weeks)	1183 (8.2)	753 (8.5)	430 (7.8)	
	Full-term (>37 weeks)	13,215 (91.8)	8149 (91.5)	5066 (92.2)	
Pregnancy, n (%)					.09[b]
	Singleton	16,166 (92.4)	10,099 (92.5)	6067 (92.2)	
	Twin	1299 (7.4)	795 (7.3)	504 (7.7)	
	Triplet	30 (0.2)	24 (0.2)	6 (0.1)	
Neonatal characteristics					
Gender, n (%)					.002[b]
	Female	8381 (48.4)	5328 (49.3)	3053 (46.9)	
	Male	8926 (51.6)	5471 (50.7)	3455 (53.1)	
Birth weight (kg), median (IQR)		3.1 (2.7-3.4)	3.1 (2.6-3.4)	3.1 (2.7-3.4)	.70[a]

[a]Wilcoxon rank sum test.

[b]Pearson chi-square test.

Table . Study results.

	Total	Before intervention	After intervention	Odds ratio	*P* value
Combined mortality rate, % (95% CI)	8.9 (8.4 - 9.3)	8.7 (8.2 - 9.3)	9.1 (8.4-9.8)	0.98 (0.86 to 1.12)	.81
Maternal mortality rate, event rate per 100,000 births (95% CI)	131.5 (85.3 - 200.6)	146.5 (86.7 - 243.6)	106.4 (46.6 - 229.7)	0.80 (0.27 to 2.32)	.68
Perinatal mortality rate, event rate per 1000 births (95% CI)	87.6 (83.5 - 91.9)	86.3 (81.1 - 91.7)	89.8 (83.1 - 97.1)	0.99 (0.87 to 1.13)	.87
Births by vacuum extraction, % (95% CI)	2.3 (2.1 - 2.6)	2.43 (2.2 - 2.8)	2.15 (1.8 - 2.5)	1.00 (0.76 to 1.33)	.99
Births by cesarean section, % (95% CI)	26.6 (25.9 - 27.2)	26 (25.2 - 26.8)	27.5 (26.4 - 28.6)	1.06 (0.94 to 1.2)	.33
Weighted Adverse Outcome Score (WAOS), median score (IQR)	39.6 (0 - 282.6)	39.1 (0 - 280.8)	40.5 (0 - 285.5)	−0.59 (−5.22 to 4.04)[a]	.80

[a]Difference.

Discussion

Principal Results

This train-the-trainer model for providing technology-enhanced, mono-professional, simulation-based obstetric training to SHOs did not result in changes to maternal and perinatal mortality outcomes. The training program also had no impact on the number of instrumental births, the number of cesareans, or the WAOS.

Strengths and Limitations

A strength of this study was the use of a randomized stepped-wedge trial design, enabling the training for all SHOs in one of the busiest labor wards worldwide. Unlike 1 group pretest-posttest designs commonly used in prior research on obstetric simulation-based training, this approach minimized bias from natural changes in health care outcomes because it could eliminate systematic period effects. Another strength was the use of the train-the-trainer model. Research has shown that training delivered by local trainers results in greater improvements in knowledge and skill acquisition [17]. Furthermore, simulation-based learning is likely to be more effective when tailored to the local context and culture [18]. A third strength of the evaluated training program was the inclusion of teamwork skills. Teamwork skills are increasingly recognized as a critical factor in reducing preventable, substandard care and are viewed as an essential competence for hospital teams [19]. These skills had not been part of previous SHO training programs at Mulago National Referral Hospital.

Limitations of our study should also be considered. First, the intended stepped-wedge design was altered during the study, as clusters requested earlier training. Although this deviation interfered with the planned study design, it was deemed unethical to withhold training further. The change in design did not affect the statistical analysis because the mean features of the stepped wedge design were maintained. Other studies have highlighted ethical concerns with the stepped-wedge design, including justifying the delayed rollout of the intervention to the control group, which is inherent to this design [20,21]. A potential solution in the future could involve using different hospitals as clusters. This approach would also address the challenge of maintaining fixed clusters of individuals during working hours, which was one of the difficulties we encountered. Interaction between trained and nontrained SHOs during shifts in examination and holiday periods may have introduced bias into the results, but our analysis was chosen conservatively, making sure that such biases would affect the treatment effect negatively. Another challenge in the work schedules of the fixed clusters was the repetition of sessions, which led to some sessions not being scheduled according to the protocol. Establishing fixed clusters of multidisciplinary obstetric team members within 1 hospital is anticipated to be even more challenging. Including different hospitals in the design could eliminate these issues and allow all health care providers involved in maternity and neonatal care at a single hospital to be trained within a short period. However, it may be difficult to include hospitals with comparable levels of care and delivery volumes, which should be accounted for in the intracluster correlation coefficient and the statistical analysis.

Comparison With Prior Work

The results of our study on maternal and perinatal outcomes do not align with those of recent studies on simulation-based emergency obstetric training in sub-Saharan Africa. Since the start of this study, 6 studies reported improvements in maternal outcomes, mainly related to postpartum hemorrhage and mortality [22-27], and 7 studies showed improvements in neonatal or perinatal outcomes after simulation-based obstetric training [25,28-33]. One of these studies showed that initial improvements declined over time [29].

When comparing our simulation-based training program to others that were effective, we want to highlight the mono-professional nature of our program. While previous research showed that simulation-based team leader training improved teamwork and communication during clinical resuscitations, our study found that training only SHOs as the leaders during obstetric emergencies did not improve patient outcomes. This aligns with the findings of Siassakos et al [34], who noted that units showing improvements had trained nearly 100% of their staff and implemented training programs within their own units. Additionally, all but 2 of the previously described effective studies were multi-professional training programs [22-31]. An exception to justify a mono-professional training program can be when the focus is on a specific technical task performed by a single health care provider, such as repairing an episiotomy. In such cases, the focus on a specific task allows for deliberate practice, where the trainee improves the task through immediate feedback, problem-solving, evaluation, and repeated performance. However, when the task involves teamwork, the training approach should shift toward a multi-professional model. In conclusion, our results, alongside the literature, suggest that future research should consider multi-professional team training in obstetrics, involving all staff within their units.

Another difference between our training program and others is that ours was a stand-alone program, while simulation-based training as part of an integrated package may be more effective in improving health outcomes [4,8,28]. Integrated packages often include equipment, maternal death reviews, health information system improvements, modified protocols, supportive supervision, mobile mentoring, and peer-assisted learning. Although we included a train-the-trainer model, restructured local protocols, and created posters with flowcharts for obstetric emergencies, the intensity and, ultimately, the training frequency of our intervention may not have been sufficient to impact maternal and neonatal outcomes.

Another notable variance between evaluated simulation-based training programs is the location of training. Our study used an off-site medical simulation center, while on-site training may be more beneficial, as it reaches more staff and generates more suggestions for organizational changes. Sorensen et al [35] found no significant differences in knowledge, patient safety attitude, motivation, or stress between on-site and off-site training, but the on-site group suggested more organizational changes. In low-resource settings, these changes may be more

valuable, although on-site training could be disrupted by clinical situations. Further research comparing on-site versus off-site training in low-resource settings would be valuable.

A further difference in previous studies on simulation-based obstetric training is the definition of mortality ratios. Some studies, including the mortality ratios in the introduction, used maternal mortality per live birth, while others used maternal mortality per number of births [8,22,26,36,37], making comparisons difficult. Additionally, the World Health Organization defines maternal and perinatal mortality ratios with different denominators, complicating statistical analyses of a combined mortality ratio. As a result, we analyzed the combined mortality rate and reported the maternal and perinatal mortality ratios separately. Moreover, the ratios are based on live births, so improving perinatal care can also affect maternal mortality outcomes. A standardized approach to mortality ratios could improve the comparability of future training programs.

A final note should be made regarding improvements in data administration. It took considerable time to manually collect, verify, and process all the data. In some cases, determining the time and cause of death was challenging, potentially leading to the inclusion of more macerated babies and higher perinatal mortality rates. This study's setting in a national underresourced referral hospital may also explain the high perinatal mortality rate. Options to address some data challenges include the development of digital data registration systems and active surveillance of data. While these solutions require initial investments of both time and money, they offer significant potential benefits in terms of efficiency and accuracy. Additionally, digital data registration and continuous data monitoring can be enhanced by the use of dashboards, which provide clinicians with an overview of current practices and can help identify deviations from targets early on [38]. This approach not only benefits the evaluation of obstetric simulation-based training but can also inform ongoing training sessions, contributing to continuous learning and improvements in obstetric training.

Given the complexity of simulation-based obstetric training implementation and evaluation in low-resource settings, future studies should consider conducting implementation and action research. This approach would be valuable for identifying barriers to effective implementation, refining training programs, and ensuring that improvements are successfully integrated into the local health care system.

Conclusions

This train-the-trainer model for providing technology-enhanced, mono-professional, simulation-based training in obstetrics to SHOs did not change maternal and perinatal mortality outcomes in a national referral hospital in a low-resource setting. This study, along with existing literature, suggests that future research should consider conducting and evaluating multi-professional team training in obstetrics, involving all staff within their units.

Acknowledgments

We thank Rob Steinweg, MSc, for all the support from Medsim, a research and training center for medical care in the Netherlands. This study was supported by Máxima Medical Center, Veldhoven, the Netherlands, and Rotary Clubs in the Netherlands and Uganda. They both supported this study by providing training supplies. The Rotary Foundation was not involved in this study's design or interpretation of study results.

Conflicts of Interest

None declared.

Checklist 1
CONSORT checklist. CONSORT: Consolidated Standards of Reporting Trials.
[PDF File, 71 KB - mededu_v11i1e54911_app1.pdf]

References

1. WHO, UNICEF, UNFPA, World Bank Group and the UN. Trends in maternal mortality: 2000 to 2017. World Bank Group. 2019. URL: https://data.worldbank.org/indicator/SH.STA.MMRT?locations=UG [accessed 2025-04-16]
2. Geleto A, Chojenta C, Mussa A, Loxton D. Barriers to access and utilization of emergency obstetric care at health facilities in sub-Saharan Africa-a systematic review protocol. Syst Rev 2018 Apr 16;7(1):60. [doi: 10.1186/s13643-018-0720-y] [Medline: 29661217]
3. Bergh AM, Baloyi S, Pattinson RC. What is the impact of multi-professional emergency obstetric and neonatal care training? Best Pract Res Clin Obstet Gynaecol 2015 Nov;29(8):1028-1043. [doi: 10.1016/j.bpobgyn.2015.03.017] [Medline: 25937554]
4. Ameh CA, Mdegela M, White S, van den Broek N. The effectiveness of training in emergency obstetric care: a systematic literature review. Health Policy Plan 2019 May 1;34(4):257-270. [doi: 10.1093/heapol/czz028] [Medline: 31056670]
5. Chou WK, Ullah N, Arjomandi Rad A, et al. Simulation training for obstetric emergencies in low- and lower-middle income countries: a systematic review. Eur J Obstet Gynecol Reprod Biol 2022 Sep;276:74-81. [doi: 10.1016/j.ejogrb.2022.07.003] [Medline: 35820293]
6. Msemo G, Massawe A, Mmbando D, et al. Newborn mortality and fresh stillbirth rates in Tanzania after helping babies breathe training. Pediatrics 2013 Feb;131(2):e353-e360. [doi: 10.1542/peds.2012-1795] [Medline: 23339223]

7. Andreatta P, Gans-Larty F, Debpuur D, Ofosu A, Perosky J. Evaluation of simulation-based training on the ability of birth attendants to correctly perform bimanual compression as obstetric first aid. Int J Nurs Stud 2011 Oct;48(10):1275-1280. [doi: 10.1016/j.ijnurstu.2011.03.001] [Medline: 21450290]
8. Dumont A, Fournier P, Abrahamowicz M, et al. Quality of care, risk management, and technology in obstetrics to reduce hospital-based maternal mortality in Senegal and Mali (QUARITE): a cluster-randomised trial. Lancet 2013 Jul 13;382(9887):146-157. [doi: 10.1016/S0140-6736(13)60593-0] [Medline: 23721752]
9. Sorensen BL, Rasch V, Massawe S, Nyakina J, Elsass P, Nielsen BB. Advanced life support in obstetrics (ALSO) and post-partum hemorrhage: a prospective intervention study in Tanzania. Acta Obstet Gynecol Scand 2011 Jun;90(6):609-614. [doi: 10.1111/j.1600-0412.2011.01115.x] [Medline: 21388368]
10. van Tetering AAC, van Meurs A, Ntuyo P, et al. Study protocol training for life: a stepped wedge cluster randomized trial about emergency obstetric simulation-based training in a low-income country. BMC Pregnancy Childbirth 2020 Jul 28;20(1):429. [doi: 10.1186/s12884-020-03050-3] [Medline: 32723330]
11. Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. BMC Med Res Methodol 2006 Nov 8;6:54. [doi: 10.1186/1471-2288-6-54] [Medline: 17092344]
12. Nielsen PE, Goldman MB, Mann S, et al. Effects of teamwork training on adverse outcomes and process of care in labor and delivery: a randomized controlled trial. Obstet Gynecol 2007 Jan;109(1):48-55. [doi: 10.1097/01.AOG.0000250900.53126.c2] [Medline: 17197587]
13. van Tetering AAC, Segers MHM, Ntuyo P, et al. Evaluating the instructional design and effect on knowledge, teamwork, and skills of technology-enhanced simulation-based training in obstetrics in Uganda: stepped-wedge cluster randomized trial. JMIR Med Educ 2021 Feb 5;7(1):e17277. [doi: 10.2196/17277] [Medline: 33544086]
14. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. Contemp Clin Trials 2007 Feb;28(2):182-191. [doi: 10.1016/j.cct.2006.05.007] [Medline: 16829207]
15. Woertman W, de Hoop E, Moerbeek M, Zuidema SU, Gerritsen DL, Teerenstra S. Stepped wedge designs could reduce the required sample size in cluster randomized trials. J Clin Epidemiol 2013 Jul;66(7):752-758. [doi: 10.1016/j.jclinepi.2013.01.009] [Medline: 23523551]
16. de Hoop E, Woertman W, Teerenstra S. The stepped wedge cluster randomized trial always requires fewer clusters but not always fewer measurements, that is, participants than a parallel cluster randomized trial in a cross-sectional design. J Clin Epidemiol 2013 Dec;66(12):1428. [doi: 10.1016/j.jclinepi.2013.07.008]
17. Chong W, Walsh M, Zannat F, et al. Effectiveness of foreign versus local trainers on skilled birth attendants' knowledge and skills acquisition after simulation-based training in rural Rwanda. Int J Gynecol Obstet 21st FIGO World Congr Gynecol Obstet BC CanadaConference Publ 2015;131(SUPPL. 5):E486.
18. Chung HS, Dieckmann P, Issenberg SB. It is time to consider cultural differences in debriefing. Simul Healthc 2013 Jun;8(3):166-170. [doi: 10.1097/SIH.0b013e318291d9ef] [Medline: 23702587]
19. Gordon M, Darbyshire D, Baker P. Non-technical skills training to enhance patient safety: a systematic review. Med Educ 2012 Nov;46(11):1042-1054. [doi: 10.1111/j.1365-2923.2012.04343.x] [Medline: 23078681]
20. Joag K, Ambrosio G, Kestler E, Weijer C, Hemming K, Van der Graaf R. Ethical issues in the design and conduct of stepped-wedge cluster randomized trials in low-resource settings. Trials 2019 Dec 19;20(Suppl 2):703. [doi: 10.1186/s13063-019-3842-1] [Medline: 31852547]
21. Prost A, Binik A, Abubakar I, et al. Logistic, ethical, and political dimensions of stepped wedge trials: critical review and case studies. Trials 2015 Aug 17;16(1):351. [doi: 10.1186/s13063-015-0837-4] [Medline: 26278521]
22. Pattinson RC, Bergh AM, Ameh C, et al. Reducing maternal deaths by skills-and-drills training in managing obstetric emergencies: a before-and-after observational study. S Afr Med J 2019 Mar 29;109(4):241-245. [doi: 10.7196/SAMJ.2019.v109i4.13578] [Medline: 31084689]
23. Nelissen E, Ersdal H, Mduma E, et al. Clinical performance and patient outcome after simulation-based training in prevention and management of postpartum haemorrhage: an educational intervention study in a low-resource setting. BMC Pregnancy Childbirth 2017 Sep 11;17(1):301. [doi: 10.1186/s12884-017-1481-7] [Medline: 28893211]
24. Hanson C, Atuhairwe S, Lucy Atim J, Marrone G, Morris JL, Kaharuza F. Effects of the helping mothers survive bleeding after birth training on near miss morbidity and mortality in Uganda: a cluster-randomized trial. Int J Gynaecol Obstet 2021 Mar;152(3):386-394. [doi: 10.1002/ijgo.13395] [Medline: 32981091]
25. Evans CL, Bazant E, Atukunda I, et al. Peer-assisted learning after onsite, low-dose, high-frequency training and practice on simulators to prevent and treat postpartum hemorrhage and neonatal asphyxia: a pragmatic trial in 12 districts in Uganda. PLoS ONE 2018;13(12):e0207909. [doi: 10.1371/journal.pone.0207909] [Medline: 30557350]
26. Chang OH, Levy B, Lytle H, et al. Implementation of the Alliance for Innovation on Maternal Health Program to reduce maternal mortality in Malawi. Obstet Gynecol 2019 Mar;133(3):507-514. [doi: 10.1097/AOG.0000000000003108] [Medline: 30741809]
27. Zongo A, Dumont A, Fournier P, Traore M, Kouanda S, Sondo B. Effect of maternal death reviews and training on maternal mortality among cesarean delivery: post-hoc analysis of a cluster-randomized controlled trial. Eur J Obstet Gynecol Reprod Biol 2015 Feb;185:174-180. [doi: 10.1016/j.ejogrb.2014.12.023] [Medline: 25590501]

28. Walker D, Otieno P, Butrick E, et al. Effect of a quality improvement package for intrapartum and immediate newborn care on fresh stillbirth and neonatal mortality among preterm and low-birthweight babies in Kenya and Uganda: a cluster-randomised facility-based trial. Lancet Glob Health 2020 Aug;8(8):e1061-e1070. [doi: 10.1016/S2214-109X(20)30232-1]
29. Rule ARL, Maina E, Cheruiyot D, Mueri P, Simmons JM, Kamath-Rayne BD. Using quality improvement to decrease birth asphyxia rates after "Helping Babies Breathe" training in Kenya. Acta Paediatr 2017 Oct;106(10):1666-1673. [doi: 10.1111/apa.13940] [Medline: 28580692]
30. Mduma ER, Ersdal H, Kvaloy JT, et al. Using statistical process control methods to trace small changes in perinatal mortality after a training program in a low-resource setting. Int J Qual Health Care 2018 May 1;30(4):271-275. [doi: 10.1093/intqhc/mzy003]
31. Mduma E, Ersdal H, Svensen E, Kidanto H, Auestad B, Perlman J. Frequent brief on-site simulation training and reduction in 24-h neonatal mortality--an educational intervention study. Resuscitation 2015 Aug;93:1-7. [doi: 10.1016/j.resuscitation.2015.04.019] [Medline: 25957942]
32. Gomez PP, Nelson AR, Asiedu A, et al. Accelerating newborn survival in Ghana through a low-dose, high-frequency health worker training approach: a cluster randomized trial. BMC Pregnancy Childbirth 2018 Mar 22;18(1):72. [doi: 10.1186/s12884-018-1705-5] [Medline: 29566659]
33. Eblovi D, Kelly P, Afua G, Agyapong S, Dante S, Pellerite M. Retention and use of newborn resuscitation skills following a series of helping babies breathe trainings for midwives in rural Ghana. Glob Health Action 2017;10(1):1387985. [doi: 10.1080/16549716.2017.1387985] [Medline: 29058568]
34. Siassakos D, Crofts JF, Winter C, Weiner CP, Draycott TJ. The active components of effective training in obstetric emergencies. BJOG An Int J Obstet Gynaecol 2009 Jul;116(8):1028-1032. [doi: 10.1111/j.1471-0528.2009.02178.x] [Medline: 19438497]
35. Sørensen JL, van der Vleuten C, Rosthøj S, et al. Simulation-based multiprofessional obstetric anaesthesia training conducted in situ versus off-site leads to similar individual and team outcomes: a randomised educational trial. BMJ Open 2015 Oct 6;5(10):e008344. [doi: 10.1136/bmjopen-2015-008344] [Medline: 26443654]
36. Fransen AF, van de Ven J, Banga FR, Mol BWJ, Oei SG. Multi-professional simulation-based team training in obstetric emergencies for improving patient outcomes and trainees' performance. Cochrane Database Syst Rev 2020 Dec 16;12(12):CD011545. [doi: 10.1002/14651858.CD011545.pub2] [Medline: 33325570]
37. Fransen AF, van de Ven J, Schuit E, van Tetering A, Mol BW, Oei SG. Simulation-based team training for multi-professional obstetric care teams to improve patient outcome: a multicentre, cluster randomised controlled trial. BJOG An Int J Obstet Gynaecol 2017 Mar;124(4):641-650. [doi: 10.1111/1471-0528.14369] [Medline: 27726304]
38. Crofts J, Moyo J, Ndebele W, Mhlanga S, Draycott T, Sibanda T. Adaptation and implementation of local maternity dashboards in a Zimbabwean hospital to drive clinical improvement. Bull World Health Organ 2014 Feb 1;92(2):146-152. [doi: 10.2471/BLT.13.124347] [Medline: 24623908]

Abbreviations

OR: odds ratio
SHO: senior house officer
WAOS: Weighted Adverse Outcome Score

Edited by A Bahattab; submitted 27.11.23; peer-reviewed by CI Sartorão Filho, D Ellwood; revised version received 05.01.25; accepted 30.01.25; published 09.05.25.

Please cite as:
van Tetering AAC, de Vries EL, Ntuyo P, van den Heuvel ER, Fransen AF, van der Hout-van der Jagt MB, Namagembe I, Byamugisha J, Oei SG
Mono-Professional Simulation-Based Obstetric Training in a Low-Resource Setting: Stepped-Wedge Cluster Randomized Trial
JMIR Med Educ 2025;11:e54911
URL: https://mededu.jmir.org/2025/1/e54911
doi:10.2196/54911

Implementation Outcomes of Reusable Learning Objects in Health Care Education Across Three Malaysian Universities: Evaluation Using the RE-AIM Framework

Hooi Min Lim[1]; Chin Hai Teo[1]; Yew Kong Lee[1]; Ping Yein Lee[2]; Kuhan Krishnan[3]; Zahiruddin Fitri Abu Hassan[4]; Phelim Voon Chen Yong[5]; Wei Hsum Yap[5]; Renukha Sellappans[6]; Enna Ayub[7]; Nurhanim Hassan[8]; Sazlina Shariff Ghazali[9]; Nurul Amelina Nasharuddin[10]; Puteri Shanaz Jahn Kassim[9]; Faridah Idris[11]; Klas Karlgren[12]; Natalia Stathakarou[12]; Petter Mordt[13]; Stathis Konstantinidis[14]; Michael Taylor[14]; Cherry Poussa[14]; Heather Wharrad[14]; Chirk Jenn Ng[1,15,16]

[1]Department of Primary Care Medicine, Faculty of Medicine, Universiti Malaya, Lembah Pantai, Kuala Lumpur, Malaysia
[10]Department of Multimedia, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Selangor, Malaysia
[11]Department of Pathology, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Selangor, Malaysia
[12]Department of Learning, Informatics, Management and Ethics, Karolinska Institutet, Stockholm, Sweden
[13]NettOp, Department of E-Learning Development, University of Stavanger, Stavanger, Norway
[14]School of Health Sciences, University of Nottingham, Nottingham, United Kingdom
[15]Centre for Population Health Research and Implementation, SingHealth Regional Health System, Singapore, Singapore
[16]Duke-NUS Medical School, Singapore, Singapore
[2]UMeHealth Unit, Faculty of Medicine, Universiti Malaya, Kuala Lumpur, Malaysia
[3]Dean's Office, Faculty of Medicine, Universiti Malaya, Kuala Lumpur, Malaysia
[4]Department of Building Surveying, Faculty of Built Environment, Universiti Malaya, Kuala Lumpur, Malaysia
[5]School of Biosciences, Faculty of Health & Medical Sciences, Taylor's University, Selangor, Malaysia
[6]School of Pharmacy, Faculty of Health & Medical Sciences, Taylor's University, Selangor, Malaysia
[7]Taylor's Digital, Taylor's University, Selangor, Malaysia
[8]Learning Innovation and Development, Centre for Future Learning, Taylor's University, Selangor, Malaysia
[9]Department of Family Medicine, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Serdang, Malaysia

Corresponding Author:
Hooi Min Lim
Department of Primary Care Medicine, Faculty of Medicine, Universiti Malaya, Lembah Pantai, Kuala Lumpur, Malaysia

Abstract

Background: Current e-learning evaluation focuses on learners' knowledge gain, satisfaction, perceptions, and attitudes; few assess the implementation outcomes of e-learning resources in teaching and learning.

Objective: In this study, we used the RE-AIM (Reach, Effectiveness, Adoption, Implementation, and Maintenance) framework to systematically evaluate the implementation outcomes of reusable learning objects (RLOs) in the context of health care education.

Methods: This study is a part of the Advancing Co-creation of RLOs to Digitise Healthcare Curriculum (ACoRD) project, wherein we developed and implemented 23 RLOs across 3 Malaysian universities for medical, pharmacy, and biomedical curricula. Implementation and dissemination strategies were employed. Data were collected using a self-administered web-based questionnaire and Google Analytics.

Results: This study reports a cumulative RLO access of 7622 users from 48 countries (reach). Users rated RLOs as very helpful (1452/2071, 70.1%) or helpful (601/2071, 29.1%). Preassessments and postassessments showed a significant improvement in the knowledge score (21 RLOs, $P<.05$) and confidence level (17 RLOs, $P<.05$) (effectiveness). All 3 Malaysian universities adopted RLOs in the fields of professional development, primary care medicine, medicine, pediatrics, nursing, pharmacy, and biomedicine (adoption). The percentage of users who completed RLOs ranged from 5.6% (10/179) to 85% (78/92), with nonbounced users (users who viewed more than one page) ranging from 16.3% (165/1014) to 88.5% (370/418) (implementation). In the 4 months following the completion of the ACoRD project, a total of 2107 users accessed RLOs (maintenance).

Conclusions: We systematically evaluated the implementation of e-learning resources by using the RE-AIM framework, informing future strategies to integrate e-learning innovations in real-world teaching and learning practices.

(JMIR Med Educ 2025;11:e63882) doi:10.2196/63882

KEYWORDS

e-learning; RE-AIM; implementation; dissemination; reusable learning objects; medical education; reach; effectiveness; adoption; maintenance

Introduction

The use of technology in medical education and health sciences is on the rise, aiming to enhance the knowledge, skills, and practice of medical students. e-Learning, the process of acquisition and use of knowledge distributed and facilitated by electronic means [1], has been shown to be effective in not only improving knowledge but also increasing students' satisfaction in medical education [2,3]. The application of e-learning objects in medical education promotes self-directed and personalized learning, serving as a complement to the conventional didactic teaching method in medical schools [4]. An e-learning object refers to a collection of digital materials structured in a meaningful way, aligned with a specific learning objective, and designed as independent, self-contained units of instructional materials [4].

Although institutions are devoting substantial resources to the development of e-learning, successful implementation of e-learning remains challenging and is often not systematically measured [5]. A systematic review showed that the evaluation of e-learning tends to focus on the assessment of learners' knowledge, satisfaction, perceptions, and attitudes, and there is a lack of rigorous evaluation of its implementation outcomes in the real-world environment [6]. Implementation outcomes are the effects of deliberate and purposive action to implement a new intervention or practice [7]. In the context of e-learning, implementation outcomes include the acceptability, adoption, appropriateness, feasibility, fidelity, and sustainability of e-learning resources in teaching and learning [8]. When e-learning objects are released as open content, reach and discoverability become important.

To date, several models have been used to evaluate the effectiveness and implementation of e-learning resources. The Kirkpatrick model, an outcome-focused model, has been widely employed to evaluate the effectiveness of an educational program across 4 levels: reaction, learning, behavior, and results [9]; however, it does not evaluate the implementation outcomes. The Context-Input-Process-Product model [10] is designed to evaluate both the process and product to determine the success of an educational program. Nevertheless, this model primarily focuses on defining the contextual factors to improve performance [11]. The Analysis, Design, Development, Implementation, and Evaluation model [12], an instructional design model, offers a structured approach to guide the development and implementation of educational programs but falls short of providing metrics for measuring implementation. Therefore, there is a need for an e-learning implementation outcome framework that provides a comprehensive and objective evaluation of the implementation of e-learning resources.

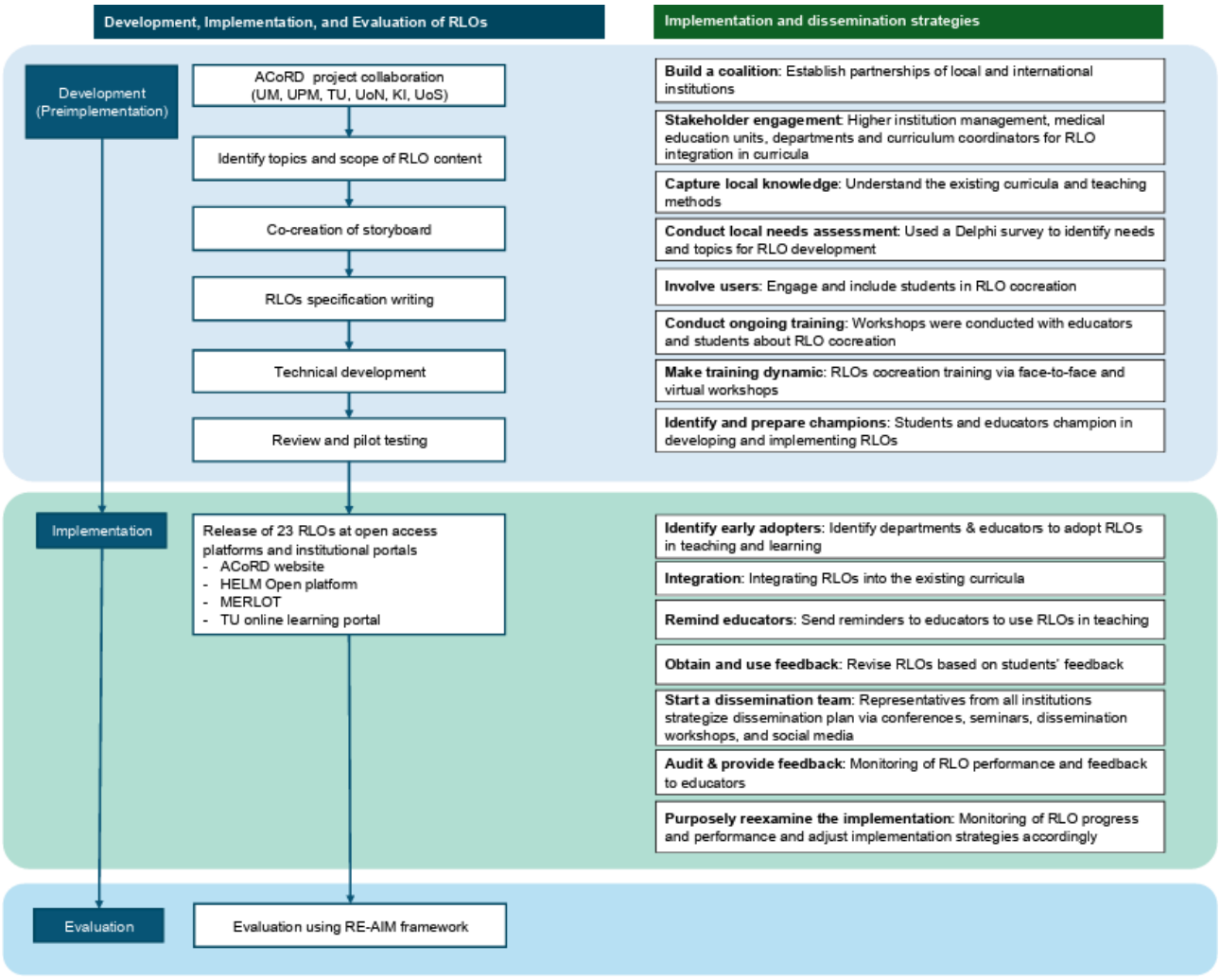
RE-AIM (Reach, Effectiveness, Adoption, Implementation, and Maintenance) is an established implementation science framework used in the health care setting to measure how effectively an evidence-based intervention is implemented [13]. It has been used in several studies to evaluate the implementation outcomes of health care provider education programs [14,15]. In this study, we aimed to use the RE-AIM framework to evaluate the implementation outcomes of reusable learning objects (RLOs), which are bite-sized, stand-alone interactive web-based resources with multimedia that focus on a single learning objective [16]. The findings from this study would provide evidence on the feasibility of using the RE-AIM framework in the context of health care education, describe the implementation strategies used in this project, and propose recommendations to improve the evaluation of e-learning resource implementation in higher educational institutions.

Methods

Study Design

This study is a part of the Advancing Co-creation of RLOs to Digitise Healthcare Curriculum (ACoRD) project, which was funded by the European Union Erasmus+ project [17]. It is a capacity building project involving Universiti Malaya, Universiti Putra Malaysia, Taylor's University, University of Nottingham, Karolinska Institutet, and University of Stavanger. The ACoRD project aimed to introduce innovative digital pedagogy methods by developing, evaluating, and disseminating high-quality, peer-reviewed RLOs that benefit health care and biomedical science learners in Malaysia. RLOs were developed, implemented, and evaluated over 3 phases (Figure 1). The implementation strategies of RLOs were developed and performed in both the development (preimplementation phase) and implementation phases.

Figure 1. Development, implementation, and evaluation of reusable learning objects in the ACoRD project. ACoRD: Advancing Co-creation of Reusable Learning Objects to Digitise Healthcare Curriculum; HELM: Health e-Learning and Media; KI: Karolinska Institutet; MERLOT: Multimedia Educational Resource for Learning and Online Teaching; RE-AIM: Reach, Effectiveness, Adoption, Implementation, and Maintenance; RLO: reusable learning object; TU: Taylor's University; UM: Universiti Malaya; UoN: University of Nottingham; UoS: University of Stavanger; UPM: Universiti Putra Malaysia.



RLO Development Phase (Preimplementation)

RLOs were developed following the ASPIRE (Aim, Storyboarding, Populating, Implementing, Release, and Evaluation) framework [18]. The development phase began by establishing a coalition between international institutions within the project. We evaluated the process and challenges of transnational partnerships and knowledge transfer by using a qualitative study, providing insights on guiding effective partnership in e-learning development in the future [17]. This was followed by stakeholder engagement at 3 Malaysian institutions, involving higher education institution management, medical education units, respective departments, and curriculum coordinators. To identify the gaps and topics for RLO development, the ACoRD team reviewed the existing curricula and determined how RLOs can be integrated into the existing topics, learning activities, and teaching delivery methods in the curricula. We also conducted a needs assessment by using a Delphi survey to compare the differences between educators and learners in prioritizing topics for RLO development [19]. After determining the topics and scope of 23 RLOs, we engaged the learners in RLO cocreation, especially in the storyboard development step. We evaluated learners' knowledge, confidence, and satisfaction of the storyboarding session by using a pre-post survey and explored their perception and experience of the cocreation process qualitatively [20]. We provided RLO specification writing training to the educators who subsequently wrote the content for RLOs. Once RLO specifications were finalized, a content review was conducted before moving on to technical development. A small pilot test was conducted with student champions on RLO usability before full implementation.

RLO Implementation Phase

The 23 RLOs (Table 1) were made available through the open-access ACoRD repository website [21], Health e-Learning and Media-Open under the University of Nottingham [22], Multimedia Educational Resource for Learning and Online Teaching [23], and the institutional online learning portal and curriculum guidebook.

The focus was on the actual integration of RLO resources into the health sciences curricula. At Universiti Malaya, RLOs were incorporated into various block postings for medicine and an undergraduate semester course for nursing. In Universiti Putra

Malaysia, RLOs were used in a professionalism module block, and at Taylor's University, they were embedded as part of semester courses in pharmacy and biomedical science. Various strategies were implemented to improve the usage of RLOs, including identifying early adopters (department and educators) to use RLOs in teaching, integrating RLOs into the existing curricula, sending reminders to educators, receiving feedback from students, establishing an effective dissemination team, monitoring RLO performance and feedback to educators, and periodic re-examining of the implementation strategies.

Table . The list of reusable learning objects developed for undergraduate students in health care professions.

University, course	RLO[a] titles (n=23)
Universiti Malaya, Kuala Lumpur, Malaysia	
Medicine	• RLO 1: Prescription Writing—Back to Basics • RLO 2: Treatment of Acute Illness • RLO 3: Principles of Family Medicine • RLO 4: Factors Affecting Nutrition in the Older Person • RLO 5: Growth Faltering in Children • RLO 6: Identify Challenging Behavior in Health Care Settings • RLO 7: How to Conduct a Literature Search
Universiti Putra Malaysia, Selangor, Malaysia	
Medicine	• RLO 8: Confidentiality • RLO 9: Breaking Bad News • RLO 10: Doctor-Patient Relationship • RLO 11: Consent • RLO 12: Verbal and Nonverbal Skills • RLO 13: Counseling Skills • RLO 14: Ethical Reasoning • RLO 15: Social Media Professionalism
Taylor's University, Selangor, Malaysia	
Pharmacy	• RLO 16: Using Nicotine Gum for Smoking Cessation • RLO 17: Using Nicotine Patches for Smoking Cessation • RLO 18: Using Varenicline for Smoking Cessation
Biomedical science	• RLO 19: Body Metabolism • RLO 20: DNA Repair • RLO 21: DNA Replication • RLO 22: Cardiac Output • RLO 23: Nervous Regulation of the Heart

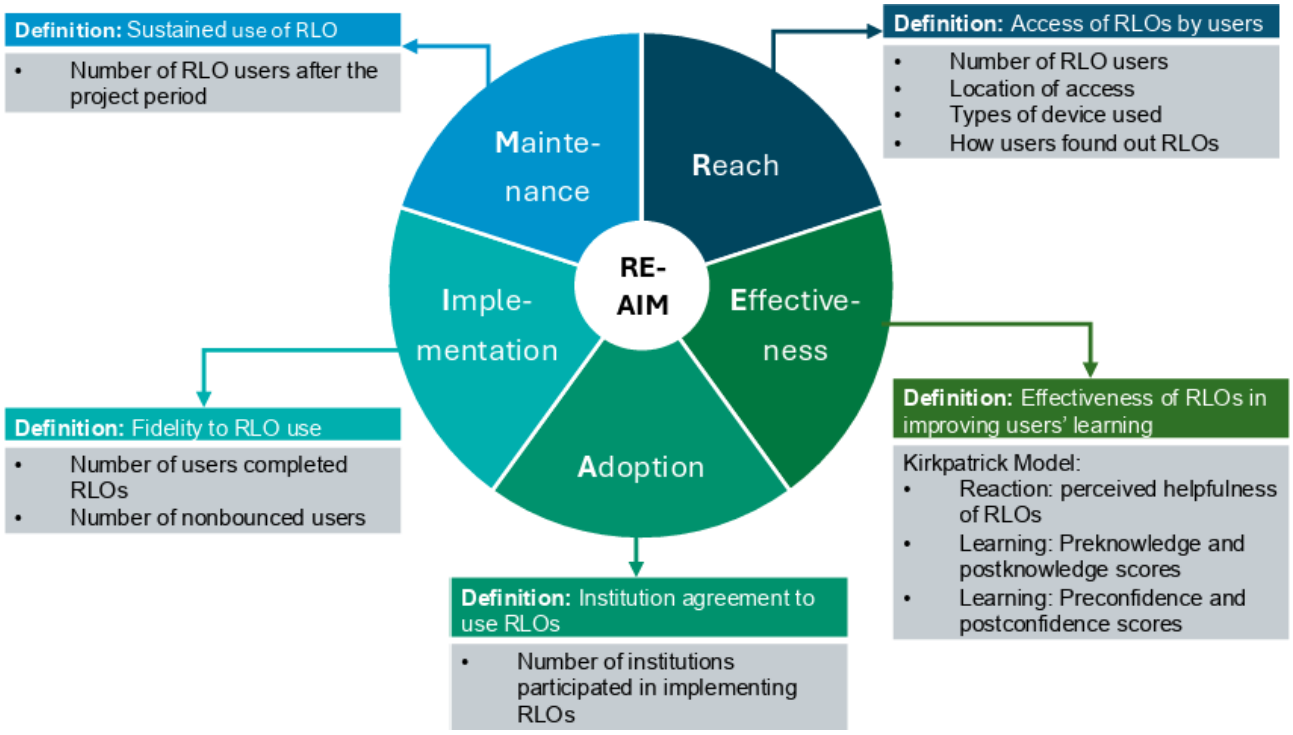
[a]RLO: reusable learning object.

Evaluation Phase

This study was performed across 3 Malaysian institutions (Universiti Malaya, Universiti Putra Malaysia, and Taylor's University). Participants in this evaluation were users who used RLOs during the implementation period from May 1, 2020, to February 24, 2022. We used a universal sampling method for data collection.

We evaluated the implementation of RLOs by using the RE-AIM framework (Figure 2). The reach domain is defined as the access of RLOs by the users. We assessed reach by measuring the absolute number of RLO users, location of access, types of devices used by the users, and how users found out about our RLOs.

Figure 2. Definitions and outcomes measures of the RE-AIM framework. RE-AIM: Reach, Effectiveness, Adoption, Implementation, and Maintenance; RLO: reusable learning object.



For the effectiveness domain, we applied the 4-level Kirkpatrick model: Reaction, Learning, Behavior, and Results [24]. It is a model that is commonly used in medical education to evaluate the effectiveness of a training program [25]. Reaction was assessed by gauging learners' perceived helpfulness of RLOs, and learning was measured using pre-RLO and post-RLO knowledge and confidence scores. Behavior measures whether learners are applying the newly acquired knowledge from RLOs in clinical practice, while results measure the ultimate impact of training using RLOs, which includes impact at an organizational level. We did not measure the behavior and results domains in this study due to limitations in the study design.

Adoption refers to the number of institutions that agreed to use RLOs in teaching and learning activities. We captured information on the institutions, departments, and existing modules in which RLOs were implemented. The implementation domain refers to users' fidelity in using RLOs for their learning. We measured the number of users who completed RLOs and the number of nonbounced users. Nonbounced users are those who are engaged with the site either by clicking on links, visiting multiple pages, or triggering events that indicate interaction, while bounced users are those who visited the RLO web page but only viewed a single page without interacting further before exiting.

The maintenance domain refers to the extent to which the sustained use of RLOs becomes institutionalized or part of the routine teaching and learning practice. We measured the number of users who visited the RLO web page 4 months after the project ended.

Research Instruments and Data Collection

We used 2 instruments to capture the RLO implementation outcomes: questionnaires and Google Analytics (Table 2). The questionnaires were designed to gather user feedback and usage patterns. We used 2 types of questionnaires: (1) a web-based questionnaire was administered at the completion of each RLO to assess users' perception of its usefulness (using a 4-point Likert scale: 1-very unhelpful, 2-unhelpful, 3-helpful, and 4-very helpful) and how they discovered RLOs, and (2) a confidence Likert scale and RLO-specific knowledge questionnaire were administered before and immediately after RLO usage. For the knowledge score, RLOs 16, 17, and 18 (using nicotine gum, nicotine patches, and varenicline for smoking cessation) were assessed using 1 set of knowledge questions, while RLOs 20 and 21 (DNA repair and replication) were assessed using another set of knowledge questions related to specific RLO topics. We used GAMM1-Google Analytics in Universiti Malaya and Universiti Putra Malaysia and Moodle monitoring in Taylor's University to capture the patterns of access (location, number of users, number of completing users).

Table . RE-AIM (Reach, Effectiveness, Adoption, Implementation, and Maintenance) outcome measures and assessment methods.

RE-AIM measures	Assessment methods
Reach	
Number of RLO[a] users	Google Analytics
Location of access	Google Analytics
Types of devices used	Google Analytics
How do users find out RLOs	Questionnaire: multiple-choice
Effectiveness	
Perceived helpfulness of RLOs in learning (Kirkpatrick level 1: reaction)[b]	Questionnaire: 4-point Likert scale
Knowledge and confidence (Kirkpatrick level 2: learning)	Questionnaire 1: pre- and post-RLO knowledge score; Questionnaire 2: pre- and post-RLO confidence score
Adoption	
Number of institutions who adopted RLOs in teaching and learning	N/A[c]
Implementation	
Number of users who completed RLOs	Google Analytics
Number of nonbounced users[d]	Google Analytics
Maintenance	
Number of RLO accesses 4 months after the ACoRD[e] project implementation (from February 24 to June 22, 2022)	Google Analytics

[a]RLO: reusable learning object.

[b]Perceived helpfulness is measured using a 4-point Likert scale (1-very unhelpful, 2-unhelpful, 3-helpful, and 4-very helpful).

[c]N/A: not applicable.

[d]Nonbounced users are defined as users who viewed more than one page.

[e]ACoRD: Advancing Co-creation of Reusable Learning Objects to Digitise Healthcare Curriculum.

Data Analysis

For the questionnaire items, categorical data were reported descriptively using proportion and percentage. The mean difference in the pre-RLO and post-RLO scores on knowledge and confidence were analyzed using Mann-Whitney *U* test, as the data were skewed. Statistical analysis was considered significant when $P<.05$. Google Analytics data were reported descriptively using proportion and percentage. All the data were analyzed using SPSS (version 27; IBM Corp) and Microsoft Excel.

Ethics Approval

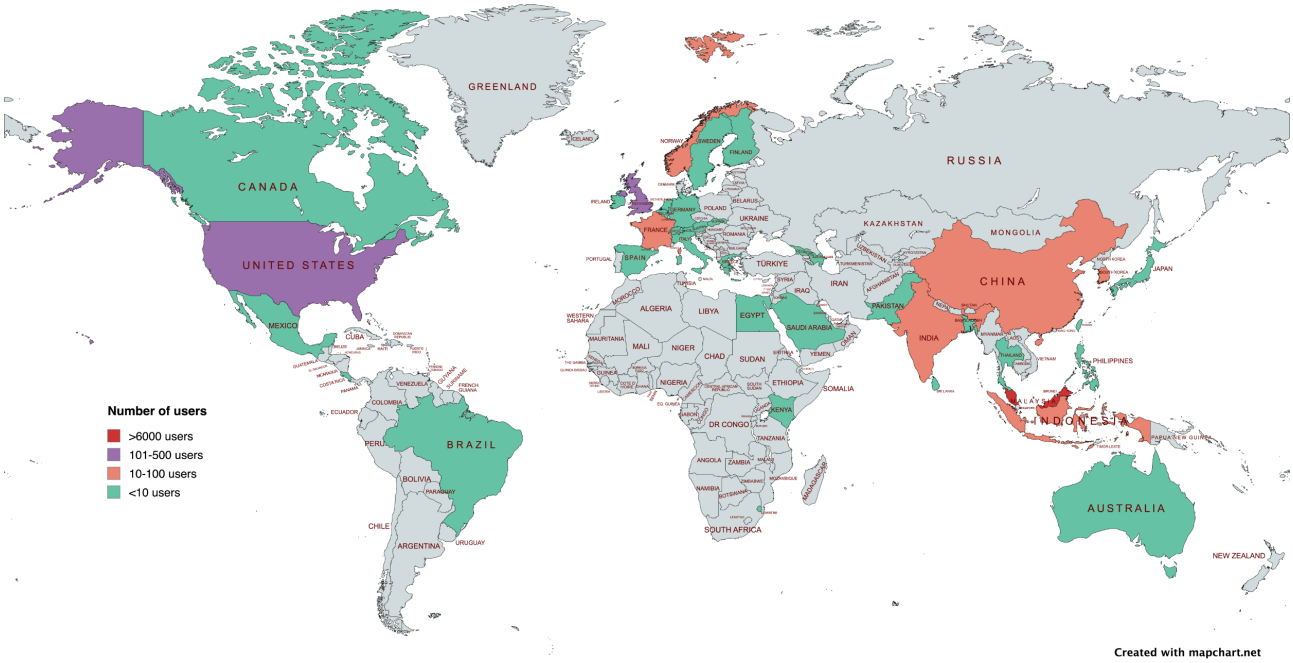
Ethics approval was granted by the Universiti Malaya Research Ethics Committee (reference UM.TNC2/UMREC-997). All participants provided their written consent. All participant information has been anonymized and deidentified. There was no financial compensation provided to the participants.

Results

Overview

From the launch of the first RLO on May 1, 2020, until February 24, 2022, Google Analytics recorded a cumulative 7622 users from 48 countries (Figure 3). We have reported the results according to the RE-AIM framework.

Figure 3. Geographical distribution of the reusable learning object users from 48 countries. There was no country with users between 500 and 6000.



Reach

Among the 7622 global users, the majority (n=6817, 89.4%) were from Malaysia, while 10.6% (n=805) were from other countries. Desktop computers were the predominant devices used to access RLOs (6045/7622, 79.3%); 20.7% (1577/7622) used portable devices such as mobile phones or tablets. The number of users for each RLO varied from 92 to 1014 (Table 3). The RLOs attracting the highest number of users were RLO 7 (How to Conduct Literature Research, n=1014), RLO 1 (Prescription Writing: Back to Basics, n=530), and RLO 17 (Using Nicotine Patches for Smoking Cessation, n=511). RLOs with fewer than 100 users were RLO 10 (Doctor-Patient Relationship) and RLO 23 (Nervous Regulation of the Heart).

Out of 2118 respondents to the questionnaire, the majority used RLOs because they were part of the course learning resource (n=1901, 89.8%), while others received a recommendation from peers, colleagues, or lecturers (n=87, 8.8%). A small proportion of the respondents (30/2118, 1.5%) found RLOs through the Health e-Learning and Media open database, open educational catalogs, or general internet searches.

Table . Number of users for each reusable learning object during project implementation.

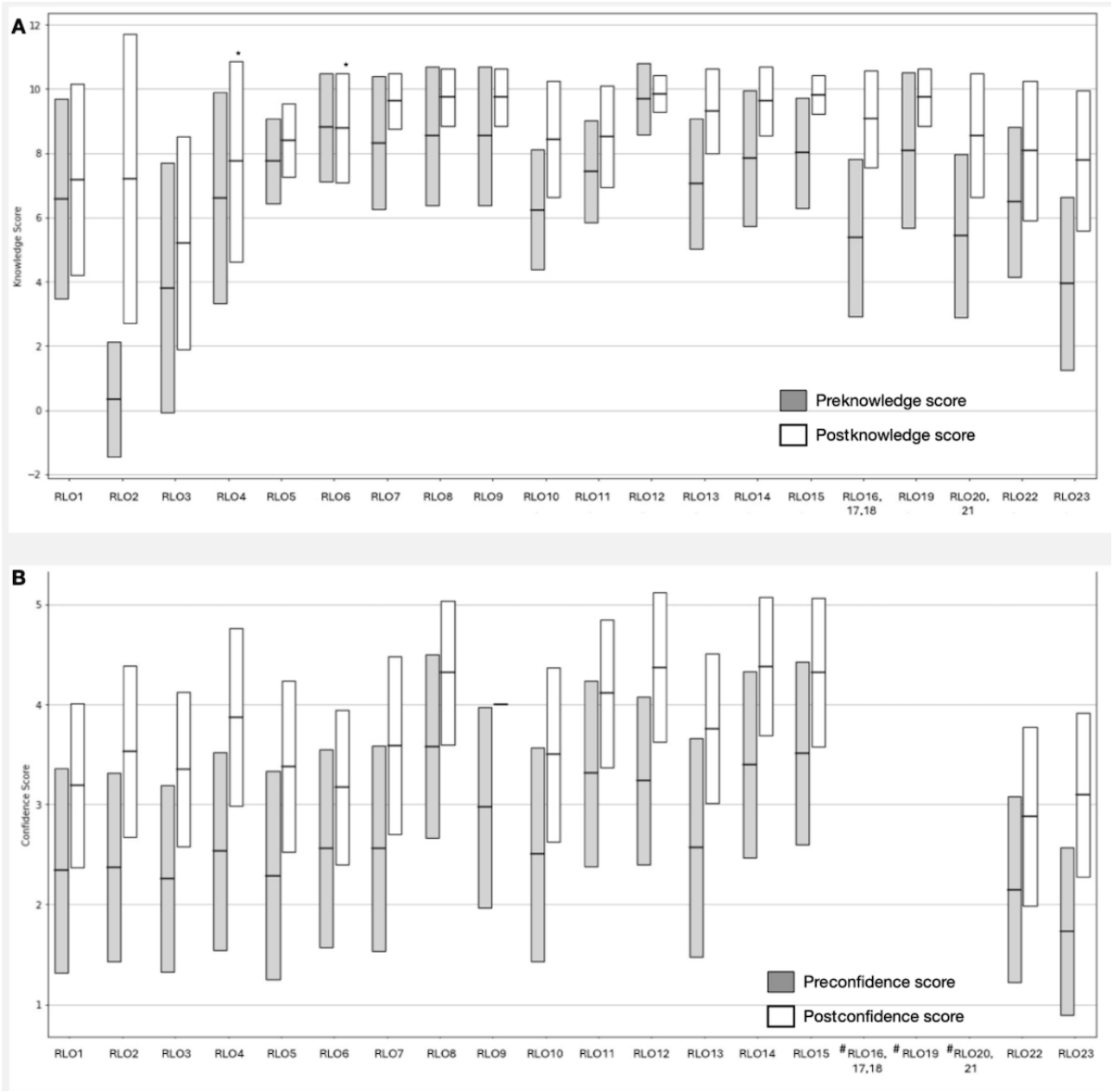
RLO[a] type	Users, n
RLO 1	530
RLO 2	366
RLO 3	467
RLO 4	310
RLO 5	363
RLO 6	184
RLO 7	1014
RLO 8	234
RLO 9	187
RLO 10	97
RLO 11	212
RLO 12	336
RLO 13	187
RLO 14	335
RLO 15	388
RLO 16	432
RLO 17	511
RLO 18	418
RLO 19	191
RLO 20	365
RLO 21	179
RLO 22	224
RLO 23	92

[a]RLO: reusable learning object.

Effectiveness

Most users rated RLOs as very helpful (1452/2071, 70.1%) and helpful (601/2071, 29%). We evaluated the preknowledge and postknowledge scores across all RLOs. The mean preknowledge score ranged from 0.33 to 9.6, while the postknowledge score ranged from 5.20 to 9.85. A significant improvement in the knowledge scores was reported for all RLOs after their use (P<.05) (Figure 4), except for RLO 4 and RLO 6, which had high mean baseline preknowledge scores of 6.61 and 8.80, respectively (Table S1 in Multimedia Appendix 1). Additionally, we assessed the confidence scores for 17 RLOs (excluding 5 RLOs, ie, RLOs 16-21 due to missing data) and found a significant improvement in the mean confidence scores after utilization for all 17 RLOs (P<.05) (Table S2 in Multimedia Appendix 1).

Figure 4. (**A**) Comparison of preknowledge and postknowledge scores for each reusable learning object. (**B**) Comparison of preconfidence and postconfidence scores for each reusable learning object. *P>.05. # indicates missing data. RLO: reusable learning object.



Adoption

All RLOs have been adopted by the 3 Malaysian universities (Universiti Malaya, Universiti Putra Malaysia, and Taylor's University). At Universiti Malaya, the departments of primary care medicine, medicine (geriatrics), pediatrics, and nursing have incorporated RLOs into their teaching and learning curricula. Universiti Putra Malaysia has used RLOs within its professional development module, while Taylor's University has implemented them in the pharmacy and biomedical undergraduate courses.

Implementation

The completion rates for RLOs varied widely, with percentages ranging from 5.6% (10/179) to 85% (78/92) as shown in Table 4. Only RLO 8 (Confidentiality), RLO 15 (Social Media Professionalism), and RLO 23 (Nervous Regulation of the Heart) had a completion rate exceeding 50%. The proportion of nonbounced users, defined as users who viewed more than one page, ranged from 16.3% (165/1014) to 88.5% (370/418). Notably, data on nonbounced users for RLOs 8-15 from Universiti Putra Malaysia were unavailable due to the absence of a tracking function on each RLO page in Google Analytics (because of technical errors).

Table . Number of users who completed reusable learning objects and number of nonbounced users during the implementation period.

RLO[a] type	Accesses, n	Users who completed RLOs, n (%)	Nonbounced[b] users, n (%)
RLO 1	530	205 (38.7)	129 (24.3)
RLO 2	366	90 (24.6)	97 (26.5)
RLO 3	467	116 (24.8)	143 (30.6)
RLO 4	310	46 (14.8)	64 (20.7)
RLO 5	363	53 (14.6)	71 (19.6)
RLO 6	184	33 (17.9)	49 (26.6)
RLO 7	1014	75 (7.4)	165 (16.3)
RLO 8	234	170 (72.7)	—[c]
RLO 9	187	68 (36.4)	—
RLO 10	97	35 (36)	—
RLO 11	212	100 (47.2)	—
RLO 12	336	164 (48.8)	—
RLO 13	187	60 (32.1)	—
RLO 14	335	116 (34.6)	—
RLO 15	388	204 (52.6)	—
RLO 16	432	121 (28)	352 (81.5)
RLO 17	511	113 (22.1)	351 (68.7)
RLO 18	418	109 (26.1)	370 (88.5)
RLO 19	191	23 (12)	68 (35.6)
RLO 20	365	61 (16.7)	271 (74.3)
RLO 21	179	10 (5.6)	111 (62)
RLO 22	224	76 (33.9)	147 (65.6)
RLO 23	92	78 (84.8)	30 (32.6)

[a]RLO: reusable learning object.

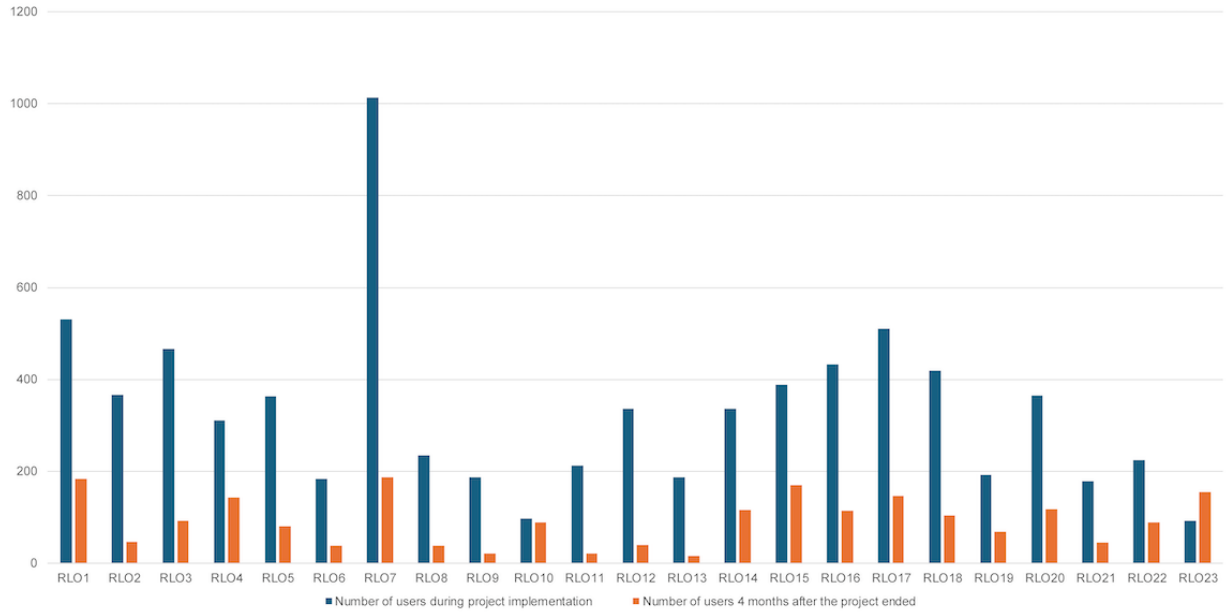
[b]Users who viewed more than one page.

[c]Not available.

Maintenance

After the ACoRD project ended on February 24, 2022, a total of 2107 users continued to access RLOs in the subsequent 4 months, ranging from 15 to 187 accesses per RLO (Figure 5).

Figure 5. Number of users who accessed reusable learning objects 4 months after the end of the project.



Discussion

Principal Findings

This study reports the successful implementation of RLOs in health care education across 3 Malaysian higher educational institutions by using the RE-AIM framework. Our findings offer a broad perspective on how the impact and outcomes of e-learning objects can be measured and evaluated systematically to inform e-learning implementation strategies.

Our study shows that RLOs were accessed by a more diverse group of learners than initially anticipated, extending beyond the Malaysian institutions, where RLOs were intended to be used. Leveraging on the unique characteristics of reusability in RLOs [26], we adopted an open educational resource approach in their deployment. This strategy significantly broadened the reach of RLOs, making them available to learners across different disciplines. Consequently, RLOs addressing more general subjects such as literature review, prescription writing, and smoking cessation attracted a larger user base compared to those focusing on discipline-specific topics.

Throughout the ACoRD project, pragmatic implementation and dissemination strategies were employed to engage a broader audience. These strategies encompassed identifying early adopters and educator champions, sending of reminders to educators about incorporating RLOs in their teaching, monitoring of RLO performance at regular intervals, and providing feedback to educators. A study on mobile health e-learning courses highlighted the advantages of iterative feedback among developers, early adopters, and end users, which aid in the refinement of existing e-learning implementation strategies and the development of new ones [27]. The ACoRD project featured a dedicated dissemination team that organized and executed activities to promote RLOs through conferences and engagement workshops, extending beyond our institutions. Establishing collaborative partnerships with other higher educational institutions is an important strategy to boost visibility and ensure the long-term use of e-learning resources [28].

In our study, users perceived RLOs as beneficial for their learning, with most RLOs demonstrating an improvement in users' knowledge and confidence levels. These findings are aligned with previous studies that have reported improvements in knowledge and understanding following RLO usage [16,29,30]. However, evaluating the effectiveness of e-learning resources remains a challenge. Most studies measured learners' reactions and learnings such as knowledge, usefulness, confidence, satisfaction, and motivation [31]. Some studies examining the effectiveness of RLOs demonstrated behavior change in relation to prescribing behavior and hearing aid use [32,33]. Measuring higher levels of learning such as behavior change and impact of learning is complex due to the time and cost involved [34]. We propose that the evaluation of RLO effectiveness should go beyond individual learning outcomes to assess their impact within the wider teaching and learning ecosystem. This could include evaluating RLOs' effectiveness from educators' perspectives and examining how RLOs contribute to student empowerment in self-directed learning [6].

In the adoption domain of RE-AIM, RLOs developed by the ACoRD project were adopted by all 3 Malaysian institutions. Our study underscores the significance of planning and initiating implementation strategies during the development phase (preimplementation) with stakeholders and institutional engagement. Engaging institutional faculty, students, and stakeholders is critical for the successful adoption of e-learning [35]. The cocreation process and involvement of stakeholders throughout the ASPIRE process facilitates the feeling of ownership of the materials produced, thereby leading to use and reuse. To ensure the widespread adoption of e-learning innovation, it is crucial to engage with and consider the

perspectives of diverse stakeholders, as the implementation of e-learning necessitates changes in teaching, learning, management, and infrastructure within the institution [36].

In our study, although RLOs reached a significant number of users, the proportion of users who completed RLOs (5.6% - 84.4%) and nonbounced users (16.3% - 81.5%) was relatively low for most RLOs. Chen et al [37] reported a similar finding, with a low nonbounced rate of 40% for an undergraduate online course. The wide discrepancy in the completion and nonbounced rates among RLOs may be attributed to the different levels of learning across various learner groups. RLOs integrated into the curriculum as teaching and learning materials achieved a completion rate greater than 70% (RLO 8 and RLO 23), suggesting that learners are more invested in completing these RLOs. Conversely, RLO 7 attracted the highest number of users (n=1014); yet, its completion rate was only 7.4% (75/1014). Our findings show that Google Analytics is beneficial for examining the fidelity of RLO usage, as it provides insights into learners' behaviors. Assessing the fidelity of RLO usage is crucial for educators and institutions to gauge the effectiveness of RLOs in teaching and learning [38].

Users continued to access RLOs after the conclusion of the ACoRD project, which indicates the sustainability of RLOs. The main strategies that we identified were the integration of RLOs into the existing health care curriculum and developing RLOs by using a robust cocreation methodology. Identifying the learning needs of teachers and learners prior to the development of RLOs facilitated their integration into the curriculum [19]. We also implemented the educator-as-champion strategy, which involves educators in content development and subsequently utilizing RLOs in their teaching. e-Learning champions among academicians in higher educational institutions are the key players in fostering the integration of technology into teaching and learning [39].

Limitations

Our study has several limitations. Since each RLO was developed and launched at different times, discrepancies in the implementation periods made data interpretation challenging. We used Google Analytics to capture the reach of our RLOs. However, we could not confirm whether access came from unique users or unique IP addresses. There is a possibility that some users could have accessed RLOs multiple times by using different IP addresses. As this study employs a pragmatic approach to capture evaluation outcomes in real-time, missing data were inevitable; however, it did not substantially impact the overall findings. Strategies were implemented pragmatically throughout the project, with practical consideration and changes made to address implementation issues; as a result, we were unable to identify and measure the effectiveness of each implementation strategy.

Conclusion

We employed the RE-AIM framework to systematically evaluate the implementation success of e-learning resources, identifying gaps and strategies for improvement. This study highlights the development, implementation process, and implementation outcome indicators of open-access RLOs in health care education. To enhance future e-learning implementation efforts, we recommend incorporating the RE-AIM framework into outcome evaluations to provide a more comprehensive evaluation of e-learning implementation.

Acknowledgments

We thank Irmi Zarina Ismail from Universiti Putra Malaysia for her contribution to the Advancing Co-creation of Reusable Learning Objects to Digitise Healthcare Curriculum (ACoRD) project. This study was funded by the European Union Erasmus+ Program under the ACoRD project (reference 598935-EPP-1-2018-1-UK-EPPKA2-CBHE-JP). The funders have no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability

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

Conflicts of Interest

None declared.

Multimedia Appendix 1
Preknowledge and postknowledge and confidence scores for each reusable learning object.
[DOCX File, 31 KB - mededu_v11i1e63882_app1.docx]

References

1. Wentling TL, Waight C, Gallaher J, La Fleur J, Wang C, Kanfer A. E-learning: a review of literature. ResearchGate. 2000 Sep. URL: https://www.researchgate.net/publication/331938876_E-learning_A_review_of_literature [accessed 2025-05-21]
2. Wang ZY, Zhang LJ, Liu YH, et al. The effectiveness of e-learning in continuing medical education for tuberculosis health workers: a quasi-experiment from China. Infect Dis Poverty 2021 May 18;10(1):72. [doi: 10.1186/s40249-021-00855-y] [Medline: 34006313]

3. Sadeghi R, Sedaghat MM, Sha Ahmadi F. Comparison of the effect of lecture and blended teaching methods on students' learning and satisfaction. J Adv Med Educ Prof 2014 Oct;2(4):146-150. [Medline: 25512938]
4. Ruiz JG, Mintzer MJ, Leipzig RM. The impact of e-learning in medical education. Acad Med 2006 Mar;81(3):207-212. [doi: 10.1097/00001888-200603000-00002] [Medline: 16501260]
5. Naveed QN, Qureshi MRN, Tairan N, et al. Evaluating critical success factors in implementing e-learning system using multi-criteria decision-making. PLoS One 2020;15(5):e0231465. [doi: 10.1371/journal.pone.0231465] [Medline: 32365123]
6. Barteit S, Guzek D, Jahn A, Bärnighausen T, Jorge MM, Neuhann F. Evaluation of e-learning for medical education in low- and middle-income countries: a systematic review. Comput Educ 2020 Feb;145:103726. [doi: 10.1016/j.compedu.2019.103726] [Medline: 32565611]
7. Proctor E, Silmere H, Raghavan R, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health 2011 Mar;38(2):65-76. [doi: 10.1007/s10488-010-0319-7] [Medline: 20957426]
8. Palmerola ED. Clarification evaluation of e-learning implementation: a developmental research design. Am J Educ Technol 2024;3(2):119-127. [doi: 10.54536/ajet.v3i2.2890]
9. Kirkpatrick D, Kirpatrick JD. Evaluating training programs: the four levels. American J Eval 2011;19(2):259-261. [doi: 10.1016/S1098-2140(99)80206-9]
10. Stufflebeam D. The CIPP model for evaluation. In: Evaluation Models: Springer Dordrecht; 2002:279-317. [doi: 10.1007/0-306-47559-6]
11. Gandomkar R. Comparing Kirkpatrick's original and new model with CIPP evaluation model. J Adv Med Educ Prof 2018 Apr;6(2):94-95. [Medline: 29607338]
12. Peterson C. Bringing ADDIE to life: instructional design at its best. ERIC. 2003. URL: https://eric.ed.gov/?id=EJ822355 [accessed 2025-05-19]
13. Holtrop JS, Estabrooks PA, Gaglio B, et al. Understanding and applying the RE-AIM framework: clarifications and resources. J Clin Transl Sci 2021;5(1):e126. [doi: 10.1017/cts.2021.789] [Medline: 34367671]
14. Gisondi MA, Keyes T, Zucker S, Bumgardner D. Teaching LGBTQ+ health, a web-based faculty development course: program evaluation study using the RE-AIM framework. JMIR Med Educ 2023 Jul 21;9:e47777. [doi: 10.2196/47777] [Medline: 37477962]
15. Golden RE, Sanders AM, Frayne SM. RE-AIM applied to a primary care workforce training for rural providers and nurses: the Department of Veterans Affairs' Rural Women's Health Mini-Residency. Front Health Serv 2023;3:1205521. [doi: 10.3389/frhs.2023.1205521] [Medline: 38028946]
16. Bath-Hextall F, Wharrad H, Leonardi-Bee J. Teaching tools in evidence based practice: evaluation of reusable learning objects (RLOs) for learning about meta-analysis. BMC Med Educ 2011 May 4;11(1):18. [doi: 10.1186/1472-6920-11-18] [Medline: 21542905]
17. Lim HM, Ng CJ, Wharrad H, et al. Knowledge transfer of e-learning objects: lessons learned from an intercontinental capacity building project. PLoS ONE 2022;17(9):e0274771. [doi: 10.1371/journal.pone.0274771] [Medline: 36126036]
18. Wharrad H, Windle R, Taylor M. Chapter 3: designing digital education and training for health. In: Konstantinidis ST, Bamidis PD, Zary N, editors. Digital Innovations in Healthcare Education and Training: Academic Press; 2021:31-45. [doi: 10.1016/B978-0-12-813144-2.00003-9]
19. Lim HM, Ng CJ, Teo CH, et al. Prioritising topics for developing e-learning resources in healthcare curricula: a comparison between students and educators using a modified Delphi survey. PLoS ONE 2021;16(6):e0253471. [doi: 10.1371/journal.pone.0253471] [Medline: 34166432]
20. Lim HM, Teo CH, Hong WH, Lee YK, Lee PY, Ng CJ. Empowering students in co-creating e-learning resources through a virtual hackathon. TAPS 2024;9(4):84-87. [doi: 10.29060/TAPS.2024-9-4/CS3263]
21. The ACoRD Project. 2018. URL: https://acord.my [accessed 2025-05-19]
22. HELM Open. University of Nottingham. URL: https://www.nottingham.ac.uk/helmopen [accessed 2025-05-19]
23. MERLOT multimedia educational resource for learning and online teaching. MERLOT. URL: https://www.merlot.org/merlot/viewMaterial.htm?id=85072 [accessed 2025-05-19]
24. Kirkpatrick DL. The four levels of evaluation. In: Brown SM, Seidner CJ, editors. Evaluating Corporate Training: Models and Issues: Springer Netherlands; 1998:95-112. [doi: 10.1007/978-94-011-4850-4_5]
25. Ragsdale JW, Berry A, Gibson JW, et al. Evaluating the effectiveness of undergraduate clinical education programs. Med Educ Online 2020 Dec;25(1):1757883. [doi: 10.1080/10872981.2020.1757883] [Medline: 32352355]
26. Windle RJ, Wharrad H, McCormick D, Laverty H, Taylor MG. Sharing and reuse in OER: experiences gained from open reusable learning objects in health. JIME 2010(1):4. [doi: 10.5334/2010-4]
27. Philpot LM, Ahrens DJ, Eastman RJ, et al. Implementation of e-learning solutions for patients with chronic pain conditions. Digit Health 2023;9:20552076231216404. [doi: 10.1177/20552076231216404] [Medline: 38033514]
28. Gallagher S, Murphy P. Implementing inter-institutional lifelong sustainability education: the UNI-ECO e-learning case study. Ir J Acad Pract 2024;11(2). [doi: 10.21427/D7QP44]

29. Hardie P, Donnelly P, Greene E, et al. The application of reusable learning objects (RLOs) in preparation for a simulation laboratory in medication management: an evaluative study. Teaching Learning Nurs 2021 Oct;16(4):301-308. [doi: 10.1016/j.teln.2021.05.002]
30. Redmond C, Davies C, Cornally D, et al. Using reusable learning objects (RLOs) in wound care education: undergraduate student nurse's evaluation of their learning gain. Nurse Educ Today 2018 Jan;60:3-10. [doi: 10.1016/j.nedt.2017.09.014] [Medline: 28987896]
31. de Leeuw R, de Soet A, van der Horst S, Walsh K, Westerman M, Scheele F. How we evaluate postgraduate medical e-learning: systematic review. JMIR Med Educ 2019 Apr 5;5(1):e13128. [doi: 10.2196/13128] [Medline: 30950805]
32. Lymn JS, Bath-Hextall F, Wharrad HJ. Pharmacology education for nurse prescribing students - a lesson in reusable learning objects. BMC Nurs 2008 Jan 23;7(1):2. [doi: 10.1186/1472-6955-7-2] [Medline: 18215261]
33. Ferguson M, Brandreth M, Brassington W, Leighton P, Wharrad H. A randomized controlled trial to evaluate the benefits of a multimedia educational program for first-time hearing aid users. Ear Hear 2016;37(2):123-136. [doi: 10.1097/AUD.0000000000000237] [Medline: 26565785]
34. El Nsouli D, Nelson D, Nsouli L, et al. The application of Kirkpatrick's evaluation model in the assessment of interprofessional simulation activities involving pharmacy students: a systematic review. Am J Pharm Educ 2023 Aug;87(8):100003. [doi: 10.1016/j.ajpe.2023.02.003] [Medline: 37597909]
35. Vovides Y, Chale SB, Gadhula R, et al. A systems approach to implementation of e-learning in medical education: five MEPI schools' journeys. Acad Med 2014 Aug;89(8 Suppl):S102-S106. [doi: 10.1097/ACM.0000000000000347] [Medline: 25072558]
36. de Souza Rodrigues MA, Chimenti P, Nogueira ARR. An exploration of e-learning adoption in the educational ecosystem. Educ Inf Technol 2021 Jan;26(1):585-615. [doi: 10.1007/s10639-020-10276-3]
37. Chen Y, Deng X, Huang Q, Luo H. Patterns and trends in online learning behaviors: evidence from Google analytics. Presented at: 2021 IEEE International Conference on Engineering, Technology & Education (TALE); Dec 5-8, 2021; Wuhan, Hubei Province, China. [doi: 10.1109/TALE52509.2021.9678689]
38. Mudawi NA, Pervaiz M, Alabduallah BI, et al. Predictive analytics for sustainable e-learning: tracking student behaviors. Sustainability 2023;15(20):14780. [doi: 10.3390/su152014780]
39. Gachago D, Morkel J, Hitge L, van Zyl I, Ivala E. Developing e-learning champions: a design thinking approach. Int J Educ Technol High Educ 2017 Dec;14(1):30. [doi: 10.1186/s41239-017-0068-8]

Abbreviations

ACoRD: Advancing Co-creation of Reusable Learning Objects to Digitise Healthcare Curriculum
ASPIRE: Aim, Storyboarding, Populating, Implementing, Release, and Evaluation
RE-AIM: Reach, Effectiveness, Adoption, Implementation, and Maintenance
RLO: reusable learning object

Edited by D Chartash; submitted 02.07.24; peer-reviewed by L Philpot, N Mora; revised version received 21.04.25; accepted 25.04.25; published 23.07.25.

Please cite as:
Lim HM, Teo CH, Lee YK, Lee PY, Krishnan K, Abu Hassan ZF, Yong PVC, Yap WH, Sellappans R, Ayub E, Hassan N, Shariff Ghazali S, Nasharuddin NA, Jahn Kassim PS, Idris F, Karlgren K, Stathakarou N, Mordt P, Konstantinidis S, Taylor M, Poussa C, Wharrad H, Ng CJ
Implementation Outcomes of Reusable Learning Objects in Health Care Education Across Three Malaysian Universities: Evaluation Using the RE-AIM Framework
JMIR Med Educ 2025;11:e63882
URL: https://mededu.jmir.org/2025/1/e63882
doi:10.2196/63882

Resident Physician Recognition of Tachypnea in Clinical Simulation Videos in Japan: Cross-Sectional Study

Kiyoshi Shikino[1,2*], MD, MHPE, PhD; Yuji Nishizaki[3*], MD, MPH, PhD; Sho Fukui[4], MD, MPH; Koshi Kataoka[3], PhD; Daiki Yokokawa[2], MD, PhD; Taro Shimizu[5], MD, MPH, PhD; Yu Yamamoto[6], MD; Kazuya Nagasaki[7], MD, PhD; Hiroyuki Kobayashi[8], MD, PhD; Yasuharu Tokuda[9,10], MD, MPH, PhD

[1]Department of Community-Oriented Medical Education, Graduate School of Medicine, Chiba University, 1-8-1, Inohana, Chu-ou-ku, Chiba, Japan

[10]Tokyo Foundation for Policy Research, Tokyo, Japan

[2]Department of General Medicine, Chiba University Hospital, Chiba, Japan

[3]Division of Medical Education, School of Medicine, Juntendo University, Tokyo, Japan

[4]General Medicine, School of Medicine, Kyorin University, Tokyo, Japan

[5]Department of Diagnostic and Generalist Medicine, Dokkyo Medical University, Mibu, Japan

[6]Division of General Medicine, Center for Community Medicine, Jichi Medical University, Shimotsuke, Japan

[7]University of Tsukuba, Tsukuba, Japan

[8]Department of Internal Medicine, Mito Kyodo General Hospital, Mito, Japan

[9]Muribushi Okinawa Center for Teaching Hospitals, Okinawa, Japan

[*]these authors contributed equally

Corresponding Author:
Kiyoshi Shikino, MD, MHPE, PhD
Department of Community-Oriented Medical Education, Graduate School of Medicine, Chiba University, 1-8-1, Inohana, Chu-ou-ku, Chiba, Japan

Abstract

Background: Traditional assessments of clinical competence using multiple-choice questions (MCQs) have limitations in the evaluation of real-world diagnostic abilities. As such, recognizing non-verbal cues, like tachypnea, is crucial for accurate diagnosis and effective patient care.

Objective: This study aimed to evaluate how detecting such cues impacts the clinical competence of resident physicians by using a clinical simulation video integrated into the General Medicine In-Training Examination (GM-ITE).

Methods: This multicenter cross-sectional study enrolled first- and second-year resident physicians who participated in the GM-ITE 2022. Participants watched a 5-minute clinical simulation video depicting a patient with acute pulmonary thromboembolism, and subsequently answered diagnostic questions. Propensity score matching was applied to create balanced groups of resident physicians who detected tachypnea (ie, the detection group) and those who did not (ie, the non-detection group). After matching, we compared the GM-ITE scores and the proportion of correct clinical simulation video answers between the two groups. Subgroup analyses assessed the consistency between results.

Results: In total, 5105 resident physicians were included, from which 959 pairs were identified after the clinical simulation video. Covariates were well balanced between the detection and non-detection groups (standardized mean difference <0.1 for all variables). Post-matching, the detection group achieved significantly higher GM-ITE scores (mean [SD], 47.6 [8.4]) than the non-detection group (mean [SD], 45.7 [8.1]; mean difference, 1.9; 95% CI, 1.1 - 2.6; *P*=.041). The proportion of correct clinical simulation video answers was also significantly higher in the detection group (39.2% vs 3.0%; mean difference, 36.2%; 95% CI, 32.8 - 39.4). Subgroup analyses confirmed consistent results across sex, postgraduate years, and age groups.

Conclusions: Overall, this study revealed that detecting non-verbal cues like tachypnea significantly affects clinical competence, as evidenced by higher GM-ITE scores among resident physicians. Integrating video-based simulations into traditional MCQ examinations enhances the assessment of diagnostic skills by providing a more comprehensive evaluation of clinical abilities. Thus, recognizing non-verbal cues is crucial for clinical competence. Video-based simulations offer a valuable addition to traditional knowledge assessments by improving the diagnostic skills and preparedness of clinicians.

(JMIR Med Educ 2025;11:e72640) doi:10.2196/72640

KEYWORDS

clinical competence; clinical simulation video; general medicine in-training examination; non-verbal information; tachypnea

Introduction

Detecting non-verbal cues is an essential skill that resident physicians must acquire during their clinical training. However, accurately assessing detection ability is challenging [1]. In traditional text-based examinations, non-verbal information, such as respiratory rate or breathing patterns, is written, making true recognition of these cues clear.

The Japan Institute for Advancement of Medical Education Program (JAMEP) created the General Medicine In-Training Examination (GM-ITE) to evaluate resident physicians' clinical knowledge across Japan [2]. This 2-hour exam comprising 80 multiple-choice questions (MCQs) [2] provides practical feedback on training programs by identifying areas for improvement through objective and reliable clinical competency assessment [3]. The GM-ITE is a validated examination that effectively assesses core medical knowledge and skills. However, its text-based design makes it difficult to evaluate the examinee's ability to detect non-verbal cues [4]. A sample GM-ITE question illustrating the typical MCQ format and content is provided in Multimedia Appendix 1 [5].

To address this gap, we developed and integrated a clinical simulation video into the GM-ITE [6], offering an effective assessment tool by providing contextualized real-world scenarios that require resident physicians to apply knowledge dynamically, mirroring actual patient interactions [7,8]. By incorporating non-verbal information with verbal cues, video simulation offers a more authentic evaluation of clinical practice. This simulation involved a medical interview and a physical examination scenario, requiring participants to actively identify critical non-verbal cues, such as tachypnea. The clinical simulation video successfully assessed multiple domains of clinical competencies [8]. However, the association between the ability to detect non-verbal cues and clinical competency remains unknown.

Herein, we explored whether the ability to detect non-verbal cues, like tachypnea, in the clinical simulation video correlates with higher GM-ITE scores and correct clinical simulation video answers. We hypothesized that resident physicians who could accurately identify tachypnea would demonstrate superior clinical competence, yielding higher GM-ITE scores.

Methods

Study Design, Setting, and Participants

We conducted a multicenter cross-sectional study of GM-ITE for the academic year 2022 (GM-ITE 2022). GM-ITE employs a methodology similar to that of internal medicine (IM)-ITE. In Japan, the GM-ITE 2022 includes many residency programs in teaching hospitals, with other hospitals allowing voluntary participation. Consequently, 9011 resident physicians from 662 Japanese hospitals participated between January 17 and January 30, 2023. After the final question, participants were asked to voluntarily watch the clinical simulation video to provide a diagnosis, followed by a questionnaire to collect information on demographics and residency programs. This study included all resident physicians who participated in GM-ITE 2022 at one of the study sites, excluding those who did not participate in or answer the clinical simulation video question, or did not provide informed consent for study participation.

Ethical Considerations

This study was approved by the Ethical Review Committee of the Japan Organization for Advancing Medical Education (approval no. 23‐26). All participants provided informed consent before participating. The study was conducted in accordance with the ethical standards and principles of the Declaration of Helsinki. Informed consent was obtained from all participants for the publication of identifying information in an online open-access publication. All participants read and signed an informed consent form before participating. To ensure confidentiality, all data were anonymized prior to analysis. No compensation was provided for involvement in the study. In accordance with ethical standards and journal policies, we obtained explicit informed consent from all actors appearing in the video material associated with this study. The actors acknowledged and agreed that the videos would be published as part of the study.

Procedures

Innovative Examination Using High-Quality Video Simulation

This study used a clinical simulation video developed and introduced in a prior study [9]. In the 5-minute video filmed from a resident physician's perspective, the resident physician conducted a medical interview and physical examination, with the camera capturing the patient's and family members' verbal and non-verbal responses. The video depicts a 40-year-old male patient with a recent history of lower limb fracture who presented to the emergency room with fainting as the chief concern. The correct diagnosis was acute pulmonary thromboembolism. This scenario was designed to challenge residents' clinical reasoning skills. The patient exhibited physical signs, including tachypnea, jugular venous distension, and a right lower limb fracture, with actual heart sounds presented during auscultation. Crucially, tachypnea (28 breaths per minute) and respiratory patterns were not highlighted in the text, but presented as non-verbal cues, requiring participants to recognize these signs through observation, without explicit textual focus.

The video was produced by a professional television production company under the guidance of three JAMEP medical supervisors and three authors who ensured the medical accuracy and educational relevance. The use of professional actors and their added effects, such as heart sounds, enhanced the scenario realism of the scenario and aligned it with the objectives of clinical residency training in Japan.

After watching the clinical simulation video, participants answered the following clinical simulation video questions [10]:

Q1. Please state the most likely diagnosis for this patient (free text).

Q2. Please state the three pertinent positive clinical information related to the most likely diagnosis (free text).

For Q1, multiple disease name patterns indicating "acute pulmonary thromboembolism" (eg, acute pulmonary embolism or pulmonary thromboembolism) were accepted as correct. For Q2, the pertinent positive clinical findings included tachypnea, jugular venous distension, and a history of right lower limb fracture. Answers were evaluated based on whether key clinical signs were correctly identified, further contributing to diagnostic ability assessment. Two authors independently assessed, discussed, identified, and agreed upon the answers. The interrater reliability was measured using the κ coefficient, which indicated almost perfect agreement for Q1 (κ=0.91) and substantial agreement for Q2 (κ=0.82) [11].

Data Collection

This study used information from the GM-ITE score and questionnaire, including sex, postgraduate year (PGY 1 or 2), rotation in the general medicine department, duration of rotation, number of monthly night shifts, average number of assigned inpatients, self-study time per day, and categorized weekly duty hours.

Exposure and Outcome Measurement Data

The exposure of interest was whether the resident physicians detected tachypnea in the innovative examination clinical simulation video. This non-verbal cue was critical for the correct diagnosis. The primary outcome was the total GM-ITE score, which ranged from 0 to 80 points. The secondary outcome included the results of the clinical simulation video (correct or incorrect).

Statistical Analyses

Initially, we described participant-level information by groups that detected tachypnea (ie, tachypnea-detection group) and did not detect tachypnea (ie, tachypnea-non-detection group) in the clinical simulation video using appropriate summary statistics.

To adjust for potential confounders, we conducted propensity score matching analysis. Propensity scores for the probability of detecting tachypnea were calculated using a logistic regression model incorporating sex, PGY, rotation in the general medicine department, rotation duration, number of night shift duties per month, average number of assigned inpatients, self-study time, and weekly duty hours. We applied a greedy matching algorithm using a logit of propensity scores with calipers equal to 0.2 of the standard deviation of logit-propensity scores [12]. We assessed the balance of covariates between the groups before and after matching using the standardized mean difference (SMD) [13]. Covariates were considered well-balanced if the absolute SMD was <0.1.

The GM-ITE scores and clinical simulation video results were compared before and after propensity score matching. T-tests were applied to assess differences in the GM-ITE scores, while the χ^2 test was used for the clinical simulation video results. Stratified subgroup analyses by gender (male or female), grade (PGY-1 or 2), and age (unknown, <27 years, and ≥27 years) were conducted to evaluate robustness.

All analyses were conducted using the Statistical Analysis System version 9.4 for Windows (SAS Institute Inc., Cary, NC, USA) and JMP version 17 for Windows (SAS Institute Inc., Cary, NC, USA) according to the Strengthening the Reporting of Observational Studies in Epidemiology guidelines. A complete case analysis was performed. The mean differences and 95% CI were reported for all analyses.

Results

Baseline Characteristics

Overall, 5105 first- and second-year resident physicians participated in this study (Figure 1). There were no missing data for any covariates or outcome variables. Resident physicians were divided based on their ability to detect tachypnea, the critical non-verbal cue (Table 1). Among the non-verbal cues included in the clinical simulation video, tachypnea exhibited the largest difference in the identification rates between correct and incorrect responders (Multimedia Appendix 2). Overall, tachypnea was overlooked and detected in 4146 and 959 residents, respectively. Before matching, several covariates demonstrated notable imbalances between the tachypnea non-detection and detection groups. Specifically, the proportion of male residents was higher in the non-detection group (69.2% vs 64.7%; SMD=0.219), while the proportion of first-year postgraduate residents was higher in the non-detection group (51.7% vs 45.3%; SMD=0.309). Age distribution also showed discrepancies, with younger residents, particularly those aged 26 years, being more common in the detection group (34.1% vs 31.1%; SMD=0.312). These imbalances were addressed through propensity score matching identifying 959 participants to achieve well-balanced covariates across groups (SMD <0.1 for all covariates).

Figure 1. Flowchart of the study. CSV-IE: innovative examination clinical simulation video; GM-ITE: General Medicine In-Training Examination.

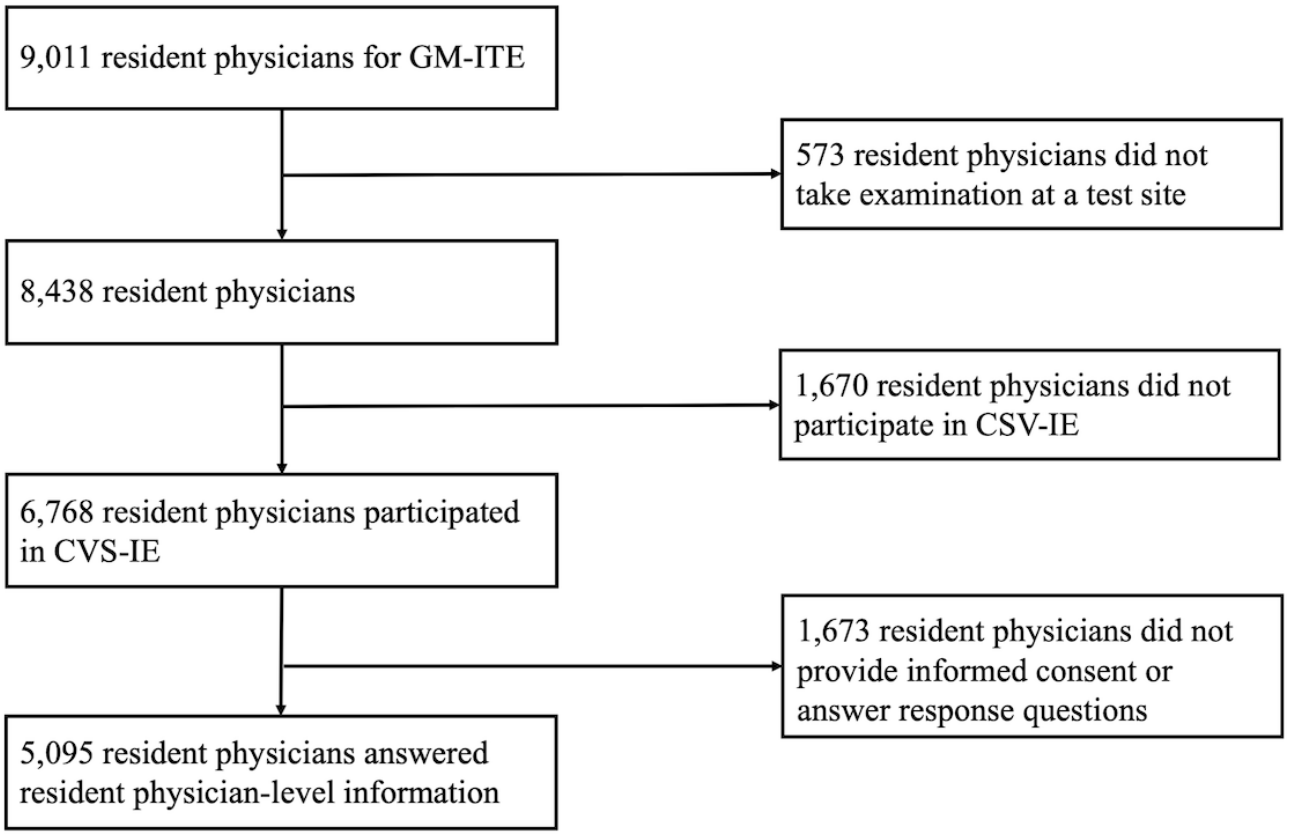


Table . Baseline characteristics of the participants included before and after propensity score matching.

Characteristic	All, n (%) (N=5105)	Before propensity score matching			After propensity score matching		
		Tachypnea non-detection group, n (%), (n=4146)	Tachypnea detection group, n (%), (n=959)	SMD	Tachypnea non-detection group, n (%), (n=959)	Tachypnea detection group, n (%), (n=959)	SMD
Sex				0.219			0.015
Men	3488 (68.3)	2867 (69.2)	621 (64.7)		614 (64.0)	621 (64.7)	
Women	1617 (31.7)	1279 (30.8)	338 (35.3)		345 (36.0)	338 (35.3)	
PGY[a]				0.309			0.036
1	2576 (50.5)	2142 (51.7)	434 (45.3)		451 (47.0)	434 (45.3)	
2	2529 (49.5)	2004 (48.3)	525 (54.7)		508 (53.0)	525 (54.7)	
Age (years)				0.312			0.079
24	166 (3.3)	138 (3.3)	28 (2.9)		31 (3.2)	28 (2.9)	
25	1024 (20.1)	816 (19.7)	208 (21.7)		205 (21.4)	208 (21.7)	
26	1618 (31.7)	1291 (31.1)	327 (34.1)		343 (35.8)	327 (34.1)	
27	1060 (20.8)	860 (20.8)	200 (20.9)		208 (21.7)	200 (20.9)	
28	497 (9.7)	419 (10.1)	78 (8.1)		62 (6.8)	78 (8.1)	
29	211 (4.1)	183 (4.4)	28 (2.9)		24 (2.5)	28 (2.9)	
≥30	482 (9.4)	401 (9.7)	81 (8.5)		78 (8.1)	81 (8.5)	
Unknown	47 (0.9)	38 (0.9)	9 (0.9)		8 (0.8)	9 (0.9)	
General medicine rotation				0.118			0.017
No	2341 (45.9)	1860 (44.9)	481 (50.2)		473 (49.3)	481 (50.2)	
Yes	2764 (54.1)	2286 (55.1)	478 (49.8)		486 (50.8)	478 (49.8)	
Internal Medicine rotation (months)				0.237			0.058
0 - 5	1362 (26.7)	1117 (26.9)	245 (25.5)		259 (27.0)	245 (25.5)	
6 - 10	3149 (61.7)	2552 (61.5)	597 (62.3)		597 (62.3)	597 (62.3)	
11 - 15	523 (10.2)	418 (10.1)	105 (11.0)		94 (9.8)	105 (11.0)	
16 - 20	53 (1.0)	44 (1.1)	9 (0.9)		6 (0.6)	9 (0.9)	
≥21	18 (0.4)	15 (0.4)	3 (0.3)		3 (0.3)	3 (0.3)	
ED[b] duty per month				0.103			0.089
0	131 (2.6)	109 (2.6)	22 (2.3)		19 (2.0)	22 (2.3)	
1 - 2	841 (16.5)	703 (17.0)	138 (14.4)		129 (13.5)	138 (14.4)	
3 - 5	3635 (71.2)	2953 (71.2)	682 (71.1)		710 (74.0)	682 (71.1)	
≥6	480 (9.4)	366 (8.8)	114 (11.9)		95 (9.9)	114 (11.9)	
Unknown	18 (0.3)	15 (0.4)	3 (0.3)		6 (0.6)	3 (0.3)	
Average number of assigned inpatients				0.145			0.063
0 - 4	1956 (38.3)	1578 (38.0)	378 (39.4)		386 (40.3)	378 (39.4)	
5 - 9	2540 (49.8)	2077 (50.1)	463 (48.3)		466 (48.6)	463 (48.3)	
10 - 14	385 (7.5)	306 (7.4)	79 (8.2)		79 (8.2)	79 (8.2)	

Characteristic	All, n (%) (N=5105)	Before propensity score matching			After propensity score matching		
		Tachypnea non-detection group, n (%), (n=4146)	Tachypnea detection group, n (%), (n=959)	SMD	Tachypnea non-detection group, n (%), (n=959)	Tachypnea detection group, n (%), (n=959)	SMD
≥15	113 (2.2)	87 (2.1)	26 (2.7)		19 (2.0)	26 (2.7)	
Unknown	111 (2.2)	98 (2.4)	13 (1.4)		9 (0.9)	13 (1.4)	
Self-study time per day (minutes)				0.198			0.026
0	2312 (45.3)	1882 (45.4)	430 (44.8)		435 (45.4)	430 (44.8)	
1 - 30	1965 (38.5)	1574 (38.0)	391 (40.8)		382 (39.8)	391 (40.8)	
31 - 60	580 (11.4)	486 (11.7)	94 (9.8)		94 (9.8)	94 (9.8)	
61 - 90	156 (3.0)	126 (3.0)	30 (3.1)		32 (3.3)	30 (3.1)	
≥91	92 (1.8)	78 (1.9)	14 (1.5)		16 (1.7)	14 (1.5)	
Duty (hours per week)				0.089			0.074
Category 1 (<60)	2565 (50.2)	2101 (50.7)	464 (48.4)		464 (48.4)	464 (48.4)	
Category 2 (60–79)	1817 (35.6)	1456 (35.1)	361 (37.6)		361 (37.6)	361 (37.6)	
Category 3 (≥80)	723 (14.2)	589 (14.2)	134 (14.0)		134 (14.0)	134 (14.0)	

[a]PGY: postgraduate year.

[b]ED: emergency department.

GM-ITE Scores

Before matching (Table 2), resident physicians in the tachypnea-detection group exhibited significantly higher GM-ITE scores (mean [SD] 47.6 [8.4] points vs 45.3 [8.0] points), with a mean difference of 2.3 points (95% CI 1.7 - 2.8). After propensity score matching, the tachypnea-detection group maintained a higher GM-ITE score (mean [SD] 47.6 [8.4] points vs 45.7 [8.1] points; mean difference 1.9 points [95% CI 1.1 - 2.6]). These results were consistent across all subgroup analyses.

Table . Comparison of General Medicine In-Training Examination (GM-ITE) scores between the tachypnea non-detection and detection groups after propensity score matching.

Characteristic	GM-ITE score, mean (SD)		
	Tachypnea non-detection group	Tachypnea detection group	Difference in the mean value (95% CI)
Before matching			
Total participants	45.3 (8.0)	47.6 (8.4)	2.3 (1.7 to 2.8)
After matching			
Total participants	45.7 (8.1)	47.6 (8.4)	1.9 (1.1 to 2.6)
Subgroup analyses after matching			
Sex			
Male	46.0 (8.4)	48.1 (8.8)	2.1 (1.2 to 3.1)
Female	45.2 (7.6)	46.6 (7.5)	1.4 (0.2 to 2.5)
Grade			
PGY-1[a]	45.2 (7.8)	46.5 (7.9)	1.9 (1.1 to 2.6)
PGY-2	46.2 (8.4)	48.5 (8.6)	2.3 (1.3 to 3.4)
Age (years)			
<27	47.2 (7.8)	48.6 (8.1)	2.6 (1.4 to 3.8)
≥27	43.5 (8.2)	46.2 (8.5)	1.0 (-2.5 to 4.5)
Unknown	42.1 (5.8)	47.8 (8.7)	5.7 (-2.0 to 13.3)

[a]PGY: postgraduate year.

Clinical Simulation Video Results

The proportion of correct answers within the tachypnea-detection group was significantly high both prior to matching (39.2% vs 3.1%; difference 36.1% [95% CI 33.0 - 39.2]) and following matching (39.2% vs 3.0%; differences 36.2% [95% CI: 32.8 - 39.4]; Multimedia Appendix 3). The proportion of correct clinical simulation video answers was consistently greater in the tachypnea detection group across subgroup analyses. Formal interaction tests were conducted to evaluate the association between tachypnea and the correct answer rate divided by sex, PGY level, and age (Multimedia Appendix 4). The interaction tests revealed no statistically significant difference between male and female participants (P=.692), indicating that sex did not substantially influence the relationship between the tachypnea detection performance and the correct answer rate. In contrast, a significant interaction was identified between PGY-1 and PGY-2 participants (P=.003). Furthermore, a significant interaction was observed between participants aged <27 years and those aged ≥27 years (P=.021).

Discussion

Principal Findings and Comparison With Previous Works

This study highlights the importance of recognizing non-verbal cues, such as tachypnea, in clinical reasoning. Resident physicians who identified tachypnea achieved significantly higher GM-ITE and correct clinical simulation video answers. Although identifying tachypnea as a non-verbal cue in the clinical simulation video was important, it was not the sole determinant of correct diagnosis. Achieving a correct response required integrating multiple findings, including jugular venous distension, cardiac murmur, and limb injury. Therefore, tachypnea detection was associated with—but not equivalent to—correctly answering the clinical simulation video question. Furthermore, the primary analysis examined the relationship between tachypnea detection and overall GM-ITE scores, which are independent of the video-based assessment and reflect broader clinical knowledge. Consistent results in propensity score-matched and stratified subgroup analyses emphasized the direct relationship between the ability to detect non-verbal information and clinical competence, underlining the need to train resident physicians to observe and interpret non-verbal cues during patient interactions. These skills are essential for accurate diagnosis and effective patient care [14-17].

The substantial GM-ITE score difference between the two groups demonstrates the value of non-verbal cue recognition in clinical competency. Video-based simulations provide a unique opportunity to address this gap, offering realistic scenarios that mirror actual clinical encounters [18]. By integrating non-verbal with verbal information, video simulations enhance diagnostic capabilities potentially underdeveloped through text-based evaluations alone.

Our findings further indicated that clinical experience, as reflected in PGY, significantly influence the ability to recognize non-verbal cues. PGY-2 residents outperformed PGY-1 residents, indicating that experiential learning plays a pivotal role in developing these perceptual skills [19]. Beyond the PGY level, analysis of the unmatched cohort revealed additional factors associated with improved cue recognition. For example,

the detection group more frequently reported a strong interest in general medicine and prior experience with video-based clinical reasoning, indicating that both motivational and educational factors may enhance observational accuracy. Furthermore, residents with greater monthly patient exposure demonstrated higher detection rates, indicating that regular clinical exposure helps refine perceptual skills. Together, these findings highlight several modifiable factors that could be targeted through future educational interventions. Further research should explore how medical training and curricula can be optimized to foster the development of diagnostic observation skills early in clinical education.

Including video-based assessments in the GM-ITE reinforces its potential to complement traditional MCQs. Video simulations enable a more comprehensive evaluation of diagnostic abilities by integrating complex, real-world scenarios [20]. They assess clinical knowledge and challenge resident physicians to process subtle cues, bridging the gap between theoretical understanding and practical application [21-23]. This also reinforces the value of integrating video-based assessments with traditional MCQ examinations to comprehensively evaluate a physician's diagnostic capabilities [21,22]. Overall, this finding suggests potential areas for further investigation into whether perceptual abilities, training experiences, or other underlying factors contribute to differences [24]. The significant proportion of second-year resident physicians in the detection group indicates the impact of clinical experience on non-verbal cue recognition, emphasizing the importance of developing skills during medical training.

Although our results showed that resident physicians who detected tachypnea also had higher GM-ITE scores; this association does not indicate redundancy between video- and text-based assessments. Rather, it suggests that the ability to recognize non-verbal cues is an important component of clinical reasoning that complements traditional knowledge-based assessments. GM-ITE MCQs primarily evaluate declarative knowledge and structured reasoning, whereas the clinical simulation video captures real-time perceptual and interpretive skills, such as visual observation and contextual integration. Therefore, this video-based format offers added value by assessing diagnostic competencies that may not be fully captured through text-based questions alone.

Using video-based simulations like clinical simulation videos significantly enhances traditional MCQ examinations by enabling the assessment of resident physicians' ability to integrate verbal and non-verbal information, thereby replicating real-world clinical scenarios, and improving the assessment of clinical competence [24]. Our findings support the inclusion of video-based simulations in medical education assessments [21]. Video simulations can better prepare resident physicians for the complexities of patient care by mimicking actual clinical situations [25,26]. However, future studies should investigate how repeated exposure to video-based assessments impacts diagnostic accuracy and confidence, as well as their influence on clinical outcomes.

The consistency in GM-ITE scores and clinical simulation video results after propensity score matching and subgroup stratification indicated the robustness of our findings [27]. Furthermore, our findings advocate for the broader adoption of video simulations in medical education, as they offer an innovative and effective means to improve diagnostic skills and preparedness among future clinicians [20,28].

In addition to repeated exposure to video-based examinations, future research should explore methods of training human raters to reliably assess non-verbal cue detection. Standardized rater-training programs or calibration protocols should ensure consistent evaluations in video-based objective structured clinical examination-type assessments. Furthermore, advances in artificial intelligence and machine learning could enable the automated recognition of visual and auditory cues in clinical simulations, supporting scalable and objective assessment of non-verbal diagnostic skills.

Limitations

This study has certain limitations. First, the cross-sectional design limited our ability to infer causality between the detection of non-verbal cues like tachypnea and higher GM-ITE scores. Future research should consider longitudinal or interventional study designs, to more definitively evaluate causality. Specifically, intervention studies could assess whether targeted training in non-verbal cue recognition—such as simulation-based educational programs—facilitates measurable improvements in the diagnostic accuracy and GM-ITE performance. Longitudinal approaches may clarify whether enhanced cue recognition skills contribute to sustained improvements in clinical competence. Second, the study relied on self-reported data from resident physicians, which may be subject to response bias. Although we used objective measures such as the GM-ITE scores, the accuracy of the self-reported demographic and training information could not be independently verified. Third, this study was conducted in Japan, where medical education and training systems may differ from other countries. Consequently, these findings may not be generalizable to resident physicians in different healthcare systems with varying educational structures. However, the highlighted diagnostic challenges—particularly the recognition or interpretation of subtle clinical cues and prevention of diagnostic errors—are likely to be relevant across a broad range of training contexts [29]. Therefore, while the GM-ITE context is unique, our findings offer insights that may resonate with global generalist training programs. Fourth, although propensity score matching was applied to control potential confounders, unmeasured variables may have influenced the results. While propensity score matching strengthened the validity of our findings by controlling for potential confounders [30], unmeasured factors, such as prior training in video simulation or baseline observational skills, may still influence the results [31]. A deeper understanding of these variables could provide further insights into the mechanisms underlying the differences in GM-ITE scores. Finally, the innovative examination using video simulations was a relatively new GM-ITE addition, and the novelty of the format might have influenced participants' performance. Ongoing evaluation of the reliability and validity of video-based assessments is essential to ensure their effectiveness as tools for measuring clinical competence.

Conclusions

The ability to recognize non-verbal cues, such as tachypnea, is a critical determinant of clinical competence among resident physicians, as evidenced by higher GM-ITE scores. Video-based simulations exhibit a transformative addition to traditional MCQ examinations, enabling a comprehensive evaluation of diagnostic abilities by integrating real-world complexities and non-verbal information.

Acknowledgments

We wish to thank the members of JAMEP for their valuable assistance and Dr. Kentaro Sakamaki from the Faculty of Health Data Science at Juntendo University for his statistical guidance. We also thank Editage for the English language review. This work was supported by the Health, Labor, and Welfare Policy Grants of Research on Region Medical (21IA2004, 24IA2016) from Japan's Ministry of Health, Labor, and Welfare. The funder had no role in the design or conduct of the study, the collection, management, analysis, and interpretation of the data, the preparation, review, or approval of the manuscript, or the decision to submit the manuscript for publication.

Conflicts of Interest

JAMEP was involved in collecting and managing data as a GM-ITE administrative organization. It did not participate in designing and conducting the study, data analysis and interpretation, preparation, review, approval of the manuscript, or the decision to submit the manuscript for publication. YN received an honorarium from JAMEP as GMITE project manager. YT is the director of JAMEP. HK received an honorarium from JAMEP as a speaker of JAMEP lecture. KS received an honorarium from JAMEP as a reviewer for the GMITE. KS, TS, and YY received honoraria from JAMEP as exam preparers for the GMITE. The authors declare no competing interests.

Multimedia Appendix 1
Sample multiple-choice question from the General Medicine In-Training Examination (GM-ITE).
[DOCX File, 18 KB - mededu_v11i1e72640_app1.docx]

Multimedia Appendix 2
Detection rates of non-verbal clinical findings in the simulation video based on answer correctness.
[DOCX File, 22 KB - mededu_v11i1e72640_app2.docx]

Multimedia Appendix 3
Comparison of the correct clinical simulation video answers between the tachypnea non-detection and detection groups after propensity score matching.
[DOCX File, 26 KB - mededu_v11i1e72640_app3.docx]

Multimedia Appendix 4
Relationship between tachypnea detection and correct clinical simulation video answer, interaction tests.
[DOCX File, 30 KB - mededu_v11i1e72640_app4.docx]

References

1. Silverman J, Kinnersley P. Doctors'non-verbal behaviour in consultations: look at the patient before you look at the computer. Br J Gen Pract 2010 Feb;60(571):76-78. [doi: 10.3399/bjgp10X482293] [Medline: 20132698]
2. Nagasaki K, Nishizaki Y, Nojima M, et al. Validation of the general medicine in-training examination using the professional and linguistic assessments board examination among postgraduate residents in Japan. Int J Gen Med 2021;14:6487-6495. [doi: 10.2147/IJGM.S331173] [Medline: 34675616]
3. Kinoshita K, Tsugawa Y, Shimizu T, et al. Impact of inpatient caseload, emergency department duties, and online learning resource on general medicine in-training examination scores in Japan. Int J Gen Med 2015;8:355-360. [doi: 10.2147/IJGM.S81920] [Medline: 26586961]
4. Daniel M, Rencic J, Durning SJ, et al. Clinical reasoning assessment methods: a scoping review and practical guidance. Acad Med 2019 Jun;94(6):902-912. [doi: 10.1097/ACM.0000000000002618] [Medline: 30720527]
5. Nishizaki Y, Shimizu T, Shinozaki T, et al. Impact of general medicine rotation training on the in-training examination scores of 11, 244 Japanese resident physicians: a nationwide multi-center cross-sectional study. BMC Med Educ 2020 Nov 13;20(1):426. [doi: 10.1186/s12909-020-02334-8] [Medline: 33187497]
6. Nickl M, Huber SA, Sommerhoff D, Codreanu E, Ufer S, Seidel T. Video-based simulations in teacher education: the role of learner characteristics as capacities for positive learning experiences and high performance. Int J Educ Technol High Educ 2022;19(1):45. [doi: 10.1186/s41239-022-00351-9] [Medline: 36065455]

7. Cleary TJ, Battista A, Konopasky A, Ramani D, Durning SJ, Artino AR Jr. Effects of live and video simulation on clinical reasoning performance and reflection. Adv Simul 2020 Dec;5(1):17. [doi: 10.1186/s41077-020-00133-1]
8. Engström H, Andersson Hagiwara M, Backlund P, et al. The impact of contextualization on immersion in healthcare simulation. Adv Simul 2016 Jan;1(1):8. [doi: 10.1186/s41077-016-0009-y]
9. Shikino K, Nishizaki Y, Fukui S, et al. Development of a clinical simulation video to evaluate multiple domains of clinical competence: cross-sectional study. JMIR Med Educ 2024 Feb 29;10:e54401. [doi: 10.2196/54401] [Medline: 38421691]
10. Yokokawa D, Shikino K, Nishizaki Y, Fukui S, Tokuda Y. Evaluation of a computer-based morphological analysis method for free-text responses in the general medicine in-training examination: algorithm validation study. JMIR Med Educ 2024 Dec 5;10:e52068. [doi: 10.2196/52068] [Medline: 39637351]
11. Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics 1977 Jun;33(2):363-374. [doi: 10.2307/2529786] [Medline: 884196]
12. Austin PC, Mamdani MM. A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. Stat Med 2006 Jun 30;25(12):2084-2106. [doi: 10.1002/sim.2328] [Medline: 16220490]
13. Yang D, Dalton JE. A unified approach to measuring the effect size between two groups using SAS. 2012 Presented at: SAS Publishing Global Forum URL: https://support.sas.com/resources/papers/proceedings12/335-2012.pdf [accessed 2024-12-30]
14. Watari T, Nishizaki Y, Houchens N, et al. Medical resident's pursuing specialty and differences in clinical proficiency among medical residents in Japan: a nationwide cross-sectional study. BMC Med Educ 2023 Jun 22;23(1):464. [doi: 10.1186/s12909-023-04429-4] [Medline: 37349724]
15. Griffith CH III, Wilson JF, Langer S, Haist SA. House staff nonverbal communication skills and standardized patient satisfaction. J Gen Intern Med 2003 Mar;18(3):170-174. [doi: 10.1046/j.1525-1497.2003.10506.x] [Medline: 12648247]
16. DiMatteo MR, Taranta A, Friedman HS, Prince LM. Predicting patient satisfaction from physicians' nonverbal communication skills. Med Care 1980 Apr;18(4):376-387. [doi: 10.1097/00005650-198004000-00003] [Medline: 7401698]
17. Marcinowicz L, Konstantynowicz J, Godlewski C. Patients' perceptions of GP non-verbal communication: a qualitative study. Br J Gen Pract 2010 Feb;60(571):83-87. [doi: 10.3399/bjgp10X483111] [Medline: 20132701]
18. Morgado M, Botelho J, Machado V, Mendes JJ, Adesope O, Proença L. Full title: Video-based approaches in health education: a systematic review and meta-analysis. Sci Rep 2024 Oct 10;14(1):23651. [doi: 10.1038/s41598-024-73671-7] [Medline: 39384592]
19. Willett LL, Palonen K, Allison JJ, et al. Differences in preventive health quality by residency year. Is seniority better? J Gen Intern Med 2005 Sep;20(9):825-829. [doi: 10.1111/j.1525-1497.2005.0158.x] [Medline: 16117750]
20. Kotwal S, Fanai M, Fu W, et al. Real-world virtual patient simulation to improve diagnostic performance through deliberate practice: a prospective quasi-experimental study. Diagnosis (Berl) 2021 Nov 25;8(4):489-496. [doi: 10.1515/dx-2020-0127] [Medline: 33675203]
21. Chimea TL, Kanji Z, Schmitz S. Assessment of clinical competence in competency-based education. Can J Dent Hyg 2020 Jun 1;54(2):83-91. [Medline: 33240368]
22. Fu Y, Zhang W, Zhang S, Hua D, Xu D, Huang H. Applying a video recording, video-based rating method in OSCEs. Med Educ Online 2023 Dec;28(1):2187949. [doi: 10.1080/10872981.2023.2187949] [Medline: 36883331]
23. Saleem M, Khan Z. Healthcare simulation: an effective way of learning in health care. Pak J Med Sci 2023;39(4):1185-1190. [doi: 10.12669/pjms.39.4.7145] [Medline: 37492303]
24. Farris C, Treat TA, Viken RJ, McFall RM. Perceptual mechanisms that characterize gender differences in decoding women's sexual intent. Psychol Sci 2008 Apr;19(4):348-354. [doi: 10.1111/j.1467-9280.2008.02092.x] [Medline: 18399887]
25. Saeed S, Khan MH, Siddiqui MMU, Dhanwani A, Hussain A, Ali MM. Hybridizing video-based learning with simulation for flipping the clinical skills learning at a university hospital in Pakistan. BMC Med Educ 2023 Aug 21;23(1):595. [doi: 10.1186/s12909-023-04580-y] [Medline: 37605200]
26. Elendu C, Amaechi DC, Okatta AU, et al. The impact of simulation-based training in medical education: a review. Medicine (Baltimore) 2024 Jul 5;103(27):e38813. [doi: 10.1097/MD.0000000000038813] [Medline: 38968472]
27. Wang J. To use or not to use propensity score matching? Pharm Stat 2021 Jan;20(1):15-24. [doi: 10.1002/pst.2051] [Medline: 32776719]
28. Singh H, Upadhyay DK, Torretti D. Developing health care organizations that pursue learning and exploration of diagnostic excellence: an action plan. Acad Med 2020 Aug;95(8):1172-1178. [doi: 10.1097/ACM.0000000000003062] [Medline: 31688035]
29. Mamede S, Goeijenbier M, Schuit SCE, et al. Specific disease knowledge as predictor of susceptibility to availability bias in diagnostic reasoning: a randomized controlled experiment. J Gen Intern Med 2021 Mar;36(3):640-646. [doi: 10.1007/s11606-020-06182-6] [Medline: 32935315]
30. Teng SW, Su YC, Pallantla R, et al. Can a propensity score matching method be applied to assessing efficacy from single-arm proof-of-concept trials in oncology? CPT Pharmacometrics Syst Pharmacol 2023 Sep;12(9):1347-1357. [doi: 10.1002/psp4.13014] [Medline: 37528543]

31. Desai RJ, Bradley MC, Lee H, et al. A simulation-based bias analysis to assess the impact of unmeasured confounding when designing nonrandomized database studies. Am J Epidemiol 2024 Nov 4;193(11):1600-1608. [doi: 10.1093/aje/kwae102] [Medline: 38825336]

Abbreviations

GM-ITE: general medicine in-training examination
IM-ITE: internal medicine residency examination
JAMEP: Japan Institute for Advancement of Medical Education Program
MCQs: multiple-choice questions
PGY: postgraduate year

Edited by M Montagna; submitted 14.02.25; peer-reviewed by N Mungoli, T Mitsunaga; revised version received 01.04.25; accepted 04.07.25; published 31.07.25.

Please cite as:
Shikino K, Nishizaki Y, Fukui S, Kataoka K, Yokokawa D, Shimizu T, Yamamoto Y, Nagasaki K, Kobayashi H, Tokuda Y
Resident Physician Recognition of Tachypnea in Clinical Simulation Videos in Japan: Cross-Sectional Study
JMIR Med Educ 2025;11:e72640
URL: https://mededu.jmir.org/2025/1/e72640
doi:10.2196/72640

Awareness and Attitude Toward Artificial Intelligence Among Medical Students and Pathology Trainees: Survey Study

Anwar Rjoop[1], MBBS, MD; Mohammad Al-Qudah[1,2], MBBS, MD; Raja Alkhasawneh[3], MBBS, MD; Nesreen Bataineh[4], MBBS, MD; Maram Abdaljaleel[5], MBBS, MD; Moayad A Rjoub[6], MBBS, MD; Mustafa Alkhateeb[7]; Mohammad Abdelraheem[7]; Salem Al-Omari[7]; Omar Bani-Mari[7]; Anas Alkabalan[7]; Saoud Altulaih[7]; Iyad Rjoub[7], MBBS, MD; Rula Alshimi[7]

[1]Department of Pathology and Microbiology, Faculty of Medicine, Jordan University of Science and Technology, Irbid, Jordan

[2]Department of Microbiology, Pathology and Forensic Medicine, Faculty of Medicine, The Hashemite University, Zarqa, Jordan

[3]Department of Pulmonary Medicine, King Hussain Medical Center, Royal Medical Services, Amman, Jordan

[4]Department of Basic Medical Sciences, Faculty of Medicine, Yarmouk University, Irbid, Jordan

[5]Department of Pathology, Microbiology, and Forensic Medicine, School of Medicine, The University of Jordan, Amman, Jordan

[6]Department of General Surgery and Urology, Faculty of Medicine, Jordan University of Science and Technology, Irbid, Jordan

[7]Faculty of Medicine, Jordan University for Science and Technology, Irbid, Jordan

Corresponding Author:
Anwar Rjoop, MBBS, MD
Department of Pathology and Microbiology, Faculty of Medicine, Jordan University of Science and Technology, Irbid, Jordan

Abstract

Background: Artificial intelligence (AI) is set to shape the future of medical practice. The perspective and understanding of medical students are critical for guiding the development of educational curricula and training.

Objective: This study aims to assess and compare medical AI-related attitudes among medical students in general medicine and in one of the visually oriented fields (pathology), along with illuminating their anticipated role of AI in the rapidly evolving landscape of AI-enhanced health care.

Methods: This was a cross-sectional study that used a web-based survey composed of a closed-ended questionnaire. The survey addressed medical students at all educational levels across the 5 public medical schools, along with pathology residents in 4 residency programs in Jordan.

Results: A total of 394 respondents participated (328 medical students and 66 pathology residents). The majority of respondents (272/394, 69%) were already aware of AI and deep learning in medicine, mainly relying on websites for information on AI, while only 14% (56/394) were aware of AI through medical schools. There was a statistically significant difference in awareness among respondents who consider themselves tech experts compared with those who do not (P=.03). More than half of the respondents believed that AI could be used to diagnose diseases automatically (213/394, 54.1% agreement), with medical students agreeing more than pathology residents (P=.04). However, more than one-third expressed fear about recent AI developments (167/394, 42.4% agreed). Two-thirds of respondents disagreed that their medical schools had educated them about AI and its potential use (261/394, 66.2% disagreed), while 46.2% (182/394) expressed interest in learning about AI in medicine. In terms of pathology-specific questions, 75.4% (297/394) agreed that AI could be used to identify pathologies in slide examinations automatically. There was a significant difference between medical students and pathology residents in their agreement (P=.001). Overall, medical students and pathology trainees had similar responses.

Conclusions: AI education should be introduced into medical school curricula to improve medical students' understanding and attitudes. Students agreed that they need to learn about AI's applications, potential hazards, and legal and ethical implications. This is the first study to analyze medical students' views and awareness of AI in Jordan, as well as the first to include pathology residents' perspectives. The findings are consistent with earlier research internationally. In comparison with prior research, these attitudes are similar in low-income and industrialized countries, highlighting the need for a global strategy to introduce AI instruction to medical students everywhere in this era of rapidly expanding technology.

(JMIR Med Educ 2025;11:e62669) doi:10.2196/62669

KEYWORDS

artificial intelligence; AI; deep learning; medical schools; pathology; Jordan; medical education; awareness; attitude; medical students; pathology trainees; national survey study; medical practice; training; web-based survey; survey; questionnaire

Introduction

Artificial intelligence (AI) is the capability of machines to simulate intelligence by exhibiting human-like traits such as understanding, deductive reasoning, and problem-solving abilities [1]. In medicine, many specialties have already used AI in clinical practice, such as oncology for cancer detection and grading [2]. AI, particularly deep learning, has garnered a lot of attention in medical education and pathology in recent years [3,4]. These techniques have been mostly used for visual activities such as consultations, seminar presentations, board exams, and archiving [5]. Even before this achievement, many experts predicted that specialized algorithms capable of reading images as well as or better than human observers would dominate the future of medicine. As a result, residents and undergraduate students are becoming increasingly concerned that pursuing medicine training may be an insecure career path [6]. However, the long-awaited expansion of AI in pathology is still underway, and the area of pathology is changing at a far slower rate than other domains (eg, radiology) [7].

The recent approvals of whole slides imaging scanners by the Food and Drug Administration for primary diagnosis, as well as the approval of the prostate AI algorithm, have cleared the road for the first steps in incorporating this new technology for use in pathologic primary diagnosis. AI solutions can serve as a unique platform for breakthroughs and advancements in anatomical and clinical pathology practice [8].

Commercially available AI systems have recently become the focus of a more concentrated evaluation. However, it is unclear whether medical students and pathology residents are afraid that AI may replace pathologists or other clinicians [9]. There is limited knowledge about medical students' perspectives toward AI and deep learning, particularly in underdeveloped nations. To investigate this topic, we conducted a multicenter national survey of undergraduate medical students and pathology residents to assess their perceptions of AI in medicine in general and pathology in particular, as well as their concern that machines will replace pathologists or other physicians soon.

The perspective and understanding of medical students are critical for guiding the development of educational curricula and training. We investigate major medical AI-related attitudes among medical students in general and in pathology in Jordan, with an emphasis on the components that should be included in the medical curriculum.

Methods

Study Design and Population

A randomized, web-based, cross-sectional study was conducted among medical students and pathology residents in Jordan from 5 public universities—Jordan University of Science and Technology (JUST), The University of Jordan, Yarmouk University, Hashemite University, and Mutah University, including about 19,000 medical students—in addition to 4 pathology residency programs: JUST, King Hussein Cancer Center, Jordan University Hospital, and Royal Medical Services, including about 80 pathology residents. The survey was published over a 10-day period from March 4 to March 14, 2024. The required sample size was estimated using the web-based Raosoft sample size calculator [10]. The proposed sample size was 318, with a 5% margin of error, a 99% confidence, a 30% response rate, and a population size of 20,000. Three hundred ninety-four students completed the questionnaire. Students were encouraged to disseminate the questionnaire among their colleagues to create a snowball sample.

Ethical Considerations

The research has been approved by the institutional review board of JUST, Irbid, Jordan (number: 3/167/2024; date: February 13, 2024). All individuals consented to participate. All methods met with the applicable standards and regulations. Participants provided informed consent at the beginning of the survey and had the ability to opt out at any time. Data were anonymized and no compensation was provided to the participants. The results of this study are original, have not been previously published, are not under review elsewhere, and have received approval from all authors. All the authors have approved final version of the manuscript.

The data were obtained using a self-administered open web-based questionnaire produced with Google Forms. To evaluate the questionnaire's applicability and forward validity, our research team translated the questions into Arabic and conducted a pilot survey with 15 randomly selected participants to examine question comprehension and language clarity. The questionnaire was circulated through medical student groups, social media forums (Facebook, WhatsApp, Telegram, and Instagram), and through announcements in lectures. Participation was entirely voluntary and unrelated to the students' educational curriculum. The students consented to participate by completing the survey. Respondent anonymity was assured by design.

Questionnaire Structure

The questionnaire items were adopted from a previously validated study [11] and amended by 2 expert pathologists (AR and MA) at 2 academic centers to apply the questions to the discipline of pathology. The questionnaire was divided into sections, each of which addressed a different issue (Table S1 in Multimedia Appendix 1). The first section of the questionnaire inquired about demographic data and self-reported technological expertise. The second portion inquired about AI and deep learning applications in medicine. The third part evaluated sources to AI in general. The fourth part focused on AI applications in medicine. The fifth section evaluated emotions and perspectives toward AI and deep learning in medicine and pathology. The sixth part inquired about the expected effects of AI on medical education and specific components that should be implemented in medical education (an open-ended question), followed by a question regarding whether basic AI knowledge should be provided in official medical courses (a Yes/No). Finally, the prospective applications of AI in pathology were considered.

Statistical Analysis

Following the completion of questionnaire submissions, the findings were converted to a comma-separated value file. To

simplify statistical analysis, the categories “disagree entirely” and “rather disagree” were summarized as disagreement, while “rather agree” and “agree entirely” were summarized as agreement. Nominal categorical variables were analyzed using the Pearson chi-square test or Fisher exact test, whereas ordinal data were analyzed using Spearman correlation. The statistical analysis was performed using SPSS version 26.0 (IBM Corp) [12], and *P* values of <.05 were considered statistically significant.

Results

Overview

After 10 days of opening the survey, 394 participants completed the questionnaire (328 medical students and 66 pathology residents). Of these respondents, 49% (193/394) were males and 51% (201/394) were females. The median age is 20 (IQR 20 - 21) years. Most medical students surveyed were in their junior years (1, 2, and 3), accounting for around 85% (279/394) of the sample, while approximately 15% (49/394) were in clinical training years (4, 5, and 6). Most pathology residents surveyed were postgraduate year 1 and postgraduate year 2 (63/66, 96%). Of the total, 44.9% (177/394) of the respondents regarded themselves as technological experts (Table 1).

Table . Demographics and self-reported technical expertise (N=394).

		Participants
I consider myself a tech expert person, n/N (%)		
	Agree entirely	46/394 (12)
	Rather agree	131/394 (33.2)
	Rather disagree	73/394 (19)
	Disagree entirely	16/394 (4)
	__[a]	128/394 (32.5)
Age (years)		
	Median	20
	IQR	20-21
	Minimum/maximum	18/34
Gender, n/N (%)		
	Male	193/394 (49)
	Female	201/394 (51)

[a]Not available (or no response).

Awareness of AI and Deep Learning in Medicine

The vast majority of respondents (272/394, 69%) were previously aware of the medical community’s discussion on AI and deep learning. There was a statistically significant difference in awareness among respondents who consider themselves tech experts compared with those who do not (*P*=.03), but no significant differences were found between males and females, medical students, and pathology residents. Furthermore, 67.5% (266/394) of respondents reported having a basic awareness of the technologies used in these fields (Table 2).

Table . First part of the questionnaire—artificial intelligence and deep learning in medicine (N=394).

Questionnaire items		Yes, n/N (%)	No, n/N (%)	*P* value[a]
“Deep learning” and “artificial intelligence” are currently being broadly discussed in the medical field.				
	Were you already aware of these topics?	272/394 (69)	122/394 (31)	.03/.46
	Do you have a basic understanding of the technologies used in these topics?	266/394 (67.5)	128/394 (32.5)	.89/.94

[a]Tech expert versus non–tech expert/medical students versus pathology residents.

The Sources for the Topic of AI

Websites were the primary sources of knowledge on AI, with 73.4% (289/394) of respondents reporting awareness via this source. In contrast, fewer students heard about AI from friends and colleagues (139/394, 35.3%), webinars (59/394, 15%), and medical school lectures (56/394, 14%). The majority of respondents (310/394, 78.6%) acknowledged that they did not learn about AI through formal AI courses (Table 3).

Table . Second part of the questionnaire—different sources of exposure to artificial intelligence as a topic in general (N=394).

Questionnaire items		Yes, n/N (%)	No, n/N (%)	*P* value[a]
Other applications we use in daily life already use artificial intelligence (eg, speech-/text-recognition). Were you aware of this from?				
	Websites	289/394 (73.4)	105/394 (26.6)	<.001/.66
	Social friends and colleagues	139/394 (35.3)	255/394 (64.7)	.20/.74
	Medical school lectures	56/394 (14)	338/394 (85.8)	.90/.28
	Webinars	59/394 (15)	335/394 (85)	.80/.37
	Training (eg, courses) in artificial intelligence	22/394 (6)	372/394 (94.4)	.15/.99
	No answer	—[b]	—	.001/.85

[a]Tech expert vs non–tech expert/medical students vs pathology residents.

[b]Not applicable (ie, total proportion of no answers: 62/394, 16%).

The Applications of AI in Medicine

More than half of respondents thought that AI might be used to diagnose disease in patients automatically (213/394, 54.1% agreement vs 82/394, 21% disagreement), and medical students agreed more than pathology residents on this topic *(P*=.04). Furthermore, 39.5% (156/394) agreed to the use of AI in automated diagnosis. Moreover, three-quarters agreed that AI might automatically indicate appropriate investigations (318/394, 80.7% agreement vs 19/394, 5% disagreement). Table 4 provides more detailed results.

Table . Third part of the questionnaire—applications for artificial intelligence in medicine (N=394).

Questionnaire items		Agree entirely, n/N (%)	Rather agree, n/N (%)	Rather disagree, n/N (%)	Disagree entirely, n/N (%)	N/A[a], n/N (%)	*P* value[b]
What potential applications for AI in medicine do you see?							
	Automated detection of disease	55/394 (14)	158/394 (40.1)	69/394 (16)	13/394 (3)	99/394 (25)	.58/.001
	Automated diagnosis of patients	39/394 (10)	117/394 (29.7)	107/394 (27.2)	19/394 (5)	112/394 (28)	.016/.000
	Automated indication of appropriate investigations (radiological, laboratory, etc)	113/394 (28.7)	205/394 (52)	17/394 (4)	2/394 (1)	57/394 (15)	<.001/.60

[a]N/A: not applicable.

[b]Tech expert vs non–tech expert/medical students vs pathology residents.

Overall Emotions and Attitudes About AI and Deep Learning in Both Medicine and Pathology

Regarding overall feelings and attitudes toward AI and deep learning in medicine and pathology, most respondents agreed that AI will revolutionize medicine in general (321/394, 81.4% agreement) and pathology in particular (312/394, 79.2% agreement), while a sizable proportion disagreed that human doctors in general (291/394, 73.9% disagreement) and pathologists (248/394, 63% disagreement) could be replaced in the near term. Furthermore, more than one-third of respondents expressed fear about recent AI developments (167/394, 42.4% agreed).

On the other hand, roughly half said that these breakthroughs make pathology or medicine more intriguing to them (184/394, 46.7% and 222/394, 56.4%, respectively), and for those specific questions, tech expert respondents were considerably more likely to say “yes” than non–tech expert respondents (*P*=.002 and *P*=.03, respectively).

Nonetheless, most respondents believed that the adoption of AI would benefit pathology (302/394, 76.7% agreement) and the entire field of medicine (305/394, 77.4% agreement). Notably, two-thirds of respondents disagreed that their medical schools or hospitals had educated them about AI and its uses (261/394, 66.2% disagreed), whereas 46.2% (182/394) expressed an

interest in learning the principles of AI and its applications in medicine. Table S1 in Multimedia Appendix 1 summarizes feelings and attitudes toward AI and deep learning in medicine and pathology.

What Specific Aspects Should Be Implemented in Medical Education?

Students were asked to select from a list of options, each of which allowed for several responses. The majority of respondents (231/394, 58.6%) expressed an interest in learning about AI's applications, potential hazards, and legal and ethical implications. Females were more likely than males to exhibit an interest in learning about AI's possible hazards and legal implications (P=.038 and P=.001, respectively). In addition, 51.8% (204/394) of the students felt that medical education should cover current AI systems and their technical foundations. However, 64.2% (253/394) of respondents expressed no interest in learning about the classification of AI reliability in medical education (Table 5).

Table . Specific aspects that should be implemented in medical education (multiple responses possible) (N=394)[a].

	Yes, n/N (%)	No, n/N (%)	P value[b]
Areas of application	231/394 (58.6)	163/394 (41.4)	.04/.95
Possible risks	231/394 (58.6)	163/394 (41.4)	.25/.83
Technical basics	204/394 (51.8)	190/394 (48.2)	.73/.74
Current AI[c] systems	204/394 (51.8)	190/394 (48.2)	.63/.52
Modes of operation	200/394 (50.8)	194/394 (49.2)	.30/.21
Legal aspects, ethics	198/394 (50.3)	196/394 (49.7)	.094/.35
Potential future developments	195/394 (49.5)	199/394 (51.5)	.45/.36
Classification of AI reliability	141/394 (35.8)	253/394 (64.2)	.77/.23
Basic AI knowledge should be provided in university courses	357/394 (90.6)	37/394 (9.4)	.21/.16

[a]The vast majority of respondents believed that university curricula should cover fundamental AI concepts (357/394, 90.6% said yes).

[b] Tech expert vs non–tech expert/medical students versus pathology residents.

[c]AI: artificial intelligence.

What Potential Applications for AI Do You See in Pathology?

In terms of pathology-specific questions, 3 quadrants (297/394, 75.4%) of respondents agreed that AI could be used to identify pathologies in slides examinations automatically, and more than half (222/394, 56.3%) agreed that AI could be used to diagnose pathologies in slides examinations and indicate appropriate further stains needed. There was a statistically significant difference between medical students, who are more inclined to agree, and pathology residents (P=.02/P<.001 and P<.001/P=.61, respectively). In addition, 79.1% (312/394) of respondents felt that AI might be used to automatically identify appropriate special studies and immunohistochemistry stains in slide examinations (Table 6).

Table . Potential applications for artificial intelligence in pathology (N=394).

	Totally agree, n/N (%)	Agree, n/N (%)	Disagree, n/N (%)	Totally disagree, n/N (%)	Neutral, n/N (%)	P value[a]
Automated detection of pathologies in slides examinations	91/394 (23)	206/394 (52.3)	26/394 (7)	5/394 (1)	66/394 (17)	.56/.001
Automated diagnosis in slides examinations	58/394 (15)	164/394 (41.6)	50/394 (13)	11/394 (3)	111/394 (28.2)	.02/<.001
Automated indication of appropriate special studies and immunohistochemical stains in slides examinations	103/394 (26.1)	209/394 (53)	11/394 (3)	2/394 (1)	69/394 (18)	<.001/.61

[a]Tech expert vs non–tech expert/medical students vs pathology residents.

Discussion

Principal Findings

Findings revealed that a significant majority of medical students were already aware of the ongoing discussion around AI and deep learning in the medical community. This awareness was significantly higher among those who described themselves as tech savvy. The survey results support the conclusion that the majority of respondents did not learn about AI throughout medical school.

The majority of respondents thought that AI had the potential to alter both medicine in general and pathology in particular. However, a significant proportion raised concerns about the potential displacement of human physicians by AI in the near future. This sentiment is consistent with other research revealing worries among medical students and pathology residents regarding the expanding role of AI in medicine. For example, in Lebanon and Kuwait, there was an agreement that AI would not replace doctors but rather significantly transform health care practices [13,14]. Notably, the majority of medical students who responded were in their junior years (279/394, 85%) of the total sample, which may reflect their interest in the subject compared with senior students. Although this allowed us to compare junior medical students' perceptions with those of postgraduate pathology students, it is regarded as a restriction for evaluating senior medical students' perspectives.

The general agreement was that the use of AI would benefit both pathology and the larger field of medicine [8]. Further comparison with prior work is discussed in the section "Comparison With Prior Work." This study provides useful insights into medical students' perspectives and attitudes regarding AI, which are critical for guiding the development of educational curricula and training.

The rapid growth of AI has the ability to change the face of a variety of medical specialties. Specifically in the visual medical disciplines, such as radiology, pathology, ophthalmology, and dermatology, AI has generated significant interest and will particularly affect the developments of these fields due to the visual nature of their occupations [3]. Deep learning applications are autonomously trained to execute certain tasks in response to the availability of large digital datasets. Visual activities include consultations, seminar presentations, board exams, and archiving [5]. Medical students' perspectives are crucial for shaping medical education, especially in rapidly changing professions. This study collects students' opinions on AI in medicine in order to better understand their requirements and expectations from medical schools, as well as add to the dataset so that data from various countries (both low and high income) may be compared to analyze future medical education plans. Learners in the digital era differ from past generations. They are growing increasingly technologically savvy and socially conscious. Pathology and radiology are 2 visual fields that have seen significant advancements in AI technology. Many studies have been conducted in radiology [11], but to our knowledge, this is the first to analyze the attitudes and awareness of pathology residents.

AI-Related Attitudes of Medical Students in Comparison With Pathology Residency Trainees

To the best of our knowledge, this is the first study to discuss pathology residents' awareness and attitudes toward AI in the field of pathology, in addition to comparing it with medical students' perspectives. In Jordan, digital pathology is a relatively new concept that is primarily used for research purposes. A digital scanner is available at one site in Jordan (at JUST), with a focus on research and consulting purposes.

In general, there was no major difference in the responses between undergraduate students and postgraduate pathology residents. Except on 2 occasions, medical students agreed more than pathology residents that AI can be used to diagnose disease in patients automatically (P=.04). There was also a statistically significant difference between medical students, who are more likely to agree than pathology residents regarding AI's ability to diagnose pathologies in slide examinations and indicate appropriate additional stains needed (P=.02/P<.001 and P<.001/P=.61, respectively). The Food and Drug Administration's recent support of whole-slide imaging scanners for primary diagnosis, together with the approval of prostate AI algorithms, marks the beginning of introducing AI technology into primary diagnostics. AI can provide a unique platform for promoting innovation and breakthroughs in anatomical and clinical pathology practice [9].

Otherwise, there was general agreement between medical students and pathology residents that, for example, the adoption of AI would benefit pathology and the entire field of medicine and that they needed to learn about AI's applications, potential hazards, and legal and ethical implications.

In summary, the findings underscore the imperative to integrate AI education into the medical field to adeptly equip future physicians for AI-augmented health care. Subsequent investigations should concentrate on assessing the efficacy of embedding AI education into medical training programs, both undergraduate and postgraduate, and probing the determinants influencing medical students' attitudes toward AI. AI may also be used in pathology to diagnose cancer, predict survival, modify molecular structures, and forecast treatments. More effort is required to navigate these applications. Furthermore, it is worth investigating the relationship between the extent of digital pathology implementation and AI awareness.

Limitations

Some limitations include the sample strategy's use of social media as well as the web-based approach, which limits randomization and generalization. Also, the majority of medical students who responded were in their junior years (279/394, 85%). Although this allowed us to compare junior medical students' perceptions to those of postgraduate pathology students, it is regarded as a restriction for evaluating senior medical students' perspectives.

Comparison With Prior Work

Compared with published work in this regard, similar results were identified in neighboring countries, including Lebanon [13]. In Kuwait, there was also an agreement that AI would not

replace doctors but rather significantly transform health care practices [14]. In the United Arab Emirates, there was a lack of acquaintance with AI found in a study that called for the introduction of specific education and training in medical schools [15]. In addition, identical outcomes were observed in industrialized countries. In Germany, two-thirds of students (539/838, 64.4%) believed that they were not well informed on AI in medicine, and 57.4% (463/807) thought that AI has beneficial applications in medicine, such as drug research, but less so for clinical use [16]. In the United States, a study showed that 91% (353/387) reported receiving no formal education related to AI [17].

Conclusions

Our study highlights a generally favorable attitude toward AI between medical students and pathology residents. A significant number of participants are already familiar with the ideas of AI and deep learning in medicine while the majority sees potential in AI for automated detection of pathologies and indication of appropriate investigations, which is a warning about replacing physicians and pathologists in the nearest future.

The study found that being a tech expert influenced respondents' awareness and attitudes toward AI. This indicates a potential gap in the current medical education system, with a large proportion of respondents expressing interest in learning more about AI and its applications in medicine.

In pathology, there is a prominent agreement on the potential applications of AI, particularly in the automated detection of pathologies in slide examinations and the automated indication of appropriate special studies and immunohistochemical stains. Medical students show that they are more enthusiastic about the integration of AI in pathology than pathology residents.

While most medical students and pathology residents acknowledge the potential of AI to revolutionize medicine and improve the pathology field, there is a clear need for integrating AI education into medical curricula and addressing concerns about its ethical and legal aspects.

Authors' Contributions

AR contributed to writing—original draft, methodology, formal analysis, data curation, conceptualization, and project administration. MAQ, RA, NB, and M Abdaljaleel contributed to writing—review and editing, supervision, methodology, and conceptualization. MAR, M Alkhateeb, M Abdelraheem, SAO, OBM, AA, SA, IR, and RA contributed to data curation, writing—review and editing, methodology, and conceptualization.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Fourth part of the questionnaire—feelings and attitudes toward artificial intelligence and deep learning in medicine and pathology (N=394).
[DOCX File, 17 KB - mededu_v11i1e62669_app1.docx]

References

1. Enholm IM, Papagiannidis E, Mikalef P, Krogstie J. Artificial intelligence and business value: a literature review. Inf Syst Front 2022 Oct;24(5):1709-1734. [doi: 10.1007/s10796-021-10186-w]
2. Londhe VY, Bhasin B. Artificial intelligence and its potential in oncology. Drug Discov Today 2019 Jan;24(1):228-232. [doi: 10.1016/j.drudis.2018.10.005]
3. Carlos RC, Kahn CE, Halabi S. Data science: big data, machine learning, and artificial intelligence. J Am Coll Radiol 2018;15(3 Pt B):497-498. [doi: 10.1016/j.jacr.2018.01.029]
4. Kohli M, Prevedello LM, Filice RW, Geis JR. Implementing machine learning in radiology practice and research. AJR Am J Roentgenol 2017 Apr;208(4):754-760. [doi: 10.2214/AJR.16.17224] [Medline: 28125274]
5. Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. Nat Rev Clin Oncol 2019 Nov;16(11):703-715. [doi: 10.1038/s41571-019-0252-y] [Medline: 31399699]
6. @AndrewYNg. Posting with permission. AI people: what would you say to him? I will share my thoughts later. Twitter. 2017 Jul 11. URL: https://mobile.twitter.com/andrewyng/status/884810469575344128?lang=de [accessed 2024-05-25]
7. Reis-Filho JS, Kather JN. Overcoming the challenges to implementation of artificial intelligence in pathology. J Natl Cancer Inst 2023 Jun 8;115(6):608-612. [doi: 10.1093/jnci/djad048] [Medline: 36929936]
8. Shafi S, Parwani AV. Artificial intelligence in diagnostic pathology. Diagn Pathol 2023 Oct 3;18(1):109. [doi: 10.1186/s13000-023-01375-z] [Medline: 37784122]
9. Kulkarni S, Seneviratne N, Baig MS, Khan AHA. Artificial intelligence in medicine: where are we now? Acad Radiol 2020 Jan;27(1):62-70. [doi: 10.1016/j.acra.2019.10.001] [Medline: 31636002]
10. Sample size calculator. Raosoft, Inc. URL: http://www.raosoft.com/samplesize.html [accessed 2025-01-03]

11. Pinto dos Santos D, Giese D, Brodehl S, et al. Medical students' attitude towards artificial intelligence: a multicentre survey. Eur Radiol 2019 Apr;29(4):1640-1646. [doi: 10.1007/s00330-018-5601-1]
12. Arbuckle JL. Amos (Version 26.0) [computer program]. IBM SPSS. 2019. URL: https://www.ibm.com/support/pages/node/878803 [accessed 2025-01-03]
13. Doumat G, Daher D, Ghanem NN, Khater B. Knowledge and attitudes of medical students in Lebanon toward artificial intelligence: a national survey study. Front Artif Intell 2022;5:1015418. [doi: 10.3389/frai.2022.1015418]
14. Buabbas AJ, Miskin B, Alnaqi AA, et al. Investigating students' perceptions towards artificial intelligence in medical education. Healthcare (Basel) 2023 May 1;11(9):1298. [doi: 10.3390/healthcare11091298]
15. Boillat T, Nawaz FA, Rivas H. Readiness to embrace artificial intelligence among medical doctors and students: questionnaire-based study. JMIR Med Educ 2022 Apr 12;8(2):e34973. [doi: 10.2196/34973] [Medline: 35412463]
16. McLennan S, Meyer A, Schreyer K, Buyx A. German medical students´ views regarding artificial intelligence in medicine: a cross-sectional survey. PLOS Digit Health 2022 Oct;1(10):e0000114. [doi: 10.1371/journal.pdig.0000114] [Medline: 36812635]
17. Liu DS, Sawyer J, Luna A, et al. Perceptions of US medical students on artificial intelligence in medicine: mixed methods survey study. JMIR Med Educ 2022 Oct 21;8(4):e38325. [doi: 10.2196/38325] [Medline: 36269641]

Abbreviations

AI: artificial intelligence
JUST: Jordan University of Science and Technology

Edited by B Lesselroth; submitted 28.05.24; peer-reviewed by A Khamees, Y Chong; revised version received 21.06.24; accepted 23.11.24; published 10.01.25.

Please cite as:
Rjoop A, Al-Qudah M, Alkhasawneh R, Bataineh N, Abdaljaleel M, Rjoub MA, Alkhateeb M, Abdelraheem M, Al-Omari S, Bani-Mari O, Alkabalan A, Altulaih S, Rjoub I, Alshimi R
Awareness and Attitude Toward Artificial Intelligence Among Medical Students and Pathology Trainees: Survey Study
JMIR Med Educ 2025;11:e62669
URL: https://mededu.jmir.org/2025/1/e62669
doi:10.2196/62669

Original Paper

Reviewing Mobile Apps for Teaching Human Anatomy: Search and Quality Evaluation Study

Guadalupe Esmeralda Rivera García[1*], PhD; Miriam Janet Cervantes López[2*], PhD; Juan Carlos Ramírez Vázquez[1*], PhD; Arturo Llanes Castillo[2*], PhD; Jaime Cruz Casados[2*], MAS

[1]Tecnológico Nacional de México, Instituto Tecnológico Superior de Pánuco, Pánuco, Veracruz, Mexico

[2]Facultad de Medicina de Tampico "Dr. Alberto Romo Caballero" de la Universidad Autónoma de Tamaulipas, Tampico, Tamaulipas, Mexico

[*]all authors contributed equally

Corresponding Author:
Guadalupe Esmeralda Rivera García, PhD
Tecnológico Nacional de México, Instituto Tecnológico Superior de Pánuco
Av. Artículo Tercero Constitucional, Colonia Solidaridad
Pánuco, Veracruz, 93990
Mexico
Phone: 52 846 1064254
Email: esmeralda.rivera@itspanuco.edu.mx

Abstract

Background: Mobile apps designed for teaching human anatomy offer a flexible, interactive, and personalized learning platform, enriching the educational experience for both students and health care professionals.

Objective: This study aimed to conduct a systematic review of the human anatomy mobile apps available on Google Play, evaluate their quality, highlight the highest scoring apps, and determine the relationship between objective quality ratings and subjective star ratings.

Methods: The Mobile App Rating Scale (MARS) was used to evaluate the apps. The intraclass correlation coefficient was calculated using a consistency-type 2-factor random model to measure the reliability of the evaluations made by the experts. In addition, Pearson correlations were used to analyze the relationship between MARS quality scores and subjective evaluations of MARS quality item 23.

Results: The mobile apps with the highest overall quality scores according to the MARS (ie, sections A, B, C, and D) were Organos internos 3D (anatomía) (version 4.34), Sistema óseo en 3D (Anatomía) (version 4.32), and VOKA Anatomy Pro (version 4.29). To measure the reliability of the MARS quality evaluations (sections A, B, C, and D), the intraclass correlation coefficient was used, and the result was "excellent." Finally, Pearson correlation results revealed a significant relationship (r=0.989; P<.001) between the quality assessments conducted by health care professionals and the subjective evaluations of item 23.

Conclusions: The average evaluation results of the selected apps indicated a "good" level of quality, and those with the highest ratings could be recommended. However, the lack of scientific backing for these technological tools is evident. It is crucial that research centers and higher education institutions commit to the active development of new mobile health apps, ensuring their accessibility and validation for the general public.

(JMIR Med Educ 2025;11:e64550) doi:10.2196/64550

KEYWORDS

anatomy; Google Play; mobile health; mHealth; Mobile App Rating Scale; MARS

Introduction

Background

At present, there is a wealth of research on mobile apps focused on various aspects and areas of health, such as musculoskeletal injuries [1]; chronic disease management [2]; pediatric disease care [3]; medication management [4]; oral hygiene [5]; asthma [6]; pediatric ear, nose, and throat surgery [7]; low back pain [8]; neurodegenerative disorders [9]; coronary arteries [10]; neurorehabilitation [11]; nutrition, anemia, and preeclampsia [12]; cancer [13,14]; cerebrovascular diseases [15]; childhood obesity [16]; diabetes [17]; tuberculosis [18]; fibromyalgia [19]; dementia [20]; chronic kidney disease [21]; and epilepsy [22],

among others. These mobile health management apps have transformed the way people access and manage information about their well-being, enabling everything from vital signs monitoring to chronic disease management. This advancement in mobile apps not only benefits patients but also opens new possibilities in health education. In particular, mobile apps for teaching human anatomy have become valuable resources that complement the learning and understanding of the structures and functions of the human body. Similar to personal health apps, these tools are designed with interactive features such as 3D models and detailed simulations that enhance the educational experience for medical students and health care professionals.

The justification for this study is based on three important approaches: (1) the technological approach regarding the use of mobile apps for teaching human anatomy, (2) the pedagogical approach, and (3) the quality evaluation approach of the apps. These approaches together provide a solid foundation to justify the importance of the study.

Technological Approach

The first approach involved reviewing and analyzing scientific publications on the use of mobile apps in teaching human anatomy, with the aim of understanding their results. This approach highlights how mobile technology has revolutionized access to knowledge, enabling students to learn in an accessible and practical manner. In teaching human anatomy, mobile apps with 3D models and simulations facilitate immersive and effective learning, complementing and even enhancing traditional methods in health sciences. One of the studies presented the software Road to Birth, developed by the University of Newcastle, which was designed to teach midwifery students at a Midwestern US university about the dynamic concepts of maternal anatomy and physiology during an obstetrics module. The students used Road to Birth, and 66% of them reported an increase in their knowledge, valuing the software as a useful and practical learning resource [23]. Another study used the mobile app AR in Anatomy, developed by the authors; which allows users to dynamically explore various parts of the human body in 3D, enhancing the educational experience [24]. Similarly, apps such as Anat_Hub, developed by faculty and researchers from the Departments of Computer Science and Medical Sciences at the University of the Western Cape; a mobile app with augmented reality (AR) to improve learning about the musculoskeletal system's anatomy, received positive evaluations. User results indicated that the anatomy system could effectively enhance student engagement and retention of anatomical concepts [25]. In addition, another study conducted with undergraduate health sciences students at the University of Cape Town analyzed the impact of an AR mobile app on learning motivation. The study included 78 students, evaluating motivation levels before and after using the app. The results showed that its use increased motivation, improving aspects such as attention, satisfaction, and confidence [26]. In the field of neuroanatomy, the use of mobile AR facilitated the understanding of complex concepts, increasing academic performance and reducing cognitive load among students [27]. Similarly, HuMAR, developed by researchers affiliated with Murdoch University and Universiti Utara Malaysia; an AR-based prototype for learning skeletal structure, demonstrated high satisfaction among students, highlighting the system's usability and functionality [28]. However, although cadavers remain the gold standard in anatomy teaching, there are financial, ethical, and supervisory limitations. Another study compared the effectiveness of virtual reality, AR, and tablet-based devices in teaching cranial anatomy. A total of 59 students participated who were randomly assigned to one of the 3 learning methods. The results suggest that these technologies can effectively complement anatomical teaching [29]. On the other hand, a recent study presented a human anatomy learning system based on AR using a marker on a mobile platform to capture images and merge them with data from an SQLite database. This system allows for interactive visualization of the human body or its organs in 3D. An evaluation conducted with high school and medical students demonstrated that the app facilitates anatomy learning more effectively due to its ability to provide interactive 3D representations through AR [30]. Another example is AEducaAR, an app developed by researchers affiliated with the University of Bologna, which combines AR with a 3D-printed anatomical model to improve anatomy teaching for medical students. Its effectiveness was evaluated with a group of 62 second-year students, comparing its use to traditional learning methods with anatomical atlas books. Although there were no significant differences in objective test results between the two methods, students expressed enthusiasm for AEducaAR in a survey, valuing its potential to motivate learning and enhance the 3D understanding of anatomical structures. This tool could also prepare students to use advanced medical technologies in their future careers [31]. In addition, a relevant study presents 10 mobile apps for teaching human anatomy, where the results indicate that the technological designs studied exhibit a high degree of usability [32]. Another study analyzed 325 anatomy mobile apps and outlined their features to facilitate dissemination in the academic field. It showcases a broad, diverse, and affordable market for human anatomy mobile apps that can complement students' education [33].

Pedagogical Approach

Medical students frequently face challenges in understanding anatomy through the images found in textbooks [34], which are flat and lack interactivity. In contrast, mobile apps for learning human anatomy can serve as a complementary resource for learning this discipline, offering students the opportunity to interact with content more deeply than in a conventional dissection room. Although traditional cadaver-based teaching remains the preferred learning method [35], anatomy education continues to face numerous challenges, including limited practical hours for students and instructors, restricted access, and the high cost of cadavers and artificial models [36]. The use of mobile apps for learning human anatomy offers significant advantages, such as immediate and continuous access to information anytime and anywhere. These apps include interactive 3D designs that allow for a detailed exploration of the human body. Furthermore, they are often updated regularly and are generally more affordable than traditional textbooks, making them accessible to a broader audience.

Quality Evaluation Approach

The third approach focuses on evaluating the quality of human anatomy mobile apps. The importance of evaluating these apps lies in the fact that higher-rated apps can serve as academic support for medical students, health care professionals, and general users. A specific methodology is required to evaluate the quality of health mobile apps. In this context, the Mobile App Rating Scale (MARS) methodology has been used. Various studies have used the MARS to assess health apps targeting a variety of conditions, such as chronic kidney disease and end-stage renal disease [37], chronic lung diseases [38], stress management [39], psoriasis [40], gastrointestinal diseases [41], pain management [42], oral hygiene [43,44], nutrition [45], genetics and genomics [46], food allergies or intolerances [47], deafness and hearing impairment [48], low back pain [49,50], neurological conditions [51], peritoneal dialysis [52], diabetes [53], COVID-19 [54], cancer [55], anticoagulation [56], dementia [57], specialized diets [58], toric intraocular lenses [59], epilepsy [60], depression [61], coronary diseases [62], dyslexia [63], autism spectrum disorder [64], nutrition [65], and pediatric palliative care [66].

This study aimed to (1) identify human anatomy mobile apps available on the Google Play store, which uses the Android operating system, covering 70.87% of the global mobile operating system market [67]; (2) evaluate these apps using the MARS, which considers engagement, functionality, aesthetics, information, subjective quality, and app specificity; (3) present the human anatomy mobile apps with the highest ratings on the MARS; and (4) determine the correlation between the objective MARS quality rating and the subjective MARS rating by health care professionals (ie, item 23).

Methods

Overview

This study was cross-sectional, as it collected data at a single time point without follow-up over time, evaluated the correlation between variables, and provided descriptive data [68-70]. The study was conducted following the guidelines of the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) initiative, which aims to improve the communication of results from observational studies among authors, editors, and readers of scientific publications, focusing primarily on cohort, case-control, and cross-sectional studies [71].

Selection Criteria for Human Anatomy Mobile Apps

From April 1 to May 30, 2024, an extensive search for mobile apps was conducted. The search term used was *anatomy*. The inclusion criteria for mobile apps were as follows: (1) available on the Google Play store; (2) related to human anatomy; (3) available in English or Spanish; (4) user rating ≥4.3 to ensure a minimum level of acceptance and satisfaction among users; (5) free to use—an essential factor in educational contexts where students or universities may face budget constraints; and (6) download count exceeding 100,000.

The exclusion criteria were as follows: (1) duplicate apps, either because of different versions or alternative names but containing the same content; and (2) apps not updated for 2 or more years.

The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology was used to select human anatomy–teaching apps. This methodology is used in systematic reviews and meta-analyses to ensure transparency and rigor in the selection and analysis of relevant studies. The phases were identification, selection, eligibility, and inclusion.

During the identification phase, an exhaustive search for mobile apps dedicated to teaching human anatomy was conducted. In the filtering phase, duplicate apps were discarded. In the eligibility phase, the characteristics of these apps were analyzed to discard those that did not meet the previously established inclusion criteria, and an additional exclusion criterion was applied. Finally, in the inclusion phase, the apps that met all eligibility requirements were integrated for analysis and evaluation.

All selected apps were recorded in Excel (version 2016; Microsoft Corporation) with the following characteristics: app name, identification screen, languages, star rating, total downloads, developer, Android version, last update date, and features.

Evaluation of Mobile Apps

Overview

We used the MARS, developed by Stoyanov et al [72], which has been widely used to evaluate the design and usability of mobile health apps. The MARS consists of 3 dimensions. The first dimension is an objective tool based on 4 main components: engagement (section A), functionality (section B), aesthetics (section C), and information quality (section D). The second dimension assesses subjective quality (section E), and the third dimension evaluates the perceived effectiveness (section F).

Each section of the MARS has several items. An item refers to a specific element, question, or unit of evaluation within a questionnaire or survey. Section A comprises 5 items and evaluates whether the app is engaging, interesting, customizable, interactive, and targeted at a specific population. Section B comprises 4 items and focuses on the app's performance, ease of use, navigation, and gesture design. Section C comprises 3 items and examines the app's design, graphics, and visual appeal. Section D comprises 7 items and analyzes the accuracy of information description, objectives, quality and quantity of information, visual information quality, credibility, and scientific evidence base of the evaluated app. The 4 objective sections of MARS (ie, A, B, C, and D) encompass 19 items. The average scores of sections A, B, C, and D represent the overall MARS quality score [72].

Section E comprises 4 items and focuses on the evaluator's personal perception of the app and typically includes items that ask about the likelihood of recommending the app, the probability of the user continuing to use the app, and the overall perception of its quality. Finally, section F comprises 6 items and focuses on how health care professionals perceive the impact

of the app on their knowledge, attitudes, intentions, and behaviors related to health.

Evaluation Instrument Scale

To evaluate each item, a 5-point Likert scale was used, ranging from 1 to 5 (1=inadequate, 2=poor, 3=acceptable, 4=good, and 5=excellent). A MARS score of more than 3 points indicates acceptable quality.

Selection of Evaluators

A total of 10 evaluators were selected, all health care professionals from the Faculty of Medicine of Tampico Dr Alberto Romo Caballero at the Universidad Autónoma de Tamaulipas, Tampico, Tamaulipas, Mexico. The inclusion criteria were (1) being a top-performing student in the final year of medical school and (2) having a mobile phone with the Android operating system to download human anatomy teaching apps from Google Play.

Evaluation Process

Before starting the evaluation of anatomy-related apps, it was necessary to train the evaluators in the use of the MARS. For this, the authors of this study convened the 10 selected evaluators at the library of the Faculty of Medicine of Tampico Dr Alberto Romo Caballero at the Universidad Autónoma de Tamaulipas, Tampico, Tamaulipas, Mexico, to present a training video in English by Stoyanov et al [72]. Following the video presentation, a training exercise was suggested for all evaluators using an app called Anatomymaster. This app, also focused on human anatomy, was not included in the study sample, as it did not meet the requirement of having a user rating ≥4.3.

The evaluators downloaded and tested the trial app for at least 10 minutes before completing the MARS web-based questionnaire. If any individual evaluation score varied by at least 2 points, the evaluators discussed it to reach a consensus, ensuring a uniform understanding of each item.

After completing the trial exercise, the 10 selected evaluators evaluated the 18 human anatomy mobile apps during June 2024. Each evaluator was provided with a list of apps, which they downloaded and used for 10 minutes before completing a web-based evaluation instrument designed based on the MARS. Each item from the different sections was rated using a Likert scale (1-5). The collected data were initially recorded in Excel.

Statistical Analysis

Overview

Statistical analyses, including the calculation of the intraclass correlation coefficient (ICC) and Pearson correlation coefficient, were performed using SPSS (version 29.0.2.0; IBM Corp).

Intraclass Correlation Coefficient

The ICC was used using a random effects model of 2 factors with consistency to measure the overall agreement between the quantitative measurements obtained by different evaluators [73]. The ICC ranges from 0 to 1, with 0 indicating a total lack of reliability among evaluators and 1 representing perfect reliability. According to the 95% CI for ICC estimation, values below 0.5 are considered “poor” reliability, those between 0.5 and 0.75 are considered “moderate,” those between 0.75 and 0.9 are considered “good,” and those above 0.90 are rated as “excellent” [74].

The individual ICC was calculated for each of the sections A, B, C, and D. The arithmetic means of each section were used to calculate the ICC for the overall MARS quality score (ie, sections A, B, C, and D). In section A, 900 data points were considered, accounting for 10 evaluators, 5 items, and 18 mobile apps. In section B, 720 data points were used (ie, 10 evaluators, 4 items, and 18 mobile apps). In section C, 540 data points were analyzed (ie, 10 evaluators, 3 items, and 18 mobile apps). Finally, in section D, 1080 data points were considered (ie, 10 evaluators, 6 items, and 18 mobile apps). In this last section, item 19 was excluded because of missing values; therefore, only 6 items were considered instead of 7. For each MARS section (ie, A, B, C, and D), the arithmetic mean and SD were calculated.

Pearson Correlation

The statistical technique of Pearson correlation was used to evaluate the relationship between the MARS quality scores (ie, sections A, B, C, and D) and the subjective item 23 from section E. The software used was SPSS.

Ethical Considerations

This study did not involve experiments on humans, animals, or the collection of sensitive personal data. It focused exclusively on evaluating publicly available mobile applications on Google Play through a structured analysis conducted by health care professionals. No direct interaction with developers or users of the applications took place, and no private or identifiable information was accessed or stored.

This type of research, which does not involve mental health e-communities or sensitive data, does not require institutional review board approval. Furthermore, the activities were carried out following the institutional policies and local guidelines of the Universidad Autónoma de Tamaulipas, Mexico, for observational research and technological product reviews.

This decision aligns with international ethical standards and aims to ensure transparency and accountability in the evaluation of publicly available technological products.

Results

Selection Criteria for Anatomy Mobile Apps

Figure 1 shows the PRISMA diagram applied to the selection of mobile apps; a total of 724 apps were identified in the Google Play store under the search criterion *anatomy*. In total, 54 duplicate apps were eliminated either because they had the same name or different names but the same content, leaving 670 in the screening phase. In the eligibility phase, the inclusion criteria were applied, where only 75 apps met these characteristics. In the same way, 57 apps that were more than 2 years old without receiving updates from the developer were excluded, finally leaving 18 apps in the inclusion phase.

In Table 1, the names of the selected apps, their identification screens, developer names, required Android operating system

version, and the date of the last update are presented. All the listed mobile apps focus on human anatomy, are available in English or Spanish, have a user rating above 4.3, can be downloaded for free, and have more than 100,000 downloads.

Figure 1. Flowchart of the selection process.

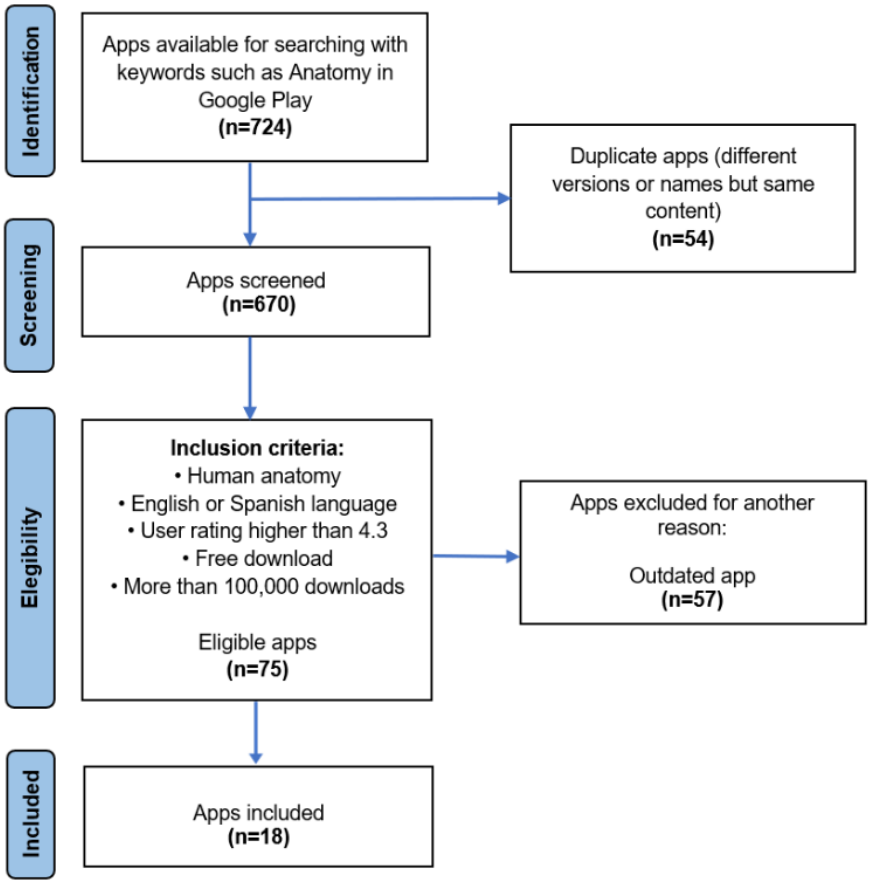


Table 1. Characteristics of selected mobile apps (Google Play, 2024).

App name	Identification screen	Developer	Android version	Last update
Anatomy Learning-Anatomía 3D		3D Medical OU	7.0 and later versions	June 16, 2024
Complete Anatomy 2024		3D4Medical from Elsevier	7.0 and later versions	May 28, 2024
Biodigital Human-3D Anatomy		BioDigital	5.0 and later versions	February 28, 2024
Anatomía–Atlas 3D		Catfish Animation Studio	8.0 and later versions	August 21, 2023
Anatomyka-Anatomy 3D		Woodoo Art s.r.o.	5.1 and later versions	November 17, 2023
VOKA 3d Anatomy and Physiology		Factory of innovations and solutions LLC	8.0 and later versions	January 27, 2024
Organos internos 3D (anatomía)		Ing Víctor Michel González Galván	5.1 and later versions	November 1, 2023
Teach Me Anatomy		TeachMeSeries Ltd	7.0 and later versions	May 22, 2024
3D Bones and Organs (Anatomy)		Education Mobile	5.1 and later versions	September 3, 2023
Esqueleto|Anatomía 3D		Catfish Animation Studio	8.0 and later versions	August 21, 2023
Visual Anatomy Lite		Education Mobile	4.4 and later versions	August 9, 2023
Gray's Anatomy-Anatomy Atlas		SEStudio	4.4 and later versions	March 1, 2023
El cuerpo humano en 3D		Mozaik Education	5.0 and later versions	May 30, 2024
e-Anatomy		IMAIOS SAS	5.0 and later versions	June 11, 2024
Sistema muscular 3D (Anatomía)		Ing. Víctor Michel González Galván	5.1 and later versions	November 27, 2023
Sistema óseo en 3D (Anatomía)		Ing. Víctor Michel González Galván	5.1 and later versions	November 6, 2023
Anatomy by Muscle & Motion		Muscle and Motion	5.0 and later versions	April 2, 2024
Flashcards de Daily Anatomy		Kenhub	5.0 and later versions	September 21, 2023

Evaluation of Mobile Apps

MARS Overall Quality Scores

The average overall MARS quality score (ie, sections A, B, C, and D) was rated as "good" (mean 4.02, SD 0.20) [75]. The 3 mobile apps with the highest overall MARS quality scores (ie, averages of sections A, B, C, and D) were Organos internos 3D (anatomía) (mean 4.34, SD 0.29), Sistema óseo en 3D (Anatomía) (mean 4.32, SD 0.28), and VOKA Anatomy Pro (mean 4.29, SD 0.28). In contrast, the apps with the lowest overall MARS quality scores were Anatomy–3D Atlas (mean 3.66, SD 0.27), Complete Anatomy 2024 (mean 3.73, SD 0.32), and Visual Anatomy Lite (mean 3.80, SD 0.28). The average overall MARS quality scores are listed in Table 2.

Table 2. Average Mobile App Rating Scale quality scores (sections A, B, C, and D).

App name	Section A—engagement (mean 3.69, SD 0.20), mean (SD)	Section B—functionality (mean 4.36, SD 0.22), mean (SD)	Section C— aesthetics (mean 4.14, SD 0.21), mean (SD)	Section D—information (mean 3.90, SD 0.23), mean (SD)	Arithmetic average A, B, C, and D (mean 4.02, SD 0.20), mean (SD)
Organos internos 3D (anatomía)	3.96 (0.19)	4.65 (0.19)	4.43 (0.23)	4.30 (0.13)	4.34 (0.29)
Sistema óseo en 3D (Anatomía)	3.98 (0.28)	4.63 (0.19)	4.47 (0.15)	4.22 (0.17)	4.32 (0.28)
VOKA 3 d Anatomy and Physiology	3.94 (0.11)	4.60 (0.24)	4.37 (0.12)	4.23 (0.05)	4.29 (0.28)
Anatomy Learning-Anatomía 3D	3.88 (0.15)	4.55 (0.13)	4.27 (0.31)	4.20 (0.24)	4.22 (0.28)
Flashcards de Daily Anatomy	3.82 (0.28)	4.55 (0.10)	4.30 (0.20)	4.05 (0.10)	4.18 (0.32)
Teach Me Anatomy	3.76 (0.24)	4.48 (0.15)	4.40 (0.20)	3.95 (0.15)	4.15 (0.35)
Anatomyka- Anatomía 3D	3.74 (0.23)	4.50 (0.14)	4.23 (0.42)	4.00 (0.11)	4.12 (0.32)
3D Bones and organs (Anatomy)	3.72 (0.33)	4.53 (0.05)	4.03 (0.12)	3.90 (0.18)	4.04 (0.35)
El cuerpo humano en 3D	3.64 (0.21)	4.38 (0.21)	4.20 (0.26)	3.92 (0.10)	4.03 (0.32)
Sistema muscular 3D (Anatomía)	3.84 (0.27)	4.33 (0.15)	4.13 (0.42)	3.83 (0.05)	4.03 (0.24)
Biodigital Human-3D Anatomy	3.66 (0.31)	4.30 (0.08)	4.10 (0.44)	3.80 (0.06)	3.97 (0.29)
Anatomy by Muscle & Motion	3.66 (0.24)	4.45 (0.13)	4.07 (0.38)	3.60 (0.13)	3.94 (0.40)
e-Anatomy	3.54 (0.36)	4.20 (0.28)	4.07 (0.15)	3.87 (0.08)	3.92 (0.29)
Gray's Anatomy-Anatomy Atlas	3.58 (0.32)	4.18 (0.29)	3.93 (0.06)	3.72 (0.08)	3.85 (0.26)
Esqueleto|Anatomía 3D	3.48 (0.13)	4.10 (0.34)	4.00 (0.10)	3.75 (0.15)	3.83 (0.28)
Visual Anatomy Lite	3.44 (0.05)	4.08 (0.35)	3.97 (0.06)	3.70 (0.06)	3.80 (0.28)
Complete Anatomy 2024	3.34 (0.09)	4.05 (0.31)	3.93 (0.06)	3.58 (0.08)	3.73 (0.32)
Anatomía–Atlas 3D	3.36 (0.05)	4.00 (0.08)	3.70 (0.30)	3.57 (0.23)	3.66 (0.27)

MARS Section Scores (Sections A, B, C, and D)

The mean (SD) for the "engagement" section (ie, section A) was 3.69 (0.20). The 3 top-rated apps in this section were Sistema óseo en 3D (Anatomía) (mean 3.98, SD 0.28), Organos internos 3D (anatomía) (mean 3.96, SD 0.19), and VOKA Anatomy Pro (mean 3.94, SD 0.19). In contrast, the 3 apps with the lowest scores in section A were Complete Anatomy 2024 (mean 3.34, SD 0.09), Anatomy–3D Atlas (mean 3.36, SD 0.05), and Visual Anatomy Lite (mean 3.44, SD 0.05).

For the "functionality" section (ie, section B), the mean (SD) was 4.36 (0.22). The top-rated apps were Organos internos 3D (anatomía) (mean 4.65, SD 0.19), Sistema óseo en 3D (Anatomía) (mean 4.63, SD 0.19), and VOKA Anatomy Pro (mean 4.60, SD 0.24). Conversely, the 3 apps with the lowest scores in this section were Anatomy–3D Atlas (mean 4.00, SD 0.08), Complete Anatomy 2024 (mean 4.05, SD 0.31), and Visual Anatomy Lite (mean 4.08, SD 0.35).

Regarding the "aesthetics" section (ie, section C), the mean (SD) was 4.14 (0.21). The highest-rated apps were Sistema óseo en 3D (Anatomía) (mean 4.47, SD 0.15), Organos internos 3D (anatomía) (mean 4.43, SD 0.23), and Teach Me Anatomy (mean 4.40, SD 0.20). In contrast, the lowest-scoring apps in this section were Anatomy–3D Atlas (mean 3.70, SD 0.30), Complete Anatomy 2024 (mean 3.93, SD 0.06), and Gray's Anatomy-Anatomy Atlas (mean 3.93, SD 0.06).

For the "information quality" section (ie, section D), the mean (SD) was 3.90 (0.23). The highest-rated apps were Organos internos 3D (anatomía) (mean 4.30, SD 0.13), VOKA Anatomy Pro (mean 4.23, SD 0.05), and Sistema óseo en 3D (Anatomía) (mean 4.22, SD 0.17). In contrast, the lowest-scoring apps in this section were Anatomy–3D Atlas (mean 3.57, SD 0.23), Complete Anatomy 2024 (mean 3.58, SD 0.08), and Anatomy by Muscle and Motion (mean 3.60, SD 0.13; Table 2).

Subjective Quality Evaluation (Section E) and Perceived Effectiveness (Section F)

The general mean (SD) for the "subjective quality" section (ie, section E) was 3.63 (0.22). The 3 top-rated mobile apps in this section were VOKA Anatomy Pro (mean 3.95, SD 0.10), Organos internos 3D (anatomía) (mean 3.93, SD 0.17), and Sistema óseo en 3D (Anatomía) (mean 3.88, SD 0.22). Conversely, the 3 apps with the lowest scores in this section were Anatomy–Atlas 3D (mean 3.25, SD 0.10), Complete Anatomy 2024 (mean 3.28, SD 0.15), and Esqueleto|Anatomía 3D (mean 3.35, SD 0.19). The average scores for section E are listed in Table 3.

Regarding the "perceived effectiveness" section (ie, section F), the recorded mean (SD) was 3.65 (0.18). The 3 top-rated apps with the highest scores were Organos internos 3D (anatomía) (mean 3.93, SD 0.10), VOKA Anatomy Pro (mean 3.90, SD 0.11), and Sistema óseo en 3D (Anatomía) (mean 3.87, SD 0.15). In contrast, the 3 apps with the lowest scores in this

section were Complete Anatomy 2024 (mean 3.37, SD 0.20), Anatomy–Atlas 3D (mean 3.40, SD 0.13), and Visual Anatomy Lite (mean 3.45, SD 0.15). The average scores for section F are listed in Table 4.

The section with the highest score was "functionality" (ie, section B), with a mean (SD) of 4.36 (0.22), followed by "aesthetics" (ie, section C), which scored a mean (SD) of 4.14 (0.21). In the third place was "information quality" (ie, section D), with a mean (SD) of 3.90 (0.22), followed by "engagement" (ie, section A), with a mean (SD) of 3.69 (0.20). The fifth position corresponded to "perceived effectiveness" (ie, section F), with a mean (SD) of 3.65 (0.18), while the sixth and final position was occupied by "subjective quality" (ie, section E), with a mean (SD) of 3.63 (0.22). It is notable that the app-specific score (ie, section F) was higher than the subjective quality score (ie, section E), although the latter was lower than the overall MARS quality score (mean 4.02, SD 0.20). This is demonstrated in Figure 2.

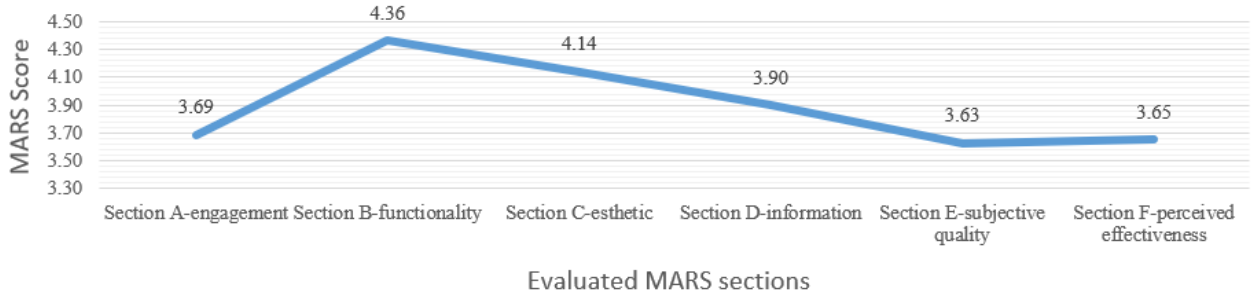
Table 3. Average score for "subjective quality" (section E).

App name	Section E—subjective quality (mean 3.63, SD 0.22), mean (SD)
VOKA 3d Anatomy and Physiology	3.95 (0.10)
Organos internos 3D (anatomía)	3.93 (0.17)
Sistema óseo en 3D (Anatomía)	3.88 (0.22)
Anatomy Learning-Anatomía 3D	3.85 (0.10)
Sistema muscular 3D (Anatomía)	3.83 (0.17)
Flashcards de Daily Anatomy	3.78 (0.10)
Teach Me Anatomy	3.75 (0.10)
Anatomyka-Anatomía 3D	3.73 (0.15)
Anatomy by Muscle & Motion	3.63 (0.15)
El cuerpo humano en 3D	3.60 (0.14)
3D Bones and organs (Anatomy)	3.58 (0.17)
Biodigital Human-3D Anatomy	3.55 (0.19)
Gray's Anatomy-Anatomy Atlas	3.53 (0.13)
e-Anatomy	3.48 (0.17)
Visual Anatomy Lite	3.40 (0.14)
Esqueleto|Anatomía 3D	3.35 (0.19)
Complete Anatomy 2024	3.28 (0.15)
Anatomía–Atlas 3D	3.25 (0.10)

Table 4. Average score for "perceived effectiveness" (section F).

App name	Section F—perceived effectiveness (mean 3.65, SD 0.18), mean (SD)
Organos internos 3D (anatomía)	3.93 (0.10)
VOKA 3d Anatomy and Physiology	3.90 (0.11)
Sistema óseo en 3D (Anatomía)	3.87 (0.15)
Anatomy Learning-Anatomía 3D	3.82 (0.21)
Flashcards de Daily Anatomy	3.80 (0.14)
Sistema muscular 3D (Anatomía)	3.78 (0.12)
3D Bones and organs (Anatomy)	3.73 (0.21)
Anatomyka-Anatomía 3D	3.72 (0.08)
Teach Me Anatomy	3.70 (0.17)
Anatomy by Muscle & Motion	3.60 (0.11)
El cuerpo humano en 3D	3.58 (0.08)
Biodigital Human-3D Anatomy	3.57 (0.15)
Gray's Anatomy-Anatomy Atlas	3.55 (0.22)
Esqueleto|Anatomía 3D	3.48 (0.35)
e-Anatomy	3.47 (0.15)
Visual Anatomy Lite	3.45 (0.15)
Anatomía–Atlas 3D	3.40 (0.13)
Complete Anatomy 2024	3.37 (0.20)

Figure 2. Average scores by the Mobile App Rating Scale (MARS) sections.



Average Scores by Section and Item

The following section details the items that received the highest and lowest scores in the various evaluated sections. In section A, item 5 concerning the target population received a mean score of 3.87 (SD 0.24), whereas item 4 related to interactivity scored a mean value of 3.52 (SD 0.20). In section B, item 6 on performance received the highest score with a mean value of 4.57 (0.14), followed by item 8 related to navigation with a mean score of 4.24 (SD 0.23). In section C, item 11 concerning graphics received a mean score of 4.37 (SD 0.28), and item 12 on visual appeal scored a mean value of 4.02 (SD 0.26). In section D, item 17 on the quality of visual information received a mean score of 3.98 (SD 0.24), whereas item 18 on credibility scored a mean value of 3.83 (SD 0.31). In section E, item 23 on overall quality received a mean score of 3.79 (SD 0.20), and item 22 on willingness to pay for the app scored a mean value of 3.52 (SD 0.24). Finally, in section F, the item with the highest overall mean score was "attitudes," with a mean value of 3.77 (SD 0.23), whereas the item "help seeking" received the lowest score (mean 3.54, SD 0.23). The mean scores with respect to the section and items are listed in Table 5.

Table 5. Mean scores by section and item.

Section and item	Scores, mean (SD)
Section A—engagement (mean 3.69, SD 0.20)	
Item 1. Entertainment	3.67 (0.35)
Item 2. Interests	3.78 (0.28)
Item 3. Customization	3.59 (0.20)
Item 4. Interactivity	3.52 (0.20)
Item 5. Target population	3.87 (0.24)
Section B—functionality (mean 4.36, SD 0.22)	
Item 6. Performance	4.57 (0.14)
Item 7. Ease of use	4.37 (0.37)
Item 8. Navigation	4.24 (0.23)
Item 9. Gestural design of the app	4.27 (0.23)
Section C—aesthetics (mean 4.14, SD 0.21)	
Item 10. Design	4.04 (0.19)
Item 11. Graphics	4,37 (0.28)
Item 12. Visual appeal	4.02 (0.26)
Section D—quality of information (mean 3.90, SD 0.23)	
Item 13. Accuracy of information description	3.86 (0.23)
Item 14. Objectives	3.95 (0.25)
Item 15. Quality of information	3.87 (0.24)
Item 16. Amount of information	3.91 (0.26)
Item 17. Quality of visual information	3.98 (0.24)
Item 18. Credibility	3.83 (0.31)
Item 19. Evidence base	__[a]
Section E—subjective quality (mean 3.63, SD 0.22)	
Item 20. Would you recommend this app?	3.58 (0.26)
Item 21. How many times would you use this app?	3.61 (0.23)
Item 22. Would you pay for this app?	3.52 (0.24)
Item 23. General qualifications	3.79 (0.20)
Section F—perceived effectiveness (mean 3.65, SD 0.18)	
Awareness	3.67 (0.27)
Knowledge	3.64 (0.22)
Attitudes	3.77 (0.23)
Intention to change	3.57 (0.22)
Help seeking	3.54 (0.23)
Behavior change	3.71 (0.17)

[a]No apps presented explicit scientific support in the descriptions and comments.

MARS Overall Quality Scores and Star Rating (Item 23)

The overall MARS quality scores were higher than the scores for item 23 (ie, subjective quality). Similarly, the overall MARS star ratings (ie, item 23) were lower than the star ratings in the Google Play store (Table 6).

Table 6. Overall MARSa quality scores, overall star ratings for item 23, and star ratings in the Google Play store.

App name	Health professionals		Users	
	MARS overall quality score (mean 4.02)	MARS (item 23; mean 3.79)	Star rating in the Google Play store (mean 4.63)	Downloads
Organos internos 3D (anatomía)	4.34	4.10	4.90	>5,000,000
Sistema óseo en 3D (Anatomía)	4.32	4.10	4.90	>1,000,000
VOKA 3d Anatomy and physiology	4.29	4.10	4.80	>100,000
Anatomy Learning-Anatomía 3D	4.22	4.00	4.80	>10,000,000
Flashcards de Daily Anatomy	4.18	3.90	4.80	>500,000
Teach Me Anatomy	4.15	3.90	4.70	>1,000,000
Anatomyka-Anatomía 3D	4.12	3.90	4.70	>500,000
3D Bones and organs (Anatomy)	4.04	3.80	4.70	>1,000,000
El cuerpo humano en 3D	4.03	3.80	4.60	>1,000,000
Sistema muscular 3D (Anatomía)	4.03	3.80	4.80	>1,000,000
Biodigital Human-3D Anatomy	3.97	3.70	4.60	>500,000
Anatomy by Muscle & Motion	3.94	3.70	4.60	>500,000
e-Anatomy	3.92	3.70	4.50	>1,000,000
Gray's Anatomy-Anatomy Atlas	3.85	3.70	4.60	>1,000,000
Esqueleto|Anatomía 3D	3.83	3.60	4.40	>1,000,000
Visual Anatomy Lite	3.80	3.60	4.40	>1,000,000
Complete Anatomy 2024	3.73	3.50	4.30	>1,000,000
Anatomía–Atlas 3D	3.66	3.40	4.30	>1,000,000

[a]MARS: Mobile App Rating Scale.

Statistical Analysis

ICC (Assessment Reliability)

The average reliability measures of the evaluation ranged from "good" to "excellent." In the engagement section (ie, section A), an ICC of 0.892 (95% CI 0.807-0.952) was obtained. In the functionality section (ie, section B), the ICC was 0.901 (95% CI 0.822-0.956). In the aesthetics section (ie, section C), an ICC of 0.866 (95% CI 0.758-0.941) was recorded. In the information quality section (ie, section D), the ICC was 0.890 (95% CI 0.804-0.951). In the subjective quality section (ie, section E), an ICC of 0.862 (95% CI 0.751-0.939) was obtained. Finally, in the app specificity section (ie, section F), an ICC of 0.868 (95% CI 0.764-0.941) was recorded. Similarly, the reliability of the overall MARS quality evaluation (ie, average of sections A, B, C, and D) was classified as "excellent," with an ICC of 0.912 (95% CI 0.820-0.963).

Pearson Correlation

For the calculation of Pearson correlation coefficient, the average MARS quality scores and the scores for subjective item 23 from section E presented earlier in Table 6 were considered. The result showed an excellent correlation (r=0.989, P<.001; Table 7). Also, the 95% CI for this correlation was 0.971 to 0.996, based on the Fisher r-to-z transformation.

Table 7. Pearson correlation results.

Correlations	MARS overall quality score	MARS (item 23)
MARS overall quality score		
Pearson correlation	1	0.989[a]
P value (bilateral)	__[b]	<.001
N	19	19
MARS (item 23)		
Pearson correlation	0.989[a]	1
P value (bilateral)	<.001	—
N	19	19

[a]The correlation is significant at the .01 level (2 sided).

[b]Not applicable.

Discussion

Overview

The primary objective of this study was to identify and assess the quality of mobile apps related to human anatomy available on Google Play using the MARS. This scale focuses on the usability and accessibility of mobile health apps, considering aspects such as engagement, functionality, aesthetics, information quality, subjective quality, and app specificity. The MARS organizes the evaluations of the apps into 3 different dimensions. The first dimension includes sections A, B, C, and D and focuses on the evaluation of the objective technical items. The evaluations in the second and third dimensions are subjective and are divided into 2 sections: section E, which considers the evaluator's personal appreciation, and section F, which focuses on the perceived effectiveness. These 3 dimensions are crucial because, while mobile health apps must meet functionality and design standards, the evaluator's perception and the app's impact are determinants for its adoption. In addition, the MARS sections cannot be considered in isolation, as they are interrelated and influence each other.

Principal Findings

In the first dimension of the MARS, the best-rated section was "functionality" with a mean score of 4.36, followed by "aesthetics" with a mean score of 4.14, "information quality" with a mean score of 3.90, and, finally, "engagement" with a mean score of 3.69, which was the least valued. Although the apps generally received a good average score, it is crucial to examine the relatively low ratings in fundamental aspects such as engagement (mean score 3.69) and look for solutions. In order to strengthen the engagement section (ie, section A), which is made up of items 1 to 5 of the MARS (ie, entertainment, interest, personalization, interactivity, and target population), specific recommendations can be applied for each item. In item 1 (ie, entertainment), it is suggested to integrate elements such as gamification, challenges, achievements, rewards, or progressive levels and use good quality graphics, attractive colors, animations, and multimedia content (eg, videos and music) to make the user experience more attractive. In item 2 (ie, interest), it is recommended to include new content and personalized reminders. For item 3 (ie, personalization), it is recommended that users be able to adjust themes, difficulty levels, colors, or display modes according to their preferences, as well as the use of artificial intelligence to offer suggestions based on the user's preferences. Regarding item 4 (ie, interactivity), it is suggested to incorporate interactive content such as questionnaires and practical activities that require active participation, as well as real-time communication to forums, live chats, or social interactions to encourage collaboration between users and provide immediate feedback to correct errors or recognize achievements. Finally, in item 5 (ie, target population), it is recommended to carry out previous studies on the characteristics and needs of the target population, such as age, educational level, and cultural context, and ensuring that the content, graphics, and design are consistent with the population for which the app was designed.

The average general quality score according to the MARS (ie, sections A, B, C, and D) was good (mean score 4.02), supported by excellent reliability with an ICC of 0.912 and a 95% CI of 0.820 to 0.963. The mobile apps that excelled in overall quality according to the MARS (ie, sections A, B, C, and D) were Organos internos 3D (anatomía) with a mean score of 4.34, Sistema óseo en 3D (Anatomía) with a mean score of 4.32, and VOKA Anatomy Pro with a mean score of 4.29. These results indicate that these apps, having received high scores and offering high-quality content, can be recommended for users interested in learning human anatomy. In the second and third dimensions of the MARS, corresponding to sections E and F, where the impact on the user is more significant, the lowest average scores were recorded: subjective quality with a mean score of 3.63 and app specificity with a mean score of 3.65. These ratings were even lower than the general quality score of the MARS, which was 4.02.

These results underscore the importance of conducting a thorough analysis of all the 3 dimensions of the MARS; otherwise, apps that are technically well developed might be overvalued, whereas those that receive better subjective ratings from users could be overlooked. This indicates that developers of human anatomy mobile apps should not only address aspects of functionality, aesthetics, engagement, and information but also actively consider user perception and the impact of their

apps. There are various practical uses of the study's results, such as a more appropriate selection of mobile apps in the student context or in medical practice, where those that obtained a higher score in the MARS evaluation are chosen, which provide greater reliability and comfort in use.

Another relevant point of discussion is that the MARS, specifically item 19 (ie, section D), which addresses "information quality," assesses whether mobile apps have scientific foundations that support their usefulness. However, the apps evaluated in this study did not present explicit scientific support in the descriptions and comments provided by the developers, which is why they lack a rating in item 19 of the MARS evaluation.

Therefore, it is crucial that research centers and universities get involved in the development of mobile health apps so that they are supported by scientific research and can be hosted in app stores to make them accessible to the general public. Collaboration among software developers, health professionals, researchers, and academics in the creation and review of educational materials for a medical mobile app would generate greater confidence in its use. In addition, conducting validation studies in real learning environments also plays an important role in assessing the quality and effectiveness of apps through various methodologies, such as the MARS framework discussed in this study.

Limitations

The main limitations are the exclusion of paid apps, apps in languages other than English or Spanish, and apps with a star ratings less than 4.3. In addition, the search was limited to apps present in the Google Play store. Although these criteria may seem restrictive, English is the predominant language in global medical education, ensuring that the evaluated apps covered a substantial portion of the app market. However, the exclusions may limit the scope, particularly by omitting paid apps, which in certain cases may offer higher-quality content that could facilitate and enhance the learning of anatomy. To address these limitations in future research, inclusion criteria could be expanded to incorporate human anatomy mobile apps available in other languages or those that are paid, creating a broader repertoire for analysis. This approach would also enable comparative studies, such as exploring potential differences between free apps and those requiring a license or payment.

Conclusions

This study provides a comprehensive and detailed analysis of apps available for teaching human anatomy, aimed at health care professionals, medical students, and interested users. For example, students and health professionals can both use a human anatomy mobile app before orthopedic surgery to consult a 3D model of the leg of a patient with a femur fracture. This would allow them to more accurately understand the location of bones, blood vessels, and muscles in the affected region, contributing to greater success in the procedure.

Overall, the evaluated apps demonstrated high quality, particularly excelling in functionality and aesthetic design. However, some apps need to improve aspects such as user engagement (ie, section A) and the quality of the information provided (ie, section D). Among the highest-rated apps according to the MARS are Organos internos 3D (anatomía), Sistema óseo en 3D (Anatomía), and VOKA Anatomy Pro.

The subjective MARS score (ie, item 23) was 3.79, in contrast to the average rating of 4.63 given by users on the Google Play store. This suggests that evaluators provided lower ratings, whereas users tend to overrate the apps. This discrepancy may stem from the fact that evaluators typically adhere to more rigorous and objective criteria, systematically assessing technical, functional, and usability aspects. Professional evaluators are often more critical regarding technical implementation and practical utility.

In contrast, users base their ratings on personal and subjective experiences, scoring according to their expectations and the level of satisfaction experienced while using the app. Both perspectives offer valuable feedback: on the one hand, an objective evaluation of quality, and on the other hand, a subjective evaluation of user satisfaction. This difference in ratings does not negatively impact the overall MARS evaluation of the apps. Instead, it provides a perspective where both developers and potential users can identify strengths and areas for improvement from complementary approaches.

This study highlights the evolving role of mobile apps as transformative tools in medical education by offering innovative solutions for accessibility and interactivity in learning. Mobile apps use advanced features such as 3D models, simulations, and dynamic interfaces, and these tools overcome the limitations of traditional methods of teaching human anatomy, such as the scarcity of cadavers and high costs of dissection laboratories. In addition, they facilitate personalized learning of topics and selection of difficulty levels. They allow continuous access, allowing students to practice and reinforce their knowledge anytime, anywhere.

To maximize the impact of mobile apps in medical education, we suggest strategies focused on design and functionality, such as the incorporation of gamification elements, challenges, and rewards to increase user motivation, as well as strengthening interactivity through real-time feedback, collaborative learning tools, and interactive clinical cases. It is essential to align the content of the apps with medical education curricula to ensure their relevance and applicability. Similarly, we recommend combining their use with traditional methods, such as face-to-face classes and laboratory practice, to offer a comprehensive learning experience. Training teachers to integrate these tools into their teaching methodologies is also essential. Finally, to guarantee both the scientific rigor and the accessibility of these mobile apps, we propose collaboration with universities and research centers to develop content based on solid scientific evidence.

Data Availability

The data used in this study are available upon request from the corresponding author.

Conflicts of Interest

None declared.

References

1. Fan K, Zhao Y. Mobile health technology: a novel tool in chronic disease management. Intell Med 2022 Feb;2(1):41-47 [FREE Full text] [doi: 10.1016/j.imed.2021.06.003]
2. Opipari-Arrigan L, Dykes DMH, Saeed SA, Thakkar S, Burns L, Chini BA, et al. Technology-enabled health care collaboration in pediatric chronic illness: pre-post interventional study for feasibility, acceptability, and clinical impact of an electronic health record-linked platform for patient-clinician partnership. JMIR Mhealth Uhealth 2020 Nov 26;8(11):e11968 [FREE Full text] [doi: 10.2196/11968] [Medline: 33242014]
3. Tabi K, Randhawa A, Choi F, Mithani Z, Albers F, Schnieder M, et al. Mobile apps for medication management: review and analysis. JMIR Mhealth Uhealth 2019 Sep 11;7(9):e13608 [FREE Full text] [doi: 10.2196/13608] [Medline: 31512580]
4. Ryan S, Ní Chasaide N, O' Hanrahan S, Corcoran D, Caulfield B, Argent R. mHealth apps for musculoskeletal rehabilitation: systematic search in app stores and content analysis. JMIR Rehabil Assist Technol 2022 Aug 01;9(3):e34355 [FREE Full text] [doi: 10.2196/34355] [Medline: 35916688]
5. Zahid T, Alyafi R, Bantan N, Alzahrani R, Elfirt E. Comparison of effectiveness of mobile app versus conventional educational lectures on oral hygiene knowledge and behavior of high school students in Saudi Arabia. Patient Prefer Adherence 2020;14:1901-1909 [FREE Full text] [doi: 10.2147/PPA.S270215] [Medline: 33116434]
6. Caponnetto P, Prezzavento G, Casu M, Polosa R, Quattropani M. Psychological factors, digital health technologies, and best asthma management as three fundamental components in modern care: a narrative review. Appl Sci 2024 Apr 16;14(8):3365 [FREE Full text] [doi: 10.3390/app14083365]
7. Dobrina R, Starec A, Brunelli L, Orzan E, De Vita C, Bicego L, et al. Applying the participatory slow design approach to a mHealth application for family caregivers in pediatric ear, nose, and throat surgery. Healthcare (Basel) 2024 Feb 08;12(4):442 [FREE Full text] [doi: 10.3390/healthcare12040442] [Medline: 38391818]
8. López-Marcos JJ, Díaz-Arribas MJ, Valera-Calero JA, Navarro-Santana MJ, Izquierdo-García J, Ortiz-Gutiérrez RM, et al. The added value of face-to-face supervision to a therapeutic exercise-based app in the management of patients with chronic low back pain: a randomized clinical trial. Sensors 2024 Jan 16;24(2):567 [FREE Full text] [doi: 10.3390/s24020567] [Medline: 38257659]
9. Giannopoulou P, Vrahatis A, Papalaskari MA, Vlamos P. The RODI mHealth app insight: machine-learning-driven identification of digital indicators for neurodegenerative disorder detection. Healthcare (Basel) 2023 Nov 19;11(22):2985 [FREE Full text] [doi: 10.3390/healthcare11222985] [Medline: 37998477]
10. Boszko M, Krzowski B, Peller M, Hoffman P, Żurawska N, Skoczylas K, et al. Impact of AfterAMI mobile app on quality of life, depression, stress and anxiety in patients with coronary artery disease: open label, randomized trial. Life (Basel) 2023 Oct 05;13(10):2015 [FREE Full text] [doi: 10.3390/life13102015] [Medline: 37895396]
11. Sánchez-Rodríguez MT, Pinzón-Bernal MY, Jiménez-Antona C, Laguarta-Val S, Sánchez-Herrera-Baeza P, Fernández-González P, et al. Designing an informative app for neurorehabilitation: a feasibility and satisfaction study by physiotherapists. Healthcare (Basel) 2023 Sep 14;11(18):2549 [FREE Full text] [doi: 10.3390/healthcare11182549] [Medline: 37761746]
12. Choudhury A, Shahsavar Y, Sarkar K, Choudhury M, Nimbarte A. Exploring perceptions and needs of mobile health interventions for nutrition, anemia, and preeclampsia among pregnant women in underprivileged Indian communities: a cross-sectional survey. Nutrients 2023 Aug 24;15(17):3699 [FREE Full text] [doi: 10.3390/nu15173699] [Medline: 37686731]
13. Martínez C, Martínez A, Uclés V, Jenkins M, Badilla A. Aplicaciones móviles para la rehabilitación de pacientes con cáncer de mama. Iberian Mag Inf Syst Technol 2022;50:310-321 [FREE Full text]
14. Wang L, Langlais CS, Kenfield SA, Chan JM, Graff RE, Allen IE, et al. mHealth interventions to promote a healthy diet and physical activity among cancer survivors: a systematic review of randomized controlled trials. Cancers (Basel) 2022 Aug 06;14(15):3816 [FREE Full text] [doi: 10.3390/cancers14153816] [Medline: 35954479]
15. Elbagoury B, Vladareanu L, Vlădăreanu V, Salem A, Travediu AM, Roushdy M. A Hybrid Stacked CNN and residual feedback GMDH-LSTM deep learning model for stroke prediction applied on mobile AI smart hospital platform. Sensors (Basel) 2023 Mar 27;23(7):350 [FREE Full text] [doi: 10.3390/s23073500] [Medline: 37050561]
16. Zarkogianni K, Chatzidaki E, Polychronaki N, Kalafatis E, Nicolaides N, Voutetakis A, et al. The ENDORSE feasibility study: exploring the use of m-Health, artificial intelligence and serious games for the management of childhood obesity. Nutrients 2023 Mar 17;15(6):1451 [FREE Full text] [doi: 10.3390/nu15061451] [Medline: 36986180]
17. Ali Sherazi B, Laeer S, Krutisch S, Dabidian A, Schlottau S, Obarcanin E. Functions of mHealth diabetes apps that enable the provision of pharmaceutical care: criteria development and evaluation of popular apps. Int J Environ Res Public Health 2022 Dec 21;20(1):64 [FREE Full text] [doi: 10.3390/ijerph20010064] [Medline: 36612402]

18. Needamangalam Balaji J, Prakash S, Park Y, Baek J, Shin J, Rajaguru V, et al. A scoping review on accentuating the pragmatism in the implication of mobile health (mHealth) technology for tuberculosis management in India. J Pers Med 2022 Sep 28;12(10):1599 [FREE Full text] [doi: 10.3390/jpm12101599] [Medline: 36294738]
19. Miró J, Lleixà-Daga M, de la Vega R, Llorens-Vernet P, Jensen M. A mobile application to help self-manage pain severity, anxiety, and depressive symptoms in patients with fibromyalgia syndrome: a pilot study. Int J Environ Res Public Health 2022 Sep 23;19(19):12026 [FREE Full text] [doi: 10.3390/ijerph191912026] [Medline: 36231327]
20. Rashid N, Chen X, Mohamad Marzuki MF, Takshe AA, Okasha A, Maarof F, et al. Development and usability assessment of a mobile app (demensia KITA) to support dementia caregivers in Malaysia: a study protocol. Int J Environ Res Public Health 2022 Sep 20;19(19):11880 [FREE Full text] [doi: 10.3390/ijerph191911880] [Medline: 36231181]
21. Teong LF, Khor BH, Radion Purba K, Gafor A, Goh BL, Bee BC, et al. A mobile app for triangulating strategies in phosphate education targeting patients with chronic kidney disease in Malaysia: development, validation, and patient acceptance. Healthcare (Basel) 2022 Mar 14;10(3):535 [FREE Full text] [doi: 10.3390/healthcare10030535] [Medline: 35327013]
22. Abreu M, Carmo A, Franco A, Parreira S, Vidal B, Costa M, et al. Mobile applications for epilepsy: where are we? Where should we go? A systematic review. Signals 2022 Feb 03;3(1):40-65 [FREE Full text] [doi: 10.3390/signals3010005]
23. Hutchcraft ML, Wallon RC, Fealy SM, Jones D, Galvez R. Evaluation of the road to birth software to support obstetric problem-based learning education with a cohort of pre-clinical medical students. Multimodal Technol Interact 2023;7(8):84 [FREE Full text] [doi: 10.3390/mti7080084]
24. Patil S, Rao A, Pardeshi N, Gavhane M, Waghole D. Augmented reality in anatomy. Int J Eng Res Technol 2021;10(4):55 [FREE Full text]
25. Boomgaard A, Fritz K, Isafiade O, Kotze R, Ekpo O, Smith M, et al. A novel immersive anatomy education system (Anat_Hub): redefining blended learning for the musculoskeletal system. Appl Sci 2022 Jun 03;12(11):5694 [FREE Full text] [doi: 10.3390/app12115694]
26. Khan T, Johnston J, Ophoff J. The impact of an augmented reality application on learning motivation of students. Adv Hum Comput Interact 2019 Feb 03;2019:1-14 [FREE Full text] [doi: 10.1155/2019/7208494]
27. Küçük S, Kapakin S, Göktaş Y. Learning anatomy via mobile augmented reality: effects on achievement and cognitive load. Anat Sci Educ 2016 Oct;9(5):411-421 [FREE Full text] [doi: 10.1002/ase.1603] [Medline: 26950521]
28. Jamali SS, Shiratuddin MF, Wong K, Oskam CL. Utilising mobile-augmented reality for learning human anatomy. Procedia Soc Behav Sci 2015 Jul;197:659-668 [FREE Full text] [doi: 10.1016/j.sbspro.2015.07.054]
29. Moro C, Štromberga Z, Raikos A, Stirling A. The effectiveness of virtual and augmented reality in health sciences and medical anatomy. Anat Sci Educ 2017 Nov;10(6):549-559 [FREE Full text] [doi: 10.1002/ase.1696] [Medline: 28419750]
30. Kurniawan M, Suharjito, Diana, Witjaksono G. Human anatomy learning systems using augmented reality on mobile application. Procedia Comput Sci 2018;135:80-88 [FREE Full text] [doi: 10.1016/j.procs.2018.08.152]
31. Cercenelli L, De Stefano A, Billi A, Ruggeri A, Marcelli E, Marchetti C, et al. AEducaAR, anatomical education in augmented reality: a pilot experience of an innovative educational tool combining AR technology and 3D printing. Int J Environ Res Public Health 2022 Jan 18;19(3):1024 [FREE Full text] [doi: 10.3390/ijerph19031024] [Medline: 35162049]
32. Martínez G, Mir F, Garcia L. Caracterización de las aplicaciones móviles para la enseñanza y el aprendizaje de la anatomía humana. Enseñanza de las ciencias. Núm. Extra 0; 2017. URL: https://ddd.uab.cat/record/184395 [accessed 2024-04-27]
33. Montaner Sanchis A, Gumbau Puchol V, Villalba Ferrer F, Eleuterio Cerveró G. Mobile learning en la anatomía humana: estudio del mercado de aplicaciones. Educ Médica 2022 Mar;23(2):100726 [FREE Full text] [doi: 10.1016/j.edumed.2022.100726]
34. Wainman B, Wolak L, Pukas G, Zheng E, Norman G. The superiority of three-dimensional physical models to two-dimensional computer presentations in anatomy learning. Med Educ 2018 Nov;52(11):1138-1146 [FREE Full text] [doi: 10.1111/medu.13683] [Medline: 30345680]
35. Havens K, Saulovich N, Saric K. A case report about anatomy applications for a physical therapy hybrid online curriculum. J Med Libr Assoc 2020 Apr;108(2):295-303 [FREE Full text] [doi: 10.5195/jmla.2020.825] [Medline: 32256241]
36. Lopes I, Teixeira B, Cortez P, de Silva G, de Sousa Neto A, de Sousa Leal N. Use of human cadavers in teaching of human anatomy in Brazilian medical faculties. Acta Sci Biol Sci 2017 May 03;39(1):1 [FREE Full text] [doi: 10.4025/actascibiolsci.v39i1.33860]
37. Siddique A, Krebs M, Alvarez S, Greenspan I, Patel A, Kinsolving J, et al. Mobile apps for the care management of chronic kidney and end-stage renal diseases: systematic search in app stores and evaluation. JMIR Mhealth Uhealth 2019 Sep 04;7(9):e12604 [FREE Full text] [doi: 10.2196/12604] [Medline: 31486408]
38. Quach S, Michaelchuk W, Benoit A, Oliveira A, Packham TL, Goldstein R, et al. Mobile heath applications for self-management in chronic lung disease: a systematic review. Netw Model Anal Health Inform Bioinform 2023;12(1):25 [FREE Full text] [doi: 10.1007/s13721-023-00419-0] [Medline: 37305790]
39. Paganini S, Meier E, Terhorst Y, Wurst R, Hohberg V, Schultchen D, et al. Stress management apps: systematic search and multidimensional assessment of quality and characteristics. JMIR Mhealth Uhealth 2023 Aug 29;11:e42415 [FREE Full text] [doi: 10.2196/42415] [Medline: 37642999]

40. Lull C, von Ahnen JA, Gross G, Olsavszky V, Knitza J, Leipe J, et al. German mobile apps for patients with psoriasis: systematic search and evaluation. JMIR Mhealth Uhealth 2022 May 26;10(5):e34017 [FREE Full text] [doi: 10.2196/34017] [Medline: 35617014]
41. Messner E, Sturm N, Terhorst Y, Sander LB, Schultchen D, Portenhauser A, et al. Mobile apps for the management of gastrointestinal diseases: systematic search and evaluation within app stores. J Med Internet Res 2022 Oct 05;24(10):e37497 [FREE Full text] [doi: 10.2196/37497] [Medline: 36197717]
42. Salazar A, de Sola H, Failde I, Moral-Munoz J. Measuring the quality of mobile apps for the management of pain: systematic search and evaluation using the mobile app rating scale. JMIR Mhealth Uhealth 2018 Oct 25;6(10):e10718 [FREE Full text] [doi: 10.2196/10718] [Medline: 30361196]
43. Sharif M, Alkadhimi A. Patient focused oral hygiene apps: an assessment of quality (using MARS) and knowledge content. Br Dent J 2019 Sep 13;227(5):383-386 [FREE Full text] [doi: 10.1038/S41415-019-0665-0]
44. Kanoute A, Carrouel F, Gare J, Dieng S, Dieng A, Diop M, et al. Evaluation of oral hygiene-related mobile apps for children in sub-Saharan Africa. Int J Environ Res Public Health 2022 Oct 01;19(19):12565 [FREE Full text] [doi: 10.3390/ijerph191912565] [Medline: 36231862]
45. Martinon P, Saliasi I, Bourgeois D, Smentek C, Dussart C, Fraticelli L, et al. Nutrition-related mobile apps in the French app stores: assessment of functionality and quality. JMIR Mhealth Uhealth 2022 Mar 14;10(3):e35879 [FREE Full text] [doi: 10.2196/35879] [Medline: 35285817]
46. Talwar D, Tseng T, Foster M, Xu L, Chen L. Genetics/genomics education for nongenetic health professionals: a systematic literature review. Genet Med 2017 Jul;19(7):725-732 [FREE Full text] [doi: 10.1038/gim.2016.156] [Medline: 27763635]
47. Mandracchia F, Llauradó E, Tarro L, Valls R, Solà R. Mobile phone apps for food allergies or intolerances in app stores: systematic search and quality assessment using the Mobile App Rating Scale (MARS). JMIR Mhealth Uhealth 2020 Sep 16;8(9):e18339 [FREE Full text] [doi: 10.2196/18339] [Medline: 32936078]
48. Romero R, Kates F, Hart M, Ojeda A, Meirom I, Hardy S. Quality of deaf and hard-of-hearing mobile apps: evaluation using the Mobile App Rating Scale (MARS) with additional criteria from a content expert. JMIR Mhealth Uhealth 2019 Oct 30;7(10):e14198 [FREE Full text] [doi: 10.2196/14198] [Medline: 31670695]
49. Escriche-Escuder A, De-Torres I, Roldán-Jiménez C, Martín-Martín J, Muro-Culebras A, González-Sánchez M, et al. Assessment of the quality of mobile applications (apps) for management of low back pain using the Mobile App Rating Scale (MARS). Int J Environ Res Public Health 2020 Dec 09;17(24):9209 [FREE Full text] [doi: 10.3390/ijerph17249209] [Medline: 33317134]
50. Scala L, Giglioni G, Bertazzoni L, Bonetti F. The efficacy of the smartphone app for the self-management of low back pain: a systematic review and assessment of their quality through the Mobile Application Rating Scale (MARS) in Italy. Life (Basel) 2024 Jun 13;14(6):760 [FREE Full text] [doi: 10.3390/life14060760] [Medline: 38929744]
51. Rendell R, Pinheiro M, Wang B, McKay F, Ewen A, Carnegie C, et al. Digital apps to improve mobility in adults with neurological conditions: a health app-focused systematic review. Healthcare (Basel) 2024 Apr 30;12(9):929 [FREE Full text] [doi: 10.3390/healthcare12090929] [Medline: 38727486]
52. Chao SM, Wang ML, Fang YW, Lin ML, Chen SF. Mobile apps for patients with peritoneal dialysis: systematic app search and evaluation. Healthcare (Basel) 2024 Mar 25;12(7):719 [FREE Full text] [doi: 10.3390/healthcare12070719] [Medline: 38610142]
53. Lee J, Kim Y. Evaluation of mobile applications for patients with diabetes mellitus: a scoping review. Healthcare (Basel) 2024 Jan 31;12(3):368 [FREE Full text] [doi: 10.3390/healthcare12030368] [Medline: 38338253]
54. Holl F, Schobel J, Swoboda W. Mobile apps for COVID-19: a systematic review of reviews. Healthcare (Basel) 2024 Jan 08;12(2):139 [FREE Full text] [doi: 10.3390/healthcare12020139] [Medline: 38255029]
55. Wasserman S, Ould Brahim L, Attiya A, Belzile E, Lambert S. An evaluation of interactive mHealth applications for adults living with cancer. Curr Oncol 2023 Jul 25;30(8):7151-7166 [FREE Full text] [doi: 10.3390/curroncol30080518] [Medline: 37622999]
56. Wang SW, Chiou CC, Su CH, Wu CC, Tsai SC, Lin TK, et al. Measuring mobile phone application usability for anticoagulation from the perspective of patients, caregivers, and healthcare professionals. Int J Environ Res Public Health 2022 Aug 16;19(16):10136 [FREE Full text] [doi: 10.3390/ijerph191610136] [Medline: 36011765]
57. Kuo HL, Chang CH, Ma WF. A survey of mobile apps for the care management of patients with dementia. Healthcare (Basel) 2022 Jun 23;10(7):1173 [FREE Full text] [doi: 10.3390/healthcare10071173] [Medline: 35885700]
58. McAleese D, Linardakis M, Papadaki A. Quality and presence of behaviour change techniques in mobile apps for the Mediterranean diet: a content analysis of android Google Play and Apple app store apps. Nutrients 2022 Mar 18;14(6):1290 [FREE Full text] [doi: 10.3390/nu14061290] [Medline: 35334947]
59. Scantling-Birch Y, Naveed H, Mukhija R, Nanavaty MA. A review of smartphone apps used for Toric intraocular lens calculation and alignment. Vision (Basel) 2022 Feb 18;6(1):13 [FREE Full text] [doi: 10.3390/vision6010013] [Medline: 35225972]
60. Mohammadzadeh N, Khenarinezhad S, Ghazanfarisavadkoohi E, Safari MS, Pahlevanynejad S. Evaluation of m-Health applications use in epilepsy: a systematic review. Iran J Public Health 2021 Mar;50(3):459-469 [FREE Full text] [doi: 10.18502/ijph.v50i3.5586] [Medline: 34178793]

61. Martín-Martín J, Muro-Culebras A, Roldán-Jiménez C, Escriche-Escuder A, De-Torres I, González-Sánchez M, et al. Evaluation of Android and Apple Store depression applications based on mobile application rating scale. Int J Environ Res Public Health 2021 Nov 27;18(23):12505 [FREE Full text] [doi: 10.3390/ijerph182312505] [Medline: 34886232]
62. Mack C, Terhorst Y, Stephan M, Baumeister H, Stach M, Messner EM, et al. "Help in a heartbeat?": A systematic evaluation of mobile health applications (apps) for coronary heart disease. Int J Environ Res Public Health 2021 Sep 30;18(19):10323 [FREE Full text] [doi: 10.3390/ijerph181910323] [Medline: 34639623]
63. Larco A, Carrillo J, Chicaiza N, Yanez C, Luján-Mora S. Moving beyond limitations: designing the Helpdys app for children with dyslexia in rural areas. Sustainability 2021 Jun 24;13(13):7081 [FREE Full text] [doi: 10.3390/su13137081]
64. Lian X, Sunar M. Mobile augmented reality technologies for autism spectrum disorder interventions: a systematic literature review. Appl Sci 2021 May 17;11(10):4550 [FREE Full text] [doi: 10.3390/app11104550]
65. Choi J, Chung C, Woo H. Diet-related mobile apps to promote healthy eating and proper nutrition: a content analysis and quality assessment. Int J Environ Res Public Health 2021 Mar 28;18(7):3496 [FREE Full text] [doi: 10.3390/ijerph18073496] [Medline: 33800531]
66. Weekly T, Walker N, Beck J, Akers S, Weaver M. A review of apps for calming, relaxation, and mindfulness interventions for pediatric palliative care patients. Children (Basel) 2018 Jan 26;5(2):16 [FREE Full text] [doi: 10.3390/children5020016] [Medline: 29373515]
67. Mobile operating system market share worldwide. StatCounter. URL: https://gs.statcounter.com/os-market-share/mobile/worldwide [accessed 2024-04-29]
68. Manterola C, Zavando D, Cartes-Velásquez R, Otzen T, Sanhueza A. Initial validation of a scale to measure methodological quality in prognosis studies. The MInCir proposal. Int J Morphol 2018 Jun;36(2):762-767 [FREE Full text] [doi: 10.4067/S0717-95022018000200762]
69. Wang X, Cheng Z. Cross-sectional studies: strengths, weaknesses, and recommendations. Chest 2020 Jul;158(1S):S65-S71 [FREE Full text] [doi: 10.1016/j.chest.2020.03.012] [Medline: 32658654]
70. Manterola C, Hernández-Leal M, Otzen T, Espinosa M, Grande L. Estudios de Corte Transversal. Un Diseño de Investigación a Considerar en Ciencias Morfológicas. Int J Morphol 2023 Feb;41(1):146-155 [FREE Full text] [doi: 10.4067/S0717-95022023000100146]
71. von Elm E, Altman D, Egger M, Pocock S, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2007 Oct 20;335(7624):806-808 [FREE Full text] [doi: 10.1136/bmj.39335.541782.AD] [Medline: 17947786]
72. Stoyanov S, Hides L, Kavanagh D, Zelenko O, Tjondronegoro D, Mani M. Mobile app rating scale: a new tool for assessing the quality of health mobile apps. JMIR Mhealth Uhealth 2015 Mar 11;3(1):e27 [FREE Full text] [doi: 10.2196/mhealth.3422] [Medline: 25760773]
73. Martínez JA, Pérez PS. Intraclass correlation coefficient. Semergen 2023;49(23):101907 [FREE Full text] [doi: 10.1007/springerreference_205392]
74. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163 [FREE Full text] [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
75. Ahmadi S, Klingelhöfer D, Erbe C, Holzgreve F, Groneberg DA, Ohlendorf D. Oral health: global research performance under changing regional health burdens. Int J Environ Res Public Health 2021 May 27;18(11):5743 [FREE Full text] [doi: 10.3390/ijerph18115743] [Medline: 34071884]

Abbreviations

AR: augmented reality
ICC: intraclass correlation coefficient
MARS: Mobile App Rating Scale
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by AH Sapci; submitted 19.07.24; peer-reviewed by D Patel, T Gladman; comments to author 05.10.24; revised version received 18.11.24; accepted 03.01.25; published 14.02.25.

Please cite as:
Rivera García GE, Cervantes López MJ, Ramírez Vázquez JC, Llanes Castillo A, Cruz Casados J
Reviewing Mobile Apps for Teaching Human Anatomy: Search and Quality Evaluation Study
JMIR Med Educ 2025;11:e64550
URL: https://mededu.jmir.org/2025/1/e64550
doi:10.2196/64550
PMID:

Original Paper

Exploring Gender Perspectives in Medical Education: Latent Semantic Analysis of Israeli First-Year Medical Students' Reflections

Rola Khamisy-Farah[1,2], MD; Raymond Farah[1,3], MD; Haneen Jabaly-Habib[1,4], MD; Yara Nakhleh Francis[1,5], MD; Nicola Luigi Bragazzi[6,7], MPH, MD, PhD

[1]Azrieli Faculty of Medicine, Bar-Ilan University, Safed, Israel

[2]Clalit Health Services, Akko, Israel

[3]Department of Internal Medicine B, Ziv Medical Center, Safed, Israel

[4]Ophthalmology Department Tzafon Medical Center, Poria, Israel

[5]Department of Obstetrics and Gynecology, Galilee Medical Center, Nahariya, Israel

[6]Laboratory for Industrial and Applied Mathematics (LIAM), Department of Mathematics and Statistics, York University, Toronto, ON, Canada

[7]Department of Computer Science, Data Science, and Information Technology, Faculty of Natural and Applied Sciences, Sol Plaatje University, Kimberley, South Africa

Corresponding Author:
Nicola Luigi Bragazzi, MPH, MD, PhD
Laboratory for Industrial and Applied Mathematics (LIAM)
Department of Mathematics and Statistics
York University
4700 Keele Street
Toronto, ON, M3J 1P3
Canada
Phone: 1 416 736 2100
Email: robertobragazzi@gmail.com

Abstract

Background: Gender is increasingly recognized as a crucial determinant of health and health care delivery. Integrating gender-sensitive content into medical education is essential for cultivating socially responsive, culturally competent, and clinically effective physicians of the future. However, limited research has examined how medical students conceptualize gender in clinical contexts, particularly through their own reflective narratives.

Objective: This study explores the thematic landscape of gender-related perceptions among first-year medical students in Israel following a mandatory course in gender medicine. Using latent semantic analysis (LSA), we examined how students reflected on gendered dimensions of health care and how these reflections varied by gender and ethnicity.

Methods: First-year medical students enrolled in the four-year path of medicine in Israel participated in a compulsory gender medicine course and were invited to submit anonymous written reflections. A total of 83 students (n=52, 63%, females; n=31, 37%, males; n=68, 82%, Jewish; and n=15, 18%, Arab) submitted responses, which were preprocessed and analyzed using LSA. The texts were lemmatized and vectorized to construct a term-document matrix, followed by singular value decomposition for dimensionality reduction. Ten latent topics were extracted, and thematic labels were assigned through an inductive, consensus-based coding procedure. Subgroup analyses were conducted by gender and ethnicity.

Results: LSA identified 10 distinct topics, accounting for 56.6% of the total variance in the overall sample. The most dominant theme was Gendered Patient-Doctor Interactions (eigenvalue=121.188; 28.1% variance; 527 terms; 75 documents), followed, in terms of variance, by Gender-Specific Diseases and Health Concerns (5.7%) and Cultural and Religious Influences on Health Care (4.3%). Reflections from female students introduced 3 unique themes: Gendered Help-Seeking and Familial Roles (2.8%), Gender and Health Education (2.5%), and Gendered Communication and Advocacy (2.2%). Male students uniquely discussed Perceived Gender Bias in Clinical and Research Settings (3.8%) and the Legal and Ethical Dimensions of Reproductive Health Care (3.3%). Among Jewish students, additional themes included Population-Level Framing of Gendered Conditions (3.7%) and Gendered Youth Expectations (2.1%). Arabic students contributed culturally specific themes, such as Modesty and Cultural Norms (8.6%), Paternal Authority and Structural Discrimination (6.3%), and Reproductive Vulnerability (3.6%).

Conclusions: Thematic patterns in student reflections suggest that gender medicine curricula are effective in fostering critical engagement with diverse gendered realities in clinical care. The emergence of culturally grounded and gender-specific themes underscores the importance of tailoring educational interventions to reflect student diversity.

(JMIR Med Educ 2025;11:e78371) doi:10.2196/78371

KEYWORDS

gender medicine; medical education; latent semantic analysis; cultural competence; reflective writing; student narratives; diversity; curricular assessment

Introduction

Sex and gender are distinct yet interrelated factors that shape health, health care access, and, ultimately, health outcomes in multiple and profound ways [1]. While sex refers to biological attributes such as chromosomes, hormones, and reproductive anatomy that typically categorize individuals as male or female, gender encompasses the socially constructed roles, behaviors, expressions, identities, and power relationships associated with being a man, woman, or gender-diverse person [2,3]. As a social determinant of health, gender intersects with biological factors as well as with other cultural, societal, and systemic constructs to influence the incidence, prevalence, and manifestation of diseases, diagnostic pathways, therapeutic decisions, clinical interactions, and patient experiences across the care *continuum* [4-6].

Despite increasing recognition of these dynamics and interwoven interplays, medical education has historically followed a reductionist biomedical model that often underrepresents or insufficiently integrates sex- and gender-sensitive perspectives [7]. In recent years, sex- and gender-based medicine has emerged as a multidisciplinary field aimed at addressing these gaps by incorporating the complexities of sex- and gender-specific differences into clinical education, research, and practice [8].

Embedding sex- and gender-based medicine into undergraduate medical curricula is essential for preparing future physicians to recognize and respond to sex- and gender-based disparities in health care delivery, enhance diagnostic accuracy, and foster more equitable patient care [9,10]. However, most medical schools do not offer a formal, integrated curriculum in sex- and gender-sensitive medicine; when included, content is often confined to reproductive health and rarely extends to other specialties such as cardiology, pharmacology, or psychiatry. Educators have expressed concerns about the lack of standardized materials, institutional support, and faculty preparedness, all of which hinder implementation. Students, while generally receptive, often report minimal exposure to these topics, and when presented, content is frequently framed in binary, biologically deterministic terms [11]. Despite these limitations, increasing evidence shows that structured educational interventions can raise awareness of implicit bias and promote more equitable clinical practice [12,13].

Yet little is known about how students internalize and reflect on sex- and gender-sensitive content when it is presented within a formal curriculum [14]. This gap underscores the need for qualitative and data-driven evaluations of student perspectives to better inform curriculum design and pedagogical strategies. Narrative reflection provides a window into students' evolving attitudes and cognitive engagement with sex and gender in clinical contexts [14,15]. It is a tool grounded in constructivist learning theory, which posits that learners actively construct knowledge through experience, reflection, and social interaction rather than passively absorbing facts [15]. Rooted in the work of Piaget, Vygotsky, and others, constructivism emphasizes that meaning-making is contextual and shaped by prior knowledge, cultural frameworks, and active engagement. In health professions education, constructivist approaches have gained traction as educators recognize the limitations of didactic instruction in preparing students for the complex, value-laden, and relational aspects of clinical care, such as those encountered in sex- and gender-based medicine. Reflective writing further enables learners to make sense of the social and ethical dimensions of medicine, including issues of power, identity, and equity [16].

By leveraging narrative reflection and a computational linguistic framework, this study aims to elucidate how students conceptualize sex and gender in medical practice and to identify thematic patterns that vary by gender and ethnicity within a diverse multicultural educational context. Through written reflections, students engage in metacognitive processing of sex- and gender-related content presented during the course. These narratives serve not only as evidence of individual meaning-making but also as artifacts that reveal the evolving cognitive, emotional, and ethical dimensions of learners' understanding. Within the constructivist paradigm, reflection functions as both a tool and a product of learning, enabling students to surface biases, grapple with complexity, and reconcile new information with existing worldviews.

Specifically, we asked (1) what thematic patterns emerge in first-year medical students' reflections following a formal course in sex- and gender-based medicine? and (2) how do these patterns differ by students' gender and ethnicity? Based on prior literature on sex and gender awareness in medical education [12,13,17,18], we hypothesized that the reflections would reveal themes capturing both clinical and sociocultural dimensions of sex and gender in medicine. These themes were expected to include recognition of sex- and gender-specific differences in disease incidence, prevalence, presentation, and outcomes; awareness of gender bias in clinical practice as presented during the course; and acknowledgment of the role of sociocultural norms in shaping health care experiences. Given the absence of prior extensive clinical exposure, we anticipated that students' narratives would primarily integrate knowledge gained from the structured curriculum with their preexisting personal,

cultural, and societal perspectives on sex, gender, and health. We further expected thematic emphases to vary by students' gender, with potential differences in the extent to which narratives addressed interpersonal, affective, and patient-centered aspects of care alongside structural, biomedical, or systems-level considerations. We also anticipated that students from different ethnic backgrounds would foreground culturally specific norms, values, and challenges in applying sex- and gender-based medicine principles. By articulating these differences, this study aims to provide medical educators with evidence to design more culturally responsive and pedagogically effective curricula in sex- and gender-sensitive medicine.

Methods

Study Population

The participants were first-year medical students enrolled in the 4-year medical program at the Azrieli Faculty of Medicine, Bar-Ilan University, Safed, Israel. They attended a mandatory sex- and gender-based medicine course as part of their curriculum. All students provided written reflections on their experiences and learning outcomes, which constituted the textual dataset for analysis.

Ethical Considerations

This study adhered to ethical guidelines for educational research. Written informed consent was obtained from all students before data collection. Institutional ethical approval was granted by the Human Subjects Institutional Review Board at Bar-Ilan University, Safed, Israel (approval number 040325496). Participation was entirely voluntary, and students' reflections were anonymized and used solely for academic and research purposes.

Course Syllabus

The course Sex- and Gender-Based Medicine was designed to introduce and deepen understanding of key concepts, theories, and practices in sex- and gender-sensitive medicine, emphasizing how sociocultural and institutional determinants shape health disparities.

Over 7 structured first-year sessions, each consisting of a 90-minute plenary lecture followed by a 90-minute small-group practice, students developed both conceptual and applied competencies for integrating a sex- and gender-aware perspective into clinical practice and medical research.

Using Bloom's taxonomy [19], the learning progression moved from remembering (defining sex and gender; describing biological, social, and cultural determinants) and understanding (explaining how sex and gender influence biological systems, disease manifestation, health-seeking behavior, and access to care) to applying (using a sex- and gender-sensitive lens to analyze case studies such as women with chest pain or men with breast cancer), analyzing (identifying bias in biomedical research design and interpretation), evaluating (assessing the impact of policy, feminist theories, and global perspectives on health care delivery), and creating (formulating strategies to address inequities and adapt medical innovations for diverse populations).

Learning activities combined plenary lectures with small-group sessions featuring case-based discussions, simulations with standardized patients, podcast analyses, and reflective exercises. The curriculum integrated national and global perspectives, traced the historical evolution of feminist thought and its influence on medicine, examined gender gaps in research, and critically analyzed the effects of bias on health care delivery and innovation.

Students were required to attend at least 80% of sessions and submit a written reflection, for instance, applying a sex- and gender-sensitive framework to a clinical scenario, thereby demonstrating mastery of course objectives across multiple levels of cognitive engagement.

Reflective Writing Assignment and Prompts

At the conclusion of the Sex- and Gender-Based Medicine course, students were required to submit a 1-page written reflection as a mandatory graded assignment. The prompt invited them to describe and critically analyze their learning experience, drawing on lectures, small-group discussions, case-based exercises, or personal experiences relevant to sex- and gender-sensitive medicine. Students were encouraged to focus either on a specific event or on their overall experience, integrating description, interpretation, and self-analysis.

The assignment was structured around the DIEP (Describe, Interpret, Evaluate, and Plan) model guiding students to (1) describe the chosen experience or case; (2) interpret the actions, motives, emotions, and contextual factors involved; (3) evaluate their own responses and the quality of their engagement; and (4) plan how to apply the insights gained to future academic, clinical, or research contexts. To support the reflective process, students were given stepwise guidance: identifying the experience, describing the learning process, engaging in introspection, recognizing shifts in perspective or values, and directing the reflection toward future practice. They were encouraged to write in the first person and connect their reflections to personal perspectives and professional aspirations. Model phrases were provided to help articulate reflective thinking (eg, "This experience made me aware of...", "Looking back at what happened..."). All submissions received individualized feedback aimed at deepening reflective skills.

Data Collection and Preprocessing

The dataset consisted of written responses submitted by students at the conclusion of the course. These reflections were collected electronically and compiled into a structured corpus for text mining. The responses varied in length and complexity, reflecting diverse perspectives on gender-related topics in medicine. Before analysis, all textual data underwent preprocessing to ensure the reliability of results. This included tokenization, stop-word removal, stemming, and lemmatization. Text was converted to lowercase, and punctuation and nonalphabetical characters were removed to standardize the dataset. Common medical terminology and gender-related terms were retained to preserve contextual relevance. Preprocessing was performed using the English stopword list and English-language stemming algorithms.

Statistical Analysis

This study used latent semantic analysis (LSA) to examine the perceptions and conceptualizations of gender medicine among Israeli first-year medical students. LSA is a computational information-processing method for humanlike, meaning-based cognitive tasks, including the extraction of conceptual meaning from large textual *corpora*. It addresses the question of "how word and passage meaning can be constructed from experience with language, that is, by what mechanisms—instinctive, learned, or both—this can be accomplished" [20]. In LSA, "words do not have meanings on their own..., words get their meanings from their mapping" [20], and it is "the underlying map (that) is (to say) the primitive substrate that gives words meaning, not vice versa" [20]. LSA recognizes that people rarely communicate using isolated words but typically speak and comprehend clusters of words, often extending beyond a single sentence. Such units of meaning are more accurately represented at the level of paragraph-length discourse.

More technically speaking, LSA relies on the "compositional constraint," which posits that "the representation of any meaningful passage must be composed as a function of the representations of the words it contains" [20]:

"LSA learns about the meaning of a word from every meeting with it and from the composition of all the passages in which it does not occur...Word meaning is latent in the evidence of experience and can be extracted from natural linguistic data" [20].

The LSA technique has been increasingly applied in health professions education to analyze reflective narratives and explore students' mental models [21,22]. By identifying patterns of word co-occurrence and reducing linguistic dimensionality, LSA detects latent themes that may not be readily apparent through manual coding alone. It thus offers a scalable, data-driven approach to evaluating the impact of educational interventions, particularly in areas involving subjective and contextually embedded content such as gender.

A term-document matrix was constructed using a bag-of-words representation with a minimum term frequency of 2 and a sparsity threshold of 0.975. Singular value decomposition was applied to reduce dimensionality and reveal key semantic relationships between words. The number of topics was determined using an optimal singular value selection approach to capture the most relevant themes without overfitting the data. LSA was first performed on the overall sample to identify global themes and semantic structures, followed by stratified analyses by gender and ethnicity.

Following LSA, the resulting topics were manually reviewed and labeled according to their most representative words. Two independent researchers (RKF and NLB) conducted the topic interpretation to ensure reliability, resolving discrepancies through discussion until consensus was reached. A hybrid coding approach was applied: an initial a priori codebook was developed based on the study's research questions and prior literature on sex- and gender-based medicine. This preliminary framework included categories expected to capture both clinical and sociocultural dimensions of student reflections. The codebook was then iteratively refined in a data-driven manner, consistent with constructivist theory, through manual review of the topic-modeling outputs and a subset of student narratives. This process allowed for the addition of new codes and the consolidation or clarification of existing ones. In doing so, the final set of codes reflected both theoretically informed expectations and emergent themes in the data, thereby enhancing the validity and reproducibility of the analysis.

The codebook, including a priori categories, their definitions, and emergent subcodes within each category, is presented in Table 1.

All LSA procedures were conducted using XLSTAT (Lumivero, LLC).

Table 1. Codebook for the thematic analysis.

A priori category	Definition	Emergent subcodes
Clinical Applications of Sex- and Gender-Based Medicine	Recognition of how sex and gender influence disease incidence and prevalence, presentation, outcomes, pharmacology, and decision-making in clinical contexts	Gender and Decision-Making in Medicine; Gender Differences in Pharmacology; Gender Differences in Pharmacology and Disease Education; Gender-Specific Diseases and Health Concerns; Reproductive Health and Feelings of Vulnerability
Sociocultural Influences on Health	Influence of cultural, religious, and social norms on health beliefs, access, and care delivery	Cultural and Religious Influences on Health Care; Cultural and Social Attitudes Toward Gender and Health; Cultural Norms in Clinical Settings; Gendered Help-Seeking and Familial Roles; Gendered Youth, Expectations, and Identity Formation; Intergenerational and Familial Perspectives on Gendered Health
Educational and Institutional Contexts	The role of medical education, institutional structures, and team dynamics in shaping awareness and practice of gender-sensitive medicine	Gender and Health Education; Gender Equality and Team Dynamics; Gender Roles in the Medical Profession; Institutional and Interpersonal Perspectives on Gender; Perceived Gender Bias in Clinical and Research Settings
Health Access and Equity Issues	Barriers and facilitators to equitable health care for diverse gender identities	Gender and Public Health Framing of Conditions; Health Care Access for Transgender Individuals; Legal and Ethical Dimensions of Gendered Health Care
Gendered Communication and Relationships	How gender influences patient-provider interactions, communication styles, and relational dynamics in care	Gendered Clinical Encounters; Gendered Communication and Assertion in Health Care Contexts; Gendered Patient-Doctor Interactions
Psychological and Emotional Dimensions of Gender and Health	Emotional, mental health, and psychosocial aspects linked to gender identity or experience	Paternal Authority and Structural Discrimination; Psychological and Emotional Dimensions of Gender and Health; Psychosocial Support and Gendered Distress

Results

Overview

All 83 students were recruited. Of these, 52 (63%) identified as female and 31 (37%) as male. In terms of ethnicity, 68 participants (82%) identified as Jewish and 15 (18%) as Arab.

In the overall sample analysis, LSA (Table 2) revealed a rich and multidimensional thematic structure, highlighting the broad spectrum of gender-related issues perceived by students in the context of medical education and clinical care.

In total, 10 distinct topics were identified. Topic 1 exhibited the highest eigenvalue (121.188), accounting for approximately 28.1% of the total variance, and demonstrated high lexical richness (527 terms) as well as strong document representation (75 documents). The primary terms within this topic indicate that students engaged critically with gendered dynamics in patient-doctor interactions, emphasizing gender as a determinant of health and its implications for clinical communication, empathy, and diagnostic equity. Beyond this dominant theme, topic 2, with an eigenvalue of 54.519 (5.68% of the explained variance), reflected students' awareness of gender-specific health concerns, particularly those disproportionately affecting women. This thematic axis reflects recognition of both the biological and social dimensions of illnesses such as breast cancer and their ramifications for patient care and support systems. Topic 3, with an eigenvalue of 47.175 and accounting for 4.26% of the overall variance, indicates that students considered how religious beliefs and cultural contexts intersect with gender identity and health care experiences, particularly for conditions such as endometriosis that are not only underdiagnosed but also socially stigmatized. This highlights the students' developing sensitivity to culturally competent care and the complexities of treating individuals within diverse socioreligious frameworks. The subsequent topics (topics 4-10) each accounted for between 3.49% and 1.90% of the variance, with eigenvalues ranging from 42.725 to 31.507. Topic 4 addressed the complex role of gender in shaping professional identities and hierarchies in medicine. This theme was further reflected in topics 7 and 8, which collectively examined gender-based dynamics in medical decision-making, team interactions, and perceived equality in clinical and academic settings. Topic 7 highlighted heightened student awareness of the gendered structures influencing teamwork, authority, and career progression within medicine, while topic 5 underscored students' growing recognition of transgender health concerns and the structural barriers to inclusive health care delivery. Topic 6 focused on gender pharmacology—particularly the often underappreciated physiological and hormonal differences that affect drug efficacy and therapeutic response. Topic 9 addressed students' reflections on interpersonal and institutional factors shaping gendered experiences within medical learning environments. Finally, topic 10 conveyed a nuanced consideration of the psychological and normative dimensions of gendered illness narratives, emphasizing students' awareness of how gender norms shape both the perception and lived experience of illness, as well as

the societal frameworks governing the provision of emotional support and care.

Table 2. Labeled topics identified through latent semantic analysis of student reflections following participation in a gender medicine course. Each topic is characterized by its corresponding eigenvalue, percentage of explained variability, key terms, and the thematically assigned label.

Topic	Label	Representative quotes	Eigenvalue	% Variability	Top words
Topic 1	Gendered Patient-Doctor Interactions	"For me, this was the first time I had to deal with questions of gender identity and the complex feelings of her and the team..." ". I intend to avoid pre-assumptions about gender identity and pay attention to how the environment may affect how patients feel..."	121.188	28.08	gender, patient, women, feel, doctor
Topic 2	Gender-Specific Diseases and Health Concerns	"We learned how much awareness is lacking about the disease among men, how male patients with breast cancer have to cope with coming to a ward where all the patients are women, and how their image is affected because the disease is perceived as a women's disease..." "In our world, the word "breast" or "breasts" is perceived as a feminine and maternal organ (for feeding a baby), and telling a man that he has breasts can be very difficult for him..."	54.519	5.68	diseas, breast, cancer, partner, didn
Topic 3	Cultural and Religious Influences on Health Care	"I spoke to my grandparents and tried to explain the situation to them in simple words, without medical terminology, I clarified that it would be documented in her medical record and that there would be no societal reason to look down on her, I reminded them that preventing treatment and causing harm to a person contradicts our religion..." "One particular issue that deeply resonated with me that was raised during the course is the prolonged time it often takes to diagnose endometriosis in women..."	47.175	4.26	peopl, comfort, religi, comprehens, endometriosi
Topic 4	Gender Roles in the Medical Profession	"I thought to myself why I was actually hurt by his behavior, maybe because I wanted to prove that I could fulfill the role no less than a male paramedic and I wanted to receive the treatment I deserved as a team leader, as a woman, as someone who spoke a different language..." "Dr. *** described to us how she participated in a discussion during one of the meetings of the heads of department, she was one of the few women, if not the only woman in such a senior and managerial role, surrounded by male colleagues..."	42.725	3.49	complex, role, profess, realiti, sick
Topic 5	Health Care Access for Transgender Individuals	"In our work as doctors, we may be required to address the unique issues of the transgender population..." "The case I chose to describe occurred several years ago and concerns my own and society's attitude toward a transgender soldier..."	40.302	3.11	base, transgend, servic, unit, serv
Topic 6	Gender Differences in Pharmacology	"I knew that drugs could affect women and men differently, but I was surprised to discover the extent of the difference..." "We were told that most drugs have been and are still being studied in men only..."	38.321	2.81	lectur, drug, dose, menstrual, gap
Topic 7	Gender and Decision-Making in Medicine	"This exposure pushed me to reflect on how gender-associated stigma influences my own life and decision-making..." "Returning to what happened in this case, it is clear that unconscious gender biases influenced medical decisions..."	36.769	2.59	student, woman, decis, separ, stori

Topic	Label	Representative quotes	Eigenvalue	% Variability	Top words
Topic 8	Gender Equality and Team Dynamics	"I had always treated all the women around me with respect and it was clear to me that men and women were equal..." "My values, based on human dignity and equality, pushed me to act in the girl's best interest..."	35.082	2.35	sister, equal, team, dr, incid
Topic 9	Institutional and Interpersonal Perspectives on Gender	"I feel that I have received significant tools for dealing with questions from patients, friends, and acquaintances on various topics..." "In addition, I learned from the case the importance and need for raising awareness of the issue and the great importance of training the staff to respect the identity of each person and allow them to receive the best medical care for them..."	34.349	2.26	subject, friend, biolog, staff, liber
Topic 10	Psychological and Emotional Dimensions of Gender and Health	"It was important to me that, during the meeting, she express herself to the doctor and raise the points that concerned her, so she could feel comfortable and overcome her fear and embarrassment..." "My choice of a gynecologist (rather than a doctor) was made mainly due to discomfort or fear of being examined by a male doctor, following unpleasant experiences and situations experienced by my friends and family..."	31.507	1.90	ill, fear, syndrom, norm, save

Subanalysis by Gender

The LSA of reflections from female medical students (Table 3) revealed that while several key topics were retained from the full-sample analysis, new themes also emerged, and some existing ones underwent lexical refinement and thematic expansion.

Topic 1 remained the most dominant, with topic 2 also consistent. Topic 3, however, demonstrated a partial thematic reorientation: while previously aligned with cultural and religious dimensions, it now emphasized communication and generational perceptions over doctrinal belief. Topic 4 incorporated a broader intersectional framing of gender complexity and patient autonomy. Similarly, topic 5 integrated educational and clinical discourse, with a particular focus on conditions such as breast cancer. Distinctively, several new themes emerged uniquely within the female-only dataset. Topic 6 introduced a previously unarticulated theme of Gendered Help-Seeking and Familial Roles, possibly reflecting caregiving norms and informal consultation practices. Topic 7 clarified the psychological and emotional dimensions of gender and health. Topic 8 suggested a pedagogical and developmental framing, while topic 9 reflected an intergenerational perspective on gendered health experiences. Finally, topic 10 revealed a discourse on self-expression, advocacy, and possible resistance to gendered dismissal or invalidation.

The LSA of reflections from male medical students (Table 4) showed that while most identified topics were retained, some exhibited lexical drift and partial thematic reorientation.

For instance, topic 6 indicated a shift toward a more biologically grounded understanding of gender. Topic 8 appeared to merge interpersonal reflections with the emotional and physical dimensions of team-based care, thereby becoming lexically broader in scope. Two new themes also emerged in the male-only analysis. Topic 9 reflected critical engagement with perceived gender bias in diagnostic processes and biomedical research. Finally, topic 10 introduced a novel legal and ethical discourse on reproductive health, suggesting that at least some male students approached gendered issues through a rights-based and normative framework.

Table 3. Labeled topics identified through latent semantic analysis of student reflections following participation in a gender medicine course in the female-only sample.

Topic	Label	Representative quotes	Eigenvalue	% Variability	Top words
Topic 1	Gendered Patient-Doctor Interactions	"The patient's actions and motivations stem from the enormous difficulty of dealing with rejection from her father and the lack of acceptance of the gender change she underwent..." "When I went to the doctors to find out the cause, they said it would pass, that I was a girl experiencing stress, or that it might be anxiety..."	101.038	31.89	patient, gender, doctor, women, medic
Topic 2	Gender-Specific Diseases and Health Concerns	"I learned a lot from my sister's case, I knew that I wanted to be a doctor even before the experience we went through as a family, but the aforementioned experience sharpened for me the importance of seeing the patient as a whole..." "At that time, I had been with my partner for about a year, I asked him to accompany me to the gynecologist due to pain I was experiencing..."	44.608	6.22	diseas, examin, sister, gynecologist, hospit
Topic 3	Cultural and Social Attitudes Toward Gender and Health	"From this experience, I realized that due to my gender and the doctor's prejudices, I received treatment that, in my opinion, was improper and disparaging..." "I think that there is at least some kind of shame and discomfort in society to talk about the subject, especially for a girl and even more challenging in an environment with boys..."	39.634	4.91	girl, peopl, opinion, talk, convers
Topic 4	Health Care Access for Transgender Individuals	"During my work as a nurse in the emergency room, I treated a patient who identified as a transgender woman..." "The patient had a psychiatric background and a past suicidal experience, which stemmed from the insults and bullying she experienced for being a transgender woman..."	36.700	4.21	respect, complex, menstrual, transgend, oper
Topic 5	Gender Differences in Pharmacology and Disease Education	"For example, we talked about breast cancer, which in most cases only affects women, and that is why most of the health system focuses on women..." "I realized that the patient experienced side effects following the dosage of the drug and that gender differences that may affect the pharmacokinetics and pharmacodynamics of the drug cannot be ignored..."	33.003	3.40	breast, cancer, lectur, student, drug
Topic 6	Gendered Help-Seeking and Familial Roles	"A few months ago, my aunt had a general check-up with her gynecologist, the results showed that she had a large polyp on her cervix..." "I felt anger toward both society and my grandparents because delaying the treatment would put my aunt's health and life at risk..."	30.062	2.82	aunt, help, step, consult, recommend
Topic 7	Psychological and Emotional Dimensions of Gender and Health	"I assume that this was due to the emotional difficulties that the syndrome brings with it..." "The doctor did not delve into the impact of the pain on my moods and did not refer me to further advice, but only gave me a prescription..."	28.420	2.52	discuss, mention, syndrom, save, delv
Topic 8	Gender and Health Education	"I plan to...continuously educate myself on gender and sex-specific health issues to improve my diagnostic skills..." "Medical education that emphasizes sex and gender medicine may improve awareness of many specific diseases, especially endometriosis, and lead to improved diagnosis and treatment..."	28.052	2.46	period, member, educ, boy, reach
Topic 9	Intergenerational and Familial Perspectives on Gendered Health	"Being nervous didn't make the exam any easier, and my mother, in her response, increased my concerns..." "This interested me because my mother had breast cancer and underwent a partial mastectomy and no doctor in the process talked about this issue..."	27.066	2.29	didn, mother, comprehens, comfort, age
Topic 10	Gendered Communication and Assertion in Health Care Contexts	"I remember trying to explain to the person in charge that I believed in that girl's pain, and that I thought it was better to leave her to rest in the room than to force her to travel while she was suffering..." "I feel like I didn't handle this situation very well because I let my emotions, the feeling that I felt attacked, manage my way of talking..."	26.382	2.17	group, explain, respond, don't, attack

Table 4. Labeled topics identified through latent semantic analysis of student reflections following participation in a gender medicine course in the male-only sample.

Topic	Label	Representative quotes	Eigenvalue	% Variability	Top words
Topic 1	Gendered Patient-Doctor Interactions	"It is possible that this was caused by unconscious gender stereotypes, as although the symptoms were typical of a heart attack, the doctors initially attributed them to stress or anxiety, delaying the correct diagnosis and appropriate treatment..." "In the course I had the experience of getting exposed to gender-based challenges in the health system..."	72.569	25.97	gender, feel, women, men, medic
Topic 2	Gender-Specific Diseases and Health Concerns	"During the Gender Medicine course we were presented with a scenario that highlighted the unique challenges faced by men who have been diagnosed with breast cancer, a condition primarily associated with women..." "For men dealing with diseases that are perceived as feminine, such as breast cancer, the lack of appropriate gender support can lead to feelings of not belonging and even avoidance of receiving essential care..."	41.085	8.32	breast, cancer, incid, –, moment
Topic 3	Health Care Access for Transgender Individuals	"The incident I chose to describe happened a few years ago and deals with the attitudes of myself and society towards a transgender soldier, the incident highlights challenges in military service..." "Indeed, the soldier arrived at my unit for a professional evaluation, while the HR personnel explained to me the complex living and service conditions that must be provided to a transgender soldier and how difficult it would be to arrange these conditions in my unit and therefore, they suggested it was not advisable for me to "deal" with such a soldier..."	39.584	7.73	base, transgend, servic, unit, train
Topic 4	Cultural and Religious Influences on Health Care	"This case highlighted the need to find a balance between respecting religious beliefs and maintaining equality and inclusion..." "This experience made me change my perspective on respecting religion and beliefs..."	34.938	6.02	student, decis, separ, belief, religi
Topic 5	Gender Equality and Team Dynamics	"There is an inconsiderate and unequal view of a woman's abilities just because she is a woman..." "The society [should] advocate for social equality without discrimination of sex and gender..."	34.212	5.77	equal, dr, manag, sister, role
Topic 6	Institutional and Interpersonal Perspectives on Gender	"The two-day course in sex and gender medicine was for me a small taste of the turbulent world of reality in our small country, which is at the heart of a clash between East and West, between religion and liberalism, between boundless freedom and compulsive conservatism..." "It also connects to Israeli reality, particularly the case of Alice Miller, who applied to become a pilot and the President of Israel, who was himself a former pilot, responded to her by saying "Just as women were meant to knit socks, flying planes is the domain of men, and one cannot change the order of the world"..."	32.122	5.09	subject, sex, realiti, world, biolog
Topic 7	Psychological and Emotional Dimensions of Gender and Health	"I connected with my partner's feelings; we both felt the disdain from the doctor and were very hurt by him..." "In general, the issue of treating a patient's body with sensitivity in front of us as medical professionals is something that is very important to me, for example, not long ago, when a woman about my age in my community called me because of chest pain that required an EKG..."	30.317	4.53	partner, topic, pain, question, long

Topic	Label	Representative quotes	Eigenvalue	% Variability	Top words
Topic 8	Gender and Decision-Making in Medicine	"During the treatment of the incident, which took place at the entrance to her home, the driver on my team, who is a paramedic by medical authority, lifted the woman's shirt to look for any other areas of rash besides what we saw..." "This incident stayed with me even later, on the next shifts I worked, and I literally promised myself that next time I would not let something like this happen when I was the team leader, and that even if I had unpleasantness in front of another team member, it was better than allowing something like this to happen..."	29.472	4.28	friend, hurt, team, bodi, high
Topic 9	Perceived Gender Bias in Clinical and Research Settings	"Meanwhile, Dr. ***'s presentation on the higher prevalence of autoimmune diseases in women and the lack of gender-oriented research in this field was eye-opening..." "Interpreting these lectures together, I see a pattern of systemic bias in health care and medical research that has long neglected women's specific needs and experiences..."	27.908	3.84	symptom, research, level, bias, underestim
Topic 10	Legal and Ethical Dimensions of Gendered Health Care	"After I got home, I searched for more information on Google and found that the law grants the woman the almost exclusive right to decide [whether to terminate or continue a pregnancy], while men, on the other hand, are not entitled to decide the issue except in very specific cases..." "The comparison to Iran, a country known for its restrictive laws toward women, expressed her feelings of oppression and coercion..."	25.898	3.31	convers, side, pregnanc, right, law

Subanalysis by Ethnicity

The LSA of reflections from Jewish medical students (Table 5) revealed that the dominant theme, topic 1, accounted for 28.06% of the variance and reinforced the central, cross-cutting theme of Gendered Patient-Doctor Interactions, which consistently emerged as the core axis across all subgroups. Other key full-sample themes—such as gender-specific diseases, transgender health care, gender differences in pharmacology, gender and decision-making, and psychosocial dimensions—were retained in both lexical structure and conceptual scope. However, 2 new themes emerged in the Jewish-only analysis. Topic 5 suggested a discourse on population-level gendered conditions, possibly linking epidemiological thinking to gendered illness patterns, while topic 10 revealed a narrative centered on early socialization, gendered expectation formation, and informal educational settings.

The LSA of reflections from Arabic medical students (Table 6) explained 92.2% of the total variance, indicating a semantically rich and cohesive thematic structure.

Topic 1, which accounted for 35.95% of the total variance (eigenvalue=58.017), reaffirmed the central theme of Gendered Patient-Doctor Interactions. Topic 2 (13.31% variance) was likewise retained. Topic 3 represented a new theme, characterized by culturally specific terms that suggested heightened sensitivity to modesty, gendered exposure, and familial expectations in clinical contexts. Similarly, topic 5 introduced a novel discourse on paternal authority and structural discrimination, with terms pointing to familial and systemic influences on gender roles and health care access—elements particularly salient in Arabic sociocultural settings. Two other themes were reoriented from their original full-sample conceptualizations. Topic 4 captured procedural aspects of care and implicitly gendered experiences during physical examination and consultation. Topic 7 aligned with the previously defined theme of gendered communication but introduced a spatial and emotional safety dimension, with cultural reframing. Additional new themes were identified in topics 9 and 10. Topic 9 pointed to concerns around reproductive health and perceived vulnerability. Finally, topic 10 captured emotional burden and coping dynamics.

Table 5. Labeled topics identified through latent semantic analysis of student reflections following participation in a gender medicine course in the Jewish-only sample.

Topic	Label	Representative quotes	Eigenvalue	% Variability	Top words
Topic 1	Gendered Patient-Doctor Interactions	"I feel that as a professional expanding my knowledge in the field of medicine in general and in ECGs specifically, I underestimated the patient due to a lack of knowledge and the routine of daily life, the treating doctor did as well, as did her husband based on his understanding..." "I think and feel that as a professional and a future doctor we must listen to our patients..."	109.760	28.06	gender, patient, feel, women, doctor
Topic 2	Gender-Specific Diseases and Health Concerns	"I was surprised, because I thought that precisely because they have little breast tissue, the diagnosis would be earlier, but most of the time they do not check for changes in the breast at all..." "During the weight loss process, my brother had fatty tissue on his chest that bothered him and looked like female breasts..."	51.474	6.17	breast, cancer, partner, didn, hurt
Topic 3	Gender and Decision-Making in Medicine	"It is clear to me that a woman's individual rights and freedom in making decisions about her own body are fundamental principles that must be protected..." "It is important to recognize that the weight of the final decisions lies with the woman, but also to be sensitive to the needs and feelings of men in these situations..."	43.140	4.34	student, opinion, sister, decis, separ
Topic 4	Health Care Access for Transgender Individuals	"This was the first time I had interviewed a transgender person..." "I chose this topic because I believe it is a complex and important issue in the field of sex and gender-aware medicine. The complexity of treating transgender patients concerns not only medical aspects, but also behavioral and psychological ones, and I think it is very important to raise awareness of the issue, create a more supportive and respectful medical environment, and improve the quality of care and life of patients who have undergone this procedure..."	41.514	4.01	base, transgend, servic, unit, train
Topic 5	Gender and Public Health Framing of Conditions	"We are required to learn about the needs of [all] populations, regardless of religion, race, gender, or age, while providing comfort and a pleasant experience to every patient..." "Everyone was shocked, as there isn't enough awareness of these conditions and the numbers, especially in this population!..."	39.559	3.65	condit, popul, month, arriv, week
Topic 6	Gender Differences in Pharmacology	"In a lecture, I learned about a drug that has a stronger effect on women, to the point that the dose needed is half that needed by men..." "It is clear that women, even if you take into account their body weight, often receive drugs in doses that are not suitable for them..."	35.990	3.02	lectur, drug, dose, menstrual, communiti
Topic 7	Gender Equality and Team Dynamics	"I always taught my paramedic trainees and colleagues along the way to listen to the patient, not to diminish their value or their feelings, to rule out and take things seriously-it's life!..." "A specific event that I remember was a shift in which I was the team leader and worked with another male paramedic from the Arab sector..."	35.558	2.95	woman, team, incid, leader, paramed
Topic 8	Institutional and Interpersonal Perspectives on Gender	"Although my mother and friends told me a little about the subject [i.e., menstruation], they never went into depth, so I didn't exactly understand what it was about until it happened to me..." "Recently, at a regular gathering among friends, I was present at a conversation, on the subject of discrimination against men in unwanted pregnancies, a topic that has come up in the past..."	33.649	2.64	subject, studi, friend, realiti, biolog
Topic 9	Psychological and Emotional Dimensions of Gender and Health	"This may have been caused by a lack of understanding of the emotional and mental effects of the syndrome..." "I assume that this was due to the emotional difficulties that the syndrome brings with it..."	31.049	2.25	social, save, syndrom, attack, mention
Topic 10	Gendered Youth, Expectations, and Identity Formation	"It prompted a deeper examination of how societal expectations around gender impact not only health care but also various aspects of daily life, from career choices to personal relationships..." "I think we have to find the power in the decisions we want to make as independent, and not as something that society expects us to do..."	29.704	2.06	girl, meet, expect, don't, camp

Table 6. Labeled topics identified through latent semantic analysis of student reflections following participation in a gender medicine course in the Arabic-only sample.

Topic	Label	Representative quotes	Eigenvalue	% Variability	Top words
Topic 1	Gendered Patient-Doctor Interactions	"Her doctor referred her to a specialist at the hospital in order to begin the treatment steps as early as possible..." "I understood that there is a price for mistakes and lack of adequate response to gender issues as an important part of the treatment..."	58.017	35.95	gender, women, medic, treatment, import
Topic 2	Gender Equality and Team Dynamics	"This incident taught me a lot about myself: I learned how hard it is for me to skip things that go against my values and don't fit my worldview, I learned how much I care about the health of the people around me and how important it is for them to feel good..." "My values, based on human dignity and equality, pushed me to act in the girl's best interest..."	35.304	13.31	dr, equal, manag, sister, valu
Topic 3	Cultural Norms in Clinical Settings	"At the time, she felt embarrassed by his reaction in front of the rest of the family and was hurt, which led to an argument between the two of them..." "After a long conversation, I was able to understand from her that she was afraid and did not know how to break the news to her parents due to embarrassment..."	28.317	8.56	aunt, society, situation, surgery, embarrass
Topic 4	Gendered Clinical Encounters	"I suggested that she postpone this step until after her consultation with the specialist, and at the end of the conversation, she asked me to accompany her to the hospital on that day..." "In my eyes, this is where the hardest part of the treatment began, because one of the steps of the surgery involved breaking the hymen to reach the polyp located on the cervix..."	25.686	7.05	pain, exam, step, consult, result
Topic 5	Paternal Authority and Structural Discrimination	"She claimed that her father and brothers did not believe her and supported her father's wife; her mother, who lived in Rahat after a divorce, did not object to the father returning her daughter home..." "Personally, the event strengthened my understanding of the importance of supporting and assisting people in distress, especially women who experience violence and discrimination..."	24.217	6.26	felt, father, explain, driver, discrimin
Topic 6	Psychological and Emotional Dimensions of Gender and Health	"This experience reinforced my understanding that mental and physical health are interconnected, and that comprehensive care must address the whole person..." "During her hospitalization, we worked to understand her experiences, her special needs as a new mother, and the impact of her mental state on her functioning as a mother..."	22.917	5.61	dose, mental, impact, clinic, provid
Topic 7	Gendered Communication and Assertion in Health Care Contexts	"The new doctor's approach was completely different, she addressed not only the physical aspects of the pain but also its emotional and social effects..." "I tried to talk to him sensitively at that moment because it was a family matter..."	20.968	4.70	emot, address, sens, room, start
Topic 8	Intergenerational and Familial Perspectives on Gendered Health	"The case ended with a compromise, in which the girl would move in with her mother..." "Following what happened, I think back to my own distant childhood and remember my mother, who is now a retired teacher and her education for me and my sister, who were the youngest children and how she educated us for equality, that whatever is fair for my sister is also fair for me..."	19.533	4.08	mother, disease, topic, face, factor
Topic 9	Reproductive Health and Feelings of Vulnerability	"First, I realized how much a lack of listening on the part of medical professionals can lead to a feeling of helplessness and loneliness in the patient..." "I chose to present this topic because it is an extremely important topic in psychiatry and medicine in general [i.e., postpartum depression], as it requires a special understanding of the physiological and emotional changes that women go through after childbirth..."	18.314	3.58	birth, touch, question, helpless, correct

Topic	Label	Representative quotes	Eigenvalue	% Variability	Top words
Topic 10	Psychosocial Support and Gendered Distress	"I learned to appreciate the importance of social and psychological support and develop a more inclusive and understanding approach towards people in complex situations..." "This includes recognizing the emotional effects of illness on a patient's life and understanding the importance of emotional support alongside medical care..."	16.941	3.07	support, girl, complex, distress, act

Discussion

Principal Findings

The findings of this study highlight the complex and multidimensional ways in which first-year medical students in Israel engage with the concept of gender in both educational and clinical contexts. LSA of student reflections revealed a dominant recurring theme—Gendered Patient-Doctor Interactions—across all analyses. This convergence aligns with prior scholarship underscoring the critical role of gender in shaping clinical communication, empathy, and trust [23]. The centrality of this theme suggests that even early exposure to gender medicine can prompt students to critically reflect on the influence of gender on therapeutic relationships and clinical outcomes [24,25]. In line with the World Health Organization's assertion that gender is a fundamental social determinant of health [5], students in this study demonstrated an emerging awareness of how gender shapes illness narratives, health care access, and diagnostic pathways. Their reflections also echoed the "gender mainstreaming" approach promoted in international medical education policy, which advocates for integrating gender considerations at all levels of health care delivery and education [5,26,27]. Notably, themes related to transgender health care, gender-specific diseases, and pharmacological differences paralleled the growing literature that underscores the need to dismantle androcentric biases in medical curricula [28,29]. These insights suggest that when given structured opportunities for reflection, learners are capable of recognizing and critically interrogating structural inequities embedded within biomedical discourse and practice.

Rather than merely acknowledging the existence of inequities, participants were challenged to examine critically the structural foundations of medical knowledge production, clinical practice, and professional formation. Many expressed dismay at the historical exclusion of female bodies from clinical research and the persistent privileging of male physiology as the normative reference point. As one student succinctly articulated: "We were presented with the historical neglect of female subjects in clinical trials in various fields, and an approach that perpetuated biased medical treatments that favor male physiology as the default standard." This statement reflected not only an awareness of bias but also a structural critique of the biomedical canon, echoing feminist analyses of knowledge systems that render women invisible or anomalous within clinical reasoning. Importantly, the student's reference to "perpetuated biased medical treatments" underscored an understanding that these are not merely vestiges of the past but ongoing epistemic and clinical injustices. The course further encouraged students to reevaluate their professional identity and the ethics of care through a gender-responsive lens. Many participants described how the training reshaped their conception of good medical practice—shifting from a mechanistic, one-size-fits-all model to a more nuanced, context-sensitive, and individualized approach. One student reflected: "I learned about the importance of listening to patients and the need to treat them as individuals, and not just as representatives of a particular gender." This shift marks a clear departure from biomedical reductionism, highlighting the importance of relational ethics and embodied listening—practices central to feminist clinical pedagogy. Importantly, the course also created space for critical emotional and identity work, particularly among male students. While some initially reacted with defensiveness or discomfort, perceiving gender-sensitive discourse as accusatory or polarizing, dialogical engagement and pedagogical scaffolding helped transform these affective responses into opportunities for deeper, transformative learning. As one male student candidly recounted: "At the beginning of the 'Sex and Gender-Aware Medicine' course, I felt, as a man, uncomfortable and even indirectly attacked...However, in small group exercises, the learning experience changed. The instructors helped me understand that the goal is not to blame, but to train us to be better doctors." This reflection illustrates how pedagogical designs grounded in openness and dialogic learning can transmute resistance into productive introspection. The student's movement from a sense of personal indictment toward professional growth signals a successful pedagogical intervention—one that shifts affective orientations while fostering ethical self-positioning. Crucially, the course did not stop at generating cognitive insights; it also catalyzed concrete commitments to behavioral change. Several students described actionable strategies for integrating sex- and gender-aware practices into their future patient encounters. One such example reads: "Following the course, I intend to ask more focused questions about the medical history of my patients, taking gender differences into account. I also intend to continue to stay up-to-date with research and literature in the field of gender medicine, in order to provide my patients with the best possible care." Here, we witness the internalization of gender mainstreaming not as an abstract principle but as a guiding framework for everyday clinical decision-making and professional formation. Some students even extended their reflections beyond the medical *curriculum*, engaging with the broader sociocultural implications of gendered behavior and ethics. One student, for example, recounted a personal experience in which a pregnant woman was ignored on public transport, using this incident as a lens to interrogate gendered expectations, moral responsibility, and social awareness. While not strictly clinical, such reflections highlight the pervasiveness of gendered norms and underscore the importance of cultivating ethical responsiveness that transcends disciplinary boundaries.

The gender-stratified analysis revealed meaningful differences in both the conceptual emphasis and emotional tone of student reflections. Female students more frequently highlighted relational and affective dimensions of gender, including intergenerational caregiving roles, informal consultation, and communicative dynamics in health care. These findings align with prior research indicating that female medical students often adopt a relational epistemology and are more inclined to reflect on their positionality and ethical responsibilities in clinical practice [30]. By contrast, male students tended to engage more with the biomedical, legal, and structural dimensions of gender, raising issues such as diagnostic bias and reproductive rights. This pattern reflects what prior literature has described as a more abstract and normative framing of social issues among male learners [30].

Moreover, this study reveals that students construct divergent understandings of gender medicine through identity lenses deeply shaped by religion, ethnicity, ideology, and sociopolitical context. One particularly compelling reflection came from a student who described the course as exposing "the tumultuous reality in our small country, which is caught in the middle of a clash between East and West, between religion and liberalism." The student reflected on the tension between their conservative upbringing and the realities of contemporary clinical practice, asking: "Is there room for the religious beliefs upon which my upbringing is based, or do I need to re-examine or abandon my personal beliefs and open up to secularism and liberalism?" These questions encapsulate the internal negotiation learners face when confronted with dissonant ideas that challenge their foundational schemas. Similarly, ethnic subgroup analyses highlighted the intersection of gender and cultural identity. Reflections from Jewish students revealed a tendency to extrapolate personal experiences to broader public health framings, often invoking themes such as epidemiological patterns and gendered health behaviors. This ability to connect micro-level experiences with macro-level perspectives aligns with what Cruess et al [31] describe as the development of the physician's "professional identity." Meanwhile, reflections from Arabic students revealed a deeper engagement with culturally embedded norms surrounding modesty, familial authority, and structural barriers to health care access—patterns consistent with scholarship on the intersection of gender, religion, and medicine in collectivist cultures [32]. Notably, these students also voiced affectively charged concerns related to reproductive health, vulnerability, and emotional burden, underscoring the cultural and emotional labor involved in reconciling conflicting expectations within clinical spaces [33,34]. Taken together, these reflections—shaped by diverse positionalities—illustrate how pedagogical encounters with gender can elicit complex and, at times, conflicting epistemic trajectories.

Overall, this study adds to the growing body of scholarship calling for the integration of gender and cultural competence into medical education. The themes identified suggest that structured gender curricula, when paired with reflective pedagogy and systematic evaluation, can cultivate critical consciousness and social responsiveness in future physicians [35].

Reflection—especially when systematically analyzed through computational methods—can serve as both a pedagogical and evaluative tool, enabling the assessment of not only cognitive but also affective and ethical dimensions of learning [36].

Implications of Findings

The findings of this study carry important implications for the design and implementation of gender medicine curricula, particularly in culturally and identity-diverse contexts. The subgroup differences observed—for instance, female students' greater emphasis on relational and affective dimensions, or Arabic students' focus on culturally embedded norms and structural constraints—demonstrate that learners do not uniformly assimilate gender medicine content. Rather than passively absorbing knowledge, students actively interpret and reframe new information through the lens of prior experiences, sociocultural background, and identity. This process reflects constructivist learning theory, which emphasizes that learning is active, situated, and coconstructed. Gender and ethnic identity often function as interpretive filters, amplifying certain issues while diminishing others. Such identity-based perspectives can enrich the learning environment by contributing diverse viewpoints, yet they may also generate tensions when personal or cultural values conflict with curricular content. From a pedagogical perspective, these findings underscore the importance of embedding structured opportunities for students to articulate and critically examine their own perspectives. Dialogical learning spaces, culturally responsive facilitation, and peer-to-peer engagement can help learners navigate potentially dissonant concepts and integrate them into their emerging professional identity. Such strategies may bridge the gap between abstract knowledge of gender medicine and its practical application across diverse clinical contexts. At the institutional level, embedding gender medicine within an explicitly intersectional framework ensures that the interwoven effects of gender, ethnicity, religion, and other social determinants are systematically addressed. This approach can strengthen both cultural competence and the ability to provide equitable, patient-centered care to heterogeneous populations.

Strengths and Limitations

Nonetheless, while the results are promising, caution is warranted. The interpretability of LSA findings is contingent on robust preprocessing, linguistic clarity within the student corpus, and careful topic labeling. Although this study benefited from certain strengths, including stratification by gender and ethnicity, the limited sample size precluded a comprehensive intersectional analysis. Furthermore, the reflections were collected at a single time point and situated within a specific institutional and sociocultural context, which may constrain the transferability of findings. Longitudinal studies are needed to assess whether these conceptual gains translate into sustained behavioral change and improved clinical competence in gender-sensitive care. In addition, comparative research across medical schools and national contexts could illuminate how curricular design, institutional culture, and sociopolitical environments shape the uptake and integration of gender medicine.

Conclusions

This study underscores the importance of embedding gender-sensitive content within medical education and demonstrates the utility of LSA as an analytic tool for examining how learners engage with complex sociomedical themes. The findings highlight the need for an intersectional, contextually attuned approach to curriculum design that actively considers students' diverse identities and lived experiences. By fostering reflexivity, empathy, and critical literacy, gender medicine education holds the potential to shape future physicians who are better equipped to provide equitable, inclusive, and socially responsive health care.

Conflicts of Interest

None declared.

References

1. Kaufman MR, Eschliman EL, Karver TS. Differentiating sex and gender in health research to achieve gender equity. Bull World Health Organ 2023 Oct 01;101(10):666-671 [FREE Full text] [doi: 10.2471/BLT.22.289310] [Medline: 37772198]
2. Johnson JL, Greaves L, Repta R. Better science with sex and gender: facilitating the use of a sex and gender-based analysis in health research. Int J Equity Health 2009 May 06;8:14 [FREE Full text] [doi: 10.1186/1475-9276-8-14] [Medline: 19419579]
3. Gualtierotti R. Bridging the gap: time to integrate sex and gender differences into research and clinical practice for improved health outcomes. Eur J Intern Med 2025 Apr;134:9-16 [FREE Full text] [doi: 10.1016/j.ejim.2025.01.030] [Medline: 39915168]
4. Hay K, McDougal L, Percival V, Henry S, Klugman J, Wurie H, Gender Equality, Norms, Health Steering Committee. Disrupting gender norms in health systems: making the case for change. Lancet 2019 Jun 22;393(10190):2535-2549 [FREE Full text] [doi: 10.1016/S0140-6736(19)30648-8] [Medline: 31155270]
5. World Health Organization. Gender mainstreaming for health managers: A practical approach. WHO Press. 2011. URL: https://apps.who.int/iris/handle/10665/44516 [accessed 2025-08-25]
6. Risberg G, Johansson EE, Hamberg K. A theoretical model for analysing gender bias in medicine. Int J Equity Health 2009 Aug 03;8:28 [FREE Full text] [doi: 10.1186/1475-9276-8-28] [Medline: 19646289]
7. Verdonk P, Benschop YWM, de Haes HCJM, Lagro-Janssen TLM. From gender bias to gender awareness in medical education. Adv Health Sci Educ Theory Pract 2009 Mar 15;14(1):135-152. [doi: 10.1007/s10459-008-9100-z] [Medline: 18274877]
8. Regitz-Zagrosek V. EMBO Rep 2012 Jun 29;13(7):596-603 [FREE Full text] [doi: 10.1038/embor.2012.87] [Medline: 22699937]
9. Sutkin G, Wagner E, Harris I, Schiffer R. What makes a good clinical teacher in medicine? A review of the literature. Acad Med 2008 May;83(5):452-466. [doi: 10.1097/ACM.0b013e31816bee61] [Medline: 18448899]
10. Wear D, Zarconi J. Can compassion be taught? Let's ask our students. J Gen Intern Med 2008 Jul;23(7):948-953 [FREE Full text] [doi: 10.1007/s11606-007-0501-0] [Medline: 18612722]
11. Khamisy-Farah R, Bragazzi NL. How to integrate sex and gender medicine into medical and allied health profession undergraduate, graduate, and post-graduate education: insights from a rapid systematic literature review and a thematic meta-synthesis. J Pers Med 2022 Apr 11;12(4):12040612 [FREE Full text] [doi: 10.3390/jpm12040612] [Medline: 35455728]
12. Brown MEL, Hunt GEG, Hughes F, Finn GM. 'Too male, too pale, too stale': a qualitative exploration of student experiences of gender bias within medical education. BMJ Open 2020 Aug 13;10(8):e039092 [FREE Full text] [doi: 10.1136/bmjopen-2020-039092] [Medline: 32792453]
13. Samuriwo R, Patel Y, Webb K, Bullock A. 'Man up': Medical students' perceptions of gender and learning in clinical practice: a qualitative study. Med Educ 2020 Feb;54(2):150-161. [doi: 10.1111/medu.13959] [Medline: 31746029]
14. Geiser E, Schilter LV, Carrier J, Clair C, Schwarz J. Reflexivity as a tool for medical students to identify and address gender bias in clinical practice: a qualitative study. Patient Educ Couns 2022 Dec;105(12):3521-3528 [FREE Full text] [doi: 10.1016/j.pec.2022.08.017] [Medline: 36075808]
15. Bada SO. Constructivism learning theory: A paradigm for teaching and learning. Journal of Research & Method in Education 2015;5(6):66-70.
16. Bleakley A. Medical humanities and medical education: How the medical humanities can shape better doctors. Routledge 2015:1-276. [doi: 10.4324/9781315771724]
17. Acosta-Martínez M, Chandran L, Cohen S, Biegon A. Design, implementation, and evaluation of an intensive course on issues in women's health and gender-based medicine. J Med Educ Curric Dev 2023;10:23821205231203783 [FREE Full text] [doi: 10.1177/23821205231203783] [Medline: 37744420]
18. Beagan B. Micro inequities and everyday inequalities: "race," gender, sexuality and class in medical school. Canadian Journal of Sociology / Cahiers canadiens de sociologie 2001;26(4):583-610. [doi: 10.2307/3341493]
19. Anderson LW, Krathwohl DR, editors. A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. New York: Longman; 2001.

20. Landauer T. LSA as a theory of meaning. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), Handbook of latent semantic analysis. Lawrence Erlbaum Associates Publishers 2007:3-34. [doi: 10.4324/9780203936399.ch23]
21. Landauer T, McNamara D, Dennis S, Kintsch W, editors. Handbook of Latent Semantic Analysis. 1st ed. Mahwah, NJ: Lawrence Erlbaum Associates; 2007.
22. Dumais ST. Latent semantic analysis. Annual Review Info Sci & Tec 2005 Sep 22;38(1):188-230. [doi: 10.1002/aris.1440380105]
23. Roter DL, Hall JA. Physician gender and patient-centered communication: a critical review of empirical research. Annu Rev Public Health 2004 Apr;25(1):497-519. [doi: 10.1146/annurev.publhealth.25.101802.123134] [Medline: 15015932]
24. Hamberg K. Gender bias in medicine. Womens Health (Lond) 2008 May;4(3):237-243. [doi: 10.2217/17455057.4.3.237] [Medline: 19072473]
25. Jenkins MR, Herrmann A, Tashjian A, Ramineni T, Ramakrishnan R, Raef D, et al. Sex and gender in medical education: a national student survey. Biol Sex Differ 2016;7(Suppl 1):45 [FREE Full text] [doi: 10.1186/s13293-016-0094-6] [Medline: 27785347]
26. Bustreo F, Ponchia AG, Rocco C, Hinton R. Strengthening the transformative potential of gender mainstreaming in global health. EClinicalMedicine 2021 Apr;34:100858 [FREE Full text] [doi: 10.1016/j.eclinm.2021.100858] [Medline: 33937728]
27. Verdonk P, Benschop YWM, De Haes JCJM, Lagro-Janssen ALM. Making a gender difference: case studies of gender mainstreaming in medical education. Med Teach 2008;30(7):e194-e201. [doi: 10.1080/01421590802213206] [Medline: 18777419]
28. Tannenbaum C, Greaves L, Graham ID. Why sex and gender matter in implementation research. BMC Med Res Methodol 2016 Oct 27;16(1):145 [FREE Full text] [doi: 10.1186/s12874-016-0247-7] [Medline: 27788671]
29. Kumagai AK, Lypson ML. Beyond cultural competence: critical consciousness, social justice, and multicultural education. Acad Med 2009 Jun;84(6):782-787. [doi: 10.1097/ACM.0b013e3181a42398] [Medline: 19474560]
30. Hsu H, Sung T. Exploring gender differences in empathy development among medical students: a qualitative analysis of reflections on juvenile correctional school visits. Med Educ Online 2025 Dec;30(1):2500556 [FREE Full text] [doi: 10.1080/10872981.2025.2500556] [Medline: 40320661]
31. Cruess RL, Cruess SR, Boudreau JD, Snell L, Steinert Y. A schematic representation of the professional identity formation and socialization of medical students and residents: a guide for medical educators. Acad Med 2015 Jun;90(6):718-725. [doi: 10.1097/ACM.0000000000000700] [Medline: 25785682]
32. Sarsour NY, Hammoud MM. Integration of Arab and Muslim health education into a medical school curriculum. MedEdPORTAL 2021 Nov;17:11188. [doi: 10.15766/mep_2374-8265.11188]
33. Tayeb HO, Tekian A, Baig M, Koenig HG, Lingard L. The role of religious culture in medical professionalism in a muslim Arab society. Perspect Med Educ 2023;12(1):56-67 [FREE Full text] [doi: 10.5334/pme.920] [Medline: 36908746]
34. Inhorn MC, Fakih MH. Arab Americans, African Americans, and infertility: barriers to reproduction and medical care. Fertil Steril 2006 Apr;85(4):844-852 [FREE Full text] [doi: 10.1016/j.fertnstert.2005.10.029] [Medline: 16580363]
35. McBee E, Ratcliffe T, Schuwirth L, O'Neill D, Meyer H, Madden SJ, et al. Context and clinical reasoning: understanding the medical student perspective. Perspect Med Educ 2018 Aug;7(4):256-263 [FREE Full text] [doi: 10.1007/s40037-018-0417-x] [Medline: 29704167]
36. Hanlon CD, Frosch EM, Shochet RB, Buckingham Shum SJ, Gibson A, Goldberg HR. Recognizing reflection: computer-assisted analysis of first year medical students' reflective writing. Med Sci Educ 2021 Feb;31(1):109-116 [FREE Full text] [doi: 10.1007/s40670-020-01132-7] [Medline: 34457870]

Abbreviations

DIEP: Describe, Interpret, Evaluate, and Plan
LSA: latent semantic analysis

Edited by B Lesselroth; submitted 01.06.25; peer-reviewed by T Bader, D Shaham; comments to author 20.06.25; revised version received 11.08.25; accepted 15.08.25; published 29.08.25.

Please cite as:
Khamisy-Farah R, Farah R, Jabaly-Habib H, Nakhleh Francis Y, Bragazzi NL
Exploring Gender Perspectives in Medical Education: Latent Semantic Analysis of Israeli First-Year Medical Students' Reflections
JMIR Med Educ 2025;11:e78371
URL: https://mededu.jmir.org/2025/1/e78371
doi:10.2196/78371
PMID:40817849

Original Paper

Enhancing Clinical Competencies Through Peer Role-Play in Oncology Graduate Students: Mixed Methods Study

Yao Wang[1], PhD; Feixiang Wang[1], MSc; Gaojie Liu[1], PhD; Yuqing luo[1], MSc; Hongjun Ba[2], PhD; Jie Long[1], MSED

[1]Guangzhou Medical University Cancer Hospital, Guangzhou, Guangdong Province, China

[2]Department of Pediatrics, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, Guangdong Province, China

Corresponding Author:
Jie Long, MSED
Guangzhou Medical University Cancer Hospital
78 Hengzhigang Road, Yuexiu District Guangzhou City
Guangzhou, Guangdong Province, 510095
China
Phone: 86 15013380071
Email: longjie07@126.com

Abstract

Background: Clinical competency is essential for oncology students to deliver high-quality patient care. However, traditional teaching methods may not fully support the development of critical skills such as communication, empathy, and clinical judgment. Peer role-play has emerged as a promising approach to bridge these gaps by enhancing interpersonal and diagnostic competencies within clinical settings.

Objective: This study aims to evaluate the effectiveness of peer role-play in developing clinical competencies among oncology graduate students during their clinical rotation.

Methods: This study involves 70 first-year oncology graduate students from Affiliated Cancer Guangzhou Medical University Cancer Hospital in a 3-month clinical rotation within the department of oncology from January 2022 to December 2023. Participants were randomly assigned to either a peer role-play group (n=35) or a traditional teaching group (n=35), ensuring balanced gender and baseline competencies. The role-play group engaged in a structured curriculum that included case presentation, classroom instruction, and weekly role-play sessions, with debriefing and feedback sessions following each role-play. The traditional teaching group adhered to a standard curriculum without role-play exercises. Assessments included a baseline oncology theory exam, Mini-Clinical Evaluation Exercise for clinical competency evaluation, and a satisfaction survey for the role-play group.

Results: Baseline theory exam scores were comparable between the 2 groups (P=.08). However, the peer role-play group demonstrated significant improvements in doctor-patient communication, medical history taking, clinical judgment, and overall clinical competence compared to the traditional teaching group (P<.05). Furthermore, students in the role-play group reported high levels of satisfaction, citing scenario realism, communication practice opportunities, and feedback quality as key benefits.

Conclusions: This study indicates that peer role-play is an effective educational approach for developing clinical competencies in oncology graduate students, particularly in communication, empathy, and clinical reasoning. Role-play provides an engaging and practical learning experience, making it a valuable addition to clinical training programs aimed at enhancing patient-centered care skills in students.

(JMIR Med Educ 2025;11:e79771) doi:10.2196/79771

KEYWORDS

peer role-play; Mini-Clinical Evaluation Exercise; clinical competencies; oncology; physician-patient communication

Introduction

The Need for Effective Communication in Oncology Education

The rapid advancements in medical knowledge and technology have reshaped oncology practice, increasing the demand for oncology graduate students to develop clinical competencies and patient-centered communication skills [1,2]. In oncology, where patients face emotionally and physically challenging conditions, the ability to communicate with empathy, clarity, and sensitivity is essential. Discussions involving complex diagnoses, treatment options, and prognoses have a profound

impact on patients and their families, requiring clinicians to balance clinical accuracy with emotional awareness. However, mastering these communication skills can be challenging for students newly entering clinical settings. Many lack the experience to navigate oncology's intense emotional landscapes, making it difficult to build rapport, demonstrate compassion, and manage challenging conversations [3].

Traditional medical education, largely based on lectures and observational learning, often falls short in fostering the interpersonal skills necessary for these interactions. There is a critical need for training approaches that provide early, practical opportunities for realistic patient engagement. Role-play, an active learning methodology, offers a promising solution by enabling students to assume both practitioner and patient roles within controlled simulations [4]. This technique allows students to practice essential clinical skills in a low-risk setting and helps them develop effective, empathetic communication with patients. By simulating real-life interactions, role-play encourages students to develop the confidence and sensitivity required to manage emotionally complex oncology conversations before they enter clinical practice [5].

Despite the recognized importance of communication in oncology, few studies have examined the specific impact of role-play on preparing students for these complex interactions. Most research on role-play in medical education focuses on general skill development or communication in less specialized fields [6], leaving a gap in understanding its potential in comprehensive oncology training. This study sought to bridge this gap by investigating the effectiveness of role-play in enhancing oncology graduate students' clinical competencies, especially in patient communication, diagnostic reasoning, and emotional intelligence.

Role-Play as a Tool for Enhancing Patient-Centered Communication

Role-play can be situated within an experiential learning and deliberate practice framework, allowing learners to iteratively rehearse complex communicative and diagnostic tasks with targeted feedback in a psychologically safe environment. In medical education, role-play has been used both informally and within structured curricula to improve communication, empathy, and professionalism, with guidance on task design, briefing, and debriefing to maximize learning [7]. Peer role-play specifically offers a low-cost, scalable alternative to standardized patients (SPs) and high-fidelity simulation while retaining core pedagogical features such as scenario authenticity, role clarity, and structured feedback [8]. Our study extends this literature by testing a systematic, oncology-specific peer role-play curriculum during a real clinical rotation.

In oncology, communication frequently involves serious illness conversations, including breaking bad news and discussing goals of care—domains supported by established frameworks such as SPIKES (setting, perception, invitation, knowledge, empathy, and strategy) and the Calgary-Cambridge guide [9,10]. Embedding these frameworks into role-play scenarios can promote patient-centered communication, empathy, and ethical clinical judgment before students encounter such conversations in the real world. While high-fidelity and SP-based simulations are valuable and increasingly used, particularly for palliative and end-of-life communication, resource-sensitive peer role-play can serve as an effective stepping stone that is feasible for routine integration in busy oncology rotations.

Through this research, we aimed to illuminate the potential of role-play to strengthen the practical and interpersonal skills necessary in oncology. By assessing student outcomes in clinical skills and communication confidence, this study sought to contribute valuable insights to the development of interactive training models that better equip future oncologists to meet the complex demands of their profession.

Methods

Study Design and Participants

This study involved 70 graduate students majoring in oncology from Guangzhou Medical University Cancer Hospital who were in their first clinical rotation in the department of oncology from January 2022 to December 2023. The participants' ages ranged from 22 to 24 years. Of these 70 graduate students, 36 (51%) were males and 34 (49%) were females. Using a simple randomization method, students were assigned to either a peer role-play group (n=35, 50%) or a traditional teaching group (n=35, 50%), ensuring an even distribution of gender, age, and baseline competencies.

Participants were randomized in a 1:1 ratio to the peer role-play or traditional teaching group using a computer-generated sequence with permuted blocks (sizes 4 and 6) stratified by gender and baseline oncology theory exam tertiles to promote balance on key variables. The randomization list was prepared by a statistician independent of the teaching and assessment teams. Allocation was concealed using sequentially numbered, opaque, sealed envelopes that were opened only after enrollment. This process yielded comparable distributions of gender, age, and baseline knowledge across groups.

Sample size was calculated based on a power analysis, which determined that a minimum of 32 students per group would be required to detect a moderate effect size (Cohen d=0.5) with 80% power and a 5% significance level. Both groups participated in a 3-month clinical rotation.

Study Setting and Curricular Context

This study was conducted at Guangzhou Medical University Cancer Hospital, a tertiary cancer center in southern China. *Oncology graduate students* refer to first-year postgraduate trainees in a 3-year Master of Medicine program who undertake supervised clinical rotations alongside coursework. Both groups participated in the standard 12-week medical oncology rotation comprising didactic lectures, case presentations and tumor boards, ward-based care, and outpatient clinics under faculty supervision. To accommodate the intervention while preserving clinical exposure, 1 weekly case discussion hour was replaced with a 60-minute peer role-play session plus structured debriefing in the role-play group; all other learning activities (lectures, case presentations, and patient care) were identical between the groups.

Before this study, structured simulation was not embedded within the oncology rotation. SPs are available at the university's skill center primarily for preclinical communication courses; however, SP-based sessions had not been routinely implemented in the oncology rotation due to timetable and cost constraints. To introduce an accessible form of simulation, we integrated peer role-play into the existing schedule by replacing 1 weekly case discussion hour with a structured role-play session followed by debriefing and feedback. This approach preserved clinical exposure while creating protected time for deliberate practice of communication, clinical reasoning, and professionalism.

Participant Characteristics and Learning Context

All participants were first-year postgraduate students in oncology who had completed a 5-year undergraduate medical degree and were beginning their initial clinical oncology rotation; none had completed specialty residency training. Instruction and clinical care were delivered primarily in Mandarin Chinese, with Cantonese used as appropriate in-patient interactions. On the basis of institutional records, students had not received formal, oncology-specific role-play or simulation training before this rotation, although all had previous exposure to lecture-based teaching and bedside observation.

Consistent with the common features of local medical education, students were accustomed to lecture-centric learning, summative written assessments, and faculty-led bedside teaching. Formative feedback opportunities existed but were less structured. Introducing weekly peer role-play with facilitated debriefing was intended to provide regular, structured opportunities for practice, feedback, and reflection aligned with experiential learning principles.

Intervention

Peer Role-Play Group Implementation

The peer role-play group participated in a structured role-play curriculum designed to enhance clinical skills and communication in oncology. The implementation of the role-play sessions involved several key steps.

Case Presentation

At the beginning of each week, the instructor conducted a case presentation session for the role-play group. During this session, the instructor introduced a real patient case, detailing the patient's medical history, physical examination findings, diagnostic results, and the clinical reasoning behind the diagnosis. This foundation enabled students to familiarize themselves with real-life oncology cases and understand the complexities involved in patient interactions.

Classroom Teaching and Role-Play Instruction

Following the case presentation session, the students attended a classroom session in which the instructor introduced the principles and significance of role-play in clinical education. The instructor explained the objectives of role-play in enhancing empathy, communication skills, and diagnostic reasoning. The instructor then demonstrated how role-play scenarios would be structured, detailing each step of the role-playing process.

Role-Play Scenario Execution

Each week, students engaged in a 60-minute role-play session. Students were divided into small groups, with each group member assigned a specific role within the scenario: health care provider, patient, or observer. The health care provider role required students to undertake patient interactions, including history taking, explanation of diagnosis, treatment planning, and handling patient emotions. Students in the patient role were given background information to simulate the patient's condition and emotional state, whereas observers provided feedback and noted areas for improvement.

Debriefing and Feedback

After each role-play session, students and instructors gathered for a structured debriefing. Instructors facilitated reflective discussions, allowing students to share their experiences and insights. Feedback was provided on communication style, empathy, clarity, and clinical decision-making. The debriefing emphasized areas in which students excelled and identified specific skills for improvement, thereby promoting a safe environment for learning and growth.

Traditional Teaching Group

The traditional teaching group followed the standard clinical rotation curriculum, which included lectures, case presentations, and observational learning. Students in this group participated in routine patient care activities supervised by attending physicians without structured role-play exercises.

Assessment and Data Collection

To evaluate the effectiveness of the intervention, assessments were conducted for both groups before and after the rotation using 3 primary measures.

Oncology Theory Exam

A standardized written test covering key oncology knowledge areas, including diagnosis, treatment protocols, and patient management principles, was used to evaluate each student's theoretical knowledge. The exam was administered only at the beginning of the clinical rotation (before the intervention) to assess students' baseline knowledge in oncology. This examination was not repeated at the end of the rotation.

Mini-Clinical Evaluation Exercise

The Mini-Clinical Evaluation Exercise (Mini-CEX), a well-recognized tool for evaluating clinical skills [11,12], was used to assess practical competencies. The Mini-CEX was conducted only at the end of the clinical rotation, after the intervention, to evaluate the practical competencies and clinical skills of the students. Students obtained medical histories from patients and conducted physical examinations. The Mini-CEX assessed students across 7 criteria using a 9-point scale:

1. History taking: accuracy in gathering patient history, responding to nonverbal cues, and demonstrating empathy
2. Physical examination: competence in conducting examinations in an organized manner, maintaining patient privacy, and managing discomfort
3. Professionalism: respect, compassion, ethical standards, and confidentiality

4. Clinical judgment: students' ability to select and execute appropriate diagnostic tests and weigh the risks and benefits of various treatment options
5. Physician-patient communication: clarity in explaining medical tests, obtaining consent, and educating patients
6. Organizational efficiency: skill in prioritizing patient care and effectively using resources
7. Overall competence: integration of clinical knowledge and overall patient care effectiveness

The Mini-CEX scoring encompassed "below expectations" (1-3 points), "meeting expectations" (4-6 points), and exceeding expectations (7-9 points). All assessments were conducted by a single evaluator who was blinded to the study condition of the students and was unaware of whether the assessment was before or after the intervention. This evaluator was not involved in other aspects of the study to avoid any potential bias and ensure consistency in the assessments.

Satisfaction Survey for the Role-Play Group

At the end of the rotation, students in the peer role-play group completed a satisfaction questionnaire. This survey measured their perceived value of the role-play activities, including aspects of engagement, realism, feedback quality, communication skill improvement, empathy development, and overall satisfaction with role-play as a learning method. The satisfaction survey comprised the following six items, each rated on a 5-point Likert scale (1="strongly disagree"; 5="strongly agree"): (1) "The role-play activities made me more involved in the learning process" (engagement), (2) "The scenarios in role-play felt realistic, helping me experience a clinical environment" (realism), (3) "I am satisfied with the quality of feedback received during the role-play activities" (feedback quality), (4) "I feel more confident in communicating with patients due to the role-play sessions" (communication skill improvement), (5) "The role-play exercises helped me better understand patients' emotions and needs" (empathy development), and (6) "I am generally satisfied with the overall effectiveness of the role-play teaching method" (overall satisfaction).

To improve the validity of the satisfaction survey, it was pilot-tested with a similar cohort to refine question clarity and response consistency.

Data Analysis

Data were analyzed to compare outcomes between the peer role-play and traditional teaching groups. Paired 2-tailed *t* tests assessed within-group improvements, and independent *t* tests compared between-group performance on the oncology theory exam and Mini-CEX. The satisfaction survey results were analyzed using descriptive statistics to summarize the feedback from the role-play group. Statistical significance was defined as $P<.05$.

Ethical Considerations

This study was conducted in accordance with the principles of the Declaration of Helsinki. Ethics approval was obtained from the ethics committee of the Affiliated Cancer Hospital of Guangzhou Medical University (number 11/01/22). 70 first-year oncology graduate students were asked to participate in the study. All of them had signed informed consent prior to participation. All data were pseudonymized before being subjected to statistical analysis. Participants did not receive any financial compensation. As an incentive, the 5 best-performing students in each group were awarded a book prize.

Results

Baseline Theory Exam Scores

At baseline, the average theory exam scores were comparable between the 2 groups. The peer role-play group achieved a mean score of 91.12 (SD 2.15), whereas the traditional teaching group scored an average of 91.33 (SD 2.16). Statistical analysis indicated no statistically significant difference between the groups' baseline theoretical knowledge ($P>.05$), confirming a similar level of theoretical understanding at the start of the rotation.

Mini-CEX Assessment

All trainees completed the Mini-CEX evaluation within an average of 36 (SD 0.5) minutes, and post-evaluation feedback required approximately 6.6 (SD 0.4) minutes per student. After the rotation, the peer role-play group performed significantly better than the traditional teaching group on the Mini-CEX domains of physician-patient communication, history taking, clinical judgment, and overall clinical competence ($P<.05$ in all cases). Detailed Mini-CEX scores for both groups are provided in Table 1.

Table 1. The distribution of the scale scores on the Mini-Clinical Evaluation Exercise assessment in the 2 groups.

Item	Peer role-play group (n=35), n (%)	Traditional teaching group (n=35), n (%)	*P* value
Medical history taking			.03
Meets expectations	13 (37)	23 (66)	
Exceeds expectations	22 (63)	12 (34)	
Clinical judgment			.02
Below expectations	0 (0)	2 (6)	
Meets expectations	15 (43)	24 (69)	
Exceeds expectations	20 (57)	9 (26)	
Physician-patient communication			.004
Meets expectations	12 (34)	25 (71)	
Exceeds expectations	23 (66)	10 (29)	
Professionalism			.03
Meets expectations	14 (40)	24 (69)	
Exceeds expectations	21 (60)	11 (31)	
Physical examination			.81
Meets expectations	18 (51)	16 (46)	
Exceeds expectations	17 (49)	19 (54)	
Organizational effectiveness			.34
Meets expectations	16 (46)	21 (60)	
Exceeds expectations	19 (54)	14 (40)	
Overall capabilities			.03
Meets expectations	12 (34)	22 (63)	
Exceeds expectations	23 (66)	13 (37)	

Satisfaction Survey Results for the Role-Play Group

At the end of the study, students in the peer role-play group completed a satisfaction survey to assess their perceptions of the role-play experience (Table 2). Results indicated high satisfaction levels, with most students agreeing that role-play was a valuable and engaging learning tool. Key areas of positive feedback included engagement (32/35, 91% agreement and "strongly agree"), the realism of the scenarios (29/35, 83% agreement and "strongly agree"), the quality of the feedback received (32/35, 91% agreement and "strongly agree"), and overall satisfaction (33/35, 94% agreement and "strongly agree").

Table 2. Students' satisfaction evaluation in the role-play group (n=35)[a].

	Strongly disagree, n (%)	Disagree, n (%)	Neutral, n (%)	Agree, n (%)	Strongly agree, n (%)
Engagement	1 (3)	0 (0)	2 (6)	25 (71)	7 (20)
Realism	2 (6)	2 (6)	2 (6)	24 (69)	5 (14)
Communication skill improvement	1 (3)	0 (0)	1 (3)	24 (69)	8 (23)
Feedback quality	0 (0)	1 (3)	1 (3)	28 (80)	4 (11)
Empathy development	0(0)	1(3)	2(6)	29(83)	3(8)
Overall satisfaction	1 (3)	1 (3)	0 (0)	26 (74)	7 (20)

[a]1=strongly disagree, 2=disagree, 3=neutral, 4=agree, and 5=strongly agree.

Discussion

This study examined the effectiveness of peer role-play versus traditional teaching methods in enhancing the clinical competencies among oncology graduate students. The findings offer valuable insights into the role of interactive learning methods, particularly in fostering communication skills and clinical decision-making.

The results indicate that, while both groups had similar theoretical knowledge at baseline, the peer role-play group

exhibited significantly higher scores in practical skills, including physician-patient communication, medical history taking, and overall clinical competence as assessed using the Mini-CEX (P<.05). This aligns with previous studies suggesting that role-play is an effective tool for enhancing soft skills in medical education, especially communication and empathy [13,14]. These improvements are crucial for oncology care, where patient interactions often involve delivering complex news and managing emotional responses.

Moreover, the peer role-play group demonstrated superior clinical decision-making skills, suggesting that role-play scenarios may foster critical thinking and integrative skills by actively engaging students in problem-solving rather than passive observation [15]. This active learning approach likely provides students with a deeper understanding of clinical workflows, which is essential for patient-centered care in high-stakes oncology settings.

The high levels of satisfaction reported by students in the role-play group underscore its acceptability and perceived value in clinical education. Students valued the realism of the scenarios and the opportunity for constructive feedback, contributing to the observed improvements in communication and clinical skills. These positive feedback results suggest that role-play was well received by students, particularly in terms of engagement and the quality of feedback. However, the results do not directly suggest an enhancement of learning outcomes or motivation, which would require further investigation.

Our scenarios deliberately targeted competencies central to oncology and palliative care communication—eliciting patient values and concerns, conveying difficult information, and negotiating shared decision-making—guided by the SPIKES framework and principles from the Calgary-Cambridge guide [9,10]. Prior work has shown that structured role-play and simulation can improve learners' confidence and observable communication behaviors in serious illness conversations, whereas ongoing feedback and deliberate practice are critical to maintaining gains. In settings in which SP programs and high-fidelity simulations are not yet routinely available in oncology rotations, peer role-play provides a pragmatic, low-cost modality that can be delivered regularly, reinforced by structured debriefing. Future work at our institution will extend scenario content to explicitly include dialogues on end of life and goals of care and compare peer role-play with SP-based approaches on performance and transfer to clinical practice.

Despite promising findings, several limitations should be acknowledged. First, this study was conducted at a single institution and focused solely on oncology graduate students, which may limit its generalizability to other medical fields. Additionally, the short rotation duration raises questions about the long-term retention of skills acquired through role-play. Another limitation is the potential for evaluator bias in Mini-CEX assessments, although the evaluator was blinded to group assignments.

This single-center study used a posttest-only design for the Mini-CEX and administered the theory exam only at baseline. Consequently, we cannot estimate within-group change or link knowledge gains to observed performance. We also relied on a single blinded assessor for Mini-CEX ratings, which may limit the generalizability of judgments. Satisfaction data were self-reported and limited to the role-play group. Finally, we did not include a dedicated empathy instrument or teamwork measure.

Future research could address these limitations by expanding sample sizes, including multiple institutions, and assessing long-term skill retention. Incorporating objective measures of patient outcomes could help assess the real-world impact of role-play on patient care quality. Additionally, comparing role-play with other interactive learning methods such as simulation-based training could deepen understanding of the most effective approaches for teaching clinical skills.

In conclusion, this study demonstrates that peer role-play is an effective and well-received method for enhancing clinical competencies in oncology education. By engaging students in realistic scenarios, role-play fosters key skills in communication, decision-making, and teamwork, which are all crucial for patient-centered care. These findings support the integration of role-play into clinical curricula as a valuable complement to traditional teaching methods, especially in fields that demand strong interpersonal and decision-making skills.

Data Availability

All datasets generated for this study were included in the manuscript.

Authors' Contributions

JL and HB conceived and designed the study. YW, FW, GL, and YL participated in data collection and processing. JL was the major contributor in organizing records and drafting the manuscript. All authors have proofread and approved the manuscript. YW and FW have contributed equally to this paper. Correspondence should be addressed to JL (longjie07@126.com) or HB (bahj3@mail.sysu.edu.cn).

Conflicts of Interest

None declared.

References

1. Moore PM, Rivera S, Bravo-Soto G, Olivares C, Lawrie T. Communication skills training for healthcare professionals working with people who have cancer. Cochrane Database Syst Rev 2018 Jul 24;7(7):CD003751 [FREE Full text] [doi: 10.1002/14651858.CD003751.pub4] [Medline: 30039853]
2. Razavi D, Delvaux N. Communication skills and psychological training in oncology. Eur J Cancer 1997 Jul;33 Suppl 6:S15-S21. [doi: 10.1016/s0959-8049(97)00195-0] [Medline: 9404235]
3. Epner DE, Baile WF. Difficult conversations: teaching medical oncology trainees communication skills one hour at a time. Acad Med 2014 Apr;89(4):578-584 [FREE Full text] [doi: 10.1097/ACM.0000000000000177] [Medline: 24556763]
4. Rønning SB, Bjørkly S. The use of clinical role-play and reflection in learning therapeutic communication skills in mental health education: an integrative review. Adv Med Educ Pract 2019 Jun;10:415-425 [FREE Full text] [doi: 10.2147/AMEP.S202115] [Medline: 31417328]
5. Bagacean C, Cousin I, Ubertini A, El Yacoubi El Idrissi M, Bordron A, Mercadie L, et al. Simulated patient and role play methodologies for communication skills and empathy training of undergraduate medical students. BMC Med Educ 2020 Dec 04;20(1):491 [FREE Full text] [doi: 10.1186/s12909-020-02401-0] [Medline: 33276777]
6. Bouaoud J, Michon L, Saintigny P. Teaching how to break bad news in oncology: in-class vs. virtual peer role-plays. Bull Cancer 2022 Jun;109(6):685-691. [doi: 10.1016/j.bulcan.2022.02.009] [Medline: 35523599]
7. Nestel D, Tierney T. Role-play for medical students learning about communication: guidelines for maximising benefits. BMC Med Educ 2007 Mar 02;7:3 [FREE Full text] [doi: 10.1186/1472-6920-7-3] [Medline: 17335561]
8. Gelis A, Cervello S, Rey R, Llorca G, Lambert P, Franck N, et al. Peer role-play for training communication skills in medical students: a systematic review. Simul Healthc 2020 Apr;15(2):106-111. [doi: 10.1097/SIH.0000000000000412] [Medline: 32168292]
9. Meitar D, Karnieli-Miller O. Twelve tips to manage a breaking bad news process: using S-P-w-ICE-S - a revised version of the SPIKES protocol. Med Teach 2022 Oct 30;44(10):1087-1091. [doi: 10.1080/0142159X.2021.1928618] [Medline: 34057007]
10. Kurtz S, Silverman J, Benson J, Draper J. Marrying content and process in clinical method teaching: enhancing the Calgary-Cambridge guides. Acad Med 2003 Aug;78(8):802-809. [doi: 10.1097/00001888-200308000-00011] [Medline: 12915371]
11. Weller JM, Nestel D, Marshall SD, Brooks PM, Conn JJ. Simulation in clinical teaching and learning. Med J Aust 2012 May 21;196(9):594. [doi: 10.5694/mja10.11474] [Medline: 22621154]
12. Yousef N, Moreau R, Soghier L. Simulation in neonatal care: towards a change in traditional training? Eur J Pediatr 2022 Apr 12;181(4):1429-1436 [FREE Full text] [doi: 10.1007/s00431-022-04373-3] [Medline: 35020049]
13. Bosse HM, Nickel M, Huwendiek S, Schultz JH, Nikendei C. Cost-effectiveness of peer role play and standardized patients in undergraduate communication training. BMC Med Educ 2015 Oct 24;15:183 [FREE Full text] [doi: 10.1186/s12909-015-0468-1] [Medline: 26498479]
14. Yamauchi K, Hagiwara Y, Iwakura N, Kubo S, Sato A, Ohtsuru T, et al. Using peer role-playing to improve students' clinical skills for musculoskeletal physical examinations. BMC Med Educ 2021 Jun 05;21(1):322 [FREE Full text] [doi: 10.1186/s12909-021-02742-4] [Medline: 34090441]
15. Baillie S, Pierce SE, May SA. Fostering integrated learning and clinical professionalism using contextualized simulation in a small-group role-play. J Vet Med Educ 2010 Sep;37(3):248-253. [doi: 10.3138/jvme.37.3.248] [Medline: 20847333]

Abbreviations

Mini-CEX: Mini-Clinical Evaluation Exercise
SP: standardized patient
SPIKES: setting, perception, invitation, knowledge, empathy, and strategy

Edited by D Chartash; submitted 27.06.25; peer-reviewed by N Philp, CC Chang; comments to author 29.07.25; revised version received 27.09.25; accepted 03.11.25; published 28.11.25.

Please cite as:
Wang Y, Wang F, Liu G, luo Y, Ba H, Long J
Enhancing Clinical Competencies Through Peer Role-Play in Oncology Graduate Students: Mixed Methods Study
JMIR Med Educ 2025;11:e79771
URL: https://mededu.jmir.org/2025/1/e79771
doi:10.2196/79771
PMID:

Effectiveness of a Fully Online Scientific Research Works Peer Support Group Model for Research Capacity Building Through Conducting Systematic Reviews Among Health Care Professionals: Retrospective Cohort Studies

Yuki Kataoka[1,2,3,4,5,6], MD, MPH, DrPH; Ryuhei So[1,7,8], MD, MPH, DrPH; Masahiro Banno[1,9], MD, PhD; Yasushi Tsujimoto[1,10,11], MD, MPH, DrPH; SRWS-PSG Mentors[1]

[1]Scientific Research WorkS Peer Support Group, 12-12 Osaka Ekimae Dai-2 Building 1-2-2 Umeda, Kita-ku, Osaka, Japan

[10]Oku Internal Medicine Clinic, Osaka, Japan

[11]Department of Health Promotion and Human Behavior, Kyoto University Graduate School of Medicine/School of Public Health, Kyoto, Japan

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

[3]Center for Medical Education, Graduate School of Medicine, Nagoya University, Aichi, Japan

[4]Department of Internal Medicine, Kyoto Min-iren Asukai Hospital, Kyoto, Japan

[5]Department of Healthcare Epidemiology, Kyoto University Graduate School of Medicine/School of Public Health, Kyoto, Japan

[6]Department of International and Community Oral Health, Tohoku University Graduate School of Dentistry, Miyagi, Japan

[7]Department of Psychiatry, Okayama Psychiatric Medical Center, Okayama, Japan

[8]CureApp, Inc., Tokyo, Japan

[9]Department of Psychiatry and Neurology, Seichiryo Hospital, Nagoya, Japan

Corresponding Author:
Yuki Kataoka, MD, MPH, DrPH
Scientific Research WorkS Peer Support Group, 12-12 Osaka Ekimae Dai-2 Building 1-2-2 Umeda, Kita-ku, Osaka, Japan

Abstract

Background: Research capacity building (RCB) among health care professionals remains limited, particularly for those working outside academic institutions. Japan is experiencing a decline in original clinical research due to insufficient RCB infrastructure. Our previous hospital-based workshops were effective but faced geographical and sustainability constraints. We developed a fully online Scientific Research Works Peer Support Group (SRWS-PSG) model that addresses geographical and time-bound constraints and establishes a sustainable economic model. Mentees use online materials, receive support from mentors via a communication platform after formulating their research question, and transition into mentors upon publication.

Objective: We evaluated whether our model's theoretical benefits translated into actual program effectiveness in RCB among health care professionals.

Methods: We conducted a retrospective cohort study of health care professionals who participated in the SRWS-PSG program between September 2019 and January 2025. Mentees progressed through a structured modular curriculum covering systematic review methodology, from protocol development to manuscript preparation, with personalized mentoring support. We evaluated manuscript submission, program discontinuation, promotion to a mentor status, and mentor response time. We collected data from program records and chat logs. Manuscript submission was defined as mentor-confirmed submission of a systematic review manuscript to a peer-reviewed journal. Program discontinuation referred to formal withdrawal before manuscript submission. Mentor promotion was defined as acceptance of an invitation to serve as a junior mentor after manuscript submission. Mentor response time was the elapsed time from a mentee's question in the chat to the first reply by an assigned mentor.

Results: Of 85 mentees analyzed, 31 (36.5%) held academic degrees (PhD or MPH), and 68 (80%) were medical doctors. During a median follow-up of 10 months, 51 (60%) submitted manuscripts and 46 (90%) became mentors. Ten mentees (12%) discontinued the program. The median mentor response time was 0.8 hours, with 90% responding within 24 hours.

Conclusions: A majority of participants of SRWS-PSG submitted manuscripts. This fully online RCB program might address geographical barriers and provides an adaptable approach for RCB across diverse health care contexts.

(JMIR Med Educ 2025;11:e78862) doi:10.2196/78862

KEYWORDS

research capacity building; online education; systematic review; peer support group; health care professionals; mentoring; e-learning

Introduction

Research capacity building (RCB) among health care professionals is essential for addressing clinical challenges and improving health care quality. Cooke's RCB framework [1] comprises four structural levels (individual, team, organizational, and network/supra-organizational) and six principles (developing skills and confidence, supporting linkages and partnerships, ensuring research close to practice, enabling appropriate dissemination, investing in infrastructure, and building sustainability and continuity). A systematic review showed that RCB strategies are interlinked and interdependent, requiring implementation through an integrated "whole of system" approach with commitment and support from all levels of leadership and management [2]. Similarly, a scoping review found that these programs are typically multifaceted, often incorporating experiential learning and mentoring, but evaluations predominantly focus on lower levels of Kirkpatrick's educational outcomes typology (eg, participant satisfaction, improved knowledge, and confidence), with few assessing objective milestones (eg, protocol completion or manuscript preparation) or broader organizational and practice impacts. The authors conclude that rigorous evaluations are needed, considering long-term outcomes and translation into clinical practice, to better bridge the research-practice gap [3].

RCB programs for practicing health care professionals outside academic institutions remain limited [2]. In Japan, the decline in original clinical research is a recognized issue [4,5]. This is largely due to insufficient RCB infrastructure, a problem that affects both academic and non-academic settings [6,7]. Like in other countries, there are often no clear expectations or structured incentives for clinicians to engage in research [8]. The situation is worsened by systemic funding shortages in academia [9]. Many Japanese clinicians are therefore motivated not by the prospect of academic promotion, but by a sense of professional responsibility—or "rectitude," a concept deeply rooted in the traditional Japanese code of conduct [10]—to address clinical questions that arise from their practice. Our program was developed to support these grass-roots community healthcare professionals, who are motivated to address their research questions [11]. The program aimed to build research capacity from the bottom up. Between 2014 and 2017, we conducted a structured in-person scholarly productive model workshop for hospital-based health care professionals. This workshop developed a curriculum and achieved academic publications from participants through team-based systematic review projects. However, this hospital-based implementation faced two constraints: limited dissemination capacity and unsustainable dependence on voluntary mentor work. In 2019, we transformed the workshop into a fully online peer support program (Scientific Research Works Peer Support Group [SRWS-PSG]) with participant fees supporting mentor compensation. This redesign addressed geographical and time-bound barriers, strengthened cross-institutional partnerships, and established a sustainable economic model for ongoing RCB activities. The program features a comprehensive curriculum with over 140 modules covering all aspects of systematic review methodology. Mentees progress through self-paced online materials and receive feedback from mentors on their protocols, manuscripts, and any questions about the systematic review process via an internet communication platform, eventually transitioning into mentors upon publication.

We previously developed a structured, in-person scholarly productive model workshop for hospital-based health care staff between 2014 and 2017 [11]. This workshop developed a curriculum and achieved academic publications from participants through team-based systematic review projects. However, this hospital-based implementation faced two constraints: limited dissemination capacity and unsustainable dependence on voluntary mentor work. In 2019, we transformed the workshop into a fully online peer support program (SRWS-PSG) with participant fees supporting mentor compensation. This redesign addressed geographical and time-bound barriers, strengthened cross-institutional partnerships, and established a sustainable economic model for ongoing RCB activities. The program features a comprehensive curriculum with over 140 modules covering all aspects of systematic review methodology. Mentees progress through self-paced online materials and receive feedback from mentors on their protocols, manuscripts, and any questions about the systematic review process via an internet communication platform, eventually transitioning into mentors upon publication.

While the need for RCB for health care professionals is clear, scalable and sustainable programs remain scarce, particularly for those outside academic institutions in Japan. This study, therefore, aimed to evaluate the effectiveness of our fully online and economically self-sustaining SRWS-PSG model. Our research question was "How effective is this program in enabling health care professionals to complete and submit systematic reviews?" To address this, we conducted a retrospective cohort study analyzing data from participants enrolled between 2019 and 2025. We hypothesized that this peer-supported, online model would demonstrate a high proportion of manuscript submission and program continuation, validating its utility as a scalable RCB solution. The findings are intended to inform medical educators, hospital administrators, and health policymakers seeking to develop and implement effective RCB programs.

Methods

Study Design

We conducted a retrospective cohort study of health care professionals who participated in the SRWS-PSG program between September 2019 and January 2025. The SRWS-PSG is designed specifically to support health care professionals from any discipline in conducting systematic reviews as their primary research output. The SRWS-PSG program is now

operated by the SRWS-PSG, a non-profit general incorporated association registered in Japan (corporate Number: 6120005024588). The mission is to increase clinical research in Japan. The members consist of health care professionals, and it is funded through membership and mentorship fees. The program provides a structured, modular curriculum covering the entire systematic review process, from protocol development to manuscript submission. Learning is facilitated through self-paced online materials and personalized, responsive mentorship via a communication platform. We applied Cooke's RCB framework [1] for the methodology and structured our findings. To report this article, we followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement [12].

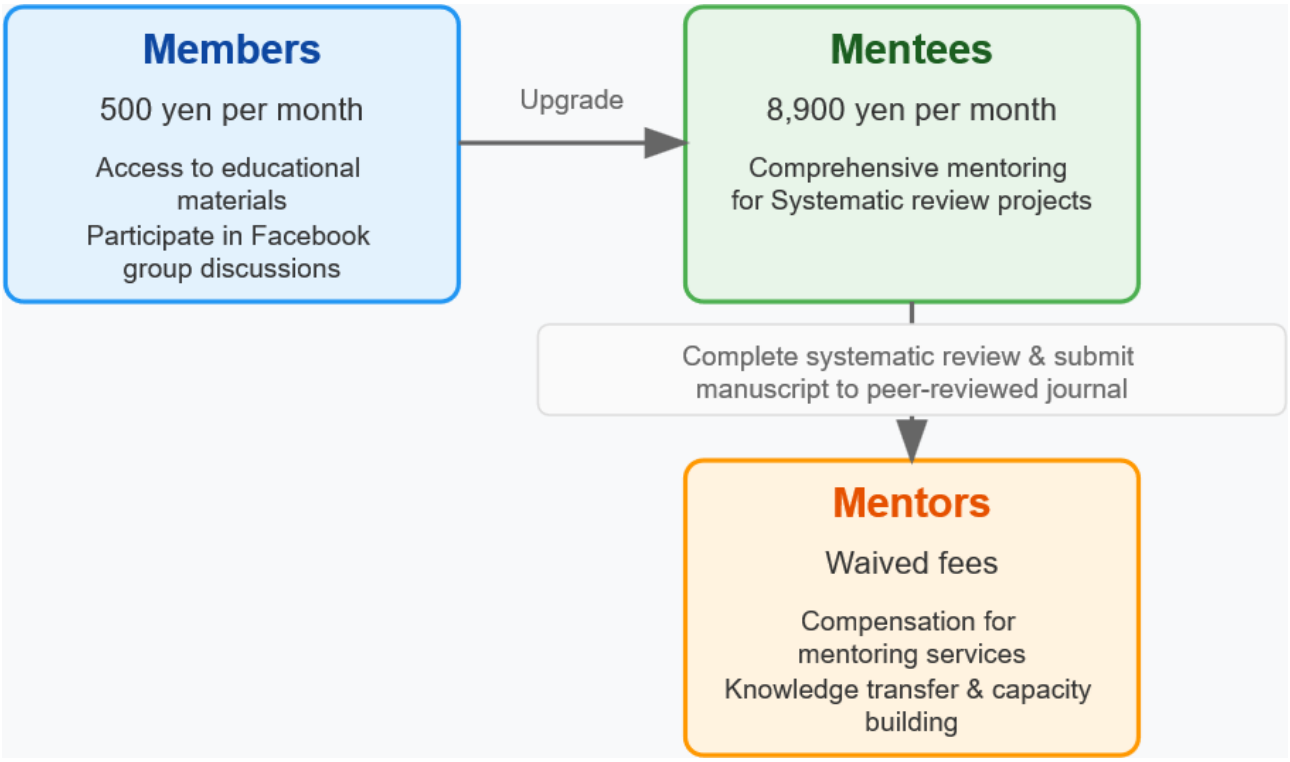
Setting

This study was conducted in a fully online environment. We recruited participants through social media, websites, academic conferences, and referrals from previous participants. The primary registration portal was CAMPFIRE, a Japanese crowdfunding platform [13]. We collected data from Slack chat logs, program participation records, and manuscript submission status. Occupations and academic degrees came from registration forms. The data cut-off was February 2025.

Participants

As of 2025, the SRWS-PSG uses a tiered membership structure with different levels of engagement and fee schedules. Entry-level "members" pay 500 yen (3 - 4 US$) per month to access educational materials and participate in research-related discussions through a Facebook group. "Mentees" pay 8900 yen (around 60 US$) per month for comprehensive mentoring support throughout their systematic review projects. Mentees who submit their manuscript to a peer-reviewed journal are invited to become mentors—with waived fees and compensation—if their two assigned mentors assess them as suitable for the role based on their aptitude for chat-based communication (Figure 1).

Figure 1. The Scientific Research Works Peer Support Group (SRWS-PSG) membership structure.



In this study, we analyzed mentees who were health care professionals from various disciplines including physicians, nurses, pharmacists, physical therapists, clinical psychologists, and registered dietitians. All mentees could communicate online, had research questions suitable for systematic review topics, and agreed to participate. None had conducted a systematic review as the lead author. We excluded those withdrew before program initiation or did not complete initial training. The cumulative incidences of manuscript submission and discontinuation of the program, stratified by university affiliation, are plotted with 95% CIs. Shaded areas represent the 95% CIs.

Curriculum

The SRWS-PSG curriculum uses outcome-based learning objectives to enable health care professionals to conduct systematic reviews and publish in peer-reviewed journals [14]. Objectives include formulating research questions, developing literature search strategies, performing critical appraisal, synthesizing evidence, evaluating certainty of evidence, and producing academic manuscripts meeting international publication standards. The curriculum follows a modular progression aligned with the systematic review process, from orientation modules through protocol development, execution, analysis, and manuscript preparation (Multimedia Appendix 1). Training materials are in Japanese. To become familiar with English academic writing, mentees without English writing

experience critically appraised an article and submitted a letter to the editor before formulating their research question. The program incorporates methodological updates including the latest Cochrane Handbook [15] and Risk of Bias 2 tool [16] to ensure current best practices.

The instructional methodology combines asynchronous self-paced learning with responsive mentoring support. Mentees first view 10 to 30 minutes of video content explaining systematic review concepts and methodologies. Mentees then draft specific sections of their research protocols—structured tasks with concrete deliverables. Each mentee receives personalized feedback from two assigned mentors—a junior and a senior mentor—through Slack, a business-oriented chat application [17]. Mentors serve as co-authors on the mentee's systematic review project. Junior mentors are recent program graduates, who are promoted to senior status after guiding three mentees to publication. Mentors respond within 24 hours, while mentees follow a "15-minute rule" (a structured guideline that requires mentees to attempt solving methodological or procedural challenges independently for at least 15 min before posting questions to mentors, designed to promote self-reliance, enhance problem-solving skills, and ensure efficient use of mentoring resources). Optional synchronous online meetings occur upon mentee request.

To enhance learning through practical application, members actively contribute to other mentees' projects. This approach leverages the systematic review methodology's requirement for independent dual assessment of article screening, data extraction, and risk of bias evaluation [18]. This collaborative structure enhances learning through practical application.

Outcomes

Overview

We selected four outcomes based on data available from our existing program records. These outcomes partially reflect key dimensions of Cooke's RCB framework [1]. Manuscript submission was chosen as an indicator of skill development and confidence building, representing participants' ability to complete a systematic review project. Program discontinuation was measured to assess individual-level sustainability and continuity. Promotion to mentor status following manuscript submission was evaluated as a marker of organizational-level sustainability and continuity. Mentor response time was analyzed as a proxy for infrastructure investment and team support quality. The outcomes were described as given below.

Manuscript Submission

Manuscript submission was defined as the time from mentor approval of the mentee's research question to start the protocol to the submission of systematic review manuscripts to academic journals. Mentees self-reported manuscript submission with mentor confirmation.

Program Discontinuation

Discontinuation was defined as when a mentee formally notified the program of their withdrawal, leading to the cessation of monthly program fee payments prior to manuscript submission.

Promotion to a Mentor

Promotion to mentor status was evaluated by determining the proportion of manuscript-submitting mentees who accepted the invitation to become a junior mentor after being assessed as suitable by their assigned mentors. This transition was documented in program administration records.

Mentor Response Time

The mentor response time was defined as the time from a mentee posting a question in a designated Slack channel to the first response from any assigned mentor in the same channel. Data were extracted from Slack chat logs.

Study Size

We included all mentees during the program operation period, without formal sample size calculation. This size was similar to that in other research capacity development program evaluation studies [19,20].

Statistical Analysis

We used descriptive statistics to summarize participant characteristics, continuation rate, and mentor response time. Manuscript submission and publication were analyzed as time-to-event outcomes. We estimated the cumulative incidence functions for both submission and publication using the Aalen-Johansen estimator [21]. Discontinuation from the program was a competing risk. Competing risk occurs when an event (manuscript submission) cannot occur because another event (discontinuation) happens first. We censored participants still active without an event at data cut-off. For subgroup analyses, we performed univariable Fine-Gray subdistribution hazard models to calculate subdistribution hazard ratios (sdHR) with 95% CIs for academic degree, profession, and university affiliation. We performed analyses using Python (version 3.11.4; Python Software Foundation) and R (version 4.3.2; R Foundation for Statistical Computing). Python packages included *lifelines* (version 0.30.0) and *scikit-survival* (version 0.24.1). R packages included *cmprsk* (version 2.2.12).

Ethical Considerations

The Ethics Committee of Kyoto Min-iren Asukai Hospital approved the study protocol (Approval number: 202502 - 3) and waived the need for individual informed consent. The waiver was granted because this retrospective study analyzed routinely collected clinical data without any direct contact or intervention, posed no more than minimal risk to participants, and obtaining consent from all eligible patients would have been impracticable. All procedures complied with the institutional ethics standards and the Declaration of Helsinki. To protect privacy and confidentiality, the analytic dataset was deidentified before analysis: direct identifiers were removed. Study data were kept on access-controlled servers.

Results

Participants

During the study period, 118 individuals registered. After excluding 33 participants (30 who failed initial training for formulating research questions, and 3 who did not start), 85

mentees entered the final analysis. Of 85 mentees, 31 (36.5%) held academic degrees (PhD: 27, MPH: 2, PhD and MPH: 2), and 44 (51.8%) were working at university-affiliated hospitals. Sixty-eight (80%) were medical doctors, and 17 (20%) represented other health care professions: physical therapists (n=8), pharmacists (n=3), clinical psychologists (n=2), nurses (n=1), registered dietitians (n=1), dentists (n=1), and unspecified (n=1). No mentees had missing follow-up data.

Outcomes

During a median follow-up of 10 months (range, 0 - 54 months), 51 mentees (60%) submitted manuscripts. Forty-six (90%) became mentors. Ten (12%) discontinued. Discontinuation reasons included work commitments (n=4), workplace change (n=2), unclear (n=2), personal reasons (n=1), and competing systematic review publication (n=1). Twenty-four mentees continue projects. Cumulative manuscript submission reached 75.2% (95% CI 62.1% - 84.3%), and cumulative discontinuation was 18% (95% CI 9.3% - 28.9%) at the end of follow-up (Figure 2). At 12 months, manuscript submission was 42.9% (95% CI 31.4% - 53.9%) and discontinuation was 2.7% (95% CI 0.5% - 8.5%). At 24 months, manuscript submission was 63.3% (95% CI 50.6% - 73.6%) and discontinuation was 5.8% (95% CI 1.9% - 13.1%). In univariable analyses, the sdHR for manuscript submission was 0.78 (95% CI 0.43 - 1.43) for academic degree, 1.00 (95% CI 0.53 - 1.86) for profession, and 1.31 (95% CI 0.77 - 2.24) for university affiliation (Figures 3-5). Of 51 submissions, 50 appeared in English peer-reviewed journals. One submission remains under peer review. Submissions included 32 pairwise meta-analyses, 7 scoping reviews, 4 network meta-analyses, 2 prognostic studies, 2 diagnostic test accuracy studies, and 3 other study types (Multimedia Appendix 2). All published articles were indexed in the Web of Science at the time of publication. Journal impact factors were available for all except *The Annals of Translational Medicine*, which was indexed in the Web of Science and PubMed at submission but later excluded from the Web of Science. Citation counts, disregarding time since publication, had a median of 6 (range 0 - 60) as of Aug 1, 2025. Twelve articles (24%) appeared in journals requiring article processing charges. Analysis of 58,952 messages showed a median mentor response time of 0.8 hours (IQR 0.1 - 6.1), with 90% responding within 24 hours.

Figure 2. Cumulative incidence of manuscript submission and discontinuation of the program are plotted with 95% CIs. Shaded areas represent the 95% CIs.

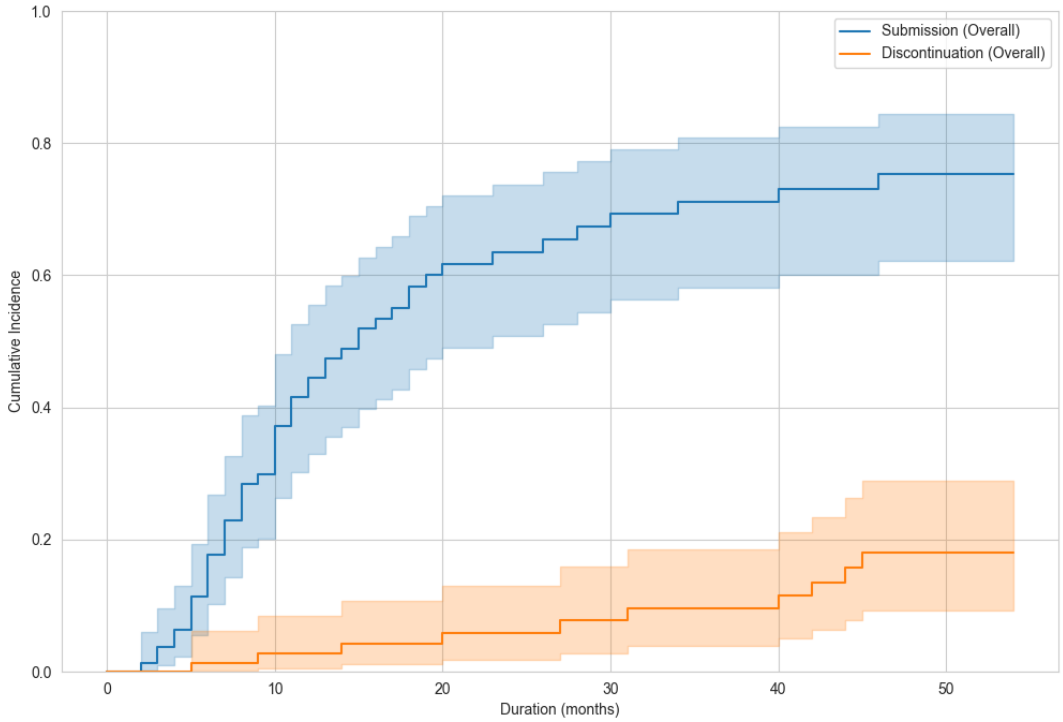


Figure 3. Cumulative incidence of manuscript submission and discontinuation of the program, stratified by the presence or absence of an academic degree (PhD or MPH), are plotted with 95% CIs. Shaded areas represent the 95% CIs.

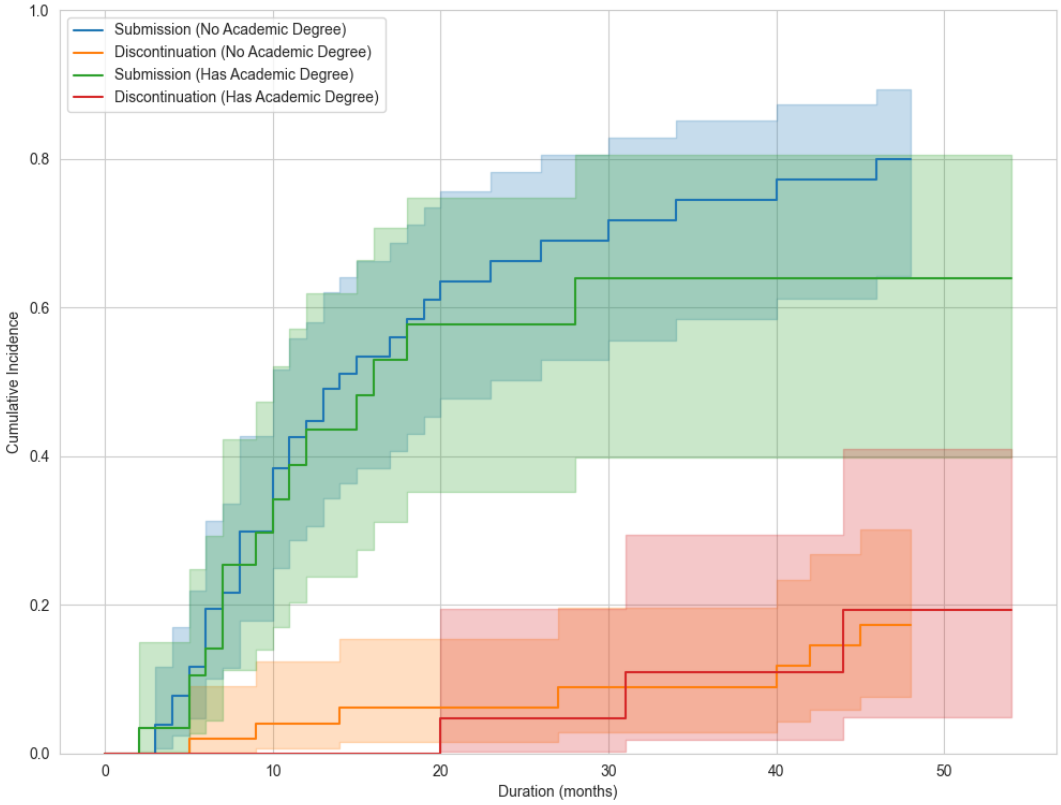


Figure 4. Cumulative incidences of manuscript submission and discontinuation of the program, stratified by profession (medical doctor or non-medical doctor), are plotted with 95% CIs. Shaded areas represent the 95% CIs.

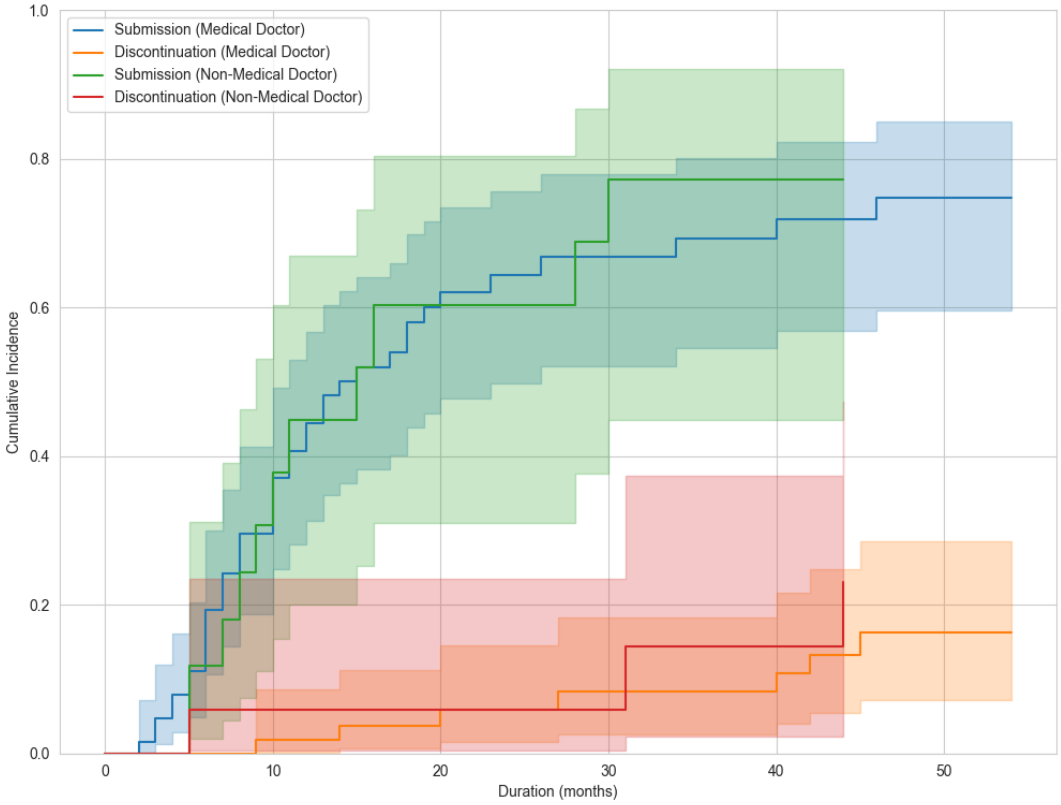
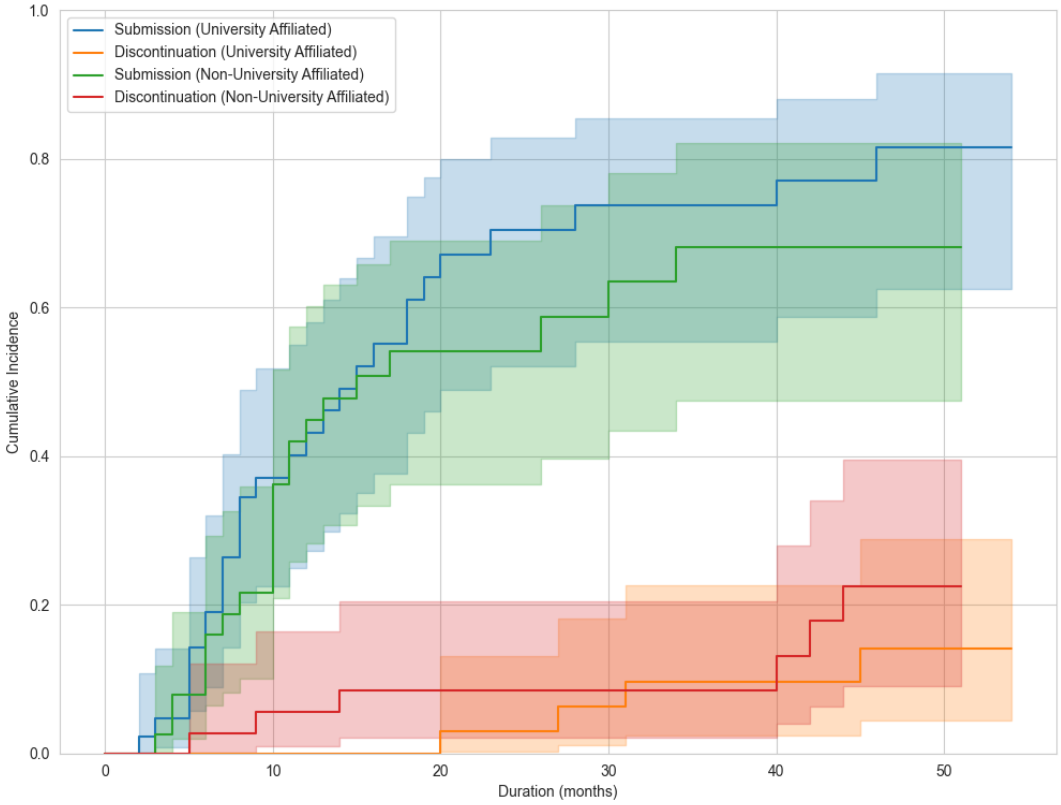


Figure 5. Cumulative incidences of manuscript submission and discontinuation of the program, stratified by university affiliation, are plotted with 95% CIs. Shaded areas represent the 95% CIs.



Discussion

Principal Results

The principal finding is that our fully online, peer-supported RCB model might contribute to actual research outputs for health care professionals. Specifically, cumulative manuscript submission reached 75.2% at the end of follow-up, while cumulative discontinuation remained at 18%. The publications appeared in journals indexed in the Web of Science, with a median citation count of 6 (range 0 - 60), indicating initial uptake. The rapid mentor responsiveness (median 0.8 hours) further contributed to program effectiveness. Ninety percent of mentees who submitted manuscripts transitioned to mentors. These quantitative outcomes provide a foundation for the subsequent discussion of the program's effectiveness through the lens of Cooke's RCB framework [1].

While we did not set specific a priori cut-offs for outcomes, we consider the observed outcomes meaningful. Of note, we view discontinuation as a particularly critical metric because it directly reflects program adequacy—participants who leave the program cannot achieve the primary outcome of manuscript submission. The reasons for discontinuation (work commitments, workplace changes, personal reasons, and competing publications) were all external factors beyond the scope of SRWS-PSG's organizational support, suggesting that the program's support structure itself was adequate.

Implications of Findings

Research-novice medical professionals were able to participate in the SRWS-PSG program. Our subgroup analyses demonstrated that manuscript submission rates were similar regardless of participants' academic degrees, professional backgrounds, or institutional affiliations. The program's fully online format addresses geographical and temporal constraints, enabling flexible participation around clinical duties. Rapid mentor feedback provides continuous support for novices navigating the systematic review process. This accessibility suggests that the SRWS-PSG model might expand RCB opportunities beyond traditional academic boundaries.

Hospital administrators seeking to foster their medical professionals' research competency could provide protected time for SRWS-PSG participation. The primary reasons for program discontinuation were external factors such as work commitments and workplace changes rather than program inadequacy. Institutional support through protected time could address these barriers and enhance completion rates [6]. The program's modest monthly fee represents a cost-effective investment for institutions.

The SRWS-PSG model's features suggest broad applicability across diverse global contexts, including low- and middle-income countries. The fully online format eliminates geographical barriers and infrastructure requirements, while the focus on systematic reviews—requiring only internet access without expensive equipment or data collection costs—makes research accessible in resource-limited settings. The self-sustaining economic model through participant fees offers an alternative to scarce grant funding. While implementation would require translation and fee adjustment to local contexts, the core elements—peer support, standardized methodology following international guidelines, and progression from mentee to mentor—address universal challenges in RCB regardless of economic or cultural setting.

Comparison to the Literature

Our findings extend previous RCB evaluations in several ways. The SRWS-PSG model's continuous feedback approach shares elements with intensive programs like writing boot camps [22] but delivers them in a sustained, online format throughout the entire systematic review process. This contrasts traditional time-limited workshops. Our approach shares goals with the Rural Research Capacity Building Program in Australia [20] but operates entirely online.

Mentees formulated research questions directly relevant to their clinical environments, bridging research and practice in ways that promote Boyer's scholarship concept of "integration" and "application" [23]. This contrasts with academic programs where research topics may be driven by funding or institutional priorities rather than clinical needs.

The self-supporting economic model offers a sustainable alternative to grant-dependent programs. Mentee fees fund mentor compensation and administrative costs. This enables 5 years of operation without external funding. Many RCB programs require institutional or funding support [2]. This approach echoes Japan's Edo period tradition of community-based education. Community mathematics teachers ("Wasan-ka") sustained themselves through teaching fees. This grass-roots system enabled people from all social classes to publish their mathematical problems at shrines as open access articles ("Sangaku") [24]. SRWS-PSG functions as a traditional citizen science model, maintaining economic sustainability through community participation rather than institutional supports.

Limitations and Future Directions

This study has several limitations. First, this retrospective observational study without a control group cannot establish causality. Although continuous program improvements pose a challenge for evaluation, a rigorously designed well-powered randomized controlled trial is the necessary next step to definitively assess the efficacy. Second, our metrics emphasized outputs such as submission but did not directly evaluate some dimensions of Cooke's framework. Specifically, we did not assess changes in individual research skills (including systematic review methodology competency beyond evidence-based medicine competency [25]), self-reported confidence, or participant motivation levels. We also could not evaluate the quality of mentor-mentee relationships, organizational-level support, and infrastructure investments, or the dynamics of online community participation and peer interactions that illuminate collaborative learning. Future studies should incorporate comprehensive assessments including skills evaluation, qualitative analyses of mentoring relationships, surveys of institutional support structures, and quantitative measures of online engagement to elucidate how these collaborative learning dynamics influence program outcomes. Third, our analysis of participant affiliation was limited to a binary variable (university-affiliated or not), which is an imperfect proxy for the actual research environment. In Japan, university affiliation does not guarantee a supportive research environment [6]. Future studies should collect more granular data on the institutional support environment. Fourth, a challenge remains in the formal assessment of mentees for mentorship roles and the structured faculty development for new mentors. While we provide web-based training on mentoring, a comprehensive evaluation system for these skills is not yet fully established and represents a key area for future development. Fifth, while we provided summary data on publication venues and citations, we did not conduct a formal analysis of article quality metrics beyond indexing status. Future evaluations should include more detailed assessments, such as independent article quality ratings, to better characterize program outputs.

Looking forward, we are exploring large language models as a "third mentor" [26,27] to improve mentor feedback speed and quality while reducing workload. We are investigating the application of large language models in the systematic review process to accelerate research timelines [28,29]. While being mindful of the potential for hallucinations in large language models [30], SRWS-PSG will improve the efficiency of medical scientific endeavors.

Conclusion

A majority of health care professionals who participated in the SRWS-PSG program submitted manuscripts with low discontinuation rates. This model might address geographical barriers and create a self-sustaining financial structure, providing an adaptable approach for RCB across health care contexts. Future research should incorporate a more comprehensive evaluation of program mechanisms and outcomes.

Acknowledgments

We would like to express our gratitude to all SRWS-PSG participants and mentors who contributed to this program over the years. Special thanks to Chisato Fujiwara for her invaluable secretarial assistance. We used Claude 4.0 Opus for English Editing. All authors reviewed the final manuscript.

Open Access publication supported by Tohoku University FY2025 APC Support Program.

Data Availability

De-identified data that support the findings of this study are available from the corresponding author upon reasonable request, subject to approval from the ethics committee.

Authors' Contributions

YK: Conceptualization, Data curation, Formal analysis, Validation, Investigation, Visualization, Methodology, Writing–original draft, Project administration. RS: Conceptualization, Investigation, Writing–review & editing. MB: Conceptualization, Data curation, Investigation, Writing – review & editing. YT: Conceptualization, Data curation, Investigation, Writing–review & editing.Data curation, Investigation, Writing – review & editing: TA, TA, EI, KK, RK, TK, SK, KK, YK, NK, RM, TN, YN, YO, MO, YS, AS, HS, YS, ST, TT, YW, JW, MY, NY, YY, and SY.

Tetsuro Aita, MD, PhD (ORCID https://orcid.org/0000-0002-1693-361X) 1. Department of General Internal Medicine and Family Medicine, Fukushima Medical University, 1 Hikarigaoka, Fukushima City, Fukushima, 960-1295, Japan 2. Scientific Research Works Peer Support Group (SRWS-PSG), Osaka, Japan.

Takashi Ariie, PT, PhD (ORCID:https://orcid.org/0000-0003-1457-3427) 1. Department of Physical Therapy, School of Health Sciences at Fukuoka, International University of Health and Welfare, 137-1 Enokizu, Okawa-shi, Fukuoka, 831-8501, Japan 2. Scientific Research Works Peer Support Group (SRWS-PSG), Osaka, Japan.

Eriya Imai, MD (ORCID: https://orcid.org/0000-0002-4511-4672) 1. Division of Anesthesia, Mitsui Memorial Hospital, Kanda-Izumi-cho-1, Chiyoda-ku, Tokyo 101-8643, Japan 2. International University of Health and Welfare Graduate School of Public Health,Tokyo, Japan. 4-1-26 Akasaka, Minato City, Tokyo 107-8402, Japan.

Kyosuke Kamijo, MD (ORCID: https://orcid.org/0000-0001-7375-7651) 1.Department of Obstetrics and Gynecology, Nagano Prefectural Shinshu Medical Center, 1332 Suzaka, Suzaka, 382-8577, Japan. 2. Scientific Research Works Peer Support Group (SRWS-PSG), Osaka, Japan m10025kk@jichi.ac.jp

Ryota Kimura, MD, PhD (ORCID https://orcid.org/0000-0002-4855-3283) 1. Department of Orthopedic Surgery, Akita University Graduate School of Medicine, 1-1-1 Hondo, Akita, Akita, 010-8543, Japan 2. Scientific Research Works Peer Support Group (SRWS-PSG), Osaka, Japan.

Takashi Kitagawa, PT, PhD 1. Department of Physical Therapy, School of Health Sciences, Shinshu University, 3-1-1 Asahi, Matsumoto, 390-8621 Japan 2. Scientific Research Works Peer Support Group (SRWS-PSG), Osaka, Japan.

Shunsuke Kondo, MD (https://orcid.org/0009-0007-3372-010X) 1. Department of Medicine, John A Burns School of Medicine, Honolulu, Hawaii, 96813, US 2. Scientific Research Works Peer Support Group (SRWS-PSG), Osaka, Japan.

Kazumasa Kotake, BS (ORCID: https://orcid.org/0000-0001-9869-1177) 1. Department of Pharmacy, Zikei Hospital/Zikei Institute of Psychiatry, 100-2 Urayasu Honmachi, Minami-Ku, Okayama-shi, Okayama, 702-8508, Japan 2. Scientific Research Works Peer Support Group (SRWS-PSG), Osaka, Japan.

Yasutaka Kuniyoshi, MD, PhD (ORCID: https://orcid.org/0000-0001-6012-0037) 1. Department of Social Services and Healthcare Management, International University of Health and Welfare, Otawara, Tochigi 324-8501, Japan 2. Scientific Research Works Peer Support Group (SRWS-PSG), Osaka, Japan.

Naoto Kuroda, MD, PhD 1. Department of Pediatrics, Wayne State University, Detroit, MI, USA 2. Scientific Research Works Peer Support Group (SRWS-PSG), Osaka, Japan.

Ryo Momosaki MD, MPH, PhD (ORCID: https://orcid.org/0000-0003-3274-3952) 1. Department of Rehabilitation Medicine, Mie University Graduate School of Medicine, Edobashi 2-174, Tsu, Mie 514-8507 Japan 2. Scientific Research Works Peer Support Group (SRWS-PSG), Osaka, Japan.

Takeshi Nakata, MD, PhD 1. Department of Endocrinology, Metabolism, Rheumatology and Nephrology, Faculty of Medicine, Oita University, 1-1 idaigaoka Yufu-shi, Oita,879-5593, Japan 2. Scientific Research Works Peer Support Group (SRWS-PSG), Osaka, Japan.

Yuki Nakashima PT PhD (ORCID:https://orcid.org/0000-0001-7838-9472) 1 Division of Rehabilitation, Department of Clinical Practice and Support, Hiroshima University Hospital, Kasumi 1-2-3, Minami-ku, Hiroshima, 734-8551 Japan 2 Scientific Research Works Peer Support Group (SRWS-PSG), Osaka, Japan.

Yasuhiro Ogura, MD (ORCID: https://orcid.org/0000-0001-8310-2069) 1. Department of Radiation Oncology, Cancer Institute Hospital of the Japanese Foundation for Cancer Research, 3-8-31 Ariake, Koto-ku, 135–8550, Tokyo, Japan 2. Division of Cellular Senescence, Cancer Institute, Japanese Foundation for Cancer Research, 3-8-31 Ariake, Koto-ku, Tokyo 135-8550, Japan 3. Department of JFCR Cancer Biology, Graduate School of Medical and Dental Sciences, Institute of Science Tokyo, 4-5-19 Yushima, Bunkyo-ku, Tokyo 113-0034, Japan 4. Scientific Research Works Peer Support Group (SRWS-PSG), Osaka, Japan.

Masatsugu Okamura, PT, PhD (ORCID: https://orcid.org/0000-0001-9136-5037) 1. Berlin Institute of Health Center for Regenerative Therapies (BCRT), Charité – Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany 2. Department of Rehabilitation Medicine, School of Medicine, Yokohama City University, 3-9 Fukuura, Kanazawa-ku, Yokohama-city, Kanagawa, 236-0004 Japan 3. Scientific Research Works Peer Support Group (SRWS-PSG), Osaka, Japan.

Yusuke Saishoji, MD, MHA(ORCID:https://orcid.org/0000-0001-7163-8229) 1.Department of General Medicine, Hakujyuji Hospital, 4-3-1, Ishimaru, NIshi-ku, Fukuoka-shi, Fukuoka, 819-8511, Japan 2. Scientific Research Works Peer Support Group (SRWS-PSG), Osaka, Japan.
Akihiro Shiroshita, MD, MPH (ORCID: https://orcid.org/0000-0003-0262-459X) 1. Division of Epidemiology, Department of Medicine, Vanderbilt University School of Medicine 2. Scientific Research Works Peer Support Group (SRWS-PSG), Osaka, Japan
Hidehiro Someko, MD (ORCID: https://orcid.org/0000-0002-7195-2055) 1 Department of Healthcare Epidemiology, Kyoto University Graduate School of Medicine / Public Health, Yoshida Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan 2 Nagoya Tokushukai General Hospital, Department of Internal Medicine, Kozojichokita 2-52, Kasugai, Aichi, 487-0016, Japan 3 Scientific Research Works Peer Support Group (SRWS-PSG), Osaka, Japan.
Yukiyoshi Sumi, MD, PhD (ORCID: https://orcid.org/0000-0001-6775-0883) 1. Department of Psychiatry, Shiga University of Medical Science, Tsukinowa-cho, Seta, Otsu-city, Shiga, 520-2192, Japan 2. Scientific Research Works Peer Support Group (SRWS-PSG), Osaka, Japan.
Shunsuke Taito PT PhD (ORCID: https://orcid.org/0000-0003-1218-4225) 1 Division of Rehabilitation, Department of Clinical Practice and Support, Hiroshima University Hospital, Kasumi 1-2-3, Minami-ku, Hiroshima, 734-8551 Japan 2 Scientific Research Works Peer Support Group (SRWS-PSG), Osaka, Japan.
Takahiro Tsuge, PT, MPH (ORCID:https://orcid.org/0000-0003-0497-944X) 1. Department of Rehabilitation, Kurashiki Medical Center, 250 Bakuro, Kurashiki, Okayama 710-8522, Japan. 2. Department of Epidemiology, Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama University, 2-5-1 Shikata-cho, Okayama 700-8558, Japan. 3. Scientific Research Works Peer Support Group (SRWS-PSG), Koraibashi, Chuo-ku 1-7-7, Osaka, 541-0043, Japan. takahiro_tsuge@yahoo.co.jp
Yoshimitsu Wada, MD (ORCID: https://orcid.org/0000-0001-5986-8726) 1. School of Public Health, The University of Tokyo, 7-3-1 Hongo, Bunkyō, Tokyo, 113-8654 Japan. 2. Scientific Research Works Peer Support Group (SRWS-PSG), Osaka, Japan
Jun Watanabe MD, PhD (ORCID: https://orcid.org/0000-0003-4477-4238) 1. Department of Surgery, Division of Gastroenterological, General and Transplant Surgery, Jichi Medical University, 3311-1 Yakushiji Shimotsuke City, Tochigi, 329-0498 Japan 2. Center for Community Medicine, Jichi Medical University, 3311-1 Yakushiji Shimotsuke City, Tochigi, 329-0498 Japan 3. Scientific Research Works Peer Support Group (SRWS-PSG), Osaka, Japan.
Mari Yamamoto, MD (ORCID: https://orcid.org/0000-0003-3927-399X) 1. Department of Rheumatology and Nephrology, Chubu Rosai Hospital, 1-10-6 Komei, Minato-ku, Nagoya, Aichi 455-8530, Japan 2. Scientific Research Works Peer Support Group (SRWS-PSG), Osaka, Japan.
Norio Yamamoto MD,PhD (ORCID:https://orcid.org/0000-0002-7902-9994) 1.Department of Orthopedic Surgery, Minato Medical Coop-Kyoritsu General Hospital, Nagoya, Aichi, 456-8611, Japan 2. Scientific Research Works Peer Support Group (SRWS-PSG), Japan.
Yuki Yoshimatsu, MD, PhD (ORCID: https://orcid.org/0000-0003-0913-3507) 1. Department of Ageing and Health, Guy's and St Thomas' NHS Foundation Trust, Great Maze Pond, London SE1 9RT, United Kingdom Centre for Exercise, Activity and Rehabilitation, University of Greenwich, Avery Hill Road, Eltham, London, SE9 2UG, United Kingdom 2. Scientific Research Works Peer Support Group (SRWS-PSG), Osaka, Japan.
Shodai Yoshihiro, MPharm (ORCID: https://orcid.org/0000-0003-2418-2323) 1. Department of pharmaceutical services, Hiroshima University Hospital. Kasumi 1-2-3, Minami-ku, Hiroshima 734-8551, Japan. 2. Scientific Research Works Peer Support Group (SRWS-PSG), Osaka, Japan.

Conflicts of Interest

All authors received research mentoring fees from SRWS-PSG.

Multimedia Appendix 1
Scientific Research Works Peer Support Group (SRWS-PSG) curriculum.
[DOCX File, 8500 KB - mededu_v11i1e78862_app1.docx]

Multimedia Appendix 2
Published articles from Scientific Research Works Peer Support Group (SRWS-PSG) mentees.
[XLSX File, 19 KB - mededu_v11i1e78862_app2.xlsx]

References

1. Cooke J. A framework to evaluate research capacity building in health care. BMC Fam Pract 2005 Oct 27;6(1):44. [doi: 10.1186/1471-2296-6-44] [Medline: 16253133]
2. Matus J, Walker A, Mickan S. Research capacity building frameworks for allied health professionals - a systematic review. BMC Health Serv Res 2018 Sep 15;18(1):716. [doi: 10.1186/s12913-018-3518-7] [Medline: 30219065]
3. King O, West E, Lee S, et al. Research education and training for nurses and allied health professionals: a systematic scoping review. BMC Med Educ 2022 May 19;22(1):385. [doi: 10.1186/s12909-022-03406-7] [Medline: 35590359]

4. Ikarashi A. Japanese research is no longer world class - here's why. Nature New Biol 2023 Nov;623(7985):14-16. [doi: 10.1038/d41586-023-03290-1] [Medline: 37880529]
5. Fukuhara S, Kataoka Y, Aoki T, Green J, Shimizu S, Toyoda N. International collaboration and commercial involvement in randomized controlled trials from 10 leading countries, 1997 through 2019. Cureus 2024 May;16(5):e61205. [doi: 10.7759/cureus.61205] [Medline: 38939267]
6. Kataoka Y, Ikegaki S, Kato D, et al. Scholarly activity support systems in internal medicine residency programs: a national representative survey in Japan. Intern Med 2019 Jul 1;58(13):1859-1864. [doi: 10.2169/internalmedicine.2312-18] [Medline: 30918184]
7. Kinoshita S, Kishimoto T. Decline in Japan's research capabilities: challenges in the medical field. Lancet 2023 Oct 7;402(10409):1239-1240. [doi: 10.1016/S0140-6736(23)01465-4] [Medline: 37805213]
8. Sanftenberg L, Stofella J, Mayr K, et al. Expectations of general practitioners on a practice based research network in Germany- a qualitative study within the Bavarian Research Practice Network (BayFoNet). BMC Prim Care 2024 Jan 2;25(1):10. [doi: 10.1186/s12875-023-02239-7] [Medline: 38166677]
9. Goda K, Igaki T, Kuhn B, et al. Japan can be a science heavyweight once more - if it rethinks funding. Nature New Biol 2025 Feb;638(8050):318-320. [doi: 10.1038/d41586-025-00394-8] [Medline: 39934325]
10. Nishigori H, Harrison R, Busari J, Dornan T. Bushido and medical professionalism in Japan. Acad Med 2014 Apr;89(4):560-563. [doi: 10.1097/ACM.0000000000000176] [Medline: 24556758]
11. Tsujimoto H, Kataoka Y, Sato Y, et al. A model six-month workshop for developing systematic review protocols at teaching hospitals: action research and scholarly productivity. BMC Med Educ 2021 Feb 10;21(1):98. [doi: 10.1186/s12909-021-02538-6] [Medline: 33568114]
12. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. The Lancet 2007 Oct;370(9596):1453-1457. [doi: 10.1016/S0140-6736(07)61602-X]
13. Scientific Research Works Peer Support Group. Increase clinical research in japan! Write a systematic review paper and become a mentor!. URL: https://community.camp-fire.jp/projects/view/187310 [accessed 2025-05-10]
14. Harden RM. AMEE Guide No. 14: Outcome-based education: Part 1-An introduction to outcome-based education. Med Teach 1999 Jan;21(1):7-14. [doi: 10.1080/01421599979969]
15. Higgins JPT, Thomas J, Chandler J, et al. Cochrane handbook for systematic reviews of interventions. URL: https://training.cochrane.org/handbook [accessed 2024-07-03]
16. Cochrane Risk of Bias 2 Tool Group. RoB 2 tool. URL: https://sites.google.com/site/riskofbiastool/welcome/rob-2-0-tool [accessed 2025-05-13]
17. Slack. What is slack. URL: https://slack.com/what-is-slack [accessed 2025-09-16]
18. Belur J, Tompson L, Thornton A, Simon M. Interrater reliability in systematic review methodology: exploring variation in coder decision-making. Sociol Methods Res SAGE Publications 2021 May;50(2):837-865. [doi: 10.1177/0049124118799372]
19. King O, West E, Lee S, et al. Research education and training for nurses and allied health professionals: a systematic scoping review. BMC Med Educ 2022;22(1):385. [doi: 10.1186/s12909-022-03406-7]
20. Webster E, Thomas M, Ong N, Cutler L. Rural research capacity building program: capacity building outcomes. Aust J Prim Health 2011;17(1):107. [doi: 10.1071/PY10060]
21. Austin PC, Ibrahim M, Putter H. Accounting for competing risks in clinical research. JAMA 2024 Jun 25;331(24):2125-2126. [doi: 10.1001/jama.2024.4970] [Medline: 38809526]
22. Duncanson K, Webster EL, Schmidt DD. Impact of a remotely delivered, writing for publication program on publication outcomes of novice researchers. Rural Remote Health 2018 May;18(2):4468. [doi: 10.22605/RRH4468] [Medline: 29793344]
23. Boyer EL. Scholarship Reconsidered: Priorities of the Professoriate: Carnegie Foundation for the Advancement of Teaching; 1990. URL: https://eric.ed.gov/?id=ED326149 [accessed 2025-09-16]
24. Hidetoshi F, Rothman T. Sacred Mathematics: Japanese Temple Geometry: Princeton University Press; 2008. URL: https://press.princeton.edu/books/hardcover/9780691127453/sacred-mathematics?srsltid=AfmBOopAzDiaMFjejvuGb2Ru1lpRhEH2hOQs3vfUqkMNX0M9sX4Pgoi2 [accessed 2025-09-16]
25. Someko H, Yamamoto R, Ariie T, et al. Validity and reliability of the Japanese version of the ACE tool for assessing evidence-based medicine competencies in medical practitioners and students: an evaluation in an online setting. Intern Med 2025 Jul 15;64(14):2136-2142. [doi: 10.2169/internalmedicine.4724-24] [Medline: 39721681]
26. Nordling L. How research managers are using AI to get ahead. Nature New Biol 2023 Dec. [doi: 10.1038/d41586-023-04160-6]
27. Burtch G, Lee D, Chen Z. The consequences of generative AI for online knowledge communities. Sci Rep 2024 May 6;14(1):10413. [doi: 10.1038/s41598-024-61221-0] [Medline: 38710885]
28. Kataoka Y, So R, Banno M, et al. Development of meta-prompts for large language models to screen titles and abstracts for diagnostic test accuracy reviews. Health Informatics. Preprint posted online on 2023. [doi: 10.1101/2023.10.31.23297818]
29. Kataoka Y, Takayama T, Yoshimura K, et al. Automating the data extraction process for systematic reviews using GPT-4o and o3. Res Synth Methods. 2025 Sep 17 p. 1-21. [doi: 10.1017/rsm.2025.10030]

30. Sun L, Han Y, Zhao Z. SciEval: a multi-level large language model evaluation benchmark for scientific research. arXiv [csCL]. Preprint posted online on Aug 25, 2023. [doi: 10.48550/arXiv.2308.13149]

Abbreviations

CI: confidence interval
RCB: research capacity building
sdHR: subdistribution hazard ratio
SRWS-PSG: Scientific Research Works Peer Support Group
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by B Lesselroth; submitted 12.06.25; peer-reviewed by DL Dubois, ME Heidari, M Beltrão; revised version received 13.08.25; accepted 09.09.25; published 02.10.25.

Please cite as:
Kataoka Y, So R, Banno M, Tsujimoto Y, SRWS-PSG Mentors
Effectiveness of a Fully Online Scientific Research Works Peer Support Group Model for Research Capacity Building Through Conducting Systematic Reviews Among Health Care Professionals: Retrospective Cohort Studies
JMIR Med Educ 2025;11:e78862
URL: https://mededu.jmir.org/2025/1/e78862
doi:10.2196/78862

Using Electronic Health Data to Deliver an Adaptive Online Learning Solution to Emergency Trainees: Mixed Methods Pilot Study

Anna Janssen[1], PhD, MA, BA; Andrew Coggins[2], MBChB; James Tadros[2], MBBS; Deleana Quinn[1], BSc; Amith Shetty[3], MBBS; Tim Shaw[1], BSc (Hons)

[1]Faculty of Medicine and Health, The University of Sydney, Charles Perkins Centre D17, Sydney, Australia

[2]Western Sydney Local Health District, Sydney, Australia

[3]NSW Ministry of Health, Sydney, Australia

Corresponding Author:
Anna Janssen, PhD, MA, BA
Faculty of Medicine and Health, The University of Sydney, Charles Perkins Centre D17, Sydney, Australia

Abstract

Background: Electronic medical records (EMRs) are a potentially rich source of information on an individual's health care providers' clinical activities. These data provide an opportunity to tailor web-based learning for health care providers to align closely with their practice. There is increasing interest in the use of EMR data to understand performance and support continuous and targeted education for health care providers.

Objective: This study aims to understand the feasibility and acceptability of harnessing EMR data to adaptively deliver a web-based learning program to early-career physicians.

Methods: The intervention consisted of a microlearning program where content was adaptively delivered using an algorithm input with EMR data. The microlearning program content consisted of a library of questions covering topics related to best practice management of common emergency department presentations. Study participants were early-career physicians undergoing training in emergency care. The study design involved 3 design cycles, which iteratively changed aspects of the adaptive algorithm based on an end-of-cycle evaluation to optimize the intervention. At the end of each cycle, an online survey and analysis of learning platform metrics were used to evaluate the feasibility and acceptability of the program. Within each cycle, participants were recruited and enrolled in the adaptive program for 6 weeks, with new cohorts of participants in each cycle.

Results: Across each cycle, all 75 participants triggered at least 1 question from their EMR data, with the majority triggering 1 question per week. The majority of participants in the study indicated that the online program was engaging and the content felt aligned with clinical practice.

Conclusions: The use of EMR data to deliver an adaptive online learning program for emergency trainees is both feasible and acceptable. However, further research is required on the optimal design of such adaptive solutions to ensure training is closely aligned with clinical practice.

(JMIR Med Educ 2025;11:e65287) doi:10.2196/65287

KEYWORDS

electronic health records; electronic medical records; digital health; emergency care; health professional education; health informatics; practice analytics; EMR; health data; online learning; pilot study; trainee; clinical activities; feasibility; acceptability; online microlearning; adaptive algorithm; adaptive solutions; adaptive program

Introduction

Background

Medical practitioners engage in a variety of formal and informal training activities to stay up to date on best practices for delivering patient care. Activities undertaken by medical practitioners in the context of professional training can take many forms, including mentorship and consultation with peers, attendance at local seminars and international conferences, journal clubs, and other activities [1]. It has also been observed that medical practitioners dedicate considerable time to undertaking education and training activities. An observational study of physicians on hospital wards found that around 7% of observed tasks involved engaging in education or supervision [2]. Another study found that medical practitioners early in their careers spend close to 2 hours every day engaging in training activities [3].

Digital technologies are increasingly being used to share knowledge and evidence between medical practitioners in the workplace [4]. Online learning as a mechanism for delivering training to the medical profession is also increasingly widespread [5]. Benefits of online learning include making training available when and where individuals would like to access it, potentially enabling more innovative approaches to teaching and making information more easily accessible [6]. In the context of training medical practitioners, flexibility or adaptivity has been emphasized as a major advantage of online education [7]. However, online learning can also have disadvantages, and there is disagreement in the literature as to whether delivering training to medical practitioners online impacts learning outcomes [5]. Another potential issue is that online learning can require more self-discipline and greater time management skills to complete than face-to-face training [8]. This phenomenon may be why online learning has been noted to have high completion rates, with 60% being a high completion rate for many online courses [9].

One strategy for strengthening online learning for medical practitioners could be to align it more closely with clinical practice. It has been noted in the literature that current approaches to medical education disconnect learning activities from practice and care delivery [10]. To date, there is limited research into how online learning could be personalized to align it more strongly with clinical practice. A recent scoping review of secondary use of data from all health information technologies (HITs) identified education as one of 4 key domains for this purpose, but data were used in this way in only 1% of the identified studies [11]. Another scoping review exploring the design of dashboards presenting data from HITs to support reflective practice found that these types of visualization tools were often designed to present data to individual medical practitioners to allow them to reflect on their practice, but such platforms rarely incorporated scaffolds that would support learning or other improvement activities [12].

The field of learning analytics has also explored the significant potential of data collected about learners in online programs for enhancing learning. Learning analytics broadly describes the collection, analysis, and reporting of data about learners and their contexts for the purpose of understanding and optimizing learning and the environments in which it occurs [13]. A recent review of the literature identified personalization as a major focus of learning analytics, particularly in the context of online learning environments where large amounts of data about the learning process are routinely collected [14]. Research into learning analytics in medical education has identified several barriers in this space, including implementation challenges, such as how to collect data to use in learning analytics; data management issues, such as governance and access to appropriate data; and outcomes challenges, such as how data can be used to assess learners, programs, and systems [15]. Furthermore, it has been noted in the wider literature on medical education that when training is delivered digitally, analytics collected about learning progress are a potentially rich source of information for informing evidence-based instruction [7].

In the context of online learning for medical practitioner education, routinely collected electronic health data such as that in electronic medical records (EMRs) may have value for aligning learning with clinical practice. While there is a wide range of HITs used in clinical practice, EMRs are a core technology that have been widely adopted in many health care organizations [16,17]. Data from EMRs have been identified as being useful for understanding the practice patterns of medical practitioners [18] and for supporting formative assessment of early-career physicians in the workplace [19]. There is also literature suggesting that health care providers across a range of disciplines are interested in EMR data being harnessed to inform learning and training activities [20]. To date, there has been limited research undertaken into the use of EMR data to support workplace learning for medical practitioners.

There is a small but growing body of research exploring the intersection between EMR data and health professional learning. A recent study investigated the use of EMR data to support medical practitioners undertaking a learning needs assessment, which was subsequently used as the basis for delivering a customized continuing medical education program for the next year [18]. In this study, medical practitioners were provided an EMR report showing their data compared to other participating practitioners, which they then used to identify personalized learning goals. Another study of this type used data from electronic prescribing systems has also been used to personalize the delivery of letters to prescribers advising them of their compliance with antimicrobial prescribing policy [21]. Researchers have also investigated the use of EMR reports to populate dashboards to educate medical practitioners about their performance. In one such study, a dashboard visualizing EHR reports was made available to medical practitioners for 6 months with the goal of reducing overprescribing of antibiotics [22], though this study was not able to demonstrate a change in practice.

Although interest is increasing in the use of EMR to support training for medical practitioners aligned with clinical practice, the data have some limitations for this application. A considerable limitation is that there can be gaps in the data and that the data are not captured for the primary purpose of education [23], and there are limitations on how these data can be used to understand performance [19]. As has already been highlighted, a further limitation is the relatively few examples in the literature demonstrating how to use EMR data to design this type of personalized education [18,22]. Furthermore, current popular approaches for presenting EMR data to medical practitioners to support reflection lack scaffolds such as self-reflection tools, goal-setting options, or learning plans that could be used to support professional learning [12].

Problem Statement

Despite the plethora of electronic data being collected by the health system and growing interest in their use to support health professional education, there is a dearth of knowledge on how to design data-driven learning interventions for professional learning. This study aimed to address that gap by investigating the feasibility and acceptability of using EMR data to adaptively deliver an online learning program. In the context of this study, feasibility included answering the following research questions:

- Could data be routinely extracted from an EMR to support data-driven learning?
- Could EMR data be linked to microlearning questions to trigger adaptive delivery to different learners?
- How engaging did learners find the data-driven learning program?

A secondary aim was to understand how to design such a program so that learners felt the content was well aligned with their clinical practice. The study investigated EMR data specifically due to the widespread adoption of this HIT in health care organizations [17] and the recognized use of EMR data for understanding clinical practice patterns [18].

Methods

Study Design

The study used 3 design cycles to iterate on the design of the adaptive online learning program. The approach was based on a design-based research approach that has been used in research education contexts to provide a systematic but flexible approach to improv practice through iterative cycles of design, development, and implementation [24]. At the end of each cycle, a mixed methodology was used to evaluate the feasibility and acceptability of the program and refine it in response to learner feedback. Within each cycle, learners were enrolled in the adaptive program for 6 weeks, with new cohorts of learners in each cycle.

Participants and Study Setting

The study was undertaken within the emergency department at 2 public metropolitan hospitals in Sydney, Australia. The 3 design cycles were run over a 12-month period between December 2018 and December 2019.

Potential participants were early-career physicians who were undergoing postgraduate training. Early-career physicians in Australia are those who have graduated from medical school and have provisional registration to practice but have not undergone enough training to be qualified to be independent practitioners. The first step of the training pathway, typically completed within 2 years of postgraduation from a medical degree, is working as an intern for 12 months. As an intern, early-career physicians undertake 10-week terms to expose them to different clinical environments. A rotation in emergency medical care is always included in this training. After working as an intern, early-career physicians may become residents for an additional year, potentially followed by being a registrar to undertake training for a specialized medical pathway [25].

Inclusion criteria for the study required participants to be interns undergoing their emergency medical care term at the study sites during the recruitment period. Participants were excluded from the study if they were not interns at the study sites or were not undertaking an emergency medical term while the study was recruiting.

All interns at the participating study sites were invited to participate in the study. Participants were recruited via flyers and an online notice with a link to sign up for the study. In design cycle 1, 22 physicians consented to participate; in design cycle 2, 36 physicians consented to participate; and in design cycle 3, 18 physicians consented to participate in the study. No physicians formally withdrew from the study.

Study Procedures

Intervention Design

The online program was delivered using the Qstream microlearning platform [26]. This is an off-the-shelf platform that sends multiple-choice questions to learners via email or a smartphone app. To reinforce a single take-home message, multiple-choice questions provided learners with detailed feedback on why the response they had entered was correct or incorrect. By default, the platform delivered learners a small bundle of 2 to 3 questions at a time, so that it only took a few minutes to respond to the bundle. The platform would then repeat questions a set number of times depending on whether the learner answered incorrectly or correctly. In the intervention, the microlearning platform was modified to alter the way questions were selected for individual learners and the frequency at which they were delivered. The modified functionality was an adaptive algorithm using data extracted from the EMR. Data were extracted from the EMR via a report that was run 2 times each week the adaptive program was running. The adaptive algorithm ingested EMR data related to the cohorts of patients an individual participant had encountered in the previous few days. The algorithm would subsequently tailor the delivery of microlearning questions for each learner based on the patients they had encountered in the reporting period. For example, if a participant had seen a patient with a heart condition, they would be sent a question on best practice for managing this patient group. If a peer of the participant had not seen a patient with a heart condition but had seen a patient presenting with shortness of breath, they would receive a question on best practice for managing patients with this condition. Multimedia Appendix 1 shows an example of the Qstream app interface and one of the cases for this intervention.

Questions were developed by a team of domain experts in emergency care and educational designers. The domain experts developed a curriculum that covered common clinical scenarios encountered in the emergency department that were considered opportunities to improve knowledge and behavior related to the best practice of potential participants. The curriculum and the questions that were built within it were developed based on the domain experts' understanding of common knowledge gaps of the potential participants. The educational designers undertook a structured review of EMR reports to understand the data that were available to trigger the delivery of curriculum content in a manner that would adapt to an individual participant's clinical encounters during the intervention. The curriculum and question set were designed to be relevant to improve clinical practice of participants as well as feasible to personalize through accessing EMR data and were developed using an evidence-based approach to developing microlearning questions [27]. The final curriculum consisted of cases on the management of care for patients presenting with the following symptoms: (1) chest pain, (2) abdominal pain (triggered by clustering 2 presenting problem codes), (3) breathlessness, (4) syncope (triggered by clustering 2 presenting problem codes), (5) headache, and (6) fever. A

seventh category was added to the curriculum for the third cohort of learners: mental health (triggered by clustering 10 presenting problem codes). A total of 45 questions were developed for the original 6 program categories. Five additional questions were developed for the mental health category made available for the third cohort of learners. Questions were not pilot-tested with representatives from the cohort they were developed for prior to use in the intervention. Table 1 presents an overview of the microlearning program curriculum and questions.

Table . Overview of the microlearning program, including the topics, the question names, and take-home messages, and the cohort that had access to the library.

Topic and question ID	Take-home message	Cohort
Abdominal pain		Cohorts 1-3
1	In females of childbearing age, ectopic pregnancy should always be the first consideration and a diagnosis of exclusion regardless of the menstrual and conception history reported.	
2	General supportive measures remain consistent in all resuscitation scenarios, especially in the bleeding patient. Warming patients may help prevent worsening coagulopathy and further bleeding.	
3	Abdominal pain in older people can be a challenging presentation. In contrast to younger patients, a broad range of life-threatening differentials should be considered.	
4	In high-stress settings, cognitive readiness may be enhanced using simple techniques that minimize autonomic hyperarousal.	
5	It is important to explore and investigate the causes of bloody or prolonged diarrhea which in this case could be due colitis. Night time defecation is suggestive of an underlying pathological cause.	
6	Pancreatitis is inflammation of the pancreas and involves activation of proteolytic enzymes that may progress to hemorrhagic necrosis of the pancreatic parenchyma. Different geographic locations may report different incidences of aetiologies but universally ethanol and gallstones are common causes.	
7	Intrarenal calculi are not, on their own, an indication for referral to urology.	
8	Examination of the testes is an essential part of the abdominal examination in young males, even if they do not report testicular pain. Torsion is the diagnosis of exclusion.	
Shortness of breath		Cohorts 1-3
1	A structured approach to x-ray interpretation is required when reviewing in the ED[a] because we do not normally have the benefit of a *formal* report from a radiologist. Subtle pneumothorax is an easily missed diagnosis.	

Topic and question ID	Take-home message	Cohort
2	BiPAP[b] NIV[c] therapy is indicated in a patient with an acute exacerbation of COPD[d] with a persistent respiratory acidosis despite appropriate initial treatment. A low level of consciousness is not an absolute contraindication to the use of NIV.	
3	The treatment of anaphylaxis requires a dose of intramuscular adrenaline. The EpiPen dose is 0.3 mg. Typically, 0.3 mg to 0.5 mg (1:1000=0.3 - 0.5 mLs) is the dose for an adult patient with anaphylaxis.	
4	In a patient with hypoxia and known COPD at risk of CO_2 retention, apply titrated oxygen first until target saturations are achieved. Positioning the patient is also an important management step.	
5	Patients with underlying cancer are at higher risk of life-threatening sepsis (associated with chemotherapy), pericardial effusions, and PE[e].	
6	A plain CXR[f] is practical, easily accessible, and universally indicated for a patient with acute shortness of breath in the ED.	
7	It is important to remember that there are noncardiorespiratory differentials to shortness of breath. Patients with a metabolic acidosis will often present with tachypnoea (Kussmaul respiration).	
Chest pain		Cohorts 1-3
1	Early diagnosis of AD[g] requires a high index of suspicion. Blood pressure treatment should target control of both heart rate and the pressure itself.	
2	Chest pain “PLUS” another symptom should trigger the thought, “Could this be a diagnosis of Aortic Dissection?” AD may mimic acute MI[h] (including ECG[i] findings).	
3	No one factor can reliably rule out ACS[j] in the ED setting. Pain radiating to the right shoulder or arm is considered more specific for ACS than pain radiating to the left arm.	
4	An ECG should be performed in anyone presenting with chest pain. While well known, S1Q3T3 is uncommon in the setting of PE. Sinus tachycardia and anterior *T* wave inversions are more common.	

Topic and question ID	Take-home message	Cohort
5	It is important to consider PE as a differential in patients with cancer. There are scoring tools available to assist you (Wells and PERC being the most commonly used in the ED).	
6	A CXR should be part of your workup for chest pain and shortness of breath. Special care should be paid to looking for pneumothorax and consolidation.	
7	MI, PE, and AD are 3 critical conditions to consider with any patient presenting with chest pain.	
8	AD should be considered in any patient with chest pain with risk factors (eg, hypertension), or chest pain with a concurrent symptom such as neurological deficits.	
Fever		Cohorts 1-3
1	The presence of severe pain (pain out of proportion to the clinical findings) in an at-risk patient should raise concerns about the diagnosis of necrotising fasciitis.	
2	It is important to note the patient's allergy to penicillin. Tazocin (piperacillin or tazobactam) is a penicillin, therefore is relatively contraindicated for febrile neutropenia, though it remains first line in guidelines.	
3	Early recognition of sepsis is paramount (new concept of qSOFA[k] rather than use of nonspecific or sensitive "SIRS[l]" criteria).	
4	Patient is MRSA[m] colonized so we should consider vancomycin as additional cover (with expert advice). Antibiotics choice should always be judiciously guided by guidelines and ID team support	
5	You must cleanse your hands before and after any patient interaction. Alcohol is generally preferable over soap and water unless the hands are soiled or if *Clostridium difficile* infection is an issue (ie, spores are not necessarily killed by alcohol rub).	
Headache		Cohorts 1-3
1	In the event of a late (>6 h) presentation with a typical SAH[n] story, further investigations are required to excluded the diagnosis. All headache cases should be discussed with a senior physician prior to discharge.	

Topic and question ID	Take-home message	Cohort
2	If a patient presents with an acute severe headache plus fever and if the CT[o] is normal, consider a lumbar puncture to exclude meningitis. Consider giving early intravenous antibiotics and antivirals.	
3	If a patient with no history of headaches presents with a sudden onset headache and classic features of SAH, and the CT scan comes back *'normal'* consider further testing. MRI[p] or MRA[q] and specialist review may be warranted under these circumstances.	
4	When a patient presents with a headache, be sure to prescribe analgesia. Patients with papilledema require ED evaluation with referral to both ophthalmology and neurology following neuroimaging.	
5	Over analgesia (especially with paracetamol, codeine, and aspirin) is associated with a paradoxical increase in headache in some patients. While analgesia in the acute setting is a mainstay of our ED management, we should discuss refractory cases with a neurologist and arrange follow-up.	
6	Typically, the CSF[r] glucose concentration is two-thirds that of the serum glucose concentration.	
7	Temporal arteritis is a sight-threatening condition that is also known as GCA[s]. Jaw claudication is a classic symptom.	
8	Low CSF headache is a distinct and familiar syndrome that is seen most frequently following lumbar puncture. Typically, the headache is orthostatic and significantly improves by lying flat.	
Syncope		Cohorts 1-3
1	Patients presenting with syncope should be risk-stratified based on their overall history, examination, and a period of observation. Risk scores (eg, San Francisco or the Rose Criteria) may have some use as an adjuvant to your thorough clinical assessment.	
2	An ECG is a very important test in a patient with syncope. It is a mandatory test in the ED for a patient with syncope.	

Topic and question ID	Take-home message	Cohort
3	Dosing of NOACs[t] can be confusing and factors such as age, weight, renal function, and indication may affect dosage. Consultation with hematology and a pharmacist regarding dosing is pertinent.	
4	Vertigo is a common ED presentation. The history given is concerning for a "*central*" cause of vertigo. Age is an independent risk factor for stroke.	
5	Early defibrillation and commencement of high-quality BLS[u] are critical to outcomes in cardiac arrest. While not contraindicated, checking for a pulse is no longer specifically recommended in assessing for signs of life or confirming arrest.	
6	Metoclopramide and prochlorperazine both worsen Parkinson disease symptoms, and therefore are contraindicated in those with Parkinson disease. Domperidone is more appropriate for these patients.	
7	If in doubt with a case of "wide complex tachycardia," call for help and assume the diagnosis is VT[v] until proven otherwise.	
8	If in doubt with a case of "wide complex tachycardia," call for help and assume the diagnosis is VT until proven otherwise.	
Mental health		Cohort 3
1	This is a possible first presentation of schizophrenia with paranoid delusions. The importance of a mental state examination and collateral history is highlighted in this case.	
2	Although there are some variations in practice, current guidelines for parental sedation have droperidol as the most appropriate first-line agent. This should be done with care not to cause further harm to the patient, both through physical and medical means. Droperidol is a commonly used agent in the ED at Westmead Hospital.	
3	Alcohol withdrawal can often present in unusual circumstances such as a change of environment limiting someone's access to alcohol. The presence of visual hallucinations, confusion, and physical complaints make a purely psychiatric diagnosis less likely.	

Topic and question ID	Take-home message	Cohort
4	A person may be treated without consent under two conditions: (1) life-threatening emergencies when a patient lacks "capacity"; and (2) under the Mental Health Act (but only for psychiatric treatments).	
5	There is no such thing as "low risk" in assessing the patient with suicidal tendencies. All patients with mental health issues in the ED have a significantly higher long-term risk of suicide than the general population.	

[a]ED: emergency department.
[b]BiPAP: bilevel positive airway pressure.
[c]NIV: noninvasive ventilation.
[d]COPD: chronic obstructive pulmonary disease.
[ee]PE: pulmonary embolism.
[f]CXR: chest x-ray.
[g]AD: aortic dissection.
[h]MI: myocardial infarction.
[i]ECG: electrocardiogram.
[j]ACS: acute coronary syndrome.
[k]qSOFA: quick sequential organ failure assessment.
[l]SIRS: systemic inflammatory response syndrome.
[m]MRSA: methicillin-resistant *Staphylococcus aureus*.
[n]SAH: subarachnoid hemorrhage.
[o]CT: computed tomography.
[p]MRI: magnetic resonance imaging.
[q]MRA: magnetic resonance angiography.
[r]CSF: cerebrospinal fluid.
[s]GCA: giant cell arteritis.
[t]NOAC: nonvitamin K antagonist oral anticoagulant.
[u]BLS: basic life support
[v]VT: ventricular tachycardia.

Intervention Delivery

During the 6-week intervention period for the online program, a report was manually extracted 2 times each week (Tuesdays and Fridays) from the EMR and fully deidentified patient information. The frequency of extraction was chosen for feasibility reasons as the extraction process was manual and we could not continually extract the data. Only structured data were included in the report as EMR data were only being used to trigger the delivery of questions, not to tailor the content in individual cases received by learners. The deidentified report was then provided to the researchers in .csv format. The EMR report was used to identify if participants had interacted with the relevant patients to populate the adaptive algorithm in the microlearning program. If the EMR data indicated a participant had encountered patient presentations related to the microlearning program, they would "trigger" a question related to managing that type of patient presentation. "Triggering" a question meant that a participant would be allocated a relevant question in the microlearning platform, and it would subsequently be pushed to them via email or the smartphone app to complete. If a participant did not see any clinical presentations that could trigger a question in the reporting period, they did not receive any questions. If participants had seen clinical presentations that could trigger a question, they were enrolled in the relevant question. Questions were pushed to each participant on the same day they were generated, so they received the case within 3 days of seeing the patient that triggered it.

The number of questions participants could receive and the topics in the program were iterated on each cycle in response to analysis of data evaluating that cycle. Participants could be re-enrolled in a question at a later point in the intervention period if they had not previously attempted to respond to it, and the EMR data indicated they had triggered it again. Modifications were made to the delivery of the algorithm for the cases in each cycle. Table 2 summarizes the iterations across all the design cycles.

Table . Summary of iterations made across the design cycles and explanation of why those modifications were made.

Summary of quantitative data	Design cycle 1	Design cycle 2	Design cycle 3
Iterations undertaken	• Cases delivered across 6 presenting problem domains. • Participants were assigned up to 3 questions each for each EMR[a] reporting period. • If only 1 patient presentation was encountered by a participant during the EMR reporting period, 3 questions related to that patient presentation would be delivered. • If no patient presentations that could trigger a question were encountered during the EMR reporting period, then no questions would be assigned.	• Cases delivered across 6 presenting problem domains. • Participants were assigned up to 3 questions each for each EMR reporting period. • If only 1 patient presentation was encountered by a participant during the EMR reporting period, only 1 question related to that patient presentation would be delivered.	• Cases delivered across 7 presenting problem domains. Mental health was added as a new presenting problem.
Justification for iterations	• N/A[b]	• Reduce the number of questions participants had to respond to in a single bundle to increase course engagement	• Increase the perceived alignment of the course content with clinical practice by using a less common, but still frequent patient presentation

[a]EMR: electronic medical record.

[b]N/A: not applicable.

During design cycle 1, each participant was assigned 3 questions chosen using the personalization algorithm based on the cases they had attended during the EMR reporting period. If only 1 clinical presentation that could trigger a question was seen by a participant, they would receive 3 different questions on that topic from the question library. For example, in a single EMR report, if a participant had seen only patients with chest pain, they would receive 3 chest pain questions; if that participant had seen 1 patient with chest pain, 1 patient with abdominal pain, and 1 patient with syncope, they would receive 1 question on each topic; if a participant had seen nonrelevant patients, they would receive zero questions.

During design cycle 2, participants who saw relevant clinical presentations in each reporting period were assigned up to 3 questions chosen using the personalization algorithm. They only received 1 question per topic triggered by the EMR data, rather than multiple questions as was the case in design cycle 1. This change was made to reduce the number of questions the participants had to respond to in a single bundle, as analysis of temporal data collected in design cycle 1 indicated some participants were completing all cases in a single bundle at the end of each week. For example, in a single EMR report, if a participant had seen only a patient with chest pain, they would receive 1 question on chest pain; if that participant had seen 1 patient with chest pain and 1 patient with syncope, they would receive 1 question on chest pain and 1 question on syncope; if that participant had seen 1 patient with chest pain, 1 patient with abdominal pain, and 1 patient with syncope, they would receive 1 question on each topic; if a participant had seen no relevant patients, they would receive zero questions.

During the final design cycle, the delivery of questions was the same as in design cycle 2; however, questions related to the topic of mental health were added to the program. Mental health cases were added to see if a less common patient presentation would increase the sense of alignment between the intervention and clinical practice.

Regardless of the design cycle, at the end of each week during an intervention period, all participants in that cycle were unenrolled from any questions they had not responded to. This was done to ensure a large backlog of questions did not accumulate for participants to answer, also to ensure participants were not receiving questions that were not aligned with their recent clinical practice. Participants could also be enrolled in the same question multiple times during the design cycle if they had not attempted it on a previous enrollment and had subsequently been unenrolled. There were two situations in which participants would not be enrolled in a new question even if they had triggered it in the EMR report. The first reason a participant was not enrolled in new questions was if they already had questions they had been enrolled in and had not finished answering. The second reason was because a participant had answered all the questions on that topic correctly and had not triggered any other topics based on the EMR report.

Intervention Evaluation

To evaluate the program, a number of data points were collected. EMR reports were used to trigger the delivery of questions during the intervention, as well as to determine the time that had elapsed between seeing a clinical scenario and completing a question. Coupled with metrics from the EMR, metrics captured by the online learning platform were extracted to understand participant engagement with the intervention, participant progress during the program, the number of questions participants were enrolled in that they completed, the accuracy of their responses, and the time that elapsed between being

allocated a question and answering it. Finally, a bespoke online survey created by the researchers was disseminated at the end of each design cycle to capture participant feedback on the program. Survey responses were anonymous. The survey consisted of a combination of Likert responses and free-text comments. Structured questions explored how participants accessed the program, the value of clinical data for learning, the relevance of the program content, the course format, and overall experience with the course. Free-text responses allowed participants to provide general comments and feedback on changes to the course. The survey questions were not pilot-tested prior to use in the study.

Data Analysis

Metrics captured by the online learning platform on participant progress during the program were descriptively analyzed to understand how adaptations to the algorithm influenced participant behaviors. These data, combined with EMR reports, were also analyzed to understand how temporal factors related to the alignment of questions with the participant's clinical practice influenced engagement with content. Reports extracted from the EMR were analyzed to understand the link between participant test ordering and allocation of questions in the online learning platform.

Structured data from the survey were descriptively analyzed. Unstructured data from free-text survey comments were analyzed to understand participant engagement with the content of the intervention, the online learning platform, and the adaptive component. The content analysis was undertaken by one researcher (AJ) who read all the comments to get a sense of the data. Additional read-throughs of the data were undertaken to code the data and subsequently classify it into categories based on the similarity of themes discussed.

Ethical Considerations

The study received ethical approval from the Western Sydney Local Health District Human Research Ethics Committee (protocol: 2019/ETH02509). All participants provided written informed consent before agreeing to participate in the study. The researchers complied with informed consent guidelines when necessary and have adhered to local, national, regional, and international law and regulations regarding the protection of personal information, privacy, and human rights. Study data was de-identified prior to analysis. Participants did not receive any compensation for participating in the study.

Results

Overview

The following sections present a description of data from each design cycle of the intervention. Table 3 shows an overview of key quantitative data compared across each design cycle, and Table 4 shows an overview of the responses to the online survey across all 3 design cycles. There was a library of 37 questions across 5 topics during design cycles 1 and 2. There was a library of 43 questions covering 6 topics in design cycle 3. If a participant encountered a patient related to one of the topics in the biweekly EMR report (2 times a week) during a design cycle, it would trigger the delivery of a question. Figure 1 shows the question topics in the program and the percentage of participants who correctly, incorrectly, or did not respond to each topic for design cycles 1-3. Figure 2 visualizes the top 20 presenting problems encountered by participants during the intervention period for design cycles 1-3.

Table . Comparison of key quantitative data points across design cycles 1-3 of the intervention.

Summary of quantitative data	Design cycle 1 (n=21)	Design cycle 2 (n=36)	Design cycle 3 (n=18)
Unique questions in cohort, mean (SD; range)	9.72 (4.62; 3-20)	9.69 (5.79; 1-22)	6.27 (4.93; 1-17)
Range of unique questions in cohort, mean (SD; range)	3 - 20 (4.62)	1 - 22 (5.79)	1 - 17 (4.93)
Participants attempting more than 50% of the questions, n (%)	18 (86)	23 (66)	7 (37)
Total patient presentations (including duplicates) for participant cohort that could have triggered a question, n	1931	4918	2961
Three most common presenting problems encountered by all participants in a cohort that could trigger a case, n	• Abdominal pain: 649 presentations • Chest pain: 484 presentations • Shortness of breath: 215 presentations	• Chest pain: 600 presentations, • Abdominal pain: 573 presentations • Shortness of breath: 335 presentations	• Abdominal pain: 357 presentations • Mental health: 159 presentations • Shortness of breath: 156 presentations
Survey responses, n	9	15	1

Table . Comparison of survey responses across design cycles 1-3 of the intervention.

Summary of survey responses	Design cycle 1 (n=9), n (%)	Design cycle 2 (n=15), n (%)	Design cycle 3 (n=1), n (%)
Participants who agreed or strongly agreed with the statement that the duration of the course suited their needs	6 (67)	12 (80)	1 (100)
Participants who agreed with the statement they would have like to have received more cases each week.	4 (44)	11 (73)	0 (0)
Participants who agreed or strongly agreed with the statement they found the online program engaging	6 (67)	12 (80)	1 (100)
Participants who agreed or strongly agreed they would recommend the course to a colleague	7 (78)	12 (80)	0 (0)
Participants who agreed or strongly agreed with the statement the program content was realistic	8 (89)	8 (89)	0 (0)
Participants who agreed or strongly agreed with the statement that the content felt aligned with clinical practice.	8 (89)	14 (93)	0 (0)
Participants who agreed or strongly agreed with the statement that the questions in the program felt aligned or linked to their clinical practice.	8 (89)	14 (93)	1 (100)
Participants who agreed or strongly agreed with the statement the program felt engaging because it used clinical data relevant to their practice.	6 (67)	10 (67)	0 (0)
Participants who agreed or strongly agreed with the statement that they would like to see clinical data used to deliver personalized professional development in future.	5 (56)	15 (100)	1 (100)

Figure 1. Questions in the program, and the percentage of participants that correctly, incorrectly, or did not respond to for design cycles 1-3.

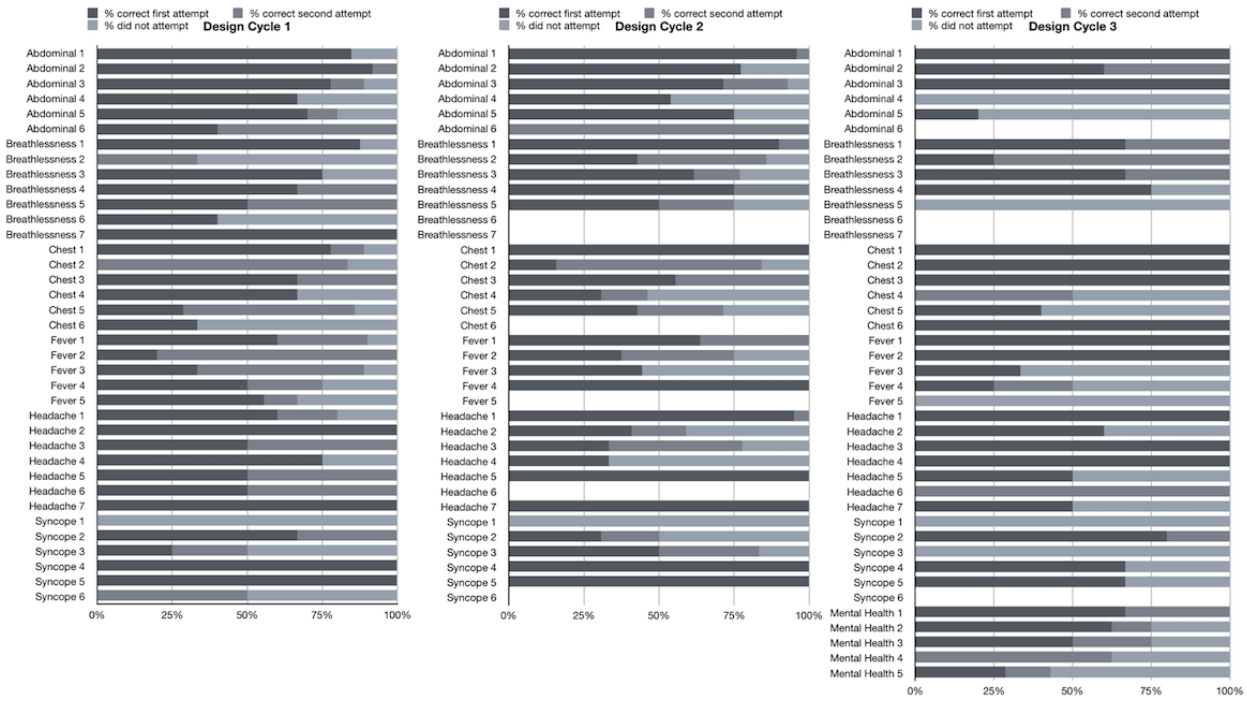


Figure 2. Visualization of the top 20 presenting problems encountered by participants during the intervention period for design cycles 1-3. (A-C) The presentations in design cycles 1-3. The presenting problems that could trigger a question in the online program are highlighted in red.

Design Cycle 1

There were 21 participants in cohort 1. Over the 6-week intervention period, the average number of unique questions participants were enrolled in was 9.72 (SD 4.62; range 3-20). A total of 9 participants in design cycle 1 responded to the postprogram survey (Table 4). The majority of respondents, 6 (67%), indicated that they completed the cases during personal time, after work, or on weekends; 2 (22%) respondents indicated they completed the cases when they had time during the workday; and the remaining 1 (11%) participant completed the program while commuting.

Design Cycle 2

There were 36 participants in cohort 2. Over the 6-week intervention period, the average number of unique questions participants were enrolled in was 9.69 (SD 5.79; range 1-22). A total of 15 participants in cohort 2 responded to the postprogram survey, but only 10 responded to the whole survey (Table 4). Of the 14 respondents who provided feedback on when they completed the program, the majority of respondents, 9 (64%), indicated they completed the cases during personal time, after work, or on weekends. In addition, 3 (21%) respondents indicated they responded to a question as soon as they received an alert, 11 (7%) respondents indicated they completed the cases when they had time during the workday, and the remaining 11 (7%) participants completed the program while commuting.

Design Cycle 3

There were 18 participants in cohort 3. Over the intervention period, the average number of unique questions participants were enrolled in was 6.27 (SD 4.93; range 1-17). One participant responded to the postprogram survey for cohort 3 (Table 4). The respondent indicated they had responded to the questions during personal time, after work, or on weekends.

Problems Encountered by Participants That Could Trigger a Question

Over the intervention period, EMR data indicated participants had a total of 2961 patient presentations that could have triggered a question. This number included duplicate instances of a participant seeing the patient presentation categories that could trigger a case. Of these, the most common presenting problem encountered by participants was abdominal pain (357 presentations), followed by mental health (159 presentations), and shortness of breath (156 presentations). Of the top 20 presenting problems participants encountered during the intervention period, all 7 presenting problem clusters that could have triggered a question in the online program were present (Figure 2).

Discussion

Principal Findings

Findings from this study indicate that EMR data can be used to personalize an online program for early-career physicians on management of common emergency presentations and link clinical practice directly with education in a timely manner. The researchers were able to extract EMR data at regular intervals during the intervention and use it to populate the adaptive algorithm that delivered personalized questions for individual participants. Regarding acceptability, findings suggested most participants found the program engaging and felt there was a level of alignment with their clinical practice. In addition, study findings indicated that early-career physicians undertaking intern training in emergency departments would like to see online programs personalized using electronic data in the future. This finding aligns with the literature, which has shown that health care providers across a range of professions are interested in seeing EMR data used in education and training [20].

Insights on EMR Data for Enabling Adaptive Learning

The potential value of using analytics related to interactions with training programs to personalize learning has been well researched in the context of learning analytics [14]. In the context of medical education, harnessing practice data about learners in digital and online education has been noted as a means to improve evidence-based instruction [7], but there are still significant gaps in our understanding of how to do this. This study presents one of the first studies demonstrating how routinely collected EMR data can be used to personalize learning. This study demonstrated that in the context of medical practitioner education, there are unique opportunities to strengthen training using analytics of data from clinical sources such as EMRs, not just learner data generated by undertaking educational interventions [15].

In the context of this study, the course curriculum was developed by first working with domain experts to identify common clinical presentations that were challenging for participants, and then refining it based on what was feasible to extract from the EMR during the intervention. Retrospective analysis of EMR data extracted during the intervention highlighted that participants had commonly encountered a number of presenting problems that were not included in the curriculum (falls, back pain, and cough). An alternative approach involving data mining of EMR data to generate a preliminary curriculum, followed by domain expert review, may have strengthened the intervention. EMR data mining has not been used in this way to date, though the approach has been shown to have some use to identify areas where medical practitioners are deviating from best practice and may benefit from undertaking specific training [28].

Furthermore, findings from this study indicated that, while data can be used to personalize training, the quantity of data collected by EMR datasets is substantial, and choosing which data to use to prompt learning can be complex. A confounding factor is that while there is an upward rise for EMR adoption in health care globally [17], real-world adoption varies considerably across countries and health care settings. This may limit the availability of EMR data for supporting the type of intervention, but there are many other clinical information systems in use in the health system [29]. Data from different clinical information has been shown to have value for understanding aspects of practitioner performance [30] and may be able to be used for this type of adaptive intervention.

Data-Driven Learning and Learner Engagement

Although it is feasible to adaptively deliver training using EMR data, findings from this study do not resolve how such an approach improves engagement with educational offerings. While study participants reported generally positive experiences with the program, there was also a notable decline in participation between the first and third cohorts. It is not clear why this decline occurred, but one explanation is that it aligned with the progression of the year. It is possible that participants undertaking the study were experiencing more exhaustion and had less capacity to engage with the intervention further along in their training year. There is some evidence to suggest that internship burnout is higher later in the year when completing postgraduate training [31].

The majority of learners in the personalized program attempted at least 50% of the program, which demonstrates a retention rate on the higher end for an online program [9]. Findings further indicated that many participants felt the intervention seemed aligned with their clinical practice. Aligning training with clinical practice may have a range of benefits for learners including increased engagement with training, better alignment with clinical practice, and more efficient delivery. Improving efficiency is important in medical practitioner education because learners are time-limited and have many competing obligations in their workloads beyond undertaking training, including completing core clinical responsibilities and administrative tasks [2,3].

Strengths, Limitations, and Future Research

A limitation of this study is that the survey response rate across all three cycles was fairly low, particularly for the third cycle. This limits our overall understanding of the intervention and its generalizability. The lack of data on the final cycle limits understanding of how changing the content may have altered the learner experience with the program in the last cycle. Although we can only speculate as to why the response rate was low, one speculation is that the evaluation was done very late in the participants' clinical term, and in some instances when they had started a new term. At this point in the term, participants had particularly busy clinical loads and may have had less capacity to provide feedback on the intervention.

Future researchers should consider further exploring how to design medical practitioner learning to be pedagogically sound and well-aligned with clinical practice. Design considerations that remain to be explored include developing a process for automated extraction of data to more closely link delivery of learning with the clinical encounter, identifying where in workflows adaptive training is optimally delivered (at point of care and after hours), and customizing content of the learning not just delivery using EMR data. It would be valuable to undertake studies that evaluate whether personalization of training using EMR data and the timing of content delivery affect learner retention rates, as well as improve the processes and outcomes of care. Beyond this, questions remain about the scalability and cost-effectiveness of adaptive learning interventions of this nature.

Conclusions

This study demonstrates that personalizing an online learning program for emergency trainees using EMR data is feasible for this group of medical practitioners, and most also found it to be an acceptable approach to align learning with workplace interactions. This opens up considerable opportunity to tailor and personalize learning for health professionals that is aligned with their practice activities and performance. However, more research is needed to understand how to deliver these types of adaptive learning interventions in a scalable and sustainable way. To do this, there is a need to develop a better understanding of how routinely collected data can be used to understand performance, as well as its suitability enabling health professionals to reflect on their performance and practice to support continuous learning. Relatedly, on the solution design side, there is a need to develop platforms that can automate the extraction, analysis, and feedback of these data to support health professional education and practice reflection in a way that is streamlined for use by health professionals. Considerable focus to date has been placed on the value of aggregating large data sets into single repositories, which represents a significant infrastructure achievement. However, moving forward, it is as important to understand why data are being collected and how they will be collected to ensure that the right information is available to provide benefits to the health system and support people's health and well-being.

Acknowledgments

AJ was undertaking a postdoctoral research fellowship funded through the Digital Health Cooperative Research Center when this research was undertaken. The Digital Health Cooperative Research Center was established and supported under the Australian Government's Cooperative Research Centres Program. The authors wish to acknowledge the support of Shannon Joyce and Shima Ghassem Pour in supporting access to fully deidentified electronic medical record reports to enable delivery of the adaptive program.

Conflicts of Interest

None declared.

Multimedia Appendix 1
An example of one of the cases used within the program, and images showing how the content was presented by the Qstream platform.

[PDF File, 213 KB - mededu_v11i1e65287_app1.pdf]

References

1. Peck C, McCall M, McLaren B, Rotem T. Continuing medical education and continuing professional development: international comparisons. BMJ 2000 Feb 12;320(7232):432-435. [doi: 10.1136/bmj.320.7232.432] [Medline: 10669451]
2. Westbrook JI, Ampt A, Kearney L, Rob MI. All in a day's work: an observational study to quantify how and with whom doctors on hospital wards spend their time. Med J Aust 2008 May 5;188(9):506-509. [doi: 10.5694/j.1326-5377.2008.tb01762.x] [Medline: 18459920]
3. Chaiyachati KH, Shea JA, Asch DA, et al. Assessment of inpatient time allocation among first-year internal medicine residents using time-motion observations. JAMA Intern Med 2019 Jun 1;179(6):760-767. [doi: 10.1001/jamainternmed.2019.0095] [Medline: 30985861]
4. Bullock A. Does technology help doctors to access, use and share knowledge? Med Educ 2014 Jan;48(1):28-33. [doi: 10.1111/medu.12378] [Medline: 24330114]
5. Ifediora CO. Online medical education for doctors: identifying potential gaps to the traditional, face-to-face modality. J Med Educ Curric Dev 2019;6(6):2382120519827912. [doi: 10.1177/2382120519827912] [Medline: 30801035]
6. Curran V, Matthews L, Fleet L, Simmons K, Gustafson DL, Wetsch L. A review of digital, social, and mobile technologies in health professional education. J Contin Educ Health Prof 2017;37(3):195-206. [doi: 10.1097/CEH.0000000000000168] [Medline: 28834849]
7. Cook DA, Triola MM. What is the role of e-learning? Looking past the hype. Med Educ 2014 Sep;48(9):930-937. [doi: 10.1111/medu.12484] [Medline: 25113119]
8. Baum S, McPherson M. The human factor: the promise & limits of online education. Daedalus 2019 Oct;148(4):235-254. [doi: 10.1162/daed_a_01769]
9. Bawa P. Retention in online courses: exploring issues and solutions—a literature review. Sage Open 2016;6(1):2158244015621777. [doi: 10.1177/2158244015621777]
10. Cutrer WB, Spickard WA III, Triola MM, et al. Exploiting the power of information in medical education. Med Teach 2021 Apr 8;43(sup2):S17-S24. [doi: 10.1080/0142159X.2021.1925234]
11. Robertson ARR, Nurmatov U, Sood HS, Cresswell K, Smith P, Sheikh A. A systematic scoping review of the domains and innovations in secondary uses of digitised health–related data. J Innov Health Inform 2016 Nov 10;23(3):611-619. [doi: 10.14236/jhi.v23i3.841] [Medline: 28059695]
12. Bucalon B, Shaw T, Brown K, Kay J. State-of-the-art dashboards on clinical indicator data to support reflection on practice: scoping review. JMIR Med Inform 2022 Feb 14;10(2):e32695. [doi: 10.2196/32695] [Medline: 35156928]
13. Lang C, Siemens G, Wise A, Gasevic D. Handbook of Learning Analytics: Society for Learning Analytics and Research; 2017.
14. Li KC, Wong BTM. Personalising learning with learning analytics: a review of the literature. In: Cheung SKS, Li R, Phusavat K, Paoprasert N, Kwok LF, editors. Blended Learning Education in a Smart Learning Environment: 13th International Conference, ICBL 2020, Bangkok, Thailand, August 24–27, 2020, Proceedings: Springer; 2020:39-48. [doi: 10.1007/978-3-030-51968-1_4]
15. Thoma B, Warm E, Hamstra SJ, et al. Next steps in the implementation of learning analytics in medical education: consensus from an International Cohort of Medical Educators. J Grad Med Educ 2020 Jun;12(3):303-311. [doi: 10.4300/JGME-D-19-00493.1] [Medline: 32595850]
16. Menachemi N, Collum TH. Benefits and drawbacks of electronic health record systems. Risk Manag Healthc Policy 2011;4:47-55. [doi: 10.2147/RMHP.S12985] [Medline: 22312227]
17. Global diffusion of eHealth: making universal health coverage achievable: report of the third global survey on eHealth. World Health Organization. 2016 Dec 15. URL: https://www.who.int/publications/i/item/9789241511780 [accessed 2025-12-10]
18. Klein D, Staples J, Pittman C, Stepanko C. Using electronic clinical practice audits as needs assessment to produce effective continuing medical education programming. Med Teach 2012;34(2):151-154. [doi: 10.3109/0142159X.2012.644826] [Medline: 22288993]
19. Sebok-Syer SS, Goldszmidt M, Watling CJ, Chahine S, Venance SL, Lingard L. Using electronic health record data to assess residents' clinical performance in the workplace: the good, the bad, and the unthinkable. Acad Med 2019 Jun;94(6):853-860. [doi: 10.1097/ACM.0000000000002672] [Medline: 30844936]
20. Shaw T, Janssen A, Crampton R, et al. Attitudes of health professionals to using routinely collected clinical data for performance feedback and personalised professional development. Med J Aust 2019 Apr;210 Suppl 6:S17-S21. [doi: 10.5694/mja2.50022] [Medline: 30927464]
21. Baysari MT, Oliver K, Egan B, et al. Audit and feedback of antibiotic use. Appl Clin Inform 2013;04(04):583-595. [doi: 10.4338/ACI-2013-08-RA-0063]
22. Linder JA, Schnipper JL, Tsurikova R, et al. Electronic health record feedback to improve antibiotic prescribing for acute respiratory infections. Am J Manag Care 2010 Dec;16(12 Suppl HIT):e311-e319. [Medline: 21322301]

23. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care 2013 Aug;51(8 Suppl 3):S30-S37. [doi: 10.1097/MLR.0b013e31829b1dbd] [Medline: 23774517]
24. Anderson T, Shattuck J. Design-based research: a decade of progress in education research. Educ Res 2012;41(1):16-25. [doi: 10.3102/0013189X11428813]
25. Pathways in medicine. Australian Medical Association. URL: https://www.ama.com.au/pathways-in-medicine#doctors-in-training [accessed 2025-10-08]
26. QStream. URL: https://qstream.com/ [accessed 2024-08-09]
27. Shaw T, Janssen A, Barnet S, et al. The CASE methodology: a guide to developing clinically authentic CASE-based scenarios for online learning programs targeting evidence-based practice. Health Educ Pract J Res Prof Learn 2018;1(1):2209-3974 [FREE Full text]
28. Fidopiastis CM, Venta KE, Baker EG, Stanney KM. A step toward identifying sources of medical errors: modeling standards of care deviations for different disease states. Mil Med 2018 Mar 1;183(suppl_1):105-110. [doi: 10.1093/milmed/usx203] [Medline: 29635597]
29. Janssen A, Donnelly C, Shaw T. A taxonomy for health information systems. J Med Internet Res 2024 May 31;26:e47682. [doi: 10.2196/47682] [Medline: 38820575]
30. Bucalon B, Whitelock-Wainwright E, Williams C, et al. Thought leader perspectives on the benefits, barriers, and enablers for routinely collected electronic health data to support professional development: qualitative study. J Med Internet Res 2023 Feb 16;25:e40685. [doi: 10.2196/40685] [Medline: 36795463]
31. Parr JM, Pinto N, Hanson M, Meehan A, Moore PT. Medical graduates, tertiary hospitals, and burnout: a longitudinal cohort study. Ochsner J 2016;16(1):22-26. [Medline: 27046399]

Abbreviations

EMR: electronic medical record
HIT: health information technology

Edited by B Lesselroth; submitted 12.08.24; peer-reviewed by A Lakdawala, I Mircheva, S Goyal; revised version received 20.05.25; accepted 23.09.25; published 17.12.25.

Please cite as:
Janssen A, Coggins A, Tadros J, Quinn D, Shetty A, Shaw T
Using Electronic Health Data to Deliver an Adaptive Online Learning Solution to Emergency Trainees: Mixed Methods Pilot Study
JMIR Med Educ 2025;11:e65287
URL: https://mededu.jmir.org/2025/1/e65287
doi:10.2196/65287

Engaging Undergraduate Medical Students With Introductory Research Training via an Educational Escape Room: Mixed Methods Evaluation

Bastien Le Guellec[1,2], MD, MS; Victoria Gauthier[1,3], PharmD; Rémi Lenain[4], MD, PhD; Alexandra Nuytten[5], MD, PhD; Luc Dauchet[1,3], MD, PhD; Brigitte Bonneau[1]; Erwin Gerard[6], PharmD; Claire Castandet[7], MD; Patrick Truffert[8], MD, PhD; Marc Hazzan[9], MD, PhD; Philippe Amouyel[1,3], MD, PhD; Raphaël Bentegeac[1,3*], MD, MPH; Aghiles Hamroun[1,3*], MD, PhD

[1]Department of Public Health, Epidemiology, Health Economics and Prevention, Centre Hospitalier Universitaire de Lille, 6 rue du professeur Laguesse, Lille, France

[2]Department of Neuroradiology, Centre Hospitalier Universitaire de Lille, Lille, France

[3]UMR1167 RID-AGE, Institut Pasteur de Lille, Inserm, Université de Lille, CHU Lille, Lille, France

[4]Department of Nephrology, Dialysis, Kidney Transplantation, and Apheresis, Centre Hospitalier Universitaire de Lille, Lille, France

[5]Department of Neonatology, Groupe Hospitalier de l'Institut Catholique de Lille, Lille, France

[6]Evaluation des technologies de santé et des pratiques médicales, Lille, France

[7]Department of Support for Pedagogy and Innovation (DAPI), Université de Lille, Lille, France

[8]Department of Neonatal Medicine, Centre Hospitalier Universitaire de Lille, Lille, France

[9]Faculty of Medicine, Université de Lille, Lille, France

[*]these authors contributed equally

Corresponding Author:
Aghiles Hamroun, MD, PhD
Department of Public Health, Epidemiology, Health Economics and Prevention, Centre Hospitalier Universitaire de Lille, 6 rue du professeur Laguesse, Lille, France

Abstract

Background: Early exposure to research methodology is essential in medical education, yet many students show limited motivation to engage with nonclinical content. Gamified strategies such as educational escape rooms may help improve engagement, but few studies have explored their feasibility at scale or evaluated their impact beyond student satisfaction.

Objective: This study aimed to assess the feasibility, engagement, and perceived educational value of a large-scale escape room specifically designed to introduce third-year medical students to the principles of diagnostic test evaluation.

Methods: We developed a low-cost immersive escape room based on a fictional diagnostic accuracy study with 6 puzzles mapped to five predefined learning objectives: (1) identifying key components of a diagnostic study protocol, (2) selecting an appropriate gold standard test, (3) defining a relevant study population, (4) building and interpreting a contingency table, and (5) critically appraising diagnostic metrics in context. The intervention was deployed to an entire class of third-year medical students across 12 sessions between March 2023 and April 2023. Each session included 60 minutes of gameplay and a 45-minute debriefing. Students completed pre- and postintervention questionnaires assessing their knowledge of diagnostic test evaluation and perceptions of research training. Descriptive statistics and 2-tailed paired *t* tests were used to evaluate score changes; univariate linear regressions assessed associations with demographics. Free-text comments were analyzed using the hierarchical classification by Reinert.

Results: Of the 530 participants, 490 (92.5%) completed the full evaluation. Many participants had had limited previous exposure to escape rooms (206/490, 42% had never participated in one), and most (253/490, 51.6%) reported low initial confidence with critical appraisal of scientific articles. Mean overall knowledge scores increased from 62 of 100 (SD 1) before to 82 of 100 (SD 2) after the activity (+32%; $P<.001$). Gains were observed across all learning objectives and were not influenced by age, sex, or previous experience. Students rated the educational escape room as highly entertaining (mean score 9.1/10, SD 1.1) and educational (mean score 8.2/10, SD 1.5). Following the intervention, 86.9% (393/452) felt more comfortable with critical appraisal of diagnostic test studies, and 79% (357/452) considered the escape room format highly appropriate for an introductory session.

Conclusions: This study demonstrates the feasibility and enthusiastic reception of a large-scale, reusable escape room aimed at teaching the fundamental principles of diagnostic test evaluation to undergraduate medical students. This approach may serve

as a valuable entry point to engage students with evidence-based reasoning and pave the way for deeper exploration of medical research methodology.

(JMIR Med Educ 2025;11:e71339) doi:10.2196/71339

KEYWORDS

escape room; undergraduate; medical students; research; engagement; gamification

Introduction

Evidence supports the need for early exposure of medical students to research and critical appraisal of scientific articles. According to the World Federation for Medical Education 2020 standards, medical curricula must include the principles of the scientific method, cover analytical and critical thinking, medical research methodology, and evidence-based medicine [1]. The ability to interpret and apply evidence-based medicine is now widely regarded as a core competency for graduating medical students, as emphasized by both the Association of American Medical Colleges and the UK Clinical Reasoning in Medical Education group [2,3]. By developing critical appraisal skills, early exposure to research in medical education favors abilities valuable for future clinical practice (analytical reasoning and communication skills).

In particular, diagnostic test studies involve key elements of Bayesian reasoning [4], such as pretest probability, likelihood ratios, and the process of updating diagnostic probabilities based on test results—all of which are central to clinical decision-making [5]. However, evidence shows that medical professionals struggle with these concepts [6]. In addition, integration of such nonclinical skills into medical curricula is arduous, especially with undergraduate students [7,8]. In particular, generating and maintaining student interest is highly challenging [9]. In France, for example, critical appraisal of scientific articles, despite its inclusion in the final undergraduate national matching exam since 2009, varies in terms of course load and content across universities, and medical students lack motivation to invest time in nonclinical skills [10]. However, as observed with the recent expansion of biomedical literature related to the COVID-19 pandemic, physicians are at the center of both scientific and societal discussions involving critical appraisal of medical literature [11]. Thus, innovative strategies are needed to engage medical students with research training and critical appraisal of scientific articles early in their curriculum [7].

Educational escape rooms (EERs) have recently garnered growing interest in health professional education [12-15]. These gamified, immersive scenarios typically involve a series of puzzles that participants solve collaboratively, each aligned with a specific learning objective. While most published studies have focused on fostering clinical reasoning or teamwork, only a few have examined the potential of EERs for teaching research methodology [16,17]. Existing evidence remains limited, often based on small-scale initiatives with variable outcomes [12,18-20]. This gap underscores the need to explore their applicability in large cohorts and in domains beyond clinical knowledge.

This study evaluated the feasibility and perceived educational value of a large-scale EER designed to introduce medical students to key aspects of scientific methodology. The pedagogical content focused specifically on the evaluation and interpretation of diagnostic test studies while also aiming to foster teamwork and engage students with a nonclinical topic early in their training.

Methods

Escape Room Design and Learning Objectives

Using guidelines from Davis et al [14], we developed an EER for third-year medical students with three main goals: (1) to introduce fundamental principles of research methodology and terminology, (2) to promote collaborative problem-solving through teamwork, and (3) to foster more positive attitudes toward research training. The design team included 4 medical doctors, 1 pharmacist, and 1 medical resident, drawing on interdisciplinary expertise in clinical medicine, public health, and pedagogy.

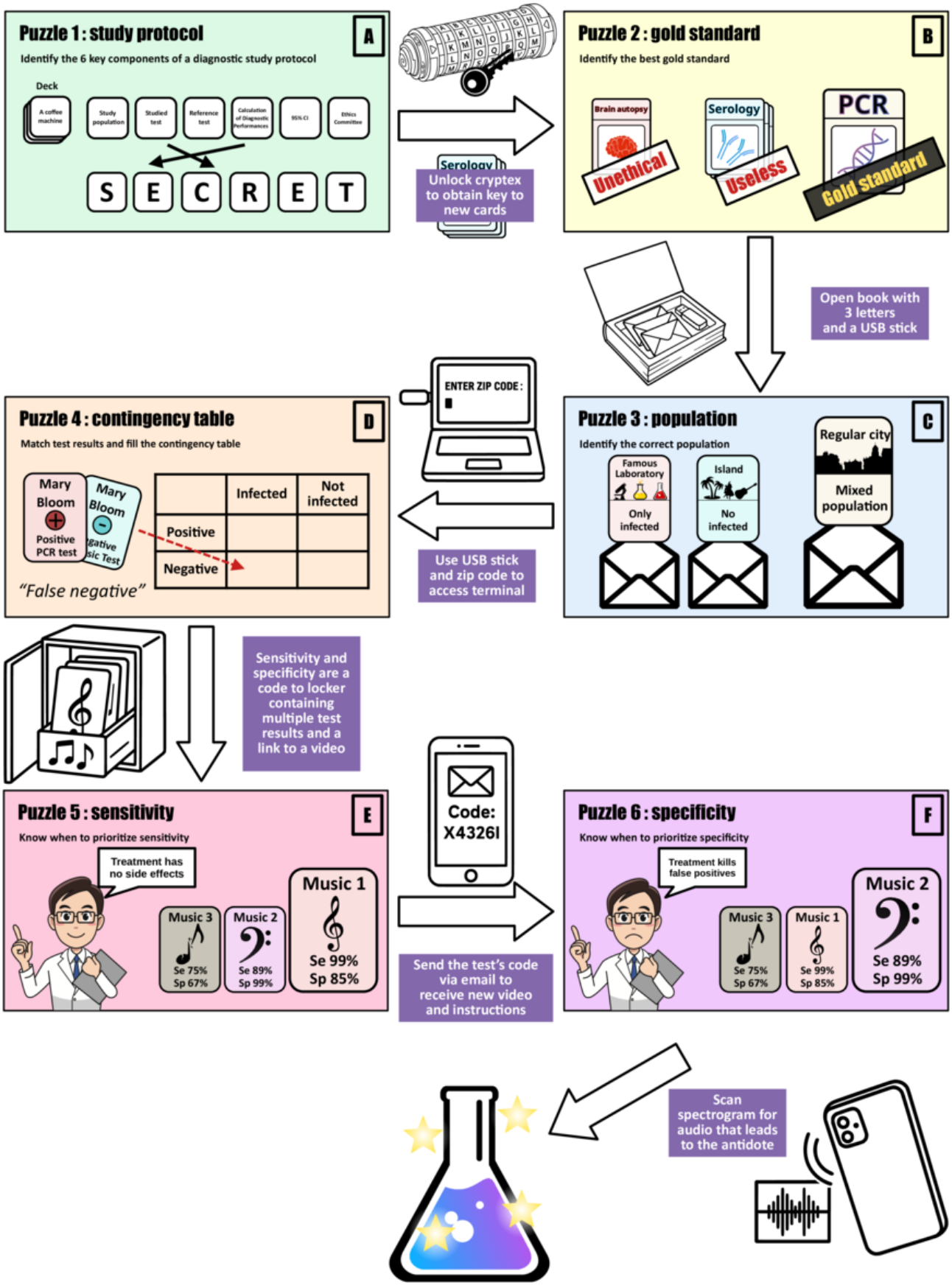
A diagnostic accuracy study was deliberately chosen as the pedagogical framework as this type of research offers an accessible and clinically meaningful entry point for undergraduate students without previous exposure to research methods. It aligns closely with diagnostic reasoning processes familiar to most learners while also serving as a structured introduction to key methodological concepts—such as reference standards, population selection, and diagnostic performance metrics (eg, sensitivity and specificity)—that remain challenging even for many practicing clinicians [5,6]. By anchoring the learning objectives in a framework that is both practical and conceptually rich, this approach facilitates early engagement with evidence-based thinking without requiring advanced statistical background.

Learning objectives were defined a priori based on recurrent misconceptions observed in upper-year students and the collective teaching experience of the faculty involved. By the end of the session and debriefing, students were expected to (1) identify the key components of a diagnostic accuracy study protocol, including relevant methodological tools; (2) recognize the characteristics of an ideal reference standard, including performance, cost, invasiveness, and availability; (3) understand the structure of a study population that includes both individuals with and without the condition of interest; (4) construct and interpret a contingency table from diagnostic test and gold standard results; and (5) critically assess the strengths and limitations of a diagnostic test based on its performance metrics and intended clinical use.

Escape Room Scenario

The escape room scenario centered on a fictional outbreak of a zombie virus disease. Students were tasked with identifying an effective diagnostic test to detect infected individuals and locate an antidote. Working in teams, they had 60 minutes to examine the research notes of a mysterious scientist who had nearly discovered a high-performing test. To succeed, students needed to solve 6 sequential puzzles, each aligned with one of the predefined learning objectives (Figure 1).

Figure 1. Narrative sequence and pedagogical content of the 6 puzzles composing the educational escape room. (A) Puzzle 1—study protocol. Participants select the 6 key components required to design a diagnostic accuracy study from a deck of 24 color-coded sticky notes. Each correct sticky note reveals 1 letter; when combined, they form a password that unlocks a cryptex. (B) Puzzle 2—gold standard. Inside the cryptex, participants find a key granting access to a new set of test cards. They must identify the most appropriate reference test. Although brain autopsy provides definitive results, it is excluded as unethical and impractical. Serology is dismissed due to poor diagnostic value. Polymerase chain reaction (PCR) is correctly selected as a feasible and valid gold standard. A clue hidden in invisible ink on the back of the PCR card leads them to the next step. (C) Puzzle 3—study population. Inside a hollow book designed as a lockable book safe, participants find a USB stick and 3 fictional letters each proposing a different study population: only infected individuals (famous laboratory), only uninfected individuals (isolated island), and a mixed population (regular municipality). Despite the laboratory's prestige, they must reject biased populations and select the mixed group to ensure valid assessment of diagnostic performance. (D) Puzzle 4—test results and matching. Using the zip code from the city letter and the USB stick, participants access a computer terminal. They learn that the studied test involves a music-based rhythm activity. They receive patient test results, one card per test per patient, and must match each pair. Recognizing that both the studied test and reference test are needed for each patient, they correctly populate the contingency table to proceed. They calculate sensitivity (Se) and specificity (Sp), which together form the code to open a locker. (E) Puzzle 5—sensitivity. Inside the locker, they find test cards associated with musical rhythms, each labeled with sensitivity and specificity values. A QR code on one card links to a researcher's video revealing that the treatment has no side effects. Participants must deduce that false positives are acceptable and, therefore, select the test with the highest sensitivity. They email the chosen test's code to proceed. (F) Puzzle 6—specificity. A second video reveals a crucial update: the treatment is actually harmful to false positives. Participants must now prioritize specificity to avoid treating uninfected individuals. They scan a spectrogram printed on the back of each card using a mobile app. The correct choice plays a final audio message and leads to the antidote and a reward.



The six puzzles included (1) classifying keywords relevant to diagnostic study protocols (*study protocol*), (2) selecting the most appropriate gold standard test from a series of candidates (*gold standard*), (3) identifying a valid study population using fictional application letters (*study population*), (4) analyzing mock laboratory test results to construct a contingency table (*contingency table*), (5) calculating diagnostic accuracy metrics based on the table (*test metrics*), and (6) interpreting these

metrics to choose the best test in context (*metric appraisal*). The terms in parentheses will be used throughout the manuscript for clarity.

Each completed puzzle provided a clue, object, or code that allowed progression to the next stage of the game. A team succeeded when they completed all 6 puzzles within the allotted time, thereby unlocking the location of the antidote. The scenario was fully autonomous: game masters were available only to provide hints or intervene in case of technical difficulties upon request. The immersive experience included props such as a cryptex, mock test description cards, a book safe, invisible ink, a digital lockbox, a custom-designed computer program, role-play videos, QR codes, a spectrogram decryption smartphone app, and purpose-built board game–style cards (Figure 2). All materials were original creations funded by the Faculty of Medicine at the University of Lille, with a total cost of €1800 (approximately US $1925) for approximately 600 students.

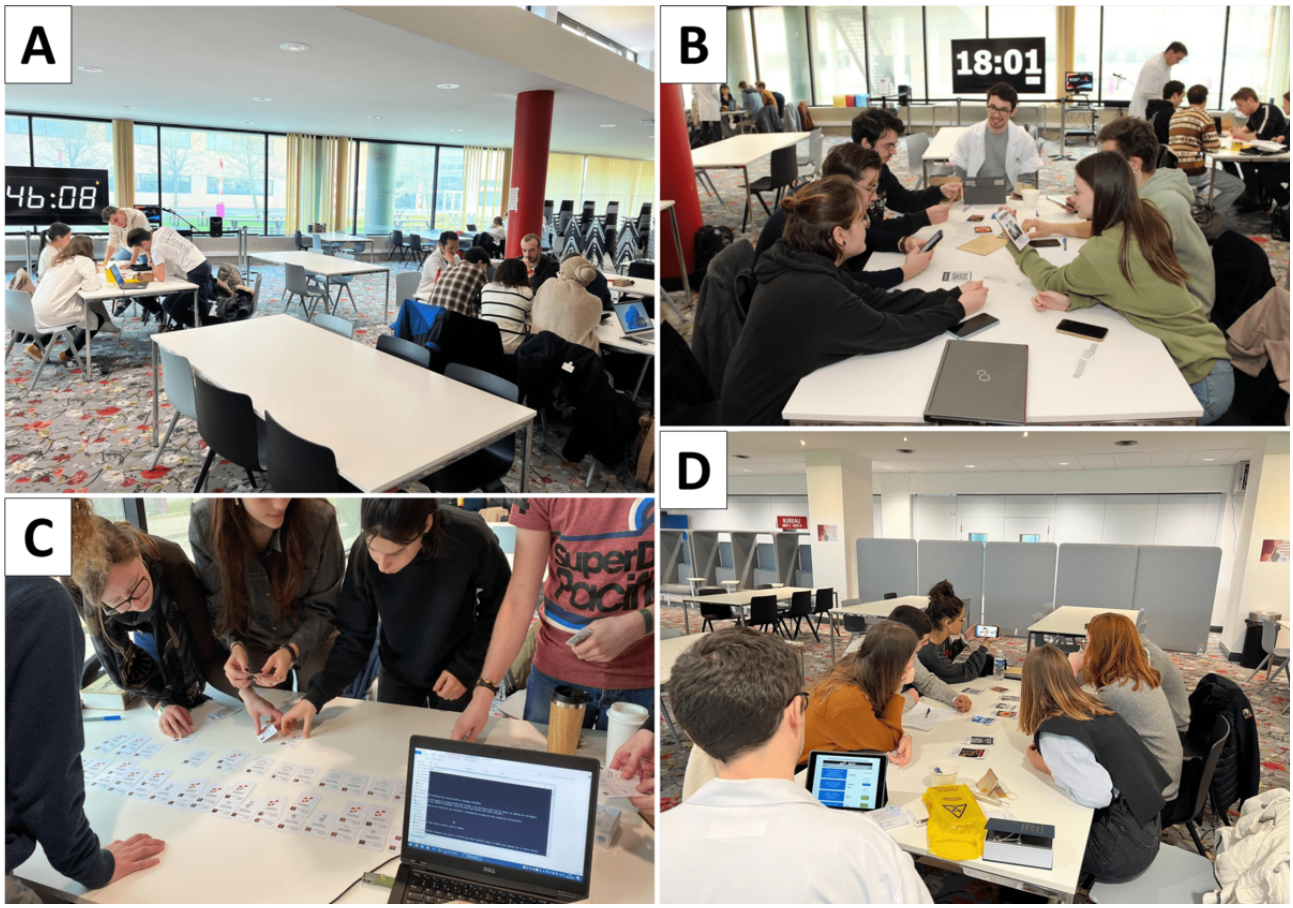
Figure 2. Accessories used throughout the educational escape room. (A) All materials are stored in a single portable bag to ensure easy deployment. (B) Keyword cards used to reconstruct the components of a diagnostic accuracy study (puzzle 1, *study protocol*). (C) Candidate reference tests from which participants must identify the most appropriate gold standard (puzzle 2, *gold standard*). (D) Fictional letters describing different populations, among which students must select the most suitable study population (puzzle 3, *study population*). (E) Decks of index and reference test result cards used to reconstitute patient data pairs (puzzle 4, *contingency table*). (F) USB flash drive containing an automated program through which students must input their reconstructed contingency table (puzzle 5, *test metrics*). (G) Test cards with various diagnostic performance metrics from which students must select the most appropriate test depending on the clinical scenario (puzzle 6, *metric appraisal*).

Setting, Materials, and Staffing

Twelve 120-minute sessions were conducted over the course of the program. Each session included a 15-minute introductory briefing, 60 minutes of gameplay, and 45 minutes of structured debriefing. Sessions were held in a large open space of approximately 150 m^2, accommodating up to 45 students per session (Figure 3). Upon arrival, students were randomly

assigned to teams of 4 to 6 using an algorithm triggered by swiping their university ID card.

Figure 3. Scenes from the educational escape room illustrating the gameplay environment and student engagement. (A) Students collaborate in real time to solve puzzles under time pressure, with a visible countdown clock reinforcing immersion. (B) Small groups work simultaneously in the same session space, each supported by a facilitator in a white coat, simulating a clinical research environment. (C) Participants match test result cards to reconstruct the correct index and reference test pairs before entering the data into a contingency table via a terminal. (D) A team discusses the interpretation of diagnostic performance metrics from QR-coded test cards guided by the game's unfolding narrative under the supervision of a dedicated faculty facilitator who performs real-time formative assessment using a tablet-based checklist. All individuals appearing in the photos provided written consent for their images to be used and published for educational and research dissemination purposes.



Eight teams participated simultaneously in each session seated at individual tables spaced apart to allow for parallel problem-solving without interference. Although certain immersive elements were shared—such as a bookshelf, storage lockers with digital codes, a whiteboard, and a wall-mounted countdown timer—each team engaged independently with dedicated materials and instructions. All essential components were either prepacked in portable game bags (Figure 2) or duplicated in 8 identical sets to ensure autonomous progression. The briefing began with a short immersive video introducing the fictional mission and setting the narrative tone, aiming to engage students from the outset and foster a collaborative atmosphere (Multimedia Appendix 1).

Each session involved 9 facilitators: 1 game master assigned to each team and a session coordinator overseeing the entire room. Facilitators ensured smooth progression and conducted real-time formative assessments. After each puzzle, they rated the team's clinical reasoning and collaborative dynamics using a tablet-based checklist with Likert scale items. This structure allowed students to work independently while maintaining consistent pedagogical oversight and documentation of performance. In total, 20 facilitators contributed across the sessions, representing a broad range of specialties including public health, radiology, nephrology, biology, biostatistics, pharmacology, pharmacy, pathology, family medicine, pediatrics, anesthesiology, and intensive care. All facilitators completed the escape room before student sessions to ensure fluency with the scenario and educational objectives.

To support scalability, all materials were designed for reuse. Game bags were reorganized after each session to allow for immediate reset. One month before implementation, the scenario was pilot-tested by 30 fourth-year medical students and 6 instructors whose feedback informed refinements to both content and logistics.

Study Setting and Design

This monocentric prospective study was conducted between March 2023 and April 2023 at the Faculty of Medicine at the University of Lille All third-year medical students enrolled in the 2022 - 2023 cohort were eligible to participate. The escape room was integrated into the mandatory curriculum as part of a course on research and critical appraisal and was a graded educational activity. Participation was required, and any student

who missed a scheduled session was expected to provide a formal justification.

Before the session, participants provided demographic information (age range and sex) and reported previous experience with recreational escape games ("How many times have you participated in a recreational Escape Game?"). They also rated their comfort with critical appraisal of scientific articles using a 5-point Likert scale ("What is your level of comfort with critical appraisal of scientific articles?") and submitted a free-text response identifying 3 words that best reflected their perception of critical appraisal ("Tell us the first three words that come to mind to describe your perception about critical appraisal of scientific articles").

They then individually completed a knowledge questionnaire covering the 5 predefined learning objectives of the session. All questions were answered using "true" or "false" and grouped by objective. The same 5 learning objectives were reassessed during the debriefing immediately after the intervention using a different but equivalent set of "true"-or-"false" questions. A score out of 100 was calculated for each objective, and relative gain was defined as the absolute change between pre- and posttest scores divided by the pretest score. The complete list of questions is available in Multimedia Appendix 2. During the debriefing, students also completed an individual postintervention survey. They were asked to provide 3 words summarizing their perception of the escape room experience and to rate both the entertainment value and the pedagogical interest of the session on a scale from 0 to 10. They were again asked to rate their comfort with critical appraisal using the same 5-point Likert scale. In addition, they were asked the following: "Do you think an escape room is a suitable format for an introductory course on the critical appraisal of scientific articles?" Responses were collected using a 5-point Likert scale ranging from "Not at all suitable" to "Absolutely suitable." Although "suitable" was not formally defined, the proximity of this question to those assessing the activity's educational and entertainment value likely guided students to interpret it in terms of relevance, clarity, and pedagogical appropriateness. Finally, students were invited to leave open-ended comments about the session, including organizational aspects, perceived strengths, limitations, and suggestions for improvement.

This monocentric prospective study was conducted between March 2023 and April 2023 at the Faculty of Medicine at the University of Lille All third-year medical students enrolled in the 2022 - 2023 cohort were eligible to participate. The escape room was integrated into the mandatory curriculum as part of a course on research and critical appraisal and was a graded educational activity. Participation was required, and any student who missed a scheduled session was expected to provide a formal justification.

Ethical Considerations

The study protocol was reviewed and approved by the Institutional Review Board of the University of Lille in February 2023. The IRB determined that formal ethical approval was not required due to the pedagogical nature of the intervention and the exclusive use of anonymized data.

All procedures were conducted in accordance with the ethical standards of the institutional and national research committees and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Upon registration on the study platform, students completed a short presession questionnaire and provided individual informed consent, including authorization to analyze their anonymized responses for research purposes. The study also ensured the privacy and confidentiality of all participants. Data were collected and stored on a secure, password-protected server, and identifying details, such as names and email addresses, were immediately deidentified and replaced with unique participant IDs to protect anonymity. No personal health information was collected.

Participants were not compensated for their involvement in the study (Multimedia Appendix 2).

Statistical Analysis

Descriptive statistics were used to summarize participant characteristics and questionnaire responses. Quantitative variables were reported as means and SDs or medians and IQRs depending on their distribution. Categorical variables were expressed as counts and percentages. To assess knowledge gains, pre- and postintervention scores were compared using paired *t* tests. Associations between participant characteristics (age, sex, previous experience with escape rooms, and self-reported ease with critical appraisal) and relative score improvement were explored using univariate linear regression models. All statistical tests were 2-sided, and a P value of $<.05$ was considered statistically significant. All analyses were conducted using the R software (version 4.3.1; R Foundation for Statistical Computing).

Qualitative Analysis of Free-Text Feedback

Students' postintervention free-text comments were analyzed using the hierarchical classification method by Reinert [21], a lexical clustering approach based on word co-occurrence patterns. The analysis was conducted using the R Interface for Multidimensional Analysis of Texts and Questionnaires [22]. Comments were anonymized and preprocessed to remove low-frequency and noninformative words. The text corpus was segmented, and descending hierarchical classification was conducted to identify stable clusters of vocabulary. Factorial correspondence analysis was used to visualize the associations between lexical classes. The number of clusters retained was selected based on thematic coherence and interpretability. Two researchers (BLG and AH) independently reviewed and labeled each cluster using representative keywords and excerpts. Discrepancies were resolved through consensus. For reporting, all illustrative quotes were translated from French into English by the authors.

Results

Study Population

Of the 560 eligible third-year medical students, 530 (94.6%) participated in the escape room activity. Students who did not attend the escape room sessions were absent for justified

personal or medical reasons. Of the 530 students who participated, 490 (92.5%) completed the preintervention questionnaire. Among them, 66.3% (325/490) were women, and 83.5% (409/490) were aged between 20 and 22 years. Most students reported limited or no previous experience with recreational escape rooms: 42% (206/490) had never participated in one, and 45.9% (225/490) had only done so once or twice. A substantial proportion expressed discomfort with critical appraisal of scientific literature: 18.8% (92/490) reported being "absolutely not comfortable," and 33% (161/490) reported being "rather not comfortable" with the task (Table 1).

Table . Description of the study population (N=490).

Characteristic	Participants, n (%)
Female sex	325 (66.3)
Age group (y)	
<20	22 (4.5)
20 - 22	409 (83.5)
22 - 25	41 (8.4)
>25	18 (3.7)
Previous experience with recreational escape rooms	
None	206 (42)
1 - 2 times	225 (45.9)
3 - 4 times	52 (10.6)
≥5 times	7 (1.4)
Reported ease with critical appraisal of scientific articles	
Absolutely not comfortable	92 (18.8)
Rather not comfortable	161 (32.9)
Neither comfortable nor uncomfortable	201 (41)
Rather comfortable	31 (6.3)
Absolutely comfortable	5 (1)

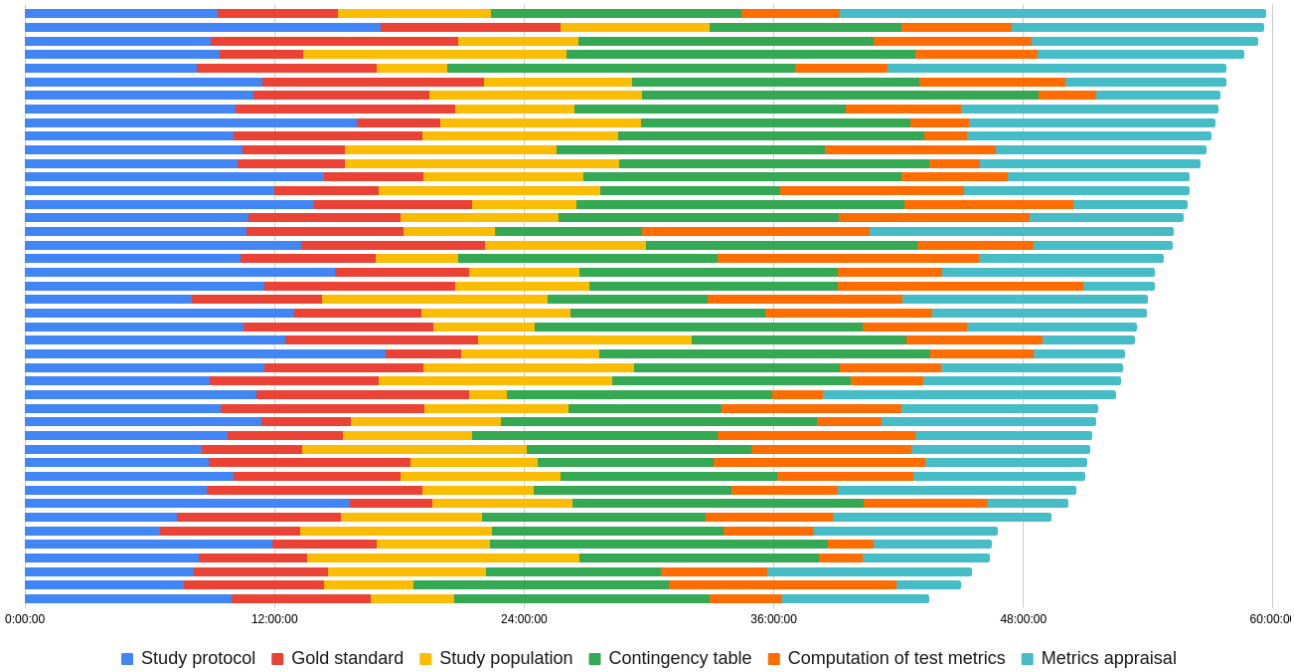
Escape Room Progression

Divided into 96 teams across 12 sessions, students advanced through the escape room with a high degree of consistency. All teams completed the mission within the allotted time, with a mean duration of 53 (SD 4) minutes. Each of the 6 puzzles required approximately 10 minutes to solve. The fastest puzzle, focused on test metrics, was completed in an average of 6 minutes and 12 seconds (SD 2 min, 49 s), whereas the most time-consuming one, involving the construction of a contingency table, took 12 minutes and 21 seconds on average (SD 2 min, 51 s; (Table 2). Team progression was largely synchronous, with most groups working on the same puzzle simultaneously (Figure 4).

Table . Table 2. Completion time of puzzles.

Puzzle	Mean completion time (Minutes:seconds)	SD of mean completion time (Minutes:seconds)
Study protocol	10:52	2:35
Gold standard	7:11	2:09
Study population	7:31	2:41
Contingency table	12:21	2:49
Computation of test metrics	6:10	2:48
Metrics appraisal	9:15	3:24

Figure 4. Puzzle progression timeline for a sample of 43 teams. Each horizontal line represents 1 team, and colors indicate the specific puzzle being solved at each moment during the session.



Pre- and Posttest Evaluations

A total of 85.3% (452/530) of the students completed the postintervention questionnaire. Student performance improved markedly following the intervention. The average score increased from 62/100 (SD 1) before the session to 82/100 (SD 2) afterward, representing a 32% gain (SD 5%; P<.001). Average subscores improved across all 5 learning objectives: from 58 (SD 3) to 71 (SD 3) for *study protocol*, from 74 (SD 2) to 89 (SD 4) for *gold standard*, from 43 (SD 1) to 83 (SD 3) for *study population*, from 71 (SD 2) to 88 (SD 5) for *contingency table*, and from 72 (SD 2) to 79 (SD 5) for *metric appraisal* (P<.001 in all cases; Figure 5). No significant associations were observed between score improvement and student characteristics (sex, age, previous escape room experience, or initial self-reported ease with critical appraisal), as shown in Table 3.

Figure 5. Pre- and posttest scores across learning objectives. Violin plots showing the distribution of student scores before and after the intervention for each of the 5 predefined learning objectives (top panels) and the overall mean score (bottom right panel). Asterisks indicate statistical significance. ***P<.001; EER: educational escape room.

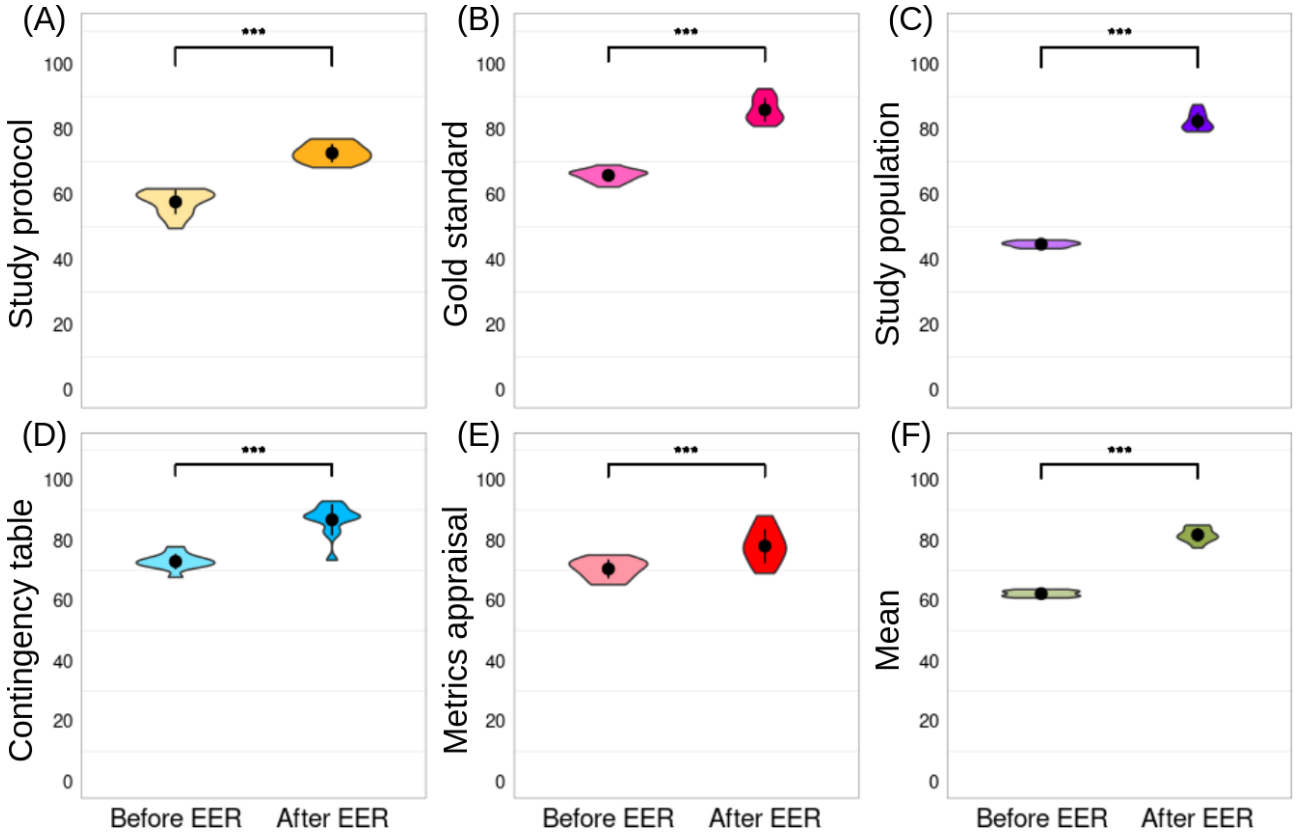


Table . Association between student characteristics and relative score improvement (univariate linear regressions).

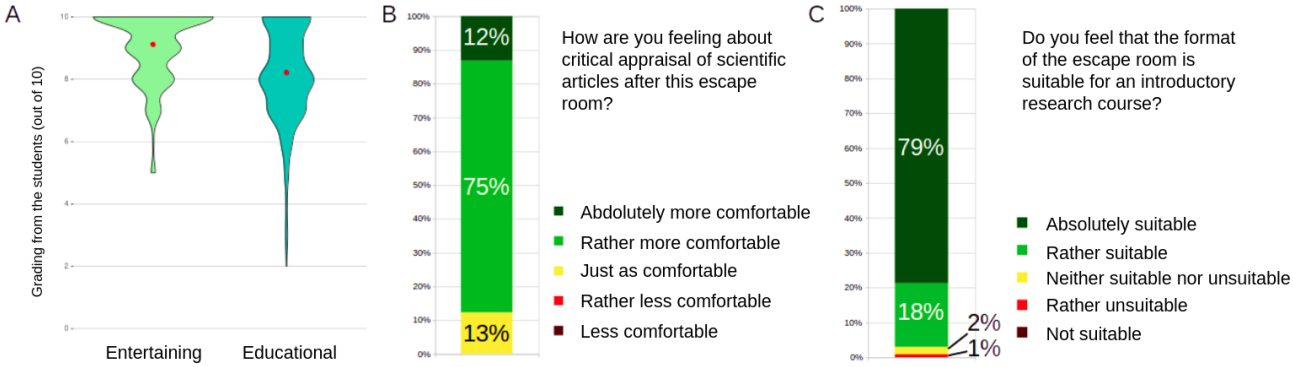
Characteristic	β (95% CI)	*P* value
Sex (female vs male)	−2.5 (−6.8 to 1.9)	.26
Age category (ordinal scale)	1.7 (−2.1 to 5.4)	.38
Previous experience with ERs[a] (ordinal scale)	0.9 (−2.8 to 4.6)	.63
Reported ease with critical appraisal of scientific articles (ordinal scale)	−0.6 (−3.3 to 2.1)	.69

[a]ER: escape room.

Students' Feedback

Feedback was strongly favorable. Students rated the session highly in terms of enjoyment (mean 9.1/10, SD 1.1) and educational value (mean 8.2/10, SD 1.5). Most participants (393/452, 86.9%) reported feeling more comfortable with the appraisal of diagnostic accuracy studies after the session. Regarding the overall suitability of escape rooms for introducing research training, 79% (357/452) rated the format as "absolutely suitable" on a 5-point Likert scale (Figure 6). The term "suitable" was not explicitly defined but followed questions about entertainment and educational value, likely guiding interpretation in terms of relevance and clarity.

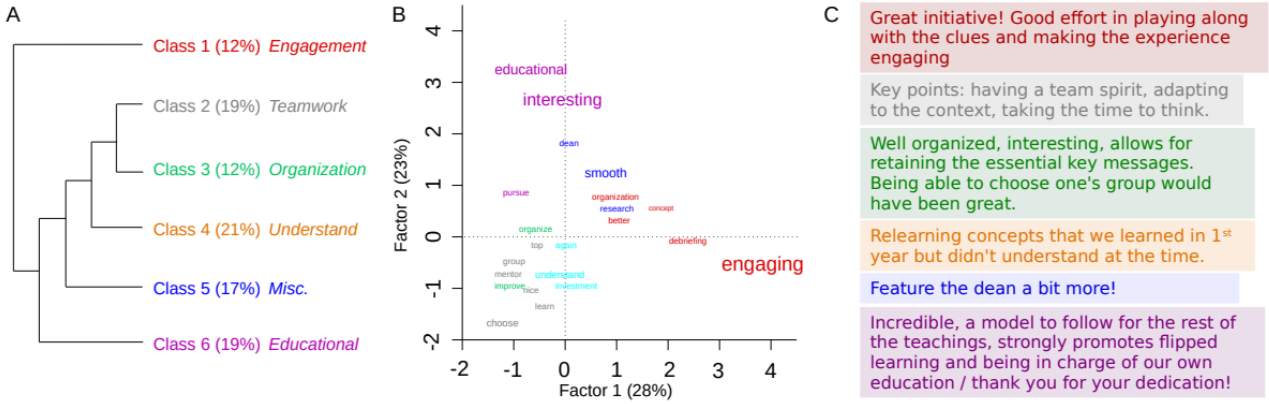
Figure 6. Students' perceptions of the educational escape room (EER). (A) Student ratings of the EER's entertainment and educational value (scale from 0 - 10); red dots indicate mean scores. (B) Proportion of students reporting increased comfort with critical appraisal of scientific articles. (C) Students' perception of the suitability of the escape room format as an introductory approach to research training.



Students' 3-word perceptions of research training shifted notably. Before the session, most descriptors were negative or reflected uncertainty (eg, "difficult," "tedious," "unknown," or "complicated"). Afterward, the most frequent terms were markedly more positive—"engaging," "entertaining," "original," and "educational" (Multimedia Appendix 3).

Thematic analysis of students' free-text comments using the hierarchical classification by Reinert [21] identified 6 primary thematic clusters, reflecting engagement, teamwork, pedagogical value, conceptual understanding, and overall organization (Figure 7 [21]).

Figure 7. Classification and thematic mapping of students' written feedback on the educational escape room. (A) The 6 semantic classes generated from a Reinert hierarchical clustering of students' open-ended comments (translated from French by the authors). Each class reflects a distinct theme: engagement, teamwork, organization, understanding, educational value, and miscellaneous remarks. (B) The factorial correspondence analysis map of the vocabulary used across classes, with words spatially projected according to their association with the underlying dimensions (factor 1: 28%; factor 2: 23%). (C) Representative anonymized student quotes from each class highlighting the diversity of perspectives on the format, group dynamics, and perceived learning outcomes.



Discussion

Our study demonstrates the feasibility of implementing a large-scale EER focused on introducing undergraduate medical students to core principles of diagnostic test evaluation as an entry point into research training. Student feedback echoed the 3 central aims of the intervention: to familiarize participants with key research concepts, foster teamwork, and improve perceptions of research-oriented courses. Pre- and postintervention assessments revealed significant immediate gains across all 5 targeted learning objectives. The activity was met with strong enthusiasm, in sharp contrast to students' initial reservations about research training.

EERs have been scarcely used in recent years for medical students, mostly to teach various clinical disciplines such as emergency medicine [23,24], radiology [25,26], surgery [16], dermatology [18], pulmonology [27], pediatrics [28], internal medicine [29], general medicine [30], infectious diseases [31], physiology [32], and patient safety [33-35] (Table 4). In this paper, we report an introductory course designed for a transversal, nonclinical skill and with no expected prerequisite among undergraduate medical students. Another study designed an EER as an introductory course involving research articles with great success, although the aim of that study was to introduce a small group of participants to technical skills in surgery [16]. A recent study by Mirshahi et al [17] described a self-developed escape room designed to assess and reinforce research competencies across a multidisciplinary group of undergraduate and postgraduate health profession students, including medical students; while promising, this initiative involved a limited number of participants and did not focus specifically on diagnostic reasoning. Most studies published to date have involved relatively small cohorts of students (Table 4) (up to 245 students over 3 years). However, to consider real-life implementation of EERs within medical curricula, it is crucial to demonstrate their compatibility with the current growing size of medical classes, often comprising hundreds of

students [36,37]. By using a parallelizable and repeatable gaming protocol, we showed the feasibility of delivering an innovative course to over 500 students while maintaining an entertaining and engaging environment. Another crucial challenge in implementing innovative pedagogic tools is the need to keep financial and human costs reasonable [38]. In our case, the allocated financial resources were constrained relative to the student count—€3.40 (US $3.95) per student. Through reusable material, this intervention can be replicated annually, refined, and adapted and enables the smoothing of the initial limited investment over the years [33]. As of July 2025, the EER has been repeated for 3 consecutive years (approximately 1800 students) without any additional financial cost. On the other hand, the human resources required for our intervention were substantial, with a total of 20 supervisors involved across the 12 sessions.

Table . Literature review of educational escape room experiments for medical students.

Study	Country	Title	Sample size	Specialty	Student level	Objective	Type of escape room	Evaluation
Zhang et al [23], 2018	United States	"Trapped as a Group, Escape as a Team: Applying Gamification to Incorporate Team-building Skills Through an 'Escape Room' Experience"	10	Emergency medicine	Residents	Team-building exercise	Commercial escape room	Postevent satisfaction survey
Backhouse and Malik [33], 2019	United Kingdom	"Escape Into Patient Safety: Bringing Human Factors to Life for Medical Students"	19	Specialty choice module	Undergraduate	Patient safety teaching	Self-created suitcase-based escape room	After-action review
Diemer et al [34], 2019	United States	"Patient Safety Escape Room: A Graduate Medical Education Simulation for Event Reporting"	120	Patient safety	Residents	Patient safety hazard reporting	Hospital case–based escape room	Postevent satisfaction survey
Kinio et al [16], 2019	Canada	"Break out of the Classroom: The Use of Escape Rooms as an Alternative Teaching Strategy in Surgical Education"	13	Surgery	Undergraduate	Motivation, engagement, and satisfaction	Escape room stations in the simulation center	Postevent satisfaction survey
Zhang et al [35], 2019	United States	"Finding the 'QR' to Patient Safety: Applying Gamification to Incorporate Patient Safety Priorities Through a Simulated 'Escape Room' Experience"	130	General medicine	Residents	Patient safety teaching	Escape room stations in the simulation center	Postevent survey
Jambhekar et al [25], 2020	United States	"Benefits of an Escape Room as a Novel Educational Activity for Radiology Residents"	164	Radiology	Residents	Team-building exercise	Self-created portable escape room	Postevent satisfaction survey
Guckian et al [18], 2020	United Kingdom	"Exploring the Perspectives of Dermatology Undergraduates With an Escape Room Game"	16	Dermatology	Undergraduate	Improving students' perceptions on the field	Self-created immersive escape room	Postevent satisfaction survey and focus groups

Study	Country	Title	Sample size	Specialty	Student level	Objective	Type of escape room	Evaluation
Liu et al [26], 2020	United Kingdom	"Feasibility of a Paediatric Radiology Escape Room for Undergraduate Education"	19	Pediatric radiology	Undergraduate	Knowledge and satisfaction	Self-created portable escape room	Pre- and posttest and satisfaction surveys
Abensur Vuillaume et al [24], 2021	France	"A Didactic Escape Game for Emergency Medicine Aimed at Learning to Work as a Team and Making Diagnoses: Methodology for Game Development"	10	Emergency medicine	Health care workers	Team-building exercise	Self-created immersive escape room	Not specified
Khanna et al [29], 2021	United States	"Escape MD: Using an Escape Room as a Gamified Educational and Skill-Building Teaching Tool for Internal Medicine Residents"	86	Internal medicine	Residents	Team work, critical thinking, and communication skills	Self-created immersive escape room	Postevent satisfaction survey
Akatsu et al [30], 2022	Japan	"Teaching 'medical interview and physical examination' from the very beginning of medical school and using 'escape rooms' during the final assessment"	140	General medicine (interview and physical examination)	Undergraduate	Course final assessment	Game-based scenarios with simulators	Postevent satisfaction survey
Dimeo et al [31], 2022	United States	"A Virtual Escape Room versus Lecture on Infectious Disease Content: Effect on Resident Knowledge and Motivation"	30	Infectious disease	Residents	Knowledge and motivation	Virtual escape room	Pretest, posttest, and motivation evaluations
Carrasco-Gomez et al [32], 2023	Spain	"Impact of a Peer-to-Peer Escape Room Activity in the Learning of Human Physiology of Medical Students From the University of Málaga"	245	Physiology	Undergraduate	Human physiology knowledge	Peer-to-peer–designed escape room	Comparative scores (vs nonparticipants) and postevent satisfaction survey

Study	Country	Title	Sample size	Specialty	Student level	Objective	Type of escape room	Evaluation
Fedorcsak [20], 2024	Norway	"Moderate Benefit of Escape Room Game on Learning Outcome in Medicine"	213	REI[a]	Undergraduate	General knowledge in REI	Self-created immersive escape room	Postevent comparative scores (vs nonparticipants)
Mirshahi et al [17], 2025	Iran	"'MORAD ESCAPE,' a Novel Research-Based Escape Room Approach for Evaluating Research Competencies of Health Professions Students"	60	Research competencies	Undergraduate and postgraduate	Research competency checklist	Self-created immersive escape room (escapED program)	Postevent evaluation of research competency and satisfaction survey

[a]REI: reproductive endocrinology and infertility.

Whether EERs meaningfully improve learning outcomes remains uncertain [12,15,19,20]. Previous studies have reported mixed results, with some suggesting potential benefits [32] and others, such as a recent controlled study by Fedorcsak [20], finding only modest improvements in declarative knowledge after a single EER session (Cohen d=0.22). In our study, we observed an immediate gain across 5 targeted learning objectives. However, in the absence of a control group and long-term follow-up, these findings should be interpreted cautiously. The structured, sequential format of the EER, aligned with the logic of a diagnostic study, may have supported knowledge acquisition, but the true impact on learning remains difficult to disentangle from engagement effects or short-term memory recall. Importantly, the intervention appeared accessible across subgroups, with no influence of previous escape room experience or self-perceived ease with critical appraisal. Beyond performance, the ability to foster interest in research methodology among a large cohort of students may in itself constitute a valuable educational outcome [12,15,19].

Our findings suggest a strong alignment between student feedback and our initial pedagogical intentions: namely, to introduce foundational research skills, foster collaborative teamwork, and demystify research training through an engaging and enjoyable format. Motivating undergraduate medical students to engage with research remains a recognized challenge [7,8,10]. In our study, students reported a marked shift in perception, from predominantly negative views on research training to positive reflections on the escape room experience. Free-text responses frequently emphasized the value of teamwork, a dimension often underrepresented in their curriculum, as well as the quality of facilitation and meaningful interactions with faculty. As observed in other EER-based interventions, we noted a high level of student cooperation and an intrinsic drive to complete the challenge within the allotted time [16,18,25,29,31]. This positive, collaborative atmosphere appeared to promote rich student-mentor interactions, an important benefit for a generation of learners shaped by recent experiences of social distancing and remote education [39].

This study has several limitations. As a pilot implementation, it was not designed to isolate the specific effects of the escape room format compared to traditional pedagogical approaches. The absence of a control group precludes any definitive conclusions about the causal impact of the intervention on student performance or attitudes. All outcome measures relied on self-reported or short-term evaluations, with no assessment of knowledge retention over time. Moreover, the educational content focused exclusively on diagnostic accuracy studies. While this choice was pedagogically motivated, it limits the generalizability of the findings to other types of research designs. This study was also conducted in a single institution, which may affect broader applicability. Future research should explore the added value of EERs in research training using comparative designs—such as cluster randomized trials—and include a broader range of research topics and longer-term follow-up to assess sustained learning outcomes. Despite these limitations, this study benefits from a high participation rate, detailed process documentation, and rich qualitative feedback, which collectively provide meaningful insights into the feasibility, perceived value, and immediate educational impact of this innovative teaching format.

In conclusion, this study demonstrates the feasibility of a large-scale, low-cost, and replicable EER to introduce undergraduate medical students to key concepts of diagnostic test evaluation. The format was well received and may offer a promising entry point into research training provided that its targeted scope and pedagogical objectives are clearly defined.

Acknowledgments

The authors would like to thank all of the faculty's services for their support in setting up the escape room experiment, from the academic services (Mrs Mathilde Moreno Garcia and Mrs Magali Lance) to the multimedia service (Mr Pierre Verquin, Mr Karim Bouadjla, and Mr Christian Delsol) and the IT department (Mr Arnaud Lecoq and Mr Jean-Christophe Alexandre) of the Faculty of Medicine at the University of Lille. They would also like to thank all the supervisors who agreed to participate in the different sessions (Dr Benoît Brassart, Dr Eole Nyangwile, Dr Gabrielle Lisembart, Dr Jean-Baptiste Gibier, Dr Julien Chapuis, Dr Mehdi Maanaoui, Dr Michaël Génin, Dr Pierre Dourlen, Dr Thavarak Ouk, Dr Victor Fages, Dr Victor Leblanc, Dr Victor Lestrade, and Professor Sébastien Aubert). This work was supported by internal educational funds from the Faculty of Medicine at the University of Lille. No external funding was received for this study.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request. In addition, all materials necessary to reproduce the educational escape room—including the full scenario, puzzles, facilitator instructions, and evaluation forms—are freely available upon request to the corresponding author for noncommercial academic use.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Immersive mission briefing shown before the start of the escape room. This introductory video was designed to immerse students in a fictional investigation scenario. It presented the mission's narrative context, introduced the main characters and storyline, and created a suspenseful atmosphere to enhance student engagement from the outset. The video served solely to set the scene and did not provide any instructions or clues related to the upcoming puzzles.
[MP4 File, 36320 KB - mededu_v11i1e71339_app1.mp4]

Multimedia Appendix 2

Content of the "true"-or-"false" questionnaire administered immediately before and after the escape room intervention. Items are grouped by learning objective (*study protocol*, *study population*, *gold standard*, *contingency table*, and *metric appraisal*). For each pretest item, the expected correct answer and the proportion of students who responded correctly at baseline are reported; identical or conceptually equivalent items were presented at the posttest time point to assess immediate knowledge gain.
[DOCX File, 36 KB - mededu_v11i1e71339_app2.docx]

Multimedia Appendix 3

Word clouds illustrating students' perceptions before and after the educational escape room (EER). The left panel reflects students' initial impressions of critical appraisal of scientific articles before the intervention, whereas the right panel captures their reactions to the EER format afterward. Comments were originally written in French and translated by the authors.
[PNG File, 106 KB - mededu_v11i1e71339_app3.png]

References

1. WFME standards for basic medical education 2020. World Federation for Medical Education. 2020. URL: https://wfme.org/download/wfme-standards-for-basic-medical-education-2020/ [accessed 2025-07-26]
2. Recommendations for preclerkship clinical skills education for undergraduate medical education. Association of American Medical Colleges. 2008. URL: https://www.stfm.org/media/1363/clinicalskills_oct09qxdpdf.pdf [accessed 2025-07-26]
3. Cooper N, Bartlett M, Gay S, et al. Consensus statement on the content of clinical reasoning curricula in undergraduate medical education. Med Teach 2021 Feb;43(2):152-159. [doi: 10.1080/0142159X.2020.1842343] [Medline: 33205693]
4. Dauchet L, Bentegeac R, Ghauss H, et al. Evaluating script concordance tests (SCTs) through the lens of Bayesian reasoning: enhancing assessment in medical education. J Epidemiol Popul Health 2025 Feb;73(1):202804. [doi: 10.1016/j.jeph.2024.202804] [Medline: 39848214]
5. Morgan DJ, Pineles L, Owczarzak J, et al. Accuracy of practitioner estimates of probability of diagnosis before and after testing. JAMA Intern Med 2021 Jun 1;181(6):747-755. [doi: 10.1001/jamainternmed.2021.0269] [Medline: 33818595]
6. Lakhlifi C, Lejeune FX, Rouault M, Khamassi M, Rohaut B. Illusion of knowledge in statistics among clinicians: evaluating the alignment between objective accuracy and subjective confidence, an online survey. Cogn Res Princ Implic 2023 Apr 20;8(1):23. [doi: 10.1186/s41235-023-00474-1] [Medline: 37081292]
7. Mlika M, Naceur A, Dziri C, et al. Critical appraisal of medical literature in undergraduate and postgraduate medical students. Front Educ 2022;7. [doi: 10.3389/feduc.2022.1036627]

8. Lee GS, Chin YH, Jiang AA, et al. Teaching medical research to medical students: a systematic review. Med Sci Educ 2021 Jan 8;31(2):945-962. [doi: 10.1007/s40670-020-01183-w] [Medline: 34457935]
9. Mai DH, Taylor-Fishwick JS, Sherred-Smith W, et al. Peer-developed modules on basic biostatistics and evidence-based medicine principles for undergraduate medical education. MedEdPORTAL 2020 Nov 24;16:11026. [doi: 10.15766/mep_2374-8265.11026] [Medline: 33274291]
10. Jegu J, Braun M, Pelaccia T. Quelle est la motivation des étudiants en médecine pour l'apprentissage de la lecture critique d'article? Pédagogie Médicale 2014 Nov 21;15(4):259-267. [doi: 10.1051/pmed/2014019]
11. Raynaud M, Goutaudier V, Louis K, et al. Impact of the COVID-19 pandemic on publication dynamics and non-COVID-19 research production. BMC Med Res Methodol 2021 Nov 22;21(1):255. [doi: 10.1186/s12874-021-01404-9] [Medline: 34809561]
12. Veldkamp A, van de Grint L, Knippels MC, van Joolingen WR. Escape education: a systematic review on escape rooms in education. Educ Res Rev 2020 Nov;31:100364. [doi: 10.1016/j.edurev.2020.100364]
13. Taraldsen LH, Haara FO, Lysne MS, Jensen PR, Jenssen ES. A review on use of escape rooms in education – touching the void. Educ Inq 2020 Dec 14;13(2):169-184. [doi: 10.1080/20004508.2020.1860284]
14. Davis K, Lo HY, Lichliter R, et al. Twelve tips for creating an escape room activity for medical education. Med Teach 2022 Apr;44(4):366-371. [doi: 10.1080/0142159X.2021.1909715] [Medline: 33872114]
15. Quek LH, Tan AJ, Sim MJ, et al. Educational escape rooms for healthcare students: a systematic review. Nurse Educ Today 2024 Jan;132:106004. [doi: 10.1016/j.nedt.2023.106004] [Medline: 37924674]
16. Kinio AE, Dufresne L, Brandys T, Jetty P. Break out of the classroom: the use of escape rooms as an alternative teaching strategy in surgical education. J Surg Educ 2019;76(1):134-139. [doi: 10.1016/j.jsurg.2018.06.030] [Medline: 30126728]
17. Mirshahi A, Khanipour-Kencha A, Keyvanpour S, Motlagh MK. "MORAD ESCAPE", a novel research-based ESCAPE room approach for evaluating research competencies of health professions students. BMC Med Educ 2025 Feb 21;25(1):289. [doi: 10.1186/s12909-025-06781-z] [Medline: 39984933]
18. Guckian J, Sridhar A, Meggitt SJ. Exploring the perspectives of dermatology undergraduates with an escape room game. Clin Exp Dermatol 2020 Mar;45(2):153-158. [doi: 10.1111/ced.14039] [Medline: 31276227]
19. Guckian J, Eveson L, May H. The great escape? The rise of the escape room in medical education. Future Healthc J 2020 Jun;7(2):112-115. [doi: 10.7861/fhj.2020-0032] [Medline: 32550277]
20. Fedorcsak P. Moderate benefit of escape room game on learning outcome in medicine. BMC Med Educ 2024 Nov 23;24(1):1353. [doi: 10.1186/s12909-024-06352-8] [Medline: 39580402]
21. Reinert A. Une méthode de classification descendante hiérarchique: application à l'analyse lexicale par contexte. Cah Anal Données 1983;8(2):187-198 [FREE Full text]
22. Souza MD, Wall ML, Thuler ADM, Lowen IMV, Peres AM. The use of IRAMUTEQ software for data analysis in qualitative research. Rev Esc Enferm USP 2018 Oct 4;52:e03353. [doi: 10.1590/S1980-220X2017015003353] [Medline: 30304198]
23. Zhang XC, Lee H, Rodriguez C, Rudner J, Chan TM, Papanagnou D. Trapped as a group, escape as a team: applying gamification to incorporate team-building skills through an "escape room" experience. Cureus 2018 Mar 2;10(3):e2256. [doi: 10.7759/cureus.2256] [Medline: 29725559]
24. Abensur Vuillaume L, Laudren G, Bosio A, Thévenot P, Pelaccia T, Chauvin A. A didactic escape game for emergency medicine aimed at learning to work as a team and making diagnoses: methodology for game development. JMIR Serious Games 2021 Aug 31;9(3):e27291. [doi: 10.2196/27291] [Medline: 34463628]
25. Jambhekar K, Pahls RP, Deloney LA. Benefits of an escape room as a novel educational activity for radiology residents. Acad Radiol 2020 Feb;27(2):276-283. [doi: 10.1016/j.acra.2019.04.021] [Medline: 31160173]
26. Liu C, Patel R, Ogunjinmi B, et al. Feasibility of a paediatric radiology escape room for undergraduate education. Insights Imaging 2020 Mar 19;11(1):50. [doi: 10.1186/s13244-020-00856-9] [Medline: 32193698]
27. Kaul V, Morris A, Chae JM, Town JA, Kelly WF. Delivering a novel medical education "escape room" at a national scientific conference: first live, then pivoting to remote learning because of COVID-19. Chest 2021 Oct;160(4):1424-1432. [doi: 10.1016/j.chest.2021.04.069] [Medline: 34029564]
28. Alejandre C, Corniero P, Claret G, Alaez C, Esteban E, Jordan I. New resident training strategy based on gamification techniques: an escape room on sepsis in children. Children (Basel) 2022 Sep 30;9(10):1503. [doi: 10.3390/children9101503] [Medline: 36291439]
29. Khanna A, Ravindran A, Ewing B, et al. Escape MD: using an escape room as a gamified educational and skill-building teaching tool for internal medicine residents. Cureus 2021 Sep 27;13(9):e18314. [doi: 10.7759/cureus.18314] [Medline: 34725586]
30. Akatsu H, Shiima Y, Gomi H, et al. Teaching "medical interview and physical examination" from the very beginning of medical school and using "escape rooms" during the final assessment: achievements and educational impact in Japan. BMC Med Educ 2022 Jan 28;22(1):67. [doi: 10.1186/s12909-022-03130-2] [Medline: 35090459]
31. Dimeo SP, Astemborksi C, Smart J, Jones EL. A virtual escape room versus lecture on infectious disease content: effect on resident knowledge and motivation. West J Emerg Med 2022 Jan 3;23(1):9-14. [doi: 10.5811/westjem.2021.12.54010] [Medline: 35060853]

32. Carrasco-Gomez D, Chao-Écija A, López-González MV, Dawid-Milner MS. Impact of a peer-to-peer escape room activity in the learning of human physiology of medical students from the university of Málaga. Front Physiol 2023 Aug 30;14:1242847. [doi: 10.3389/fphys.2023.1242847] [Medline: 37711460]
33. Backhouse A, Malik M. Escape into patient safety: bringing human factors to life for medical students. BMJ Open Qual 2019 Mar 30;8(1):e000548. [doi: 10.1136/bmjoq-2018-000548] [Medline: 31206043]
34. Diemer G, Jaffe R, Papanagnou D, Zhang XC, Zavodnick J. Patient safety escape room: a graduate medical education simulation for event reporting. MedEdPORTAL 2019 Dec 27;15:10868. [doi: 10.15766/mep_2374-8265.10868] [Medline: 32342008]
35. Zhang XC, Diemer G, Lee H, Jaffe R, Papanagnou D. Finding the "QR" to patient safety: applying gamification to incorporate patient safety priorities through a simulated "escape room" experience. Cureus 2019 Feb 5;11(2):e4014. [doi: 10.7759/cureus.4014] [Medline: 31007972]
36. Bunton SA, Salsberg E. Impact of increasing class size. Acad Med 2009 Jan;84(1):8. [doi: 10.1097/ACM.0b013e318190164c] [Medline: 19116468]
37. Sassoon EC, Craig RC, O'Brien DG. The challenges of expanding medical student numbers in the UK: a scoping review. Future Healthc J 2025 Jun 28;12(3):100278. [doi: 10.1016/j.fhj.2025.100278] [Medline: 40755486]
38. Paterson-Brown S. Improving patient safety through education. BMJ 2011 Feb 9;342:d214. [doi: 10.1136/bmj.d214] [Medline: 21307108]
39. Rose S. Medical student education in the time of COVID-19. JAMA 2020 Jun 2;323(21):2131-2132. [doi: 10.1001/jama.2020.5227] [Medline: 32232420]

Abbreviations

EER: educational escape room

Edited by B Lesselroth; submitted 16.01.25; peer-reviewed by IK Singgih, K Filippou, K McConville; revised version received 29.07.25; accepted 09.09.25; published 08.12.25.

Please cite as:
Le Guellec B, Gauthier V, Lenain R, Nuytten A, Dauchet L, Bonneau B, Gerard E, Castandet C, Truffert P, Hazzan M, Amouyel P, Bentegeac R, Hamroun A
Engaging Undergraduate Medical Students With Introductory Research Training via an Educational Escape Room: Mixed Methods Evaluation
JMIR Med Educ 2025;11:e71339
URL: https://mededu.jmir.org/2025/1/e71339
doi:10.2196/71339

Gamified Learning in a Virtual World for Undergraduate Emergency Radiology Education: Quasi-Experimental Study

Alba Virtudes Pérez-Baena[1], MD; Teodoro Rudolphi-Solero[2], MD, PhD; Rocío Lorenzo-Álvarez[3], MD, PhD; Miguel José Ruiz-Gómez[2], PhD; Francisco Sendra-Portero[2], MD, PhD

[1]Department of Radiology, Hospital Comarcal de Antequera, Antequera, Spain

[2]Department of Radiology and Physical Medicine, Facultad de Medicina, Universidad de Málaga, Bvd. Luis Pasteur, 32., Málaga, Spain

[3]Department of Emergency and Intensive Care, Hospital de la Axarquía, Vélez-Málaga, Spain

Corresponding Author:
Francisco Sendra-Portero, MD, PhD
Department of Radiology and Physical Medicine, Facultad de Medicina, Universidad de Málaga, Bvd. Luis Pasteur, 32., Málaga, Spain

Abstract

Background: Emergency radiology is essential for future doctors, who will face urgent cases requiring radiologic diagnosis. Using virtual simulations, gamified clinical scenarios, and case-based learning enhances practical understanding, develops technical and communication skills, and fosters educational innovation.

Objective: This study aimed to assess the feasibility of learning emergency radiology in the virtual world Second Life (Linden Lab) through a gamified experience by evaluating team performance in clinical case resolution, individual performance on seminar assessments, and students' perceptions of the activity.

Methods: Teams of 3 - 4 final-year medical students, during a 2-week radiology clerkship, had access to 7 clinical cases in virtual clinical stations and were randomly assigned 2 to solve and submit. They later discussed the cases in a synchronous virtual meeting and attended an emergency radiology seminar. The experience was repeated over 2 consecutive years to assess reproducibility through comparison of learning outcomes and students' perceptions. Learning outcomes were evaluated through team-based case resolution and individual seminar assessments. Students' perceptions were gathered via a voluntary questionnaire including 5-point Likert scale items, cognitive load ratings, 10-point evaluations, and open-ended comments.

Results: In total, 182 students participated in 2020 - 2021 and 170 in 2021 - 2022, demonstrating strong team-based case resolution skills with mean scores of 7.36 (SD 1.35) and 8.41 (SD 0.99), respectively (P<.001). The perception questionnaire had a 90.6% response rate. The highest cognitive load was observed in avatar editing (median 7, 95% CI 6.56 - 6.96). Case-solving cognitive load was significantly lower in 2021 - 2022 compared with 2020 - 2021 (median 6, 95% CI 5.69 - 6.21 vs 5.10 - 5.66; P<.001). The students rated the experience highly, with average scores exceeding 8.0 out of 10 across various aspects. Notably, the highest-rated aspects were the teaching staff (9.13, SD 1.15), cases (8.60, SD 1.31), project organization (8.42, SD 1.67), and virtual rooms (8.36, SD 1.62). The lowest-rated aspect was internet connectivity (6.68, SD 2.53). Despite the positive scores, all aspects were rated significantly lower in 2021 - 2022 compared with 2020 - 2021. These year-to-year comparisons in performance and perception support the reproducibility of the experience.

Conclusions: This study demonstrates that a game-based learning experience in the Second Life virtual world, combining virtual clinical scenarios and team-based tasks, is feasible and reproducible within a radiology clerkship. Students showed strong performance in case resolution and rated the experience highly, within a playful context that integrated asynchronous and synchronous activities. Lower ratings in the second year may reflect contextual differences, such as changes in COVID-19 pandemic restrictions.

(JMIR Med Educ 2025;11:e68518) doi:10.2196/68518

KEYWORDS

radiology education; medical students; computer simulation; virtual worlds; emergency radiology; game-based learning; case-based learning

Introduction

Gamification is the use of game elements in nongame environments, such as in education. This alternative educational approach promotes student motivation and active participation in the learning process [1]. Its implementation has been shown to enhance engagement, increase motivation, improve knowledge retention, and develop problem-solving skills [2-4].

The effectiveness of gamification is rooted in the fulfillment of 3 basic psychological needs [2]: (1) autonomy—the need to feel in control of one's actions; (2) competence—the desire to achieve goals and experience success [5]; and (3) relatedness—the need to feel connected to others and be part of a group [6]. Through game-based tasks, students are encouraged to experiment, interact, and collaborate with peers [7]. However, gamification alone is not sufficient to ensure meaningful learning outcomes. Its effectiveness depends on thoughtful design and alignment with the specific skills students are expected to acquire [8]. To support knowledge acquisition and consolidation, game elements must be carefully selected based on academic goals, student needs, and a pedagogically sound teaching methodology [9]. In this context, gamification proves to be a viable strategy at the university level—particularly in undergraduate programs—as a means to fulfill the educational goals outlined in academic curricula [10]. Studies have reported positive academic outcomes, high levels of student engagement, and strong adherence to gamified activities [11-14]. At the postgraduate level, gamification has also found applications in health care education. It has been used with surgical residents through the Da Vinci simulator to enhance surgical skills [15], with internal medicine residents—both individually and in teams—via online platforms to update clinical knowledge [16], with radiology residents through virtual tools for interpreting chest x-rays [17], and with otorhinolaryngology residents to assess and train new laryngoplasty techniques [18].

Game-based learning, often referred to as digital games [19,20], supports students in achieving learning objectives through immersive and engaging learning experiences. These activities have shown promising results with medical students, especially when implemented in immersive digital platforms [21,22], and their use in undergraduate education is worth exploring. These immersive platforms, known as virtual worlds, are computer-generated 3D spaces where people interact remotely through representations of themselves called avatars. The concept of virtual worlds, along with others such as virtual reality, mirror worlds, and augmented reality, fits within the broader notion of the metaverse [23], a collective virtual shared space expressed through digital media and the internet. Clinical simulation environments in such settings have been developed to train various skills, including taking anamnesis from virtual patients [24], resolving clinical situations in a pneumology ward [21], training cardiopulmonary resuscitation [25], or developing communication skills with patients [26,27].

One of the most widely used virtual worlds for health professional education is Second Life (SL; Linden Lab) [28]. Its advantages include remote access, a strong sense of presence, ease of access, user anonymity, opportunities to develop communication skills, promotion of active learning, and being free of charge. SL allows educators to design and recreate clinical training scenarios [29], including those based on the Objective Structured Clinical Examination (OSCE) format, in which specific "stations" simulate clinical cases in a standardized, reliable, and objective way [30]. It has already proven to be a valuable tool for teaching radiology in both synchronous and asynchronous online formats [31], and it has supported interactive learning focused on radiological anatomy and imaging signs [22]. However, to our knowledge, no experiences have yet incorporated radiology OSCE-like scenarios using gamified approaches in SL or similar environments.

Emergency radiology is essential for medical students, as future doctors will inevitably encounter urgent clinical scenarios requiring timely and accurate radiologic diagnosis. The ability to interpret commonly used x-ray examinations in the primary assessment of acutely unwell patients, and to recognize basic findings on emergency computed tomography (CT) scans, is considered a core competency in undergraduate medical training [32]. The European Society of Radiology highlights the importance of educating medical students in emergency radiology to ensure they understand imaging's critical role in acute care and can contribute effectively in managing clinical emergencies [33]. Furthermore, teaching emergency radiology to medical students has been shown to enhance their knowledge, promote appropriate imaging usage, support evidence-based decision-making, and increase awareness of the potential downstream effects of incidental findings [34].

This study aimed to assess the feasibility of learning emergency radiology in the virtual world SL through a gamified experience by evaluating team performance in clinical case resolution, individual performance on seminar assessments, and students' perceptions of the activity. The educational activity combined case-based problem-solving in OSCE-style stations with virtual online seminars. The activity was developed for sixth-year medical students (in their final year of medical school in our country) as part of a formal radiology clerkship.

Methods

Overview

This was a quasi-experimental study conducted over 2 academic years to evaluate a virtual, gamified radiology learning experience.

Background and Project Design

Sixth-year medical students at our university complete a 2-week radiology clerkship, consisting of 4 days of hospital practice, ten 2-hour seminars on clinical radiology, and online learning activities. Each year, 7 groups of 24 - 28 students successively complete the radiology clerkship between mid-October and early February. Eligibility criteria included enrollment in this clerkship during the 2020 - 2021 or 2021 - 2022 academic years. All students met these criteria, and no exclusions were applied. The educational purpose of the activity in this study, called "Rainbow-Game," was to provide students with reflective training on clinical situations involving medical emergencies mediated by the corresponding radiological procedures, in a playful and collaborative context, as part of the activities of the clerkship.

Each group was randomly divided into 7 teams of 3 - 4 students named with the colors of the rainbow. On the first day of the clerkship, students were briefed on the details of this experience. They had to dress their avatar in shirts of their team's color and had 9 days to visit, on demand, a virtual space with 7 OSCE

stations, each containing an emergency radiology case. Students had unrestricted access to all stations throughout this period, encouraging self-directed learning and flexibility in how they engaged with the cases. Although they were encouraged to review all 7 cases, only 2 were randomly assigned to each team on the eighth day for formal written resolution and submission via SL's internal messaging system (notecard). To support them, they received a short orientation on how to use SL and step-by-step mini-tutorials to assist with independent navigation, creating and sending notecards, and customizing their avatars with the team T-shirts. Although students used their own devices, they were advised to ensure basic technical compatibility and stable internet access. As part of the game, teams also had to send the professor original and imaginative SL screenshots showing their avatars dressed in the corresponding color. On the tenth day, there was a 2-hour meeting with the whole group in SL. First, the teams had to orally present the results of the assigned cases and discuss them with their peers. Subsequently, they received a 1-hour seminar on emergency radiology, in which the professor presented 15 clinical cases, distributed equally among head, chest, and abdominal emergencies. At the end of the seminar, 3 exam cases were arranged to be answered individually using a notecard. A 24-item checklist was used to correct answers uniformly (see Multimedia Appendix 1). This activity was repeated during 2 academic years (2020 - 2021 and 2021 - 2022) with the same contents and organization. The TREND (Transparent Reporting of Evaluations with Nonrandomized Designs) statement [35] was followed to ensure transparent and comprehensive reporting.

Outcomes and Assessment

The study evaluated three categories of outcomes: (1) team performance in clinical case resolution, (2) individual performance on seminar assessments, and (3) students' perceptions of the activity. Team performance was scored using a structured checklist (0 - 10 points per case), and individual seminar responses were graded using a 24-item checklist (Multimedia Appendix 1). In both cases, higher scores indicated better performance.

Students' perceptions were captured through a questionnaire adapted from a previously validated instrument [22] (Multimedia Appendix 2) including 5-point Likert items (ordinal), 9-point cognitive load ratings (ordinal), 10-point satisfaction ratings (continuous), and open-ended comments. Higher Likert scores reflected more positive perceptions, higher cognitive load scores indicated greater perceived mental effort, and higher satisfaction ratings indicated greater satisfaction.

Virtual Scenarios

This study was conducted at the SL location named "The Medical Master Island," designed with various academic buildings surrounded by trees and plants (Figure 1A). Seven OSCE stations were designed with access from the same distributor, imitating a real OSCE. Each station had: (1) an access door; (2) a panel on the wall describing the clinical situation and the questions to be answered; (3) a table with one or 2 monitors, showing the images of the case; and (4) x-ray or CT equipment to contextualize the place as a radiology room (Figure 1B). To minimize repetitions, 16 cases were used (Table 1), which were rotated for the 7 groups (Table 2). An example case is shown in Figure 2. The clinical situation presented to the students is that of a 65-year-old woman who went to the emergency department due to progressive dyspnea occurring with minimal effort, reporting weight loss over the last month. Her personal history includes smoking and dyslipidemia. On examination, she is conscious, oriented, and cooperative, with mild respiratory retractions. Global hypoventilation in the left lung is noted. Oxygen saturation is 85%. She is afebrile. An imaging test is performed. Tasks to be carried out included describing as accurately as possible the imaging technique used and the pathological findings, establishing a differential diagnosis, and outlining the clinical approach to the suspected condition. Indicate what other tests would be required and why. The group meeting was held in the main building's aula magna, which had monitors with images to discuss the cases and a panel of slides to present the seminar (Figure 1C). Students submitted screenshots of their teams as part of the competition (Figure 1D).

Figure 1. Various scenes during the Second Life virtual world experience. (A) Aerial view of Medical Master Island with a flying avatar in the foreground. In the background, the main building, where the synchronous seminars were held, can be seen. (B) Example of an Objective Structured Clinical Examination station set up as a radiology room, with a team of students reviewing a head computed tomography case. The description of the clinical situation and the tasks to be performed are displayed on the wall. (C) Scene during a synchronous meeting with a student group, reviewing one of the Objective Structured Clinical Examination cases. (D) Screenshot submitted by a team of students as part of the competition.

Table . Description of the 16 cases used in the Rainbow-Game.

Case	Image modality shown	Case description
1	Brain CT[a] without IV[b] contrast: 18 axial slices.	52-year-old man. Spontaneous intraparenchymal hemorrhage with a subarachnoid component.
2	Abdominal CT with IV contrast: 18 axial slices and 18 coronal slices.	39-year-old woman. Multiple renal lacerations with hemoperitoneum.
3	Chest x-ray: posteroanterior.	65-year-old woman. Complete opacification of the left hemithorax: atelectasis of left lung due to a bronchial carcinoma.
4	Brain CT without IV contrast: 18 axial slices.	79-year-old man. Acute ischemic lesion in the left cerebral hemisphere: MCA stroke with subfalcine herniation.
5	Abdominal radiography: oblique.	73-year-old man. Large bowel dilatation with coffee bean sign: Acute sigmoid volvulus with pneumoperitoneum.
6	Abdominal CT with IV contrast: 18 axial slices.	55-year-old woman. Gallbladder distention, wall thickening, mucosal hyperenhancement, and pericholecystic fat stranding: Acute cholecystitis.
7	Brain CT without IV contrast: 18 axial slices plus 18 in bone window.	40-year-old man. Parietal acute epidural hematoma with associated bone fracture.
8	Chest x-ray: posteroanterior.	59-year-old woman. Widening of the mediastinum, wide aortic contour, tracheal deviation, aortic kinking: Acute aortic dissection.
9	Abdominal CT with IV contrast: 18 axial slices.	55-year-old man. Ruptured abdominal aortic aneurysm with hemoperitoneum.
10	Brain CT without IV contrast: 18 axial slices plus 18 in bone window.	58-year-old man. Frontal intraparenchymal hemorrhage, subdural hemorrhage, and occipital skull fracture.
11	Chest x-ray: posteroanterior.	65-year-old man. Pulmonary consolidation without volume loss, air bronchogram, and silhouette sign: Lobar pneumonia.
12	Abdominal CT with IV contrast: 18 axial slices.	60-year-old man. Colonic wall thickening, pericolic fat stranding in an area of sigmoid diverticulosis: Acute diverticulitis.
13	Chest x-ray: posteroanterior.	62-year-old man. Consolidations and ground-glass opacities bilateral, peripheral, and located in the lower fields: COVID-19 pneumonia.
14	Brain CT with and without IV contrast: 2 sets of 18 axial slices.	55-year-old woman. Single cortical lesion, round, well-demarcated with enhancement and perilesional vasogenic edema: Metastatic lesion.
15	Abdominal CT with IV contrast: 18 axial slices and 18 coronal slices.	49-year-old man. Large soft-tissue mass, with internal heterogeneity: Retroperitoneal sarcoma.
16	Chest x-ray: posteroanterior and lateral.	76-year-old man. Bone osteoblastic lesions: Metastasis due to prostate carcinoma.

[a]CT: computed tomography

[b]IV: intravenous.

Table . Assignment of cases to the Objective Structured Clinical Examination stations across different groups.

Groups[a]	Station 1	Station 2	Station 3	Station 4	Station 5	Station 6	Station 7
Group 1	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7
Group 2	Case 8	Case 9	Case 10	Case 11	Case 12	Case 13	Case 14
Group 3	Case 15	Case 16	Case 1	Case 2	Case 3	Case 4	Case 5
Group 4	Case 6	Case 7	Case 8	Case 9	Case 10	Case 11	Case 12
Group 5	Case 13	Case 14	Case 15	Case 16	Case 1	Case 2	Case 3
Group 6	Case 4	Case 5	Case 6	Case 7	Case 8	Case 9	Case 10
Group 7	Case 11	Case 12	Case 13	Case 14	Case 15	Case 16	Case 1

[a] The same cases were used for the 7 consecutive groups of students in both the 2020–2021 and 2021–2022 academic years.

Figure 2. A chest x-ray corresponding to case number 3 shows complete opacification of the left hemithorax due to atelectasis of the left lung caused by bronchial carcinoma.

Qualification of the Participants

Each team was qualified based on 3 parameters, normalized to 10 points each: (1) the written response to the 2 assigned cases, using a checklist for each case (40%); (2) the originality and quality of the best screenshot submitted (20%); and (3) the average of the individual points of the team to the 3 seminar questions (40%). Students were informed that their team's score on this compulsory online activity would have no impact on their clerkship grades. However, as extrinsic motivation, the team with the highest score in each group received a bonus of one extra point in the final grade, up to 10 points.

Perception of the Experience by the Students

After the seminar, the students were asked to complete a questionnaire about the experience (see Multimedia Appendix 2), which included: (1) a dichotomous question about whether they knew SL before the experience; (2) an assessment of 13

aspects of the game with a Likert scale of 1 - 5 (from totally disagree to totally agree); (3) an assessment of the cognitive load of 6 aspects of the game, following the 9-point scale proposed by Paas and Merriënboer [36]; (4) a rating of up to 10 points on 9 aspects of the game; and (5) a space for open comments asking "anything else to add." The questionnaire contained questions constructed, worded, and validated in previous studies [22,31].

Data Analysis

Descriptive statistics were performed using Microsoft Excel 2021 to characterize the population and subpopulations of participants, and the SPSS statistical package v24 (IBM Corporation) was used for statistical analysis. The normality of the data was assessed using the Shapiro–Wilk test. Based on the distribution of each variable, unpaired 2-sample *t* tests were used for continuous variables that met the normality assumption, while Mann–Whitney *U* tests were applied to compare ordinal or nonnormally distributed variables. Statistical significance was accepted at a probability of error of $P<.05$. Reproducibility was assessed by comparing student performance (project scores and seminar results) and perception data (questionnaire responses) across the 2 academic years.

Open comments were analyzed using the systematic collaborative consensus coding by committee [37]. Comments related to aspects of the clerkship other than the Rainbow-Game were not considered. During 2 consensus meetings, a 2-layer hierarchical coding was established, with 3 codes in the first layer (advantages, disadvantages, and suggestions) and different subcodes in the second layer.

Ethical Considerations

This study received approval from the Institutional Ethics Committee for Experimentation at the University of Malaga (decision number 141 - 2022-H; approval date: January 18, 2023). Questionnaire participation was voluntary, and participants gave their explicit informed consent at the time of submission. All data collected were anonymized before analysis to ensure the privacy and confidentiality of participants. Students were informed that their participation would not affect their academic evaluation and had the opportunity to opt out without any consequences. No monetary or material compensation was provided for participation in the study.

Results

Outcome of the Rainbow-Game

Three hundred and fifty-two students participated in this project, 182 in 2020 - 2021 and 170 in 2021 - 2022. Of the 352 students, 227 (64.5%) were women and 125 (35.5%) were men. The mean age was 23.7 (SD 2.8) years, with a median of 23 and a range from 22 to 51 years.

The normality of the distributions was assessed using the Shapiro–Wilk test. OSCE case scores (2020: W=0.934, $P=.009$; 2021: W=0.891, $P<.001$) and the final experience score (2020: W=0.874, $P<.001$; 2021: W=0.944, $P=.02$) did not follow a normal distribution and were compared using the Mann–Whitney *U* test. In contrast, the results of the seminar questions (2020: W=0.966, $P=.17$; 2021: W=0.981, $P=.59$) and academic clerkship grades (2020: W=0.968, $P=.20$; 2021: W=0.967, $P=.18$) followed a normal distribution and were compared using unpaired 2-sample *t* tests. The average rating up to 10 points of the teams in 2020 - 2021 versus 2021 - 2022 was as follows: in OSCE cases, 7.36 (SD 1.35) versus 8.41 (SD 0.99*;* $P<.001$); screenshots taken in SL, 7.31 (SD 1.65) versus 7.04 (SD 1.73; $P=.35$); the seminar questions 4.98 (SD 0.83) versus 4.41 (SD 1.07; $P=.004$); and the final score of the experience, 6.40 (SD 0.86) versus 6.54 (SD 0.77; $P=.40$). Figure 3 shows the average score of the 7 groups for each year in chronological order. There was no difference in the academic grade of the clerkship between the students of both years (7.68, SD 0.44 vs 7.79, SD 0.43; $P=.19$).

In Figure 3, the final scores were obtained by considering the percentage contribution of the OSCE cases (40%), the SL screenshots (20%), and the seminar test (40%). The points represent the mean values of the 7 teams in each group. The solid line corresponds to the 2020 - 2021 academic year, and the dashed line to 2021 - 2022. The average score obtained in the 16 OSCE cases is shown in Figure 4. In 5 cases, the score obtained in 2021 - 2022 was significantly higher than in 2020 - 2021; there were no other significant differences. Considering both years, the easiest case was number 9, a ruptured abdominal aortic aneurysm with hemoperitoneum on CT (mean score 9.58, SD 0.79), and the most difficult was number 16, osteoblastic bone metastases from prostate carcinoma in a chest x-ray (5.83, SD 2.25).

Figure 3. Scoring of the different components of the Rainbow-Game educational experience. OSCE: Objective Structured Clinical Examination; SL pics: Second Life pictures.

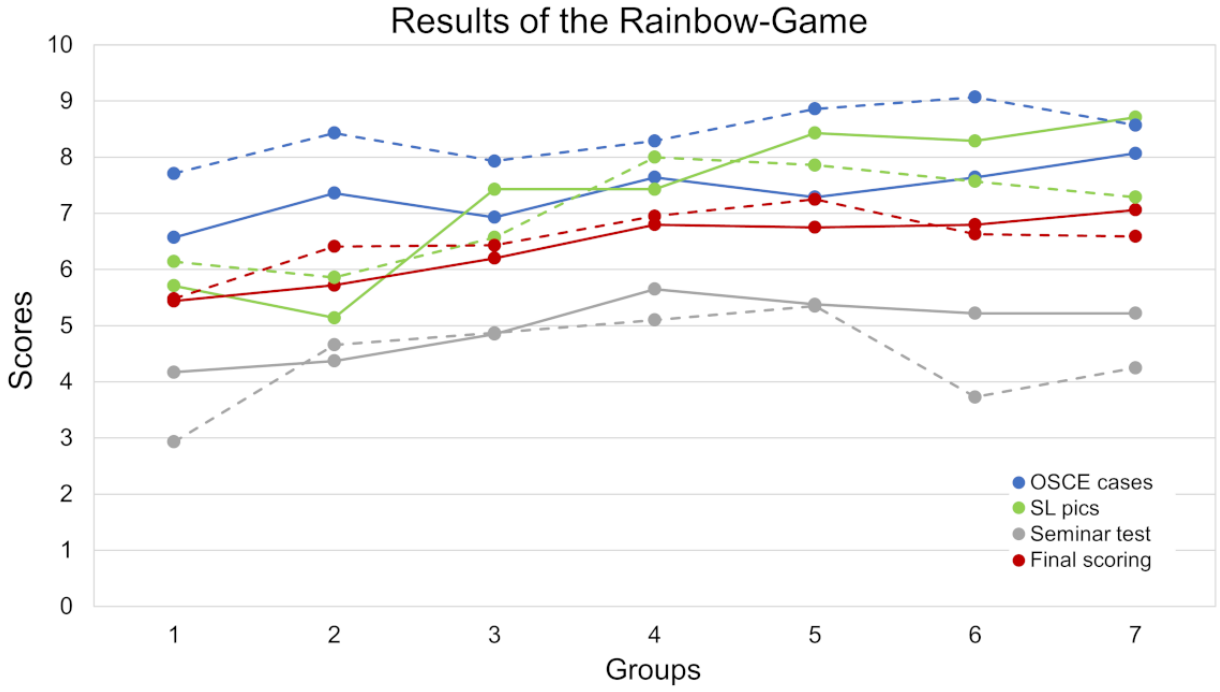
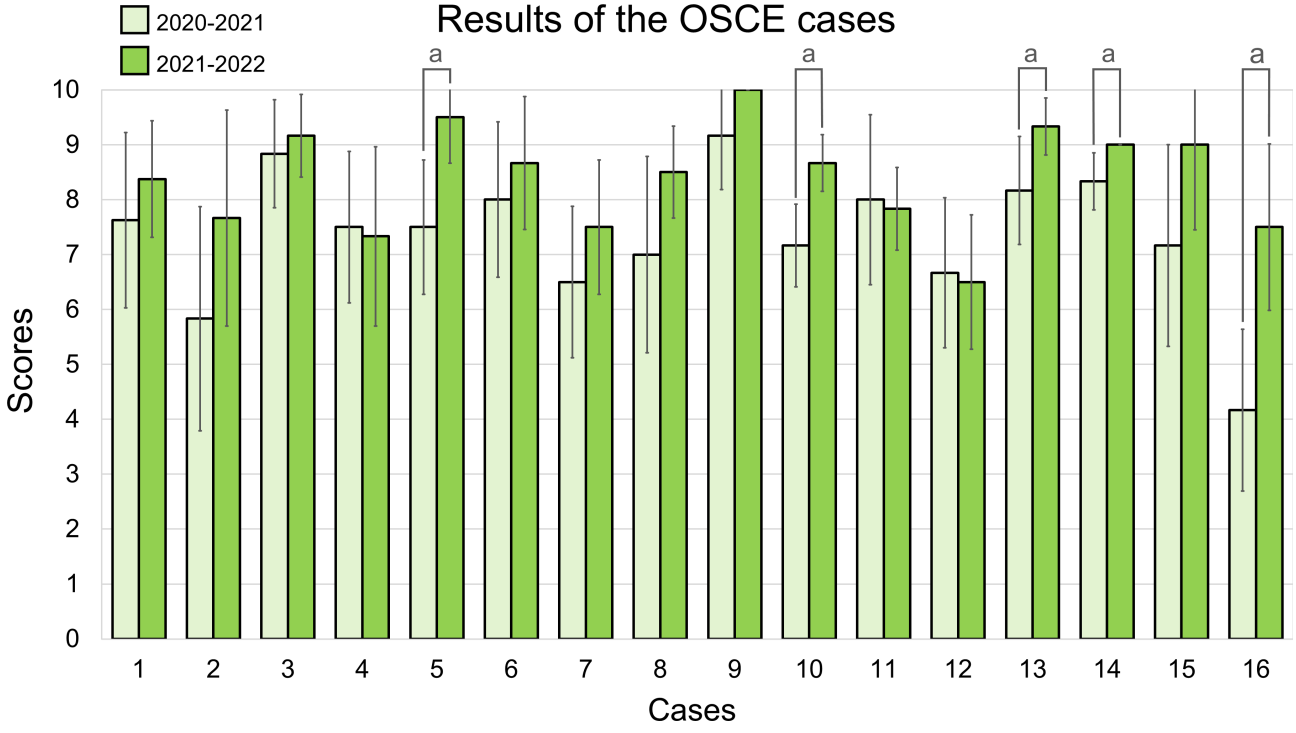


Figure 4. Bar chart with the average score obtained for each case over 2 consecutive years. Each year, 6 teams evaluated the cases, except for case 1, which was evaluated by 8 teams. Error bars represent the SD. Statistically significant differences, with $P<.05$, are identified with the letter "a." OSCE: Objective Structured Clinical Examination;



Students' Perception

Three hundred and nineteen students (90.6%) submitted the perception questionnaire, 167 (91.8%) in 2020 - 2021 and 152 (89.4%) in 2021 - 2022. Only 38 (11.9%) stated that they did not know SL previously, the rest had participated in educational activities in SL during their third year. All items from the questionnaire showed significant deviation from normality, with Shapiro–Wilk tests yielding $P<.001$ in both cohorts. Consequently, all comparisons were performed using the Mann–Whitney *U* test.

The results regarding cognitive load are summarized in Table 3. The task that required the greatest mental effort was editing and dressing the avatar, with an average score between rather high and high mental effort (median 7, 95% CI 6.56 - 6.96),

followed by solving the cases proposed in the OSCE stations, with an average score between no mental effort and rather high mental effort (median 6, 95% CI 5.48 - 5.86). The rest of the tasks required from rather low to very low mental effort. There were no significant differences in cognitive load between the 2 years except for the tasks of dressing the avatar, which required greater mental effort in 2021 - 2022, and the resolution of OSCE cases, which required less mental effort in 2021 - 2022.

Table . Results of the questionnaire about the cognitive load.

How much mental effort does it cost you to develop the following tasks?[a]	2020 - 2021		2021 - 2022		*P* value[b]	Both years	
	Mean (SD)	Median (95% CI)	Mean (SD)	Median (95% CI)		Mean (SD)	Median (95% CI)
Moving around in Second Life	3.85 (2.25)	4 (3.51 - 4.19)	3.93 (2.25)	3 (3.57 - 4.29)	.59	3.89 (2.25)	3 (3.64 - 4.14)
Communicate by written chat	2.07 (1.52)	2 (1.84 - 2.30)	2.23 (1.69)	2 (1.96 - 2.50)	.50	2.15 (1.60)	2 (1.97 - 2.33)
Communicate by voice	2.67 (2.14)	2 (2.34 - 3.00)	2.99 (2.34)	2 (2.62 - 3.36)	.19	2.83 (2.24)	2 (2.58 - 3.08)
Edit and dress your avatar	6.62 (1.81)	7 (6.34 - 6.90)	6.90 (1.85)	7 (6.61 - 7.19)	.08	6.76 (1.83)	7 (6.56 - 6.96)
Solve the proposed cases in the OSCE-RX[c] room	5.95 (1.71)	6 (5.69 - 6.21)	5.38 (1.73)	6 (5.10 - 5.66)	<.001	5.67 (1.74)	6 (5.48 - 5.86)
Follow the development of the seminar in Second Life	4.54 (2.26)	5 (4.19 - 4.89)	4.44 (1.92)	5 (4.14 - 4.74)	.33	4.49 (2.10)	5 (4.26 - 4.72)

[a]Likert scale from 1 to 9 according to: (1) Very, very low mental effort; (2) Very low mental effort; (3) Low mental effort; (4) Rather low mental effort; (5) Neither high nor low mental effort; (6) Rather high mental effort; (7) High mental effort; (8) Very high mental effort; and (9) Very, very high mental effort.

[b]*P* is the probability of error of the Mann Whitney *U* test. Statisitical significance set at $P<.05$.

[c]OSCE-RX: Objective Structured Clinical Examination: Radiology.

Overall, more than 95% of the respondents agreed or strongly agreed that the OSCE case selection was suitable for their training, the contents were appropriate, and that they worked as a team. In addition, more than 79% found the environment of the OSCE rooms attractive, the competition design appropriate, and the information provided adequate (Figure 5). Between 49% and 56% agreed that learning radiology in SL is interesting and that playing and competing is a better way to learn. Twenty-two percent of students disagreed with the suitability of their computers, and 9% disagreed with the adequacy of their internet connection for working in SL. Fifty-four percent of the 2021 - 2022 students agreed that they had fun during the experience. There was significantly lower agreement in 2021 - 2022 on 8 of the 5-point Likert scale statements.

The participants rated the experience up to 10 points, with mean scores higher than 8 points in 7 of 9 items (Table 4). The lowest rating was given for connectivity to SL. All mean scores in 2021 - 2022 were significantly lower than those in 2020 - 2021. One hundred and forty-three questionnaires (44.8%) included open comments: 82 (49.1%) in 2020 - 2021, and 61 (40.1%) in 2021 - 2022. After the initial first-layer coding, 12 second-layer subcodes were found among the advantages, 8 among the disadvantages, and 9 among the suggestions, as shown in Table 5, along with the thematic description and frequency.

Figure 5. Diagram showing the degree of agreement, expressed as a percentage, reached each year in the responses to the experience evaluation questionnaire, which used a 1 - 5 Likert scale. OSCE: Objective Structured Clinical Examination;

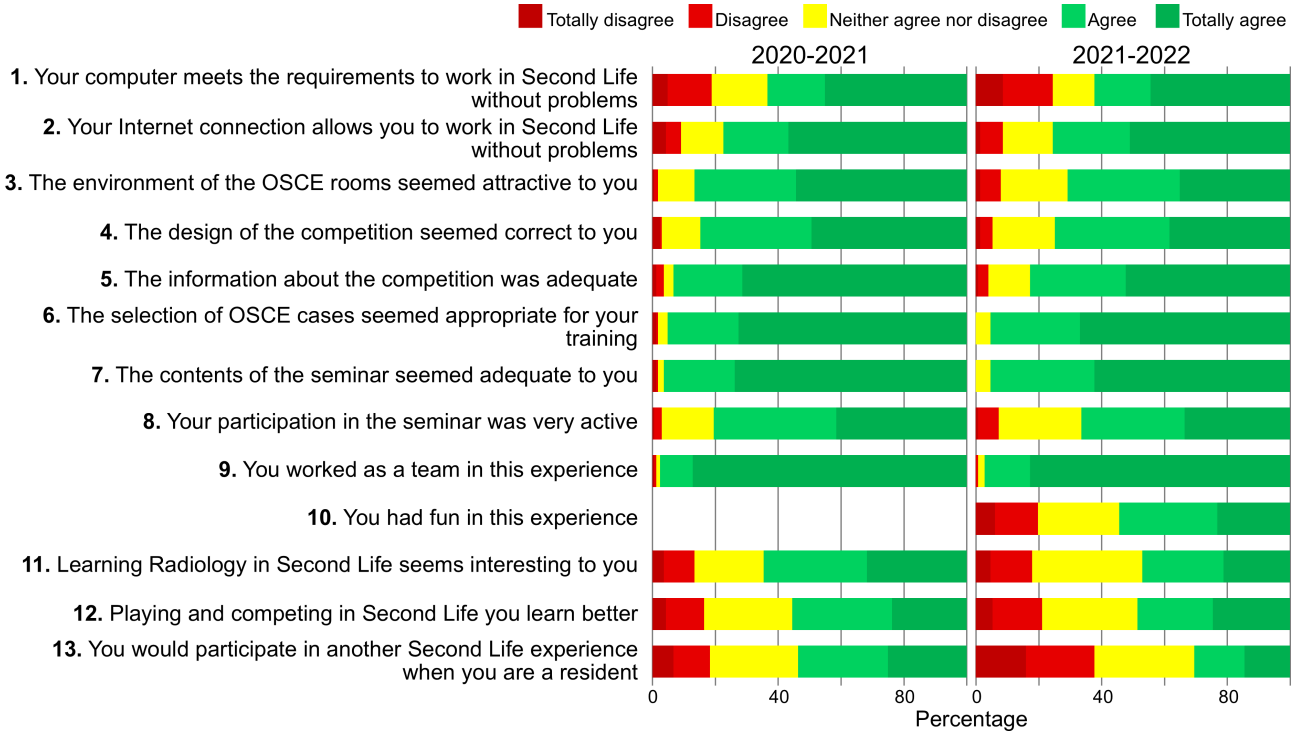


Table . Rating in a 0 - 10 points scale of various aspects of the experience.

Items	2020 - 2021, mean (SD)	2021 - 2022, mean (SD)	*P* value[a]	Both years, mean (SD)
Overall experience	7.81 (1.70)	7.07 (1.72)	<.001	7.46 (1.75)
Organization of the project	8,88 (1.45)	7.92 (1.75)	<.001	8.42 (1.67)
Environment of the OSCE[b] rooms	8.66 (1.44)	8.03 (1.74)	<.001	8.36 (1.62)
OSCE cases	8.75 (1.35)	8.43 (1.25)	.007	8.60 (1.31)
The virtual seminar	8.63 (1.66)	7.62 (2.03)	<.001	8.15 (1.91)
The teachers	9.30 (1.17)	8.93 (1.10)	<.001	9.13 (1.15)
The utility for your training	8.52 (1.59)	7.56 (1.98)	<.001	8.06 (1.85)
Interaction with peers	8.55 (1.65)	7.73 (1.99)	<.001	8.16 (1.86)
Connectivity to Second Life	7.22 (2.37)	6.09 (2.57)	<.001	6.68 (2.53)

[a]*P* is the probability of error of the Mann–Whitney *U* tests data. Statistical significance set at *P*<.05.

[b]OSCE: Objective Structured Clinical Examination

Table . Thematic codification of the open comments included in the questionnaire.

Codes and subcodes	2020 - 2021	2021 - 2022	Both
Advantages			
Appreciation: with terms like I liked, interesting, attractive, gratifying, enjoyable, positive, very cool, fantastic.	30	19	49
Acknowledgment: recognition, thanks to teachers for the effort, design, and organization.	20	9	29
Didactic: indicating that the experience is useful for learning, profitable, helpful, formative.	14	11	25
Playful: fun, entertaining, expressing that you learn by playing.	18	9	27
Innovative: also expressed with terms such as new, original, unusual, surprising, creative, different.	13	8	21
Teamwork: highlighting its importance, social contact, collaboration, coworking.	14	4	18
Cases: finding them interesting, of balanced difficulty, useful or of didactic value for active learning.	6	9	17
Seminar: emphasizing the interest in medical training, the educational value or the feeling of presence and dynamics as in the classroom.	11	4	15
COVID: a good solution or adequate for the pandemic situation, which allows them to maintain social contact.	7	3	10
T-shirt: together with the picture contest, expressing that they liked it, thought it was a creative idea, or that it favors a good atmosphere.	3	1	4
Guidelines: highlighting that they were good, useful, or detailed.	4	0	4
2D Platforms: preference over 2D platform platforms such as Zoom, Google Meet, Microsoft Teams, etc.	0	2	2
Disadvantages			
Technical problems: due to the computer or the Internet connection, the program does not run well.	22	20	42
Mild or occasional	15	15	30
Serious (prevents executing SL)[a]	7	5	12
Of them, resolved	6	4	10
Second Life (SL): running SL is complicated, even stressful. They feel that they are not used to it or that they do not handle the interface well. A learning curve is necessary	13	11	24
Dressing up: the tasks of dressing the avatar are difficult, complicated, time-consuming, or not important to them.	8	9	17
Time: this activity becomes more complicated in the last year of the degree, in which there is a lot of occupation with other activities and tasks.	1	15	16
Face-to-face: preference for face-to-face seminars in the real world.	5	6	11
Camera: problems using the camera (avatar vision) and seeing the images properly.	6	4	10
Notecards: problems sending notecards	0	2	2
2D Platforms: preference for 2D online platforms, such as Zoom, Google Meet, Microsoft Teams, etc.	1	1	2
Suggestions			

Codes and subcodes	2020 - 2021	2021 - 2022	Both
Voluntary: proposal that these activities be voluntary, even on vacation.	0	7	7
Other platforms: to be used as a resource when Second Life fails or there are connection problems.	3	1	4
More cases: include more cases, with more modalities, such as ultrasound, breast imaging, etc.	2	1	3
Training SL: provide training on handling SL in medical school, including tricks, handling avatars, etc.	2	2	4
Modifications: slight modifications to the current experience related to scheduling, team building, or information	2	2	4
SL proposals: new gamification proposals, learning strategies, and radiology repositories. Even do all the clerkship seminars in SL.	4	0	4
Patients: add virtual patients to OSCE stations	1	0	1
Computers SL: enable computers with the SL viewer in the Faculty of Medicine.	0	1	1
Ubiquity: carry out the experience with professors and students from other universities.	0	1	1

[a]SL: Second Life

Discussion

Principal Findings

Today's radiology students have different learning styles compared with previous generations, preferring to shift from the traditional lecture-based approach to more active learning methods that engage them more effectively [38]. This study presents a new educational experience for sixth-year medical students, conducted in SL during a 2-week radiology clerkship and replicated over 2 academic years. This educational experience is grounded in a constructivist and experiential learning approach [39], integrating several complementary pedagogical strategies. These include game-based learning to foster engagement and motivation [40]; case-based learning to stimulate clinical reasoning [38]; and team-based learning to promote collaboration and peer discussion [41]. The design also combines asynchronous and synchronous components to support flexibility and self-regulated learning. Simulation-based activities in a virtual world (SL) provide a safe, immersive environment for clinical skill development [42,43]. In addition, the experience draws on Self-Determination Theory [44], addressing both intrinsic motivation (interest and enjoyment) and extrinsic motivation (competition and rewards). Further pedagogical framing may be informed by Bers' Coding as Another Language approach [45], which conceptualizes digital environments as spaces for exploration, communication, and meaning-making—aligning with our use of SL for collaborative and immersive clinical learning.

Case-based learning connects theoretical knowledge with the clinical environment and encourages students to think like doctors. By reflecting on radiological images through clinical cases, students can appreciate the role of radiology in patient care [38]. The 24/7 availability of the content in SL enables the organization of asynchronous tasks, such as the 9-day period for evaluating OSCE cases in this study, adapting to students' study schedules, an essential factor for reflection and self-regulated learning [46]. Team-based learning is particularly suitable for visual topics like radiology, as it promotes and facilitates group discussion of complex, real-life radiological cases [47]. Teamwork dynamics are vital in undergraduate medical training, as they foster collaboration among students with varying levels of knowledge and experience [4,48].

Virtual world technologies, such as SL, make avatar-mediated student assessment in simulated 3D scenarios technically feasible to set up and run. This has been demonstrated in home accident scenarios in geriatric medicine [49], as well as in office or hospital settings [27], but has never been used to replicate scenarios for learning radiology. In this study, 7 OSCE stations were developed to simulate clinical scenarios involving medical emergencies, providing students with a variety of cases. The environment created allowed students to work in groups remotely at significantly lower cost than a similar scenario in the real world.

The educational activity in this study includes a significant component of competitive team gamification, a cooperative learning technique that enhances students' motivation and focus on learning tasks [48], combining group rewards with individual responsibility [50]. According to self-determination theory, motivation has 2 components [51,52]: intrinsic motivation, defined as participation in an activity because it is found to be inherently interesting and enjoyable, and extrinsic motivation, where participation is driven by external factors such as rewards, promotions, or the avoidance of academic failure. Competition is a powerful extrinsic motivator; however, it is criticized for creating high-pressure environments that reduce intrinsic motivation and hinder optimal learning [53]. Previous studies have shown that multi-user competitive games developed in SL through asynchronous activities can enhance medical students' learning of basic radiological content, such as anatomy and semiology, and that students find these games highly beneficial

for their training as physicians, both when competing individually [25,54] and in teams [55]. This study proposes a different approach to learning games, incorporating clinical content commonly found in medical practice, centered on case resolution through teamwork and supplemented by synchronous discussion and debate activities. In both courses, students recognized and appreciated the educational value of the Rainbow-Game experience.

SL is an appropriate environment to develop oral communication skills in clinical radiology through the presentation and discussion of content [56]. In this study, effective communication was fostered through group discussion and reflection on the cases. The experience concluded with a seminar on emergency radiology to reinforce students' knowledge and clinical reasoning. Seminars conducted in SL provide a strong sense of co-presence and have an educational impact comparable to those held in person, as long as the same objectives, content, and script are maintained [57].

Emergency radiology is an essential part of the undergraduate medical curriculum, ensuring that students can recognize major radiological emergencies [58]. In this study, students delved into clinical reasoning on cases commonly found in hospital emergency settings, achieving good results on the cases assigned to each team, with differences mainly related to the difficulty of the cases (Figure 4). This provides a measure of student performance on clinical cases in the final year of medical school. In contrast, the seminar exam questions were scored poorly due to limited response time and possibly an excessive number of items on the assessment checklist (see Multimedia Appendix 1). Students in 2021 - 2022 performed better on the cases than those in 2020 - 2021 and reported a lower cognitive load to solve them. This could indicate that they were better prepared for this type of tasks; however, they performed worse on the seminar questions, and there were no significant differences in the academic grades for the clerkship, so some leakage of results from one year to another cannot be ruled out. Although no traditional learning group was included, similar clerkship grades across both years suggest that the SL-based activity maintained academic performance while adding value through engagement and teamwork.

In general, the students perceived the experience as innovative and appropriate for their training. They valued the selected cases, the design and organization of the project, and the opportunity to work collaboratively in teams. In their open comments, they frequently described the activity as "fun," "interesting," and "useful for learning," highlighting its originality and the engaging nature of solving clinical cases in a virtual world. Several students also appreciated the social interaction and teamwork it encouraged, particularly under pandemic restrictions. However, some disadvantages were noted, particularly technical issues related to SL, such as interface difficulties or connectivity problems. Although students received technical support, using personal devices may have contributed to differences in their experiences. These issues may be related to the technical requirements of the platform, which must be installed locally and require specific hardware and internet conditions [59]. One student suggested enabling access to university computers with the SL viewer. The task of customizing avatars was often described as time-consuming and lacking clear educational purpose. A number of students, particularly in 2021 - 2022, felt the activity required too much time, and some suggested it should be optional or scheduled during a less demanding academic period.

These differences in perception may be attributed to the timing of both experiences during different phases of the COVID-19 pandemic. In the first semester of 2020 - 2021, students faced a stressful situation with restricted in-person academic activities, where online solutions were seen as essential for teaching. One year later, teaching had partially normalized, with some precautions still in place, but there was a certain sense of fatigue from the overuse of online activities [60]. This may have contributed to a greater perception of workload among the students and, consequently, a lower appreciation of the learning experience. Despite lower satisfaction scores in 2021 - 2022, students performed better on case resolution tasks and reported lower cognitive load. This suggests that learning outcomes do not always align with perceived satisfaction. Improved performance may reflect greater digital familiarity and self-regulation skills developed through previous remote learning. In contrast, lower ratings may have been influenced by reduced novelty, increased academic pressure, or persistent fatigue from extended online education, rather than the activity's educational value.

Comparison With Previous Work

This type of learning experience may be less appealing to those who are not drawn to gamification and technology. In fact, it has been shown that mandatory participation in gamification activities in SL can lead to a lower perception and acceptance of the game by a proportion of students [54]. Mandatory activities contribute to the extrinsic motivation of medical students, but "imposed" gamification has a counterproductive effect, described as "mandatory fun" [61]. User acceptance of 3D virtual world technology is positively influenced by their perception of the ease of use, usefulness, enjoyment, and visual appeal of a virtual learning environment, which significantly impacts student satisfaction, learning outcomes, and retention [62,63]. Most participants in the Rainbow-Game found the experience interesting and suitable for learning, and more than half found it fun. However, some students may not have perceived sufficient reward in a compulsory learning game. Therefore, the option for the best team in each group to earn an additional point on their academic grade compensates for this by providing an extrinsic motivation.

This study demonstrates how a mandatory game-based learning experience in SL can be added to the formal teaching organization of a radiology clerkship. Training in case resolution, peer discussion, and attending a seminar on emergency radiology provides added value to the educational objectives of this clerkship in the final year of the degree. The design of the cases and OSCE stations, along with the positive perception of students, is another added value. The immediate continuation of this project, which has already been completed and is pending publication, involved the development of individual, synchronous radiology OSCEs conducted over a restricted period of several minutes per station, emulating

face-to-face OSCEs, like those carried out at the end of the degree [64]. Future developments will focus on comparing the virtual world activity with a control group performing identical tasks either in real-world settings or in 2D virtual environments. This will allow for a more comprehensive evaluation of the added value of immersive 3D learning environments in radiology education.

Limitations

The main limitation of this experience, consistent with previous studies [31,57], was the presence of technical problems with the computers and the online connection needed to run SL correctly, as recognized in the open comments of 13% of the questionnaires. However, two-thirds of these issues were minor or occasional problems, and a third were resolved through simple solutions, such as borrowing a computer or using a cable connection. Since SL is developed as computer software, it requires suitable hardware that meets minimum requirements. In addition, SL is an internet-based technology with a client/server structure, which relies on stable internet connectivity. Despite the significant increase in internet access for higher education students, many still face restricted connectivity [63]. Another limitation to consider is that this study was conducted at a single institution. Although the experience has been replicated over 2 academic years, it would be valuable to see if it could be implemented elsewhere, which would require the cooperation of professors from other institutions and the inclusion of the project in the teaching guide for the corresponding courses.

Conclusions

This study presents a game-based radiology learning experience conducted in the SL virtual world, integrating simulated case-based learning in virtual OSCE stations, team-based competitive learning, and synchronous group sessions. The experience, adapted to a radiology clerkship, is feasible and reproducible, promoting clinical reasoning and teamwork among students in a playful context that they recognize and value highly. In addition, it is worth exploring the educational potential of OSCEs in 3D virtual environments for radiology and other medical disciplines.

Acknowledgments

This work was partially supported by the Educational Innovation Projects of the University of Malaga PIE19-217 and PIE22-045. Funding for open access charge was provided by Universidad de Málaga / CBUA. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors also want to acknowledge the students who participated in this study and voluntarily sent the project assessment questionnaires.

Data Availability

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

Authors' Contributions

AVP-B contributed to conceptualization, data curation, formal analysis, investigation, methodology, resources, validation, and visualization; prepared the original draft; and participated in reviewing and editing the manuscript. TR-S contributed to conceptualization, data curation, formal analysis, investigation, methodology, resources, software, validation, and visualization; and reviewed and edited the manuscript. RL-A contributed to conceptualization, data curation, investigation, resources, and visualization; and reviewed and edited the manuscript. MJR-G contributed to data curation, formal analysis, methodology, validation, and visualization; and reviewed and edited the manuscript. FS-P contributed to conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, and visualization; prepared the original draft; and reviewed and edited the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Presentation of the 16 clinical cases and the seminar exam cases, along with the checklist used for their correction.
[PDF File, 1274 KB - mededu_v11i1e68518_app1.pdf]

Multimedia Appendix 2
Perception questionnaire on the Rainbow-Game experience.
[PDF File, 211 KB - mededu_v11i1e68518_app2.pdf]

References

1. Aguado-Linares P, Sendra-Portero F. Gamification: basic concepts and applications in radiology. Radiologia (Engl Ed) 2023;65(2):122-132. [doi: 10.1016/j.rxeng.2022.10.014] [Medline: 37059578]

2. Van Roy R, Zaman B. Why gamification fails in education—and how to make it successful. In: Reiners T, Wood L, editors. Gamification in Education and Business: Springer; 2014:485-509. [doi: 10.1007/978-3-319-10208-5_22]
3. Barata G, Gama S, Jorge J, Gonçalves D. Gamification for smarter learning: tales from the trenches. Smart Learn Environ 2015 Dec;2(1):1-23. [doi: 10.1186/s40561-015-0017-8]
4. Dichev C, Dicheva D. Gamifying education: what is known, what is believed and what remains uncertain: a critical review. Int J Educ Technol High Educ 2017 Dec;14(1):9. [doi: 10.1186/s41239-017-0042-5]
5. Brigham TJ. An Introduction to gamification: adding game elements for engagement. Med Ref Serv Q 2015;34(4):471-480. [doi: 10.1080/02763869.2015.1082385] [Medline: 26496401]
6. Hense J, Klevers M, Sailer M, Horenburg T, Mandl H, Günthner W. Using gamification to enhance staff motivation in logistics. In: Meijer SA, Smeds R, editors. Frontiers in Gaming Simulation: Springer; 2014:206-213. [doi: 10.1007/978-3-319-04954-0_24]
7. Teixes Argilés F. Gamificación: Motivar Jugando, 1st edition: Editorial UOC; 2014.
8. Cuadros L, López A. Gamificación como estrategia para fortalecer la producción textual en Ciencias Naturales [Article in Spanish]. Rev Doc Univ 2020;21(1):55-79 [FREE Full text]
9. Mero J, Campuzano J, López S, Jara CH. La gamificación como estrategia para la estimulación del aprendizaje de las ciencias naturales [Article in Spanish]. Polo Conoc 2022;7(3):1335-1344. [doi: 10.23857/pc.v7i3.3795]
10. Rutledge C, Walsh CM, Swinger N, et al. Gamification in action: theoretical and practical considerations for medical educators. Acad Med 2018 Jul;93(7):1014-1020. [doi: 10.1097/ACM.0000000000002183] [Medline: 29465450]
11. Bochennek K, Wittekindt B, Zimmermann SY, Klingebiel T. More than mere games: a review of card and board games for medical education. Med Teach 2007 Nov;29(9):941-948. [doi: 10.1080/01421590701749813] [Medline: 18158669]
12. Akl EA, Pretorius RW, Sackett K, et al. The effect of educational games on medical students' learning outcomes: a systematic review: BEME Guide No 14. Med Teach 2010 Jan;32(1):16-27. [doi: 10.3109/01421590903473969] [Medline: 20095770]
13. Akl EA, Sackett KM, Erdley WS, et al. Educational games for health professionals. Cochrane Database Syst Rev 2013 Jan 31;3(1):CD006411. [doi: 10.1002/14651858.CD006411.pub3] [Medline: 23440807]
14. Janssen A, Shaw T, Goodyear P, Kerfoot BP, Bryce D. A little healthy competition: using mixed methods to pilot a team-based digital game for boosting medical student engagement with anatomy and histology content. BMC Med Educ 2015 Oct 12;15:173. [doi: 10.1186/s12909-015-0455-6] [Medline: 26459198]
15. Kerfoot BP, Kissane N. The use of gamification to boost residents' engagement in simulation training. JAMA Surg 2014 Nov;149(11):1208-1209. [doi: 10.1001/jamasurg.2014.1779] [Medline: 25229631]
16. Nevin CR, Westfall AO, Rodriguez JM, et al. Gamification as a tool for enhancing graduate medical education. Postgrad Med J 2014 Dec;90(1070):685-693. [doi: 10.1136/postgradmedj-2013-132486] [Medline: 25352673]
17. Chen PH, Roth H, Galperin-Aizenberg M, Ruutiainen AT, Gefter W, Cook TS. Improving abnormality detection on chest radiography using game-like reinforcement mechanics. Acad Radiol 2017 Nov;24(11):1428-1435. [doi: 10.1016/j.acra.2017.05.005] [Medline: 28647389]
18. Isaacson G, Ianacone DC, Soliman AMS. Ex vivo ovine model for suspension microlaryngoscopy training. J Laryngol Otol 2016 Oct;130(10):939-942. [doi: 10.1017/S0022215116008756] [Medline: 27572497]
19. Wouters P, van Oostendorp H. A meta-analytic review of the role of instructional support in game-based learning. Comput Educ 2013 Jan;60(1):412-425. [doi: 10.1016/j.compedu.2012.07.018]
20. Erhel S, Jamet E. Digital game-based learning: Impact of instructions and feedback on motivation and learning effectiveness. Comput Educ 2013 Sep;67:156-167. [doi: 10.1016/j.compedu.2013.02.019]
21. Toro-Troconis M, Kamat A, Partridge MR. Design and development of a component-based system for virtual patients in the virtual world of second life®. J Emerg Techol Web Intell 2011;3(4):308-316. [doi: 10.4304/jetwi.3.4.308-316]
22. Lorenzo-Alvarez R, Rudolphi-Solero T, Ruiz-Gomez MJ, Sendra-Portero F. Game-based learning in virtual worlds: a multiuser online game for medical undergraduate radiology education within Second Life. Anat Sci Educ 2020 Sep;13(5):602-617. [doi: 10.1002/ase.1927] [Medline: 31665564]
23. Kye B, Han N, Kim E, Park Y, Jo S. Educational applications of metaverse: possibilities and limitations. J Educ Eval Health Prof 2011;18:32. [doi: 10.3352/jeehp.2021.18.32]
24. Danforth DR, Procter M, Chen R, Johnson M, Heller R. Development of virtual patient simulations for medical education. J Virtual Worlds Res 2009;2(2):1-11. [doi: 10.4101/jvwr.v2i2.707]
25. Creutzfeldt J, Hedman L, Felländer-Tsai L. Cardiopulmonary resuscitation training by avatars: a qualitative study of medical students' experiences using a multiplayer virtual world. JMIR Serious Games 2016 Dec 16;4(2):e22. [doi: 10.2196/games.6448] [Medline: 27986645]
26. Lee J, Kim H, Kim KH, Jung D, Jowsey T, Webster CS. Effective virtual patient simulators for medical communication training: a systematic review. Med Educ 2020 Sep;54(9):786-795. [doi: 10.1111/medu.14152] [Medline: 32162355]
27. Kava BR, Andrade AD, Marcovich R, Idress T, Ruiz JG. Communication skills assessment using human avatars: piloting a virtual world objective structured clinical examination. Urol Pract 2017 Jan;4(1):76-84. [doi: 10.1016/j.urpr.2016.01.006] [Medline: 37592593]
28. Liaw SY, Carpio GAC, Lau Y, Tan SC, Lim WS, Goh PS. Multiuser virtual worlds in healthcare education: A systematic review. Nurse Educ Today 2018 Jun;65:136-149. [doi: 10.1016/j.nedt.2018.01.006]

29. Jivram T, Kavia S, Poulton E, Hernandez AS, Woodham LA, Poulton T. The development of a virtual world problem-based learning tutorial and comparison with interactive text-based tutorials. Front Digit Health 2021;3:611813. [doi: 10.3389/fdgth.2021.611813] [Medline: 34713092]
30. Staziaki PV, Sarangi R, Parikh UN, Brooks JG, LeBedis CA, Shaffer K. An objective structured clinical examination for medical student radiology clerkships: reproducibility study. JMIR Med Educ 2020 May 6;6(1):e15444. [doi: 10.2196/15444] [Medline: 32374267]
31. Lorenzo-Alvarez R, Pavia-Molina J, Sendra-Portero F. Exploring the potential of undergraduate radiology education in the virtual world Second Life with first-cycle and second-cycle Medical Students. Acad Radiol 2018 Aug;25(8):1087-1096. [doi: 10.1016/j.acra.2018.02.026] [Medline: 30782465]
32. Undergraduate radiology curriculum. The Royal College of Radiologists. 2022. URL: https://www.rcr.ac.uk/media/utcam3d4/rcr-curriculum_undergraduate-radiology-curriculum_february-2022.pdf [accessed 2025-03-22]
33. Modern radiology in medical education: chapter 20 – emergency radiology. European Society of Radiology. 2025. URL: https://www.myesr.org/app/uploads/2025/03/ESR_Modern-Radiology_eBook_Chapter_20v2.pdf [accessed 2025-03-22]
34. Leschied JR, Knoepp US, Hoff CN, et al. Emergency radiology elective improves second-year medical students' perceived confidence and knowledge of appropriate imaging utilization. Acad Radiol 2013 Sep;20(9):1168-1176. [doi: 10.1016/j.acra.2013.05.011]
35. Des Jarlais DC, Lyles C, Crepaz N, TREND Group. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. Am J Public Health 2004 Mar;94(3):361-366. [doi: 10.2105/ajph.94.3.361] [Medline: 14998794]
36. Paas FGWC, Van Merriënboer JJG. Instructional control of cognitive load in the training of complex cognitive tasks. Educ Psychol Rev 1994 Dec;6(4):351-371. [doi: 10.1007/BF02213420]
37. Saldaña J. The Coding Manual for Qualitative Researchers, 2nd edition: Sage Publications Ltd; 2013.
38. Fromke EJ, Jordan SG, Awan OA. Case-based learning: its importance in medical student education. Acad Radiol 2022 Aug;29(8):1284-1286. [doi: 10.1016/j.acra.2021.09.028] [Medline: 35835535]
39. Kolb AY, Kolb DA. Learning styles and learning spaces: enhancing experiential learning in higher education. AMLE 2005 Jun;4(2):193-212. [doi: 10.5465/amle.2005.17268566]
40. Plass JL, Homer BD, Kinzer CK. Foundations of game-based learning. Educ Psychol 2015 Oct 2;50(4):258-283. [doi: 10.1080/00461520.2015.1122533]
41. Michaelsen LK, Sweet M. The essential elements of team ‐ based learning. New Dir Teach Learn 2008 Dec;2008(116):7-27. [doi: 10.1002/tl.330]
42. Flowers MG, Aggarwal R. Second Life, a novel simulation platform for the training of surgical residents. Expert Rev Med Devices 2014 Mar;11(2):101-103. [doi: 10.1586/17434440.2014.863706] [Medline: 24308733]
43. Aebersold M, Tschannen D, Stephens M, Anderson P, Lei X. Second Life®: a new strategy in educating nursing students. Clin Simul Nurs 2012 Nov;8(9):e469-e475. [doi: 10.1016/j.ecns.2011.05.002]
44. Deci EL, Ryan RM. The "What" and "Why" of goal pursuits: human needs and the self-determination of behavior. Psychol Inq 2000 Oct;11(4):227-268. [doi: 10.1207/S15327965PLI1104_01]
45. Bers MU. Coding as another language: a pedagogical approach for teaching computer science in early childhood. J Comput Educ 2019 Dec;6(4):499-528. [doi: 10.1007/s40692-019-00147-3]
46. van Houten-Schat MA, Berkhout JJ, van Dijk N, Endedijk MD, Jaarsma ADC, Diemers AD. Self-regulated learning in the clinical context: a systematic review. Med Educ 2018 Oct;52(10):1008-1015. [doi: 10.1111/medu.13615] [Medline: 29943415]
47. Smeby SS, Lillebo B, Slørdahl TS, Berntsen EM. Express team-based learning (eTBL): a time-efficient TBL approach in neuroradiology. Acad Radiol 2020 Feb;27(2):284-290. [doi: 10.1016/j.acra.2019.04.022] [Medline: 31186155]
48. Van Gaalen AEJ, Jaarsma ADC, Georgiadis JR. Medical students' perceptions of play and learning: qualitative study with focus groups and thematic analysis. JMIR Serious Games 2021 Jul 28;9(3):e25637. [doi: 10.2196/25637] [Medline: 34319237]
49. Andrade AD, Cifuentes P, Oliveira MC, Anam R, Roos BA, Ruiz JG. Avatar-mediated home safety assessments: piloting a virtual objective structured clinical examination station. J Grad Med Educ 2011 Dec;3(4):541-545. [doi: 10.4300/JGME-D-11-00236.1] [Medline: 23205205]
50. Sánchez E. Competition and collaboration for game-based learning: a case study. In: Wouters P, van Oostendorp H, editors. Instructional Techniques to Facilitate Learning and Motivation of Serious Games Advances in Game-Based Learning: Springer; 2017. [doi: 10.1007/978-3-319-39298-1_9]
51. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 2000;55(1):68-78. [doi: 10.1037/0003-066X.55.1.68]
52. Gagné M, Deci EL. Self ‐ determination theory and work motivation. J Organ Behavior 2005 Jun;26(4):331-362. [doi: 10.1002/job.322]
53. Featherstone M, Habgood J. UniCraft: Exploring the impact of asynchronous multiplayer game elements in gamification. Int J Hum Comput Stud 2019 Jul;127:150-168. [doi: 10.1016/j.ijhcs.2018.05.006]

54. Rudolphi-Solero T, Lorenzo-Alvarez R, Ruiz-Gomez MJ, Sendra-Portero F. Impact of compulsory participation of medical students in a multiuser online game to learn radiological anatomy and radiological signs within the virtual world Second Life. Anat Sci Educ 2022 Aug;15(5):863-876. [doi: 10.1002/ase.2134] [Medline: 34449983]
55. Rudolphi-Solero T, Jimenez-Zayas A, Lorenzo-Alvarez R, Domínguez-Pinos D, Ruiz-Gomez MJ, Sendra-Portero F. A team-based competition for undergraduate medical students to learn radiology within the virtual world Second Life. Insights Imaging 2021 Jun 29;12(1):89. [doi: 10.1186/s13244-021-01032-3] [Medline: 34185165]
56. Pino-Postigo A, Domínguez-Pinos D, Lorenzo-Alvarez R, Pavía-Molina J, Ruiz-Gómez MJ, Sendra-Portero F. Improving oral presentation skills for radiology residents through clinical session meetings in the virtual world Second Life. Int J Environ Res Public Health 2023 Mar 8;20(6):4738. [doi: 10.3390/ijerph20064738] [Medline: 36981654]
57. Lorenzo-Alvarez R, Rudolphi-Solero T, Ruiz-Gomez MJ, Sendra-Portero F. Medical student education for abdominal radiographs in a 3D virtual classroom versus traditional classroom: A randomized controlled trial. AJR Am J Roentgenol 2019 Sep;213(3):644-650. [doi: 10.2214/AJR.19.21131] [Medline: 31287725]
58. Lewis PJ, Shaffer K. Developing a national medical student curriculum in radiology. J Am Coll Radiol 2005 Jan;2(1):8-11. [doi: 10.1016/j.jacr.2004.07.016] [Medline: 17411753]
59. Second Life system requirements. Second Life. URL: https://secondlife.com/system-requirements [accessed 2025-07-30]
60. de Oliveira Kubrusly Sobral JB, Lima DLF, Lima Rocha HA, et al. Active methodologies association with online learning fatigue among medical students. BMC Med Educ 2022 Feb 1;22(1):74. [doi: 10.1186/s12909-022-03143-x] [Medline: 35105362]
61. Mollick ER, Rothbard N. Mandatory Fun: Consent, Gamification and the Impact of Games at Work, 1st edition 2014. [doi: 10.2139/ssrn.2277103]
62. Ghanbarzadeh R, Hossein Ghapanchi A. Antecedents and consequences of user acceptance of three-dimensional virtual worlds in higher education. J Inf Technol Educ: Res 2020;19:855-889. [doi: 10.28945/4660]
63. Ghanbarzadeh R, Ghapanchi AH. Drivers of users' embracement of 3D digital educational spaces in higher education: a qualitative approach. Tech Know Learn 2023 Dec;28(4):1707-1744. [doi: 10.1007/s10758-022-09600-2]
64. García-Seoane JJ, Ramos-Rincón JM, Lara-Muñoz JP, el grupo de trabajo de la ECOE-CCS de la CNDFME, Miembros del grupo de trabajo de la ECOE-CCS de la CNDFME (por orden alfabético de Universidad). Changes in the Objective Structured Clinical Examination (OSCE) of University Schools of Medicine during COVID-19. Experience with a computer-based case simulation OSCE (CCS-OSCE). Rev Clin Esp 2021 Oct;221(8):456-463. [doi: 10.1016/j.rce.2021.01.004] [Medline: 33564195]

Abbreviations

CT: computed tomography
OSCE: Objective Structured Clinical Examination
SL: Second Life
TREND: Transparent Reporting of Evaluations with Nonrandomized Designs

Edited by TDA Cardoso; submitted 07.11.24; peer-reviewed by F Shojaei, M Gasmi, S Mitchell; revised version received 01.05.25; accepted 05.05.25; published 05.08.25.

Please cite as:
Pérez-Baena AV, Rudolphi-Solero T, Lorenzo-Álvarez R, Ruiz-Gómez MJ, Sendra-Portero F
Gamified Learning in a Virtual World for Undergraduate Emergency Radiology Education: Quasi-Experimental Study
JMIR Med Educ 2025;11:e68518
URL: https://mededu.jmir.org/2025/1/e68518
doi:10.2196/68518

Integrated e-Learning for Shoulder Anatomy and Clinical Examination Skills in First-Year Medical Students: Randomized Controlled Trial

Roland Koch[1], Dr med; Lena Gassner[1]; Navina Gerlach[1], MSc, MPH; Teresa Festl-Wietek[2], Dr; Bernhard Hirt[3], Prof Dr; Stefanie Joos[1], Prof Dr; Thomas Shiozawa[3], Dr med, MME

[1]Institute for General Practice and Interprofessional Care, University Hospital Tübingen, Osianderstr. 5, Tübingen, Germany

[2]TIME - Tuebingen Institute for Medical Education, Medical Faculty, University of Tuebingen, Tübingen, Germany

[3]Department of Anatomy, Institute of Clinical Anatomy and Cell Analysis, Faculty of Medicine, Eberhard Karls University of Tübingen, Tübingen, Germany

Corresponding Author:
Roland Koch, Dr med
Institute for General Practice and Interprofessional Care, University Hospital Tübingen, Osianderstr. 5, Tübingen, Germany

Abstract

Background: Applying functional anatomy to clinical examination techniques in shoulder examination is challenging for physicians at all learning stages. Anatomy teaching has shifted toward a more function-oriented approach and has increasingly incorporated e-learning. There is limited evidence on whether the integrated teaching of professionalism, clinical examination technique, and functional anatomy via e-learning is effective.

Objective: This study aimed to investigate the impact of an integrated blended learning course on the ability of first-year medical students to perform a shoulder examination on healthy volunteers.

Methods: Based on Kolb's experiential learning theory, we designed a course on shoulder anatomy and clinical examination techniques that integrates preclinical and clinical content across all 4 stages of Kolb's learning cycle. The study is a randomized, observer-blinded controlled trial involving first-year medical students who are assigned to one of two groups. Both groups participated in blended learning courses; however, the intervention group's course combined clinical examination, anatomy, and professional behavior and included a peer-assisted practice session as well as a flipped classroom seminar. The control group's course combined an online lecture with self-study and self-examination. After completing the course, participants uploaded a video of their shoulder examination. The videos were scored by 2 blinded raters using a standardized examination checklist with a total score of 40.

Results: Thirty-eight medical students were included from the 80 participants needed based on the power calculation. Seventeen intervention and 14 control students completed the 3-week study. The intervention group students scored a mean of 34.71 (SD 1.99). The control students scored a mean of 29.43 (SD 5.13). The difference of means was 5.3 points and proved to be statistically significant (P<.001; 2-sided Mann-Whitney U test).

Conclusions: The study shows that anatomy, professional behavior, and clinical examination skills can also be taught in an integrated blended learning approach. For first-year medical students, this approach proved more effective than online lectures and self-study.

Trial Registration: ISRCTN Registry ISRCTN13061552; https://www.isrctn.com/ISRCTN13061552

(JMIR Med Educ 2025;11:e62666) doi:10.2196/62666

KEYWORDS

functional anatomy; integrated learning; blended learning; undergraduate medical education; clinical examination; randomized controlled trial

Introduction

Although clinical guidelines recommend that imaging diagnostics be based on a structured clinical examination, primary care physicians frequently rely on imaging modalities such as magnetic resonance imaging when managing shoulder complaints [1-4]. This raises the broader issue of how well clinical anatomy and musculoskeletal examination skills are integrated and emphasized during medical education and training. In fact, both residents and medical students report substantial learning needs in musculoskeletal examination

techniques, except among those with a preexisting interest in musculoskeletal specialties [5-11].

Learning clinical competencies is a complex process that can be better understood through the application of a theoretical framework [12-14]. Kolb's experiential learning theory provides such a framework. It conceptualizes learning as a continuous cycle involving concrete experience, reflective observation, abstract conceptualization, and active experimentation [15]. This model is particularly relevant in traditional Flexnerian curricula, where preclinical and clinical training are distinctly delineated, with the cultivation of clinical competencies unfolding over the course of several semesters. It encompasses various dimensions—such as interpreting pathology, applying appropriate examination techniques, and demonstrating professional behavior—achieved through diverse instructional methods, all within a stressful learning environment [3,11,16-20].

Due to the fragmented way students encounter clinical skills and related knowledge, anatomy education has increasingly moved toward region- and function-based integration—though this shift remains incomplete [21-25].

The advent of the COVID-19 pandemic, along with its associated restrictions on in-person interaction, accelerated the adoption of digital teaching formats in medical education. This shift aligned with a general trend toward expanding e-learning in both clinical and anatomical instruction [26-36]. Specifically, shoulder examination techniques can be taught either face-to-face or digitally through synchronous and asynchronous methods, with face-to-face formats demonstrating superior outcomes and higher learner acceptance [9,13,14,27,37]. Blended learning approaches, such as the flipped classroom, which combine digital and in-person elements, have gained increased traction since the pandemic by effectively linking clinical and anatomical content [11,25-27,32,33,38-41]. In addition, peer teaching has been shown to enhance students' confidence in clinical examination skills and facilitate the transfer of anatomical knowledge into clinical practice [42-44].

The optimal timing for introducing clinical skills within anatomy education and how best to align these skills with examination formats remain subjects of debate [11,45-47]. Clinical skills are predominantly assessed through objective structured clinical examinations (OSCEs), which can also be conducted using video recordings [14,16,48,49].

In summary, Flexnerian curricula typically teach and assess anatomical knowledge, clinical skills, and professional behavior separately, using fragmented instructional methods that include both digital and face-to-face formats. These components correspond to distinct stages of Kolb's experiential learning cycle—for instance, peer teaching facilitates reflective observation and abstract conceptualization [42], while clinical training fosters concrete experience and active experimentation [9,37]. However, Kolb emphasizes that meaningful and deep learning requires progression through all stages of the cycle [15]. This suggests that educational approaches intentionally integrating these phases may lead to improved learning outcomes.

Although previous studies have demonstrated that anatomy and clinical skills can be taught through both face-to-face and digital modalities [36-38,50], there remains a lack of robust empirical evidence supporting integrated teaching approaches that unify preclinical and clinical domains. Our intervention was therefore designed not only to bridge preclinical and clinical content but also to combine multiple instructional methods, aligning with Kolb's model to support an effective learning process.

This study aimed to assess the effect of this integrated blended-learning course on first-year medical students' clinical performance, professionalism, and anatomical knowledge.

Methods

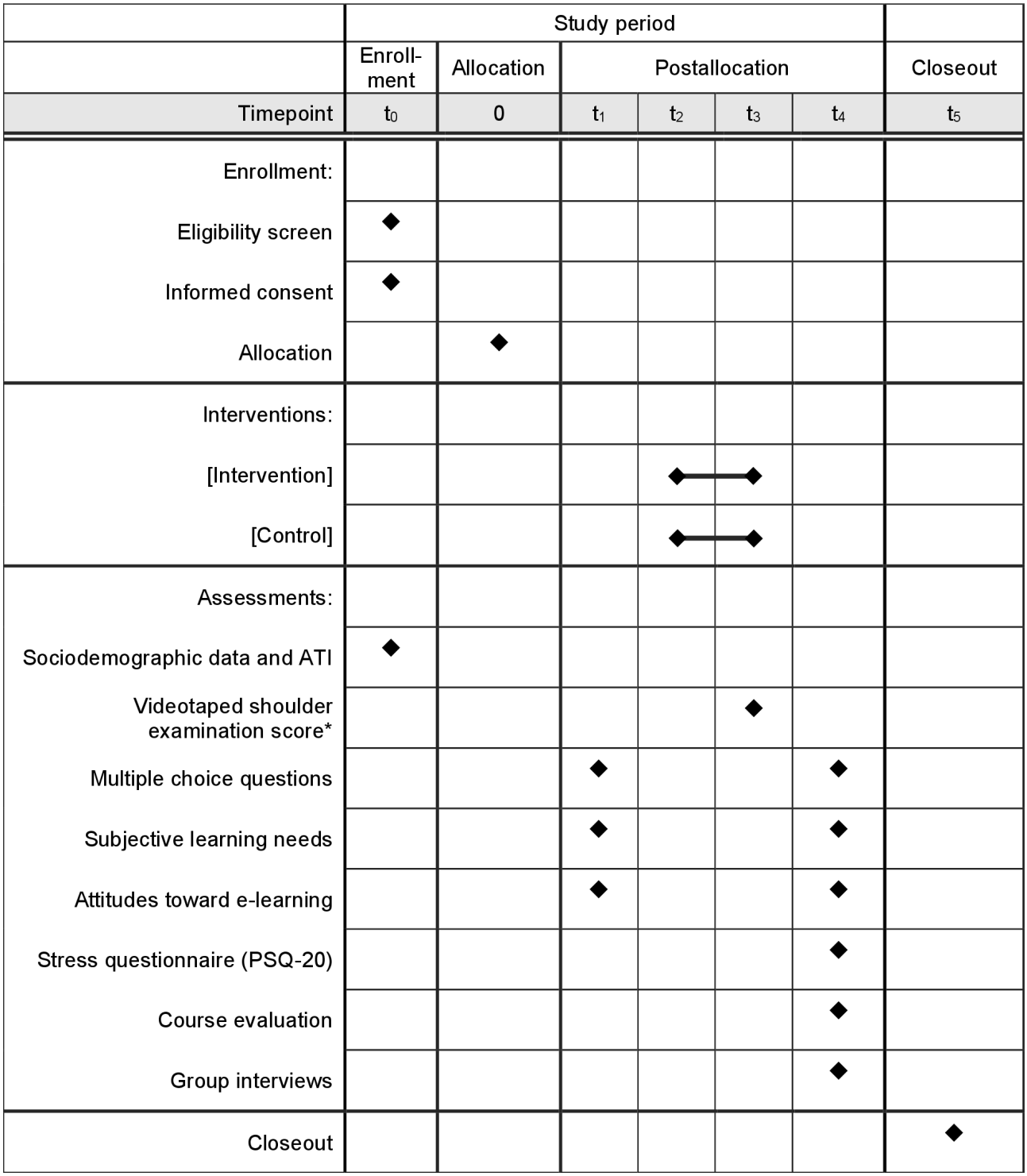
This paper follows the CONSORT (Consolidated Standards of Reporting Trials) guidelines [51].

Study Design

The study was conducted as a 2-arm, randomized, observer-blinded intervention. The study's acronym, TraceX, is derived from "TRansfer of AnatomiCal knowledge in the EXamination situation for preclinical medical students." It was conducted as part of a curricular development project with the same name. The project was funded by the Medical Faculty of Tübingen University Hospital (Universitätsklinikum Tübingen [UKT]). The study protocol underwent an external peer-review process by reviewers not involved in the project. The dean of the faculty and the faculty commission approved the project in 2020.

Its protocol is illustrated in Figure 1, based on the SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) figure template [52]. The study compared students' performance in a videotaped shoulder examination after a 3-week blended-learning course. The intervention and control groups received 2 different blended-learning modules.

Figure 1. Transfer of anatomical knowledge in the examination situation for preclinical medical students (TraceX) study design protocol. The asterisk indicates the primary outcome. ATI: affinity for technology interaction; PSQ-20: Perceived Stress Questionnaire.



Study Population and Setting

The study took place at the Medical Faculty of Eberhard Karls University Tübingen in southern Germany. It was conducted between October 2021 and March 2022 by the Institute for General Practice and Interprofessional Health Care in cooperation with the Institute of Clinical Anatomy and Cell Analysis. In Tübingen, online courses are provided using the Integrated Learning, Information, and Work Cooperation System (ILIAS) online-learning platform, which is a commonly used learning management system among German universities (ILIAS open source e-Learning e.V). It allows log-in via authenticated student email accounts.

Pilot Study

Several instruments used in the study were piloted in 2020 with 18 first-year medical students. The primary purpose of the pilot study was to check interrater reliability for the primary outcome measurement instrument and to calculate the internal reliability of the self-developed questionnaire items. The results of these checks and their impact on the development of the items are listed below for each respective instrument.

In addition, participants of the pilot study were asked to assess the online learning modules. Based on their feedback, the course content was adapted. Two of the pilot study participants provided valuable comments on course benefits and weaknesses in an additional voluntary online interview. The interview served as an extended evaluation. Furthermore, LG (who later conducted the group interviews in the main study) piloted interview guidelines and received interviewer training based on the participants' feedback.

Recruitment

Recruitment for the main study started in May 2021. First-year medical students were approached via email and in 2 online lectures. Interested students were invited to contact the Institute for General Practice's office to receive additional information (eg, type of study, goal, significance, and participant rights). Study participation was voluntary and did not impact other courses in the curriculum. For students who did not wish to participate in the study, regular curriculum activities proceeded. The teaching coordinator (responsible for course and lecture coordination at the Institute) communicated with students but did not participate in the study design or execution. Eligibility criteria included being a first-year medical student, having provided informed consent to use ILIAS, and feeling healthy enough to participate in the study.

Intervention and Control

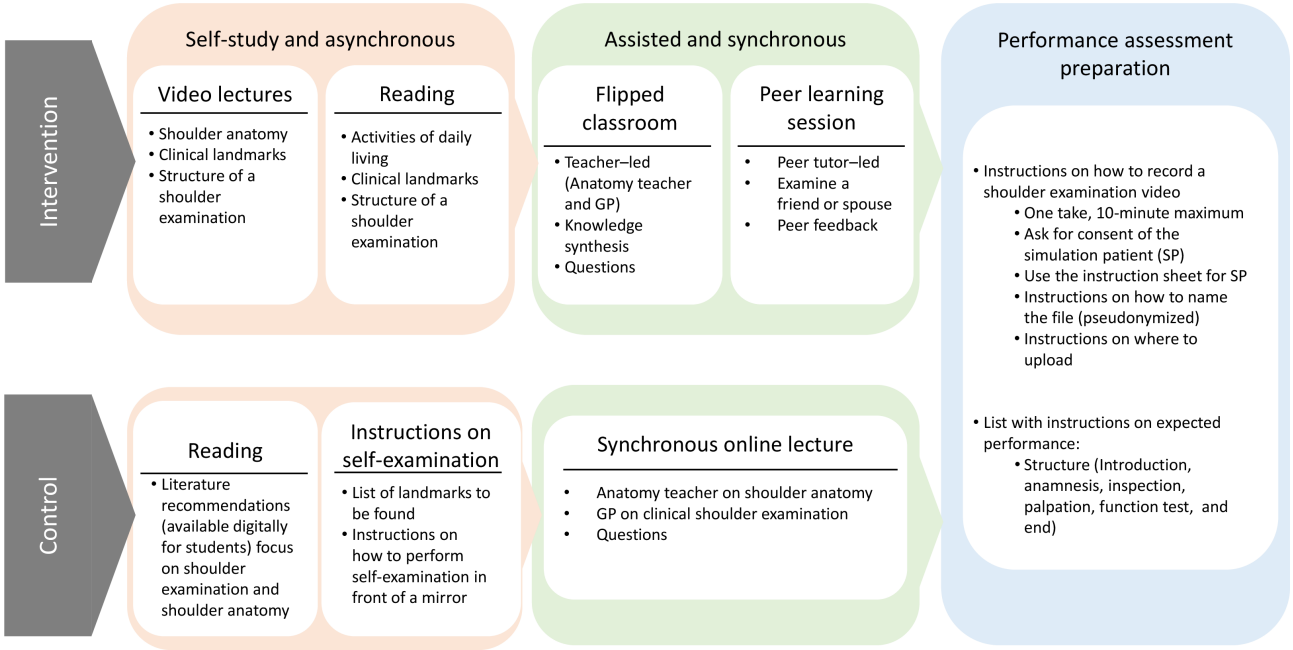
Intervention and control courses were developed collaboratively in interdisciplinary workshops involving general practitioners (GPs), a medical didactics expert, an orthopedic surgeon, a psychotherapist, medical students, a physiotherapist, and a medical psychologist.

Common learning objectives were developed first and then applied to both groups, which differed in instructional design and learning format. The intervention group received access to structured online modules, including physiotherapist-led examination videos, functional and topographic anatomy videos performed by an anatomy lecturer, and modules on daily life impact and professional conduct. Content was integrated across preclinical and clinical domains. After self-study, students joined a 90-minute flipped-classroom seminar (taught by the same anatomy lecturer and a GP), which included preparation for a peer-training session. Due to COVID-19 restrictions, this 90-minute peer-training session was held in decentralized student pairs or moderated small groups on Zoom (Zoom Communications Inc), with trained peer tutors facilitating feedback.

The control group followed a more traditional format with lectures and self-study. Students received literature for self-study and instructions for self-examination (eg, palpation in front of a mirror). In addition, they attended 2 distinct synchronous online lectures: one on anatomy (anatomy lecturer) and one on clinical examination (GP), based on the same content as the intervention videos but presented with still images, self-demonstration by the lecturers, and a greater separation of clinical and preclinical content.

In summary, the intervention emphasized integrated, interactive learning with structured feedback and more practice time, providing a learning process with all 4 steps of Kolb's learning cycle [15], while the control group reflected conventional, lecture-based learning with limited interaction and a more fragmented learning process. Please refer to Figure 2 for an overview.

Figure 2. Intervention and control design. GP: general practitioner.



Randomization and Allocation

From those students who signed a written informed consent to participate in the study, basic sociodemographic data were obtained at t0 to enable stratified randomization. Randomization was undertaken in a 1:1 ratio using lists generated with the R programming package (The R Core Team, R Foundation, and R Consortium). Randomization was stratified for gender and

occupational experience and used block sizes of 2 and 4. Individualized identification codes were generated by the participants in the surveys (t1 and t4) to allow for linking of presurvey and postsurvey data without compromising participant anonymity or disclosing group affiliation. All other data that could reveal group affiliation were concealed until grading of the final examinations and statistical analysis was completed. Participants were not blinded to their own group affiliation. Randomization and group allocation were performed by 2 researchers not involved in the assessment and development of course content.

The groups received the same instructions for exam performance and grading criteria. Finally, a postcourse questionnaire was distributed to both groups, and voluntary postcourse online group interviews were conducted (t4). After study completion, members of the control group were provided full access to all e-learning resources to preclude disadvantage.

Primary Outcome

This study hypothesizes that an integrated course based on Kolb's learning theory would improve first-year medical students' performance in a standardized, complex OSCE shoulder station requiring structured clinical examination, anatomical knowledge, and professional conduct. Students in the intervention group attended the integrated course and were expected to outperform those in the control group. The primary outcome was OSCE performance at t3, assessed through a structured examination focusing on inspection, palpation, and functional movement assessment, with emphasis on identifying anatomical landmarks rather than detecting pathology.

Because contact restrictions prevented a face-to-face OSCE setting, participants recorded videos of their shoulder examinations and uploaded them to ILIAS. Video length was limited to 10 minutes and included a volunteer (peer, spouse, or family member) acting as a simulation patient. Students were instructed to ensure their volunteers had no preexisting shoulder pain or functional impairments and to obtain consent before recording. Simulation patients were asked to follow the students' instructions and cooperate during the examination, creating a controlled environment intended to minimize bias introduced by the simulation patients.

The uploaded videos were assessed independently by a pair of blinded raters (GP and a medical student) using an examination checklist. The checklist comprised a total of 40 items, including 12 items on anatomical knowledge, such as topography and functional anatomy, 17 items on structured clinical examination skills, and 9 items on medical professionalism. It was developed in multidisciplinary workshops based on existing literature [14] and items from the OSCE used at UKT [53]. It was piloted in the pilot study. In that study, the pair of most congruent raters (a medical student and a GP) achieved an interrater validity of κ=0.573 (Cohen κ). Due to this moderate agreement, several items were operationalized for better standardization of examiners [54]. The final version of the examination score used in this study is included as Multimedia Appendix 1.

Secondary Outcomes

Participants' subjective learning needs were assessed across three domains—anatomical knowledge, clinical examination skills, and professionalism—at time points t1 and t4 using a questionnaire. This questionnaire was developed during an interdisciplinary workshop and subsequently piloted in the initial study phase. The sample of 18 pilot study participants showed a Cronbach α of 0.92. The domain "anatomical knowledge" contained 5 items (ie, visible landmarks, anatomical structure of the shoulder, function, range of motion, and association of function with activities of daily living). The domain "clinical examination" also consisted of 5 items (ie, structured examination steps, history taking, shoulder inspection, shoulder palpation, and test of function). Finally, the domain "professionalism" was covered with 4 items (ie, use of patient-centered language, autonomous handling of the examination situation, perceptiveness toward patient feedback, and addressing patient feedback during the examination). All items were operationalized on a 4-point Likert scale. This yielded sum scores on a 5- to 20-point range and a 4- to 16-point range, respectively.

Theoretical anatomical knowledge was assessed before and after the course (at t1 and t4) using the same set of 10 randomized multiple-choice questions (MCQs). Each question had 5 answer options, with only 1 correct answer, resulting in a knowledge score ranging from 0 to 10. These MCQs reflect the current standard used in the written state medical examination in Germany.

Using Likert scale rating questions, we used 4 items of the standardized Affinity for Technology Interaction Short Scale (ATI-S), rendering a sum score on a 4- to 24-point scale. The ATI-S is a reliable (Cronbach α 0.88-0.92) and validated instrument [55].

The course evaluation included 5 items adapted from the standard evaluation used by the Medical Faculty of Tübingen, covering overall assessment, contribution to personal learning, curriculum alignment, course structure, and clarity of learning goals. Cronbach α for these items was 0.741 based on pilot study data. The evaluation was conducted online via ILIAS at time point t4. In addition, several items assessing blended learning formats were incorporated at t1 and t4, aiming to assess attitudes toward blended learning before and after the intervention. Students were asked to rate whether blended learning promotes self-directed learning, encourages lesson preparation and review, supports exam readiness, enhances curricular value, enables flexible learning, and increases study satisfaction.

Participants were additionally invited to online group interviews at t4. The interviews were analyzed with thematic analysis [56]. An in-depth mixed methods evaluation that includes the interviews will be published separately.

To evaluate the possible impact of the intervention on perceived stress, the validated and standardized 20-item Perceived Stress Questionnaire (PSQ-20) was applied at t4. It contains 20 items with 4 subscales: worries, tension, joy, and demands. Cronbach α is 0.80-0.86 [57].

Statistical Analysis

General

Data were analyzed using SPSS (version 27; IBM Corp). Summary statistics on participants' demographic characteristics, baseline information, course evaluation, and primary outcome were computed using means, SDs, and proportions.

The following statistical tests were used to assess group differences in primary and secondary outcomes. For the primary outcome and exploratory analysis of subscores, 2-sided Mann-Whitney *U* tests were performed to account for nonparametric data distributions. For comparison between groups for nominal scales, chi-square (χ^2) tests were performed or, if cell size was less than 5, Fisher exact test. For the secondary outcomes on longitudinal effects preintervention and postintervention, Wilcoxon rank tests were performed.

Concerning the internal reliability of questionnaires and evaluation based on the pilot study data, Cronbach α was calculated. We also calculated κ statistics underlying the primary outcome for assessment of interrater reliability in examination grading.

The study's intention-to-treat approach considers possible contamination effects by between-group communication or student account misuse between groups.

Power

Power analysis suggested a sample size of 62 subjects to detect intergroup differences in examination scores of 16% with a SD of 20% and an estimated effect size of 0.71 when using 2-sided *t* tests for independent samples at a 5% significance level [58]. Adding a 25% buffer for dropouts, a total of 80 participants were aimed for as the total sample size.

Ethical Considerations

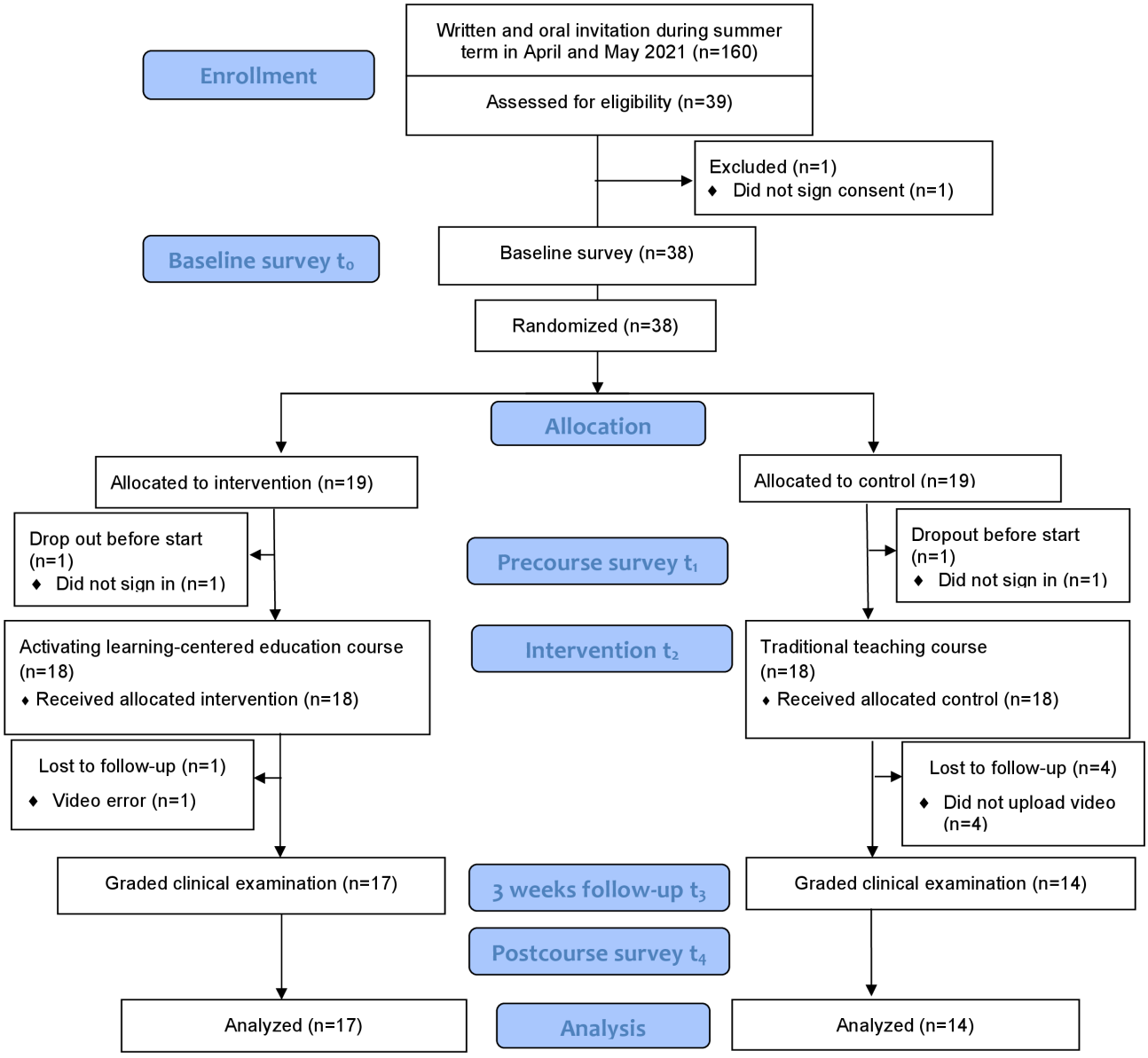
The study was assessed by the Ethics Committee of the Medical Faculty of Tübingen University on May 4, 2020, (231/2020BO1) and did not require ethical approval. Study participation was voluntary and students provided informed consent to participate in the study and access the respective course material in ILIAS. No financial inducement was given for study participation. However, all participating students received a bookstore voucher to compensate for their time commitment. Furthermore, participants who completed the study requirements were eligible to enter a lottery drawing for one iPad. Data were stored on ISO 27001–certified servers at UKT. The trial protocol is available from the authors upon request.

Results

Study Flow

Of 39 enrolled students, 38 were allocated to the intervention and control groups. In both groups, one participant, who, despite having signed informed consent, did not sign in to ILIAS. In the intervention group, one student dropped out before analysis due to an unresolvable video error. The control group had 4 dropouts who did not upload a video. In total, analyzable data from 31 study participants were obtained. Figure 3 below shows the study flow (CONSORT chart).

Figure 3. CONSORT (Consolidated Standards of Reporting Trials) study flowchart.



Sociodemographic and Occupational Characteristics

The majority of participants had no current employment. The intervention group had 2 more participants currently employed in the medical sector. Statistically, neither current nor past employment differed significantly between groups. Participant characteristics at baseline (t1) are presented in Table 1.

Table . Participant characteristics (N=18).

Characteristic	Intervention	Control	*P* value
Age (years), mean (SD)	21.78 (3.56)	21.94 (5.68)	.92[a]
Female[b], n (%)	13 (72)	14 (78)	≥.99[c]
Previous occupational experience, n (%)			
Yes	9 (50)	8 (44)	.74[d]
Currently employed, n (%)			
Yes, medical sector	4 (22)	2 (11)	.73[c]
Yes, nonmedical sector	1 (6)	2 (11)	—[e]
No	13 (72)	13 (72)	—
Missing	0 (0)	1 (6)	—

[a]Student *t* test.

[b]No participant identified as nonbinary.

[c]Fisher exact test.

[d]Chi-square test.

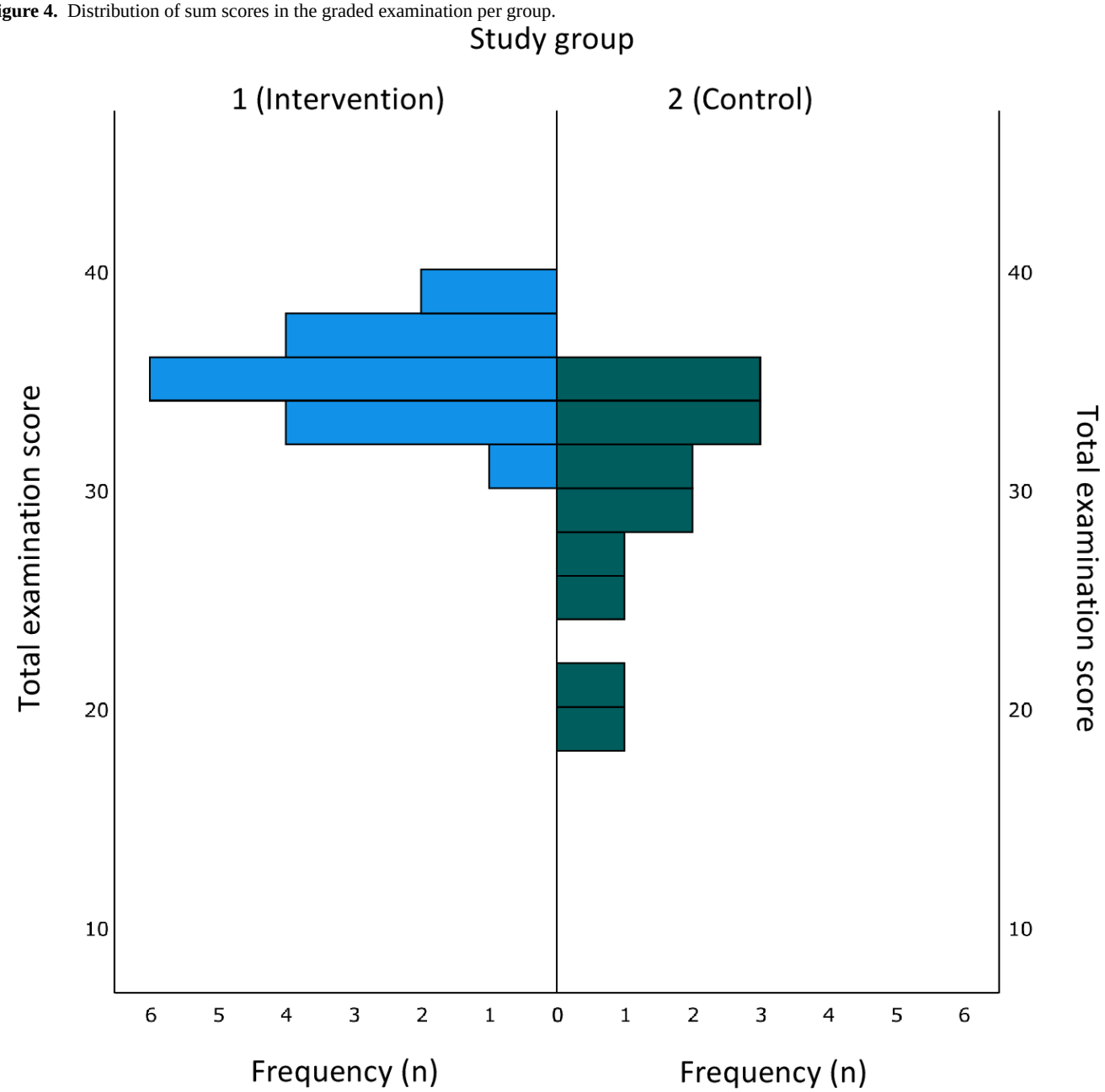
[e]Not applicable.

Primary Outcome

Overview

Intervention group participants reached higher mean performance scores compared to those in the control group (mean 34.71, SD 1.99 vs mean 29.43, SD 5.13), with differences being statistically significant at $P<.001$. Score distributions are shown in Figure 4. Interrater reliability for the shoulder examination checklist score between the two blinded raters was 0.763.

Figure 4. Distribution of sum scores in the graded examination per group.



Exploratory Analysis of Primary Outcome Subscores

We conducted a post hoc exploratory analysis of the primary outcome subscores: anatomical knowledge, clinical examination skills, and professionalism. In the anatomy subscore, the intervention group performed slightly better than the control group, with the difference reaching statistical significance. For clinical examination skills, the difference between groups was less pronounced but remained statistically significant. Table 2 summarizes the results of this analysis.

Table . Exploratory analysis of primary outcome subscores.

Variable	Intervention, mean (SD)	Control, mean (SD)	*P* value[a]
Total performance score (40 items, primary outcome)	34.71 (1.99)	29.43 (5.13)	<.001
Anatomical knowledge (12-item subscore)	9.88 (1.73)	8.21 (1.93)	.02
Clinical examination (17-item subscore)	17.88 (0.86)	15.14 (3.92)	.05
Medical professionalism (9-item subscore)	6.94 (0.90)	6.07 (1.44)	.15

[a]Mann-Whitney *U* test, exact significance.

Secondary Outcomes

Both groups showed a significant reduction in self-reported learning needs in anatomy and clinical examination. MCQ performance improved significantly in both groups after the course, with no significant differences between groups. Participants in the intervention group spent, on average, 2 hours more in the learning module than those in the control group, a statistically significant difference. No relevant group differences were found in attitudes toward blended learning. In the course evaluation, the intervention group rated the blended learning experience and overall course evaluation less favorably but gave higher ratings for the achievement of learning goals and curricular alignment compared to the control group.

Table 3 provides an overview of the secondary outcomes. Group differences at t1 were not significant and are not included.

Table . Secondary outcomes.

Measurement	Dropout (n=5), t_1, mean (SD)	Intervention (n=17), t_1, mean (SD)	Intervention, mean (SD)	Difference, t_4-t_1	*P* value[a] (interv)	Control (n=14), t_1, mean (SD)	Control, t_4, mean (SD)	Difference, t_4-t_1	*P* value[a] (control)	Difference intervention-control (time); positive: favors intervention	*P* value[b]
Affinity for technology interaction	11.2 (4.55)	12.65 (3.52)	—[c]	—	—	14 (2.96)	—	—	—	−1.35 (t_1)	.36
Time spent (hours)	—	—	6.06 (2.42)	—	—	—	4.0 (1.76)	—	—	2.06 (t_4)	.01
Perceived Stress Questionnaire score, index	—	—	0.46 (0.09)	—	—	—	0.47 (0.11)	—	—	−0.01 (t_4)	.98
Learning needs											
Subscore anatomy	17.2 (2.17)	15.76 (2.73)	9.53 (2.27)	−6.24	.00	15.57 (3.8)	10.43 (2.06)	−5.14	.00	−0.90 (t_4)	.22
Subscore professional conduct	11.6 (3.78)	8.76 (2.75)	6.82 (2.43)	−1.94	.00	9.5 (4.07)	8.07 (2.06)	−1.43	.13	−1.25 (t_4)	.19
Subscore examination skills	19 (1.73)	18.41 (1.77)	9 (2.35)	−9.41	.00	18.43 (3.52)	11.5 (2.47)	−6.93	.00	−2.50 (t_4)	.01
MCQs[e]	4.6 (2.07)	5.53 (2)	6.94 (1.43)	1.41	.01	4.71 (2.09)	6.5 (1.83)	1.79	.00	0.82 (t_4)	.63
Attitudes toward blended learning (mean item score; 1=strongly disagree, 5=strongly agree)											
Facilitate continuous self-sufficient learning	2.4 (1.14)	2.76 (1.2)	2.65 (1.32)	−0.12	.56	2.93 (0.73)	2.79 (0.89)	−0.14	.63	−0.14 (t_4)	.68
Motivate me to prepare for or follow up on classroom events	2.8 (1.3)	2.94 (0.93)	2.94 (1.02)	0.00	1	2.93 (0.83)	2.57 (0.51)	−0.36	.17	0.37 (t_4)	.32
Support me in preparing for exams	2 (1.23)	2.65 (0.93)	2.59 (0.8)	−0.06	.76	2.50 (0.86)	2.29 (0.61)	−0.21	.37	0.3 (t_4)	.23
Provide added value in learning	4.2 (1.1)	3.35 (1)	3.47 (1.07)	0.12	.6	3.43 (0.76)	3.36 (0.63)	−0.07	.74	0.11 (t_4)	.92

Measurement	Dropout (n=5), t_1, mean (SD)	Intervention (n=17), t_1, mean (SD)	Intervention, mean (SD)	Difference, t_4-t_1	*P* value[a] (interv)	Control (n=14), t_1, mean (SD)	Control, t_4, mean (SD)	Difference, t_4-t_1	*P* value[a] (control)	Difference intervention-control (time); positive: favors intervention	*P* value[b]
Allow learning at any time and from any place	1.2 (0.45)	1.59 (0.87)	1.53 (0.87)	−0.06	.65	1.14 (0.36)	1.86 (0.86)	0.72	.01	−0.33 (t_4)	.22
Increase satisfaction with my studies	2.2 (1.3)	3.18 (1.07)	2.71 (1.26)	−0.47	.05	3.57 (0.85)	3.14 (0.66)	−0.43	.11	−0.43 (t_4)	.22
Facilitate my learning	2.2 (1.1)	2.76 (1.25)	2.47 (1.18)	−0.29	.06	2.71 (0.83)	3.07 (0.73	0.36	.21	−0.6 (t_4)	.11
Evaluation (mean item score; 1=strongly disagree, 5=strongly agree)											
The blended learning module was well executed	—	—	1.94 (0.83)	—	—	—	2.86 (0.66)	—	—	−0.92 (t_4)	.01
Learning goals were clearly defined	—	—	4.24 (0.56)	—	—	—	3.21 (0.8)	—	—	1.03 (t_4)	.00
The course was clearly structured	—	—	4.59 (0.5)	—	—	—	3.57 (0.94)	—	—	1.02 (t_4)	.00
The course aligned well with the curriculum	—	—	4.35 (0.6)	—	—	—	3.57 (1.16)	—	—	0.78 (t_4)	.03
The course contributed to my learning success	—	—	4.06 (0.75)	—	—	—	3.57 (1.1)	—	—	0.49 (t_4)	.17

Measurement	Dropout (n=5), t_1, mean (SD)	Intervention (n=17), t_1, mean (SD)	Intervention, mean (SD)	Difference, t_4-t_1	*P* value[a] (interv)	Control (n=14), t_1, mean (SD)	Control, t_4, mean (SD)	Difference, t_4-t_1	*P* value[a] (control)	Difference intervention-control (time); positive: favors intervention	*P* value[b]
Course assessment (grade; 1=very good, 6=insufficient)	—	—	1.88 (0.6)	—	—	—	2.57 (0.85)	—	—	−0.69 (t_4)	.03

[a]Wilcoxon rank-sign test.

[b]Mann-Whitney *U* test.

[c]Not applicable.

[d]MCQ: multiple-choice question.

Discussion

Principal Findings

Overview

This randomized, observer-blinded trial tested whether an integrated course based on Kolb's learning theory would improve first-year medical students' OSCE performance. Students in the intervention group outperformed the control group after a 3-week course. Both groups showed reduced learning needs, though intervention students spent approximately 2 hours more in the course. The study confirmed the hypothesis but had limitations to consider.

Clinical Examination Skills

The observed difference in clinical examination performance between the 2 groups corresponded to a significant reduction in learning needs related to examination skills. These findings align with previous studies. Brewer et al [37] found that second-year students who received face-to-face instruction in clinical examination skills performed better than those in asynchronous e-learning or self-study groups, with the self-study group achieving the lowest scores. Brewer's study used simulated patients without pathology and included a broader range of clinical tests. Both Brewer et al [37] and Vivekananda-Schmidt et al [13] reported high effect sizes for OSCE shoulder performance when students supplemented regular studies with asynchronous computer-assisted learning. Vivekananda-Schmidt also noted improved student confidence with computer-assisted learning, though our study did not assess confidence due to its limited relevance to performance [7,12].

Anatomical Knowledge

Participants in the intervention group demonstrated improved clinical examination performance without a negative impact on their theoretical anatomical knowledge, as measured by MCQs. We believe that the exclusion of advanced clinical examination tasks such as the Jobe test or Hawkins sign from the learning goals contributed to the increase in anatomical knowledge demonstrated in the exam and the more pronounced reduction of anatomical learning needs in the intervention group. According to the participating preclinical and clinical teachers in the multidisciplinary workshop, these tests require extensive knowledge of clinical pathology and might confuse first-year medical students. This notion is in line with existing research that advises careful didactic planning of anatomical content [2,3,11].

Although the online instruction in our study offered only basic anatomical content—focusing on muscles, bones, and function while underrepresenting coverage of nerves and vessels—both groups had recently attended curricular lectures designed to provide deeper anatomical knowledge. After the course, participants in both groups could identify anatomical structures and assess normal findings in a structured way. The key difference was that the intervention group had the chance to apply their theoretical knowledge in practice and receive structured feedback during the peer tutoring session [11,42]. Our findings suggest that early integration of practical experience does not hinder but may, in fact, support deeper anatomical and clinical learning if implemented appropriately [3,11,25].

Professional Conduct

Neither the professional subscore in the shoulder examination nor learning needs showed significant differences between groups. Professional conduct might represent a subjective feeling of confidence rather than a measurable outcome. It may not have been adequately operationalized in this study; for example, the relative passive role of the standardized patients posed no real challenges for professionalism. In addition, too little time was allocated to student-teacher interaction and exchange about professional conduct—a common shortcoming of blended learning [32]. This shows that some aspects of learning should be performed face-to-face and that time must be allocated for this interaction [26,30,32]. Also, professional conduct is difficult to teach explicitly and is often learned from various sources over a long period of time [13,18,19,27]. Social learning elements play a crucial role in fostering professionalism even

in a predominantly remote learning environment [32]. It thrives through direct interaction between instructors and learners. This insight will inform the curricular integration of our learning module.

In summary, this study shows that clinical examination skills can successfully be taught early in the curriculum using a blended learning approach. The reduction in learning needs was not dependent on current or past employment or different levels of preexisting anatomical knowledge. We controlled for differences in gender and preoccupation through randomization. A baseline performance check using a structured examination before the study would perhaps have shown a clearer dependence on preexisting examination skills [59].

Dropout Analysis

The 5 dropouts had higher subjective learning needs at t1, scored lower in the MCQ knowledge test, and had a slightly lower technical affinity. Their attitudes toward blended learning differed from those of the other study participants. While they believed that blended learning provided added value to learning, they were less confident that it helped them prepare for exams or increase personal satisfaction with their studies. The dropout sample was too small for statistical analyses. However, there are 2 ways to interpret this: there might be a subgroup of students with such low expectations of blended learning that they might not use the method if given the choice [60], or the small study population could reflect a selection bias of motivated, stress-resilient, technology-affine, high-performing students who would do well in any learning environment [32].

Other Secondary Outcomes

The intervention group required, on average, 2 additional hours to complete the course compared to the control group. This was due to engagement with video lectures and demonstrations, as well as participation in the peer-assisted synchronous learning session, all of which required learners to allocate additional time, even though they had greater flexibility to choose when and where to engage with most of the content than the control group. Similar increases in workload have been observed in other studies on blended learning formats [36].

This interpretation is further supported by the intervention group's high ratings of course structure, clarity of learning objectives, and perceived alignment with the curriculum.

Although overall stress levels were high in both groups at t4, they were comparable to those of other medical student cohorts [20]. We therefore infer that the increased workload did not result in a heightened perception of stress among intervention group participants.

Technical affinity was equally high across both groups, consistent with findings that current medical student cohorts are generally digitally literate [27]. For our sample, this suggests that time invested in studying, rather than digital affinity, was more strongly associated with learning success [46,47].

Implications for Curricular Development

When Should Clinical Aspects in Anatomy Learning Be Integrated?

Our study supports the integration of anatomy teaching and clinical examination early in the curriculum. Both study groups, exposed to clinical examination content during the first year, reported reduced learning needs and demonstrated measurable skill acquisition without compromising anatomical knowledge. These findings suggest that early competency development is feasible and particularly effective when teaching content and assessment formats are well aligned [10,14,21-25,45].

What is the Role of Blended Learning in Integrated Clinical and Anatomy Learning?

Our study demonstrated that an intervention using synchronous peer-assisted clinical training paired with asynchronous training videos in a flipped-classroom seminar improved student outcomes compared to self-study and synchronous online lectures, aligning with previous studies [30,37]. However, despite improved performance, students reported a less satisfactory blended learning experience, highlighting the challenge of effectively implementing such approaches [27,29]. Based on our findings, key factors for successful blended learning include clear learning goals, structured delivery, and theory-guided integration of content and teaching methods, which may be more influential than the delivery format itself on student evaluation [32,33]. In addition, effective implementation must account for contextual factors such as the COVID-19 pandemic, which may have influenced student perceptions in our study [26]. These findings highlight that while blended learning can enhance outcomes, its effectiveness depends on thoughtful design, clear objectives, and contextual factors—underscoring the need for nuanced, context-aware research in this field [41].

Limitations

This study is limited by its small sample size, although a power analysis indicated a need for 80 participants; only 38 were recruited, and 31 completed the trial. While randomization controlled for gender and professional experience, the small sample size limits generalizability and prevents definitive conclusions about causality.

Recruitment was difficult despite incentives, likely due to increased student workload during the COVID-19 pandemic [26,27]. Although results were promising, the recruitment shortfall and end of funding prevented a second cohort. In addition, curriculum changes during the later stages of the pandemic affected the study environment, leading us to conclude the trial and prioritize efforts to integrate the intervention into the curriculum.

Concerning the main outcome, the intervention course provided learners with tailored modular learning material and feedback opportunities, in contrast to the control group. Importantly, both groups received teaching on shoulder anatomy and clinical skills—a more integrated approach than the regular curriculum—and identical guidance on the performance assessment. This substantially reduces the study's ability to

isolate the effect of blended versus traditional learning. However, both groups showed knowledge gains, suggesting that early integration of clinical content is effective overall. The superior performance in the intervention group likely reflects not only the learning format but also the stronger alignment between the instructional design and the assessment tasks—consistent with our hypothesis that an integrated, theory-based course would improve clinical competency [47].

A reconnaissance effect on anatomical knowledge cannot be excluded, as the same MCQs were repeated at t4. However, student performance and learning needs indicated stable or improved knowledge. Peer learning sessions and exams were intended to be conducted face-to-face but were instead held via videoconferencing due to COVID-19. Although this deviated from the protocol, comparable learning outcomes have been reported in other studies. Video-recorded OSCEs are an established method [16], with only one upload failure. Interobserver reliability was substantial and consistent with other studies [9,14,37].

Conclusions

First-year medical students' clinical performance on shoulder examinations can objectively be improved by an integrated blended-learning course in anatomy, professional conduct, and clinical examination skills. The results encourage early integration of clinical examination skills in preclinical medical education. Didactically, early clinical examination courses should focus on functional anatomy, a structured examination approach, and healthy subjects. Methodologically, the content can be delivered using blended learning that considers social aspects of learning and aligns the content to a standardized shoulder examination simulating a single OSCE station. If a face-to-face exam is not possible, videotaped examinations are a viable alternative. This study does not provide information about the sustainability of the learning. A cohort study that follows the progression through surface and deep learning phases to the transfer phase in this integrative approach is warranted [11].

Acknowledgments

We thank Anna-Jasmin Wetzel and Nadine Nicole Koch for their support in the randomization process. We also thank all participants of the pilot study and our medical student tutors who moderated the online training sessions. The generative artificial intelligence tool ChatGPT (OpenAI) was used to revise the language and style of the manuscript, which were further reviewed and revised by the study group. The authors acknowledge support from the Open Access Publication fund of the University of Tübingen. The project received intramural funds from UKT (Profil Plus funds for curricular development).

Data Availability

The datasets generated or analyzed during this study are not publicly available due to data protection regulations but are available from the corresponding author on reasonable request.

Authors' Contributions

TS and RK contributed to the conceptualization of the study, with additional input from LG and TFW. Data curation and formal analysis were performed by NG, LG, and RK. TS and RK were responsible for funding acquisition. LG, NG, and RK conducted the investigation. Methodology was developed by NG, LG, RK, and TFW. Project administration was carried out by NG, LG, and RK. SJ, BH, and TS provided resources, while RK and NG developed the software. Supervision was provided by SJ, BH, and TS. Validation was carried out by SJ, BH, and TS. RK, NG, and LG were responsible for visualization. RK, LG, and NG prepared the original draft, and BH, TS, TFW, and SJ reviewed and edited the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Translation of the checklist used for video upload scoring (primary outcome).
[DOCX File, 23 KB - mededu_v11i1e62666_app1.docx]

Checklist 1
CONSORT (Consolidated Standards of Reporting Trials) checklist.
[PDF File, 30735 KB - mededu_v11i1e62666_app2.pdf]

References

1. Kjelle E, Andersen ER, Krokeide AM, et al. Characterizing and quantifying low-value diagnostic imaging internationally: a scoping review. BMC Med Imaging 2022 Apr 21;22(1):73. [doi: 10.1186/s12880-022-00798-2] [Medline: 35448987]
2. Mitchell C, Adebajo A, Hay E, Carr A. Shoulder pain: diagnosis and management in primary care. BMJ 2005 Nov 12;331(7525):1124-1128. [doi: 10.1136/bmj.331.7525.1124] [Medline: 16282408]

3. Meder A, Stefanescu MC, Ateschrang A, et al. Evidence-based examination techniques for the shoulder joint. Z Orthop Unfall 2021 Jun;159(3):332-335. [doi: 10.1055/a-1440-2242] [Medline: 34111895]
4. Cortes A, Quinlan NJ, Nazal MR, Upadhyaya S, Alpaugh K, Martin SD. A value-based care analysis of magnetic resonance imaging in patients with suspected rotator cuff tendinopathy and the implicated role of conservative management. J Shoulder Elbow Surg 2019 Nov;28(11):2153-2160. [doi: 10.1016/j.jse.2019.04.003] [Medline: 31281001]
5. Kroop SF, Chung CP, Davidson MA, Horn L, Damp JB, Dewey C. Rheumatologic skills development: what are the needs of internal medicine residents? Clin Rheumatol 2016 Aug;35(8):2109-2115. [doi: 10.1007/s10067-015-3150-4] [Medline: 26694057]
6. Denizard-Thompson N, Feiereisel KB, Pedley CF, Burns C, Campos C. Musculoskeletal basics: the shoulder and the knee workshop for primary care residents. MedEdPORTAL 2018 Sep 15;14:10749. [doi: 10.15766/mep_2374-8265.10749] [Medline: 30800949]
7. Day CS, Yeh AC. Evidence of educational inadequacies in region-specific musculoskeletal medicine. Clin Orthop Relat Res 2008 Oct;466(10):2542-2547. [doi: 10.1007/s11999-008-0379-0] [Medline: 18636305]
8. Day CS, Yeh AC, Franko O, Ramirez M, Krupat E. Musculoskeletal medicine: an assessment of the attitudes and knowledge of medical students at Harvard Medical School. Acad Med 2007 May;82(5):452-457. [doi: 10.1097/ACM.0b013e31803ea860] [Medline: 17457065]
9. Stansfield RB, Diponio L, Craig C, et al. Assessing musculoskeletal examination skills and diagnostic reasoning of 4th year medical students using a novel objective structured clinical exam. BMC Med Educ 2016 Oct 14;16(1):268. [doi: 10.1186/s12909-016-0780-4] [Medline: 27741946]
10. Skelley NW, Tanaka MJ, Skelley LM, LaPorte DM. Medical student musculoskeletal education: an institutional survey. J Bone Joint Surg Am 2012 Oct 3;94(19):e146. [doi: 10.2106/JBJS.K.01286] [Medline: 23032597]
11. Cheung CC, Bridges SM, Tipoe GL. Why is anatomy difficult to learn? The implications for undergraduate medical curricula. Anat Sci Educ 2021 Nov;14(6):752-763. [doi: 10.1002/ase.2071] [Medline: 33720515]
12. Vivekananda-Schmidt P, Lewis M, Hassell AB, et al. Validation of MSAT: an instrument to measure medical students' self-assessed confidence in musculoskeletal examination skills. Med Educ 2007 Apr;41(4):402-410. [doi: 10.1111/j.1365-2929.2007.02712.x] [Medline: 17430286]
13. Vivekananda-Schmidt P, Lewis M, Hassell AB, ARC Virtual Rheumatology CAL Research Group. Cluster randomized controlled trial of the impact of a computer-assisted learning package on the learning of musculoskeletal examination skills by undergraduate medical students. Arthritis Rheum 2005 Oct 15;53(5):764-771. [doi: 10.1002/art.21438] [Medline: 16208642]
14. Battistone MJ, Barker AM, Beck JP, Tashjian RZ, Cannon GW. Validity evidence for two objective structured clinical examination stations to evaluate core skills of the shoulder and knee assessment. BMC Med Educ 2017 Jan 13;17(1):13. [doi: 10.1186/s12909-016-0850-7] [Medline: 28086879]
15. Kolb DA. Experiential Learning: Experience as the Source of Learning and Development: Financial Times/Prentice Hall; 2015.
16. Vivekananda-Schmidt P, Lewis M, Coady D, et al. Exploring the use of videotaped objective structured clinical examination in the assessment of joint examination skills of medical students. Arthritis Rheum 2007 Jun 15;57(5):869-876. [doi: 10.1002/art.22763] [Medline: 17530689]
17. Zabel J, Sterz J, Hoefer SH, et al. The use of teaching associates for knee and shoulder examination: a comparative effectiveness analysis. J Surg Educ 2019;76(5):1440-1449. [doi: 10.1016/j.jsurg.2019.03.006] [Medline: 30956084]
18. Cappi V, Artioli G, Ninfa E, et al. The use of blended learning to improve health professionals' communication skills: a literature review. Acta Biomed 2019 Mar 28;90(4-S):17-24. [doi: 10.23750/abm.v90i4-S.8330] [Medline: 30977745]
19. Shiozawa T, Griewatz J, Hirt B, Zipfel S, Lammerding-Koeppel M, Herrmann-Werner A. Development of a seminar on medical professionalism accompanying the dissection course. Ann Anat 2016 Nov;208:208-211. [doi: 10.1016/j.aanat.2016.07.004] [Medline: 27497713]
20. Heinen I, Bullinger M, Kocalevent RD. Perceived stress in first year medical students - associations with personal resources and emotional distress. BMC Med Educ 2017 Jan 6;17(1):4. [doi: 10.1186/s12909-016-0841-8] [Medline: 28056972]
21. Reynolds A, Goodwin M, O'Loughlin VD. General trends in skeletal muscle coverage in undergraduate human anatomy and anatomy and physiology courses. Adv Physiol Educ 2022 Jun 1;46(2):309-318. [doi: 10.1152/advan.00084.2021] [Medline: 35201918]
22. Burger A, Huenges B, Köster U, et al. 15 years of the model study course in medicine at the Ruhr University Bochum. GMS J Med Educ 2019;36(5):Doc59. [doi: 10.3205/zma001267] [Medline: 31815169]
23. Simon M, Martens A, Finsterer S, Sudmann S, Arias J. The Aachen model study course in medicine - development and implementation. Fifteen years of a reformed medical curriculum at RWTH Aachen University. GMS J Med Educ 2019;36(5):Doc60. [doi: 10.3205/zma001268] [Medline: 31815170]
24. McBride JM, Drake RL. National survey on anatomical sciences in medical education. Anat Sci Educ 2018 Jan;11(1):7-14. [doi: 10.1002/ase.1760] [Medline: 29265741]
25. Khalil MK, Giannaris EL, Lee V, et al. Integration of clinical anatomical sciences in medical education: design, development and implementation strategies. Clin Anat 2021 Jul;34(5):785-793. [doi: 10.1002/ca.23736] [Medline: 33905130]

26. Shin M, Prasad A, Sabo G, et al. Anatomy education in US Medical Schools: before, during, and beyond COVID-19. BMC Med Educ 2022 Feb 16;22(1):103. [doi: 10.1186/s12909-022-03177-1] [Medline: 35172819]
27. Xiao J, Evans DJR. Anatomy education beyond the Covid-19 pandemic: a changing pedagogy. Anat Sci Educ 2022 Nov;15(6):1138-1144. [doi: 10.1002/ase.2222] [Medline: 36066879]
28. Park H, Shim S, Lee YM. A scoping review on adaptations of clinical education for medical students during COVID-19. Prim Care Diabetes 2021 Dec;15(6):958-976. [doi: 10.1016/j.pcd.2021.09.004] [Medline: 34736876]
29. Banovac I, Katavić V, Blažević A, et al. The anatomy lesson of the SARS-CoV-2 pandemic: irreplaceable tradition (cadaver work) and new didactics of digital technology. Croat Med J 2021 Apr 30;62(2):173-186. [doi: 10.3325/cmj.2021.62.173] [Medline: 33938657]
30. McWatt SC. Responding to Covid-19: a thematic analysis of students' perspectives on modified learning activities during an emergency transition to remote human anatomy education. Anat Sci Educ 2021 Nov;14(6):721-738. [doi: 10.1002/ase.2136] [Medline: 34523241]
31. Wolniczak E, Roskoden T, Rothkötter HJ, Storsberg SD. Course of macroscopic anatomy in Magdeburg under pandemic conditions. GMS J Med Educ 2020;37(7):Doc65. [doi: 10.3205/zma001358] [Medline: 33364344]
32. Venkatesh S, Rao YK, Nagaraja H, Woolley T, Alele FO, Malau-Aduli BS. Factors influencing medical students' experiences and satisfaction with blended integrated e-learning. Med Princ Pract 2020;29(4):396-402. [doi: 10.1159/000505210] [Medline: 31801145]
33. Kuhn S, Frankenhauser S, Tolks D. Digital learning and teaching in medical education: already there or still at the beginning? Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2018 Feb;61(2):201-209. [doi: 10.1007/s00103-017-2673-z] [Medline: 29234823]
34. Tarpada SP, Hsueh WD, Gibber MJ. Resident and student education in otolaryngology: a 10-year update on e-learning. Laryngoscope 2017 Jul;127(7):E219-E224. [doi: 10.1002/lary.26320] [Medline: 27782300]
35. Van Nuland SE, Rogers KA. The skeletons in our closet: e-learning tools and what happens when one side does not fit all. Anat Sci Educ 2017 Nov;10(6):570-588. [doi: 10.1002/ase.1708] [Medline: 28575530]
36. Green RA, Whitburn LY. Impact of introduction of blended learning in gross anatomy on student outcomes. Anat Sci Educ 2016 Oct;9(5):422-430. [doi: 10.1002/ase.1602] [Medline: 26929149]
37. Brewer PE, Racy M, Hampton M, Mushtaq F, Tomlinson JE, Ali FM. A three-arm single blind randomised control trial of naïve medical students performing a shoulder joint clinical examination. BMC Med Educ 2021 Jul 21;21(1):390. [doi: 10.1186/s12909-021-02822-5] [Medline: 34284771]
38. Kyaw BM, Posadzki P, Paddock S, Car J, Campbell J, Tudor Car L. Effectiveness of digital education on communication skills among medical students: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019 Aug 27;21(8):e12967. [doi: 10.2196/12967] [Medline: 31456579]
39. Garrison DR, Kanuka H. Blended learning: uncovering its transformative potential in higher education. Internet High Educ 2004 Apr;7(2):95-105. [doi: 10.1016/j.iheduc.2004.02.001]
40. Lin DC, Bunch B, De Souza RZD, et al. Effectiveness of pedagogical tools for teaching medical gross anatomy during the COVID-19 pandemic. Med Sci Educ 2022 Apr;32(2):411-422. [doi: 10.1007/s40670-022-01524-x] [Medline: 35228893]
41. Chen F, Lui AM, Martinelli SM. A systematic review of the effectiveness of flipped classrooms in medical education. Med Educ 2017 Jun;51(6):585-597. [doi: 10.1111/medu.13272] [Medline: 28488303]
42. Baskaran R, Mukhopadhyay S, Ganesananthan S, et al. Enhancing medical students` confidence and performance in integrated structured clinical examinations (ISCE) through a novel near-peer, mixed model approach during the COVID-19 pandemic. BMC Med Educ 2023 Feb 23;23(1):128. [doi: 10.1186/s12909-022-03970-y] [Medline: 36823563]
43. Ivarson J, Hermansson A, Meister B, Zeberg H, Bolander Laksov K, Ekström W. Transfer of anatomy during surgical clerkships: an exploratory study of a student-staff partnership. Int J Med Educ 2022 Aug 31;13:221-229. [doi: 10.5116/ijme.62eb.850a] [Medline: 36049218]
44. Avonts M, Michels NR, Bombeke K, et al. Does peer teaching improve academic results and competencies during medical school? A mixed methods study. BMC Med Educ 2022 Jun 4;22(1):431. [doi: 10.1186/s12909-022-03507-3] [Medline: 35659218]
45. Biggs J. Enhancing teaching through constructive alignment. High Educ 1996 Oct;32(3):347-364. [doi: 10.1007/BF00138871]
46. Backhaus J, Huth K, Entwistle A, Homayounfar K, Koenig S. Digital affinity in medical students influences learning outcome: a cluster analytical design comparing vodcast with traditional lecture. J Surg Educ 2019;76(3):711-719. [doi: 10.1016/j.jsurg.2018.12.001] [Medline: 30833205]
47. Farkas GJ, Mazurek E, Marone JR. Learning style versus time spent studying and career choice: which is associated with success in a combined undergraduate anatomy and physiology course? Anat Sci Educ 2016;9(2):121-131. [doi: 10.1002/ase.1563] [Medline: 26301828]
48. Brannick MT, Erol-Korkmaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. Med Educ 2011 Dec;45(12):1181-1189. [doi: 10.1111/j.1365-2923.2011.04075.x] [Medline: 21988659]
49. Gormley GJ, Collins K, Boohan M, Bickle IC, Stevenson M. Is there a place for e-learning in clinical skills? A survey of undergraduate medical students' experiences and attitudes. Med Teach 2009 Jan;31(1):e6-12. [doi: 10.1080/01421590802334317] [Medline: 19253150]

50. Pereira JA, Pleguezuelos E, Merí A, Molina-Ros A, Molina-Tomás MC, Masdeu C. Effectiveness of using blended learning strategies for teaching and learning human anatomy. Med Educ 2007 Feb;41(2):189-195. [doi: 10.1111/j.1365-2929.2006.02672.x] [Medline: 17269953]
51. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010 Mar 23;340:c332. [doi: 10.1136/bmj.c332] [Medline: 20332509]
52. Chan AW, Tetzlaff JM, Altman DG, et al. SPIRIT 2013 statement: defining standard protocol items for clinical trials. Ann Intern Med 2013 Feb 5;158(3):200-207. [doi: 10.7326/0003-4819-158-3-201302050-00583] [Medline: 23295957]
53. Graf J, Smolka R, Simoes E, et al. Communication skills of medical students during the OSCE: gender-specific differences in a longitudinal trend study. BMC Med Educ 2017 May 2;17(1):75. [doi: 10.1186/s12909-017-0913-4] [Medline: 28464857]
54. Zimmermann P, Kadmon M. Standardized examinees: development of a new tool to evaluate factors influencing OSCE scores and to train examiners. GMS J Med Educ 2020;37(4):Doc40. [doi: 10.3205/zma001333] [Medline: 32685668]
55. Franke T, Attig C, Wessel D. A personal resource for technology interaction: development and validation of the Affinity for Technology Interaction (ATI) scale. Int J Hum Comput Interact 2019 Apr 3;35(6):456-467. [doi: 10.1080/10447318.2018.1456150]
56. Braun V, Clarke V, Hayfield N, Terry G. Thematic analysis. In: Liamputtong P, editor. Handbook of Research Methods in Health Social Sciences: Springer; 2019:978-981. [doi: 10.1007/978-981-10-5251-4_103]
57. Fliege H, Rose M, Arck P, et al. The Perceived Stress Questionnaire (PSQ) reconsidered: validation and reference values from different clinical and healthy adult samples. Psychosom Med 2005;67(1):78-88. [doi: 10.1097/01.psy.0000151491.80178.78] [Medline: 15673628]
58. Cohen J. Statistical Power Analysis for the Behavioral Sciences, 2nd edition: Routledge; 2013. [doi: 10.4324/9780203771587]
59. Westphale S, Backhaus J, Koenig S. Quantifying teaching quality in medical education: the impact of learning gain calculation. Med Educ 2022 Mar;56(3):312-320. [doi: 10.1111/medu.14694] [Medline: 34767274]
60. Azizi SM, Roozbahani N, Khatony A. Factors affecting the acceptance of blended learning in medical education: application of UTAUT2 model. BMC Med Educ 2020 Oct 16;20(1):367. [doi: 10.1186/s12909-020-02302-2] [Medline: 33066768]

Abbreviations

ATI-S: Affinity for Technology Interaction Short Scale
CONSORT: Consolidated Standards of Reporting Trials
GP: general practitioner
ILIAS: Integrated Learning, Information, and Work Cooperation System
MCQ: multiple-choice question
OSCE: objective structured clinical examination
PSQ-20: 20-item Perceived Stress Questionnaire
SPIRIT: Standard Protocol Items: Recommendations for Interventional Trials
TraceX: Transfer of anatomical knowledge in the examination situation for preclinical medical students
UKT: Universitätsklinikum Tübingen

Edited by B Lesselroth, S Tsuei; submitted 28.05.24; peer-reviewed by E Ogut, P Chauhan, S Boesner; revised version received 21.07.25; accepted 22.07.25; published 17.09.25.

Please cite as:
Koch R, Gassner L, Gerlach N, Festl-Wietek T, Hirt B, Joos S, Shiozawa T
Integrated e-Learning for Shoulder Anatomy and Clinical Examination Skills in First-Year Medical Students: Randomized Controlled Trial
JMIR Med Educ 2025;11:e62666
URL: https://mededu.jmir.org/2025/1/e62666
doi:10.2196/62666

E-Learning for Pediatric Emergency Department Staff in Point-of-Care Electroencephalogram Interpretation: Prospective Cohort Study

Leopold Simma[1,2], MD; Maurice Henri Schneeberger[3]; Stefanie von Felten[4], PhD; Michelle Seiler[1,2], MD; Georgia Ramantani[2,5], PhD; Bigna Katrin Bölsterli[2,5,6,7], MD

[1]Emergency Department, University Children's Hospital Zurich, Lenggstrasse 30, Zurich, Switzerland

[2]Children's Research Center, University Children's Hospital Zurich, University of Zurich, Zurich, Switzerland

[3]Master Program in Biostatistics, University of Zurich, Zurich, Switzerland

[4]Department of Biostatistics, Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland

[5]Department of Neuropediatrics, University Children's Hospital Zurich, Zurich, Switzerland

[6]Department of Pediatric Neurology, Ostschweizer Kinderspital, St. Gallen, Switzerland

[7]Child Development Center, University Children's Hospital Zurich, Zurich, Switzerland

Corresponding Author:
Leopold Simma, MD
Emergency Department, University Children's Hospital Zurich, Lenggstrasse 30, Zurich, Switzerland

Abstract

Background: Status epilepticus (SE) represents a critical pediatric emergency necessitating prompt treatment and monitoring. The diagnosis of nonconvulsive SE and the monitoring of convulsive SE require electroencephalogram (EEG) recordings. The integration of simplified point-of-care EEG may improve care in pediatric emergency departments.

Objective: This study aims to assess the efficacy of an electronic EEG self-learning module for improving the interpretation of normal cortical activity, artifacts, and seizure patterns in point-of-care EEG by pediatric emergency medicine (PEM) providers.

Methods: This prospective cohort study was conducted in a tertiary academic pediatric emergency department and primarily targeted senior medical staff while also engaging junior medical staff and registered nurses. A novel EEG e-learning module trained participants to identify normal cortical activity, artifacts, and seizure patterns. The study comprised pretest, posttest, and 3-month retention assessments to evaluate the EEG total score as its primary outcome and basic EEG knowledge and confidence measures as secondary outcomes. Outcomes were analyzed using mixed-effects proportional odds logistic regression models.

Results: Of 102 PEM providers invited, 61 individuals participated (25 senior medical staff, 15 junior medical staff, and 21 registered nurses), and 29 finished the 3-tiered study. In finishers, the EEG total score (max=12 points), indicative of accurate EEG classification, increased substantially between pretest and posttest from a median of 7 (IQR 5 - 8) to 10 (IQR 7 - 11) points, corresponding with an increase in the odds of achieving higher EEG total scores at the posttest (odds ratio 24.18, 95% CI 7.398-79.043, *P*<.001). At the retention test, the EEG total score remained elevated, although to a lesser extent (median 8 points [IQR 6 - 9]). Similar trends were observed in secondary outcomes.

Conclusions: The implementation of an e-learning EEG module improved the ability of PEM providers to interpret EEGs. This study highlights the feasibility of imparting basic EEG skills to nonexperts through targeted educational interventions. However, the sustained retention of such skills requires improvement, emphasizing the necessity for ongoing refresher training.

(JMIR Med Educ 2025;11:e69395) doi:10.2196/69395

KEYWORDS

electroencephalography; medical education; seizures; status epilepticus; critical care; pediatric emergency medicine; point-of-care systems; emergency service

Introduction

Acute central nervous system disorders are common childhood emergencies [1]. These disorders are very frequent among highly acute presentations in pediatric emergency departments (PEDs) and are the most frequent type of presentation in pediatric resuscitation bays [2-4]. Status epilepticus (SE) is a paramount medical emergency in pediatric neurology. Early identification of SE in PEDs is crucial for prompt initiation of treatment and thus for improving patient outcomes [5]. However, nonconvulsive SE (NCSE) often eludes detection, particularly in patients with altered mental status (AMS) or seemingly

controlled convulsive SE [6,7]. Standard electroencephalogram (EEG) is the gold standard for monitoring SE treatment and detecting NCSE due to its high sensitivity in identifying electrographic-only seizure activity [8].

Although electroencephalograms (EEGs) in PEDs provide valuable diagnostic information and aid in decision-making [9,10], immediate access to standard EEG may be challenging in many settings [10]. Low standard EEG availability poses a challenge, both during and after standard working hours, because standard EEG requires significant time, equipment, and specialized personnel [11]. A simplified, reduced-lead EEG, also termed point-of-care EEG (pocEEG), has emerged as a feasible option [12] to aid diagnostics [13-15] and facilitate the management of AMS, the treatment and monitoring of SE, and the detection of NCSE [15,16]. However, the interpretation of pocEEG by nonexpert PED staff in the absence of or with delayed access to EEG experts, such as pediatric neurologists, is challenging, undermining the applicability of this novel approach in the PED setting [17].

Previous research has explored efforts to train nonexperts in EEG interpretation in both adult and neonatal and pediatric acute care settings [18,19], with most studies focusing on processed EEG, such as amplitude integrated EEG or continuous EEG in intensive care units [18]. Yet, only one study has investigated the acquisition of standard EEG interpretation skills in adult emergency department (ED) physicians [20]. To date, data are lacking on the feasibility and efficacy of teaching basic EEG skills to pediatric emergency medicine (PEM) providers, even though here this approach could significantly improve patient care by facilitating AMS evaluation and expediting NCSE diagnosis and treatment [17,21]. To address this gap, we designed a novel pocEEG e-learning module to transfer basic EEG knowledge of montages, EEG signal generation, and EEG reading and interpretation skills to PEM providers. The module focuses on identifying normal cortical activity, artifacts, and seizure patterns. These 3 domains are important for distinguishing between true neurological events and artifacts, and thus for identifying and managing seizures.

Therefore, the primary aim of this study was to determine the impact of our EEG e-learning module on EEG skill acquisition by nonexpert PED staff through a 3-tiered longitudinal assessment. Secondary aims were to assess whether training led to measurable improvement in recognizing specific EEG patterns, self-assessment, and confidence among PEM providers over time.

Methods

Study Design and Setting

This study was conducted at the tertiary academic PED of the University Children's Hospital Zurich, Switzerland, which treats 50,000 patients annually. The department is staffed by diverse professional groups, including senior medical staff (SMS) such as PEM attendings and PEM fellows, junior medical staff (JMS) such as pediatric residents and pediatric surgery residents, registered nurses (RNs), and medical practice assistants.

In this prospective cohort study, we tested an e-learning module as an intervention primarily targeting the SMS, who have the final responsibility for patient evaluations. However, all PEM providers were invited to participate in a 3-visit program consisting of a baseline pretest followed by a posttest and a 3-month retention test.

Study Population

For enrollment, 102 PEM providers were invited via email using the REDCap (Research Electronic Data Capture; Vanderbilt University) survey tool [22]. We extended invitations for anonymous participation in this e-learning module to all SMS, JMS, and RNs within our PED. Our primary goal was to actively engage the majority of SMS while also providing access to EEG training and promoting the use of pocEEG in the PED by JMS and RNs. After completion of the pretest, participants were provided with a passcode to access the learning module. One hour after pretest completion, participants were provided with another survey link and prompted to complete the posttest by entering a passcode provided on module completion. The retention test was automatically scheduled 3 months following posttest completion. Up to 4 automated reminders were sent to encourage participation between 88 and 120 days after pretest. Pretest participants who failed to complete the learning module were invited to join a control group by completing the 3-month retention test.

Study Protocol

In a collaborative effort, a PEM expert and a neurophysiologist (LS and BKB) devised an innovative e-learning module aimed at facilitating the use and interpretation of pocEEG. Drawing on insights from a previous adult study [20], this module incorporates audio-guided PowerPoint (Microsoft Corp) presentations. The content covers fundamental EEG principles, including electrode placement according to the 10‐20 system, a standardized method of electrode placement, and insights into standard EEG and pocEEG findings. The module focuses on identifying normal EEG waveforms, artifacts, and seizure patterns in accordance with the American Clinical Neurophysiology Society 2021 guidelines [23].

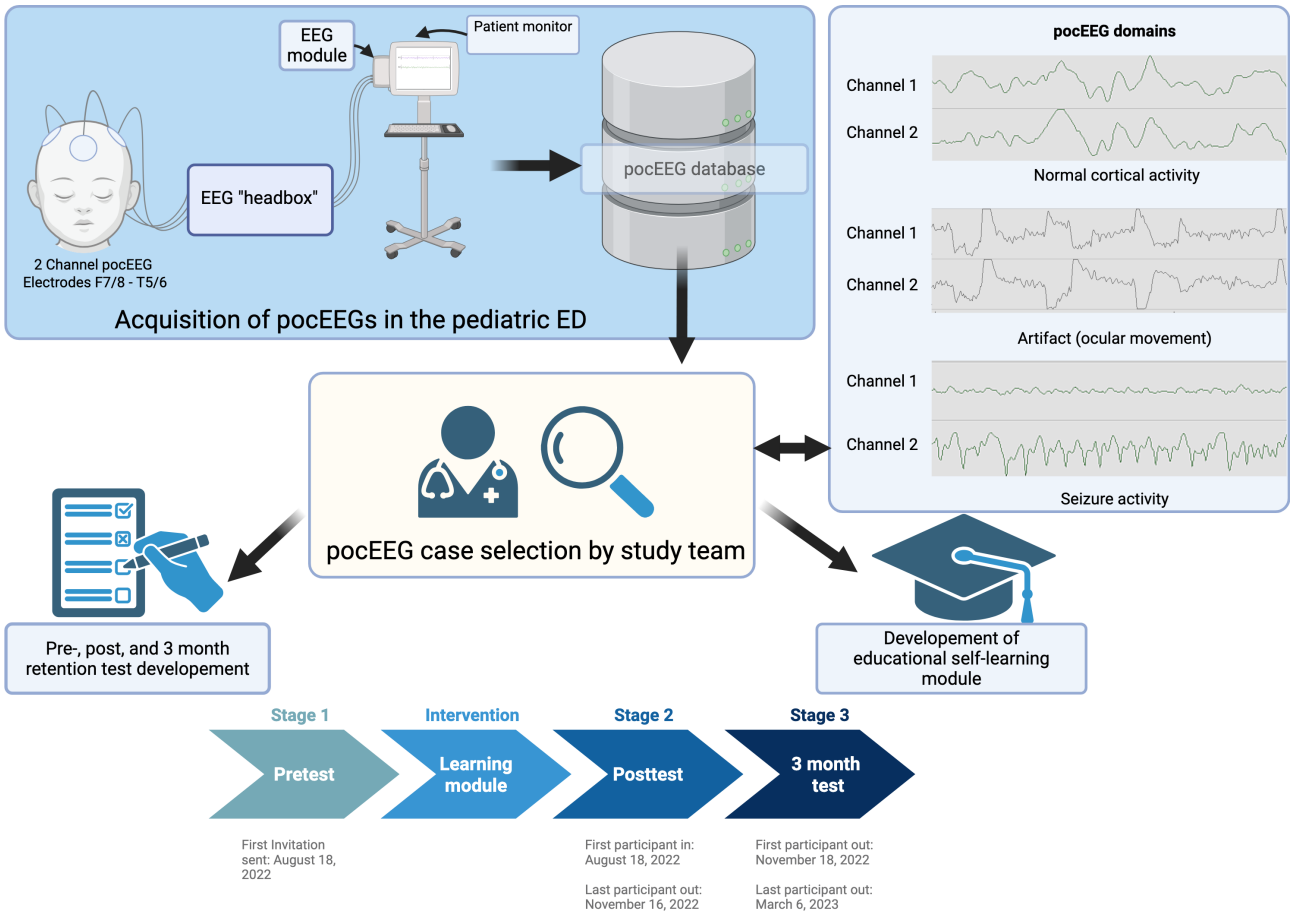
The main segment of the module, spanning about 60 minutes, is dedicated to two primary objectives: (1) providing a systematic framework for assessing cortical activity by familiarizing participants with normal EEG waveforms, and (2) educating participants about critical EEG findings such as seizure patterns in pocEEGs. Additionally, a 20-minute chapter offers technical instructions for pocEEG covering electrode placement, EEG, and optionally video recording, and troubleshooting techniques. Although this technical knowledge is not tested, it enhances participants' understanding of pocEEG procedures.

Measurements and Test Materials

Participants were assessed on their ability to distinguish normal cortical activity, artifacts, and seizure patterns by reviewing 12 anonymized pocEEGs. These pocEEGs were acquired via patient monitors from patients presenting to the PED during the introduction phase of this modality [24] and consisted of 2-channel pocEEG recordings (channels F7/8 and T5/6; 10‐20

system) as described elsewhere [24]. These recordings were processed with Vitalrecorder (version 1.8.16.4) [25], and educationally valuable EEGs were carefully selected from the pocEEG database by PEM and neurophysiology experts (LS and BKB; Figure 1).

Figure 1. pocEEG acquisition, learning module, and test development (created with BioRender.com). EEG: electroencephalogram; pocEEG: point-of-care electroencephalogram.



Each test session included basic demographic data on the participants' profession, seniority, and post board certification experience; a self-assessment of confidence; and basic EEG knowledge. The 12 EEG interpretation tasks comprised 7 pocEEG trace snapshots (10 s duration) and 5 pocEEGs with video segments lasting up to 60 seconds each. Each of the 3 domains, normal cortical activity, artifacts, and seizure patterns, was assessed with 4 pocEEGs. EEG interpretation knowledge was tested with single-choice questions, each with 5 answer options. Each answer included a statement about background symmetry, dominant activity, and whether the pocEEG included an artifact or indicated seizure activity. Additionally, participants were asked to disclose which of their answers they had guessed [26]. To minimize bias, question sequences varied across all 3 tests. A pediatric neurology fellow who was not involved in the study piloted the test material. Feedback was used to adjust item difficulty and ensure technical feasibility. Study data were collected and managed using the REDCap electronic data capture tool hosted at the University Children's Hospital Zurich [22]. We report this survey data in accordance with the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) guidelines [27], and a detailed CHERRIES checklist is provided in Checklist 1.

Outcome Measures

We assessed both the theoretical knowledge and practical skills of participants in pocEEG interpretation through a series of 3 tests, which yielded these outcomes:

- Primary Outcome
  - EEG total score: This score is calculated from the accurate classification of 12 pocEEG recordings into categories of normal cortical activity, artifacts, or seizure patterns. Scores range from 0 (no correct answers) to 12 (all correctly answered).
- Secondary Outcomes
  - EEG basic knowledge score: This score reflects the accuracy of answers to 5 basic EEG knowledge questions (single choice, each with 4 answer options) evaluated across 3 tests. Scores range from 0 (no correct answer) to 5 (all answered correctly).
  - Classification accuracy per domain: The 12 EEG recordings are categorized into 3 domains: normal activity, artifacts, and seizure patterns, each represented by 4 EEGs. The classification of each EEG recording is evaluated with a binary outcome (correct/incorrect).
  - Assurance level: This includes the total number of answers guessed per test, ranging from 0 (no guesses)

to 12 (all guessed), and the number of correct, deliberate answers per test, ranging from 0 to 12.
- Self-assessment and confidence measures: Five items measured on a Likert scale from 1 (completely agree) to 5 (completely disagree).
- Estimation of an intervention effect: To measure the effect of the e-learning module, we compare the EEG total scores of finishers and the control group in the 3-month retention test.

Statistical Analysis

Descriptive statistics of study participants were tabulated overall, by professional group, and by group depending on test completion (see below). We report frequency and percentage for categorical variables.

The ordinal primary outcome EEG total score was analyzed by mixed-effects proportional odds logistic regression (PolrME) and mixed-effects enhanced proportional odds logistic regression (ePolrME) with the PolrME and ColrME functions in the R package tramME [28]. For simplicity, we report only the ePolrME results. To account for repeated measurements, random intercepts per study participant were included in the models. To investigate changes in the EEG total score between tests in participants who completed all tests, models were fitted with test (pre-, post-, and 3 mo retention) and professional group (SMS, JMS, and RNs) as explanatory variables. The analysis primarily focused on SMS, the largest professional group. To test whether changes in EEG total score across all tests differed between professional groups, we included an interaction between test and professional group in the models and used likelihood ratio tests to assess whether interaction terms improved model fit.

Similar models were fitted for the secondary outcome EEG basic knowledge score. To investigate the effect of completing the e-learning module as an intervention on EEG total score, we fitted an ordinary proportional odds logistic regression model to participants who completed all components of the study and to a control group who only completed the pretest and retention test without engaging with the e-learning module or posttest. The EEG total score at the pretest and completion of the e-learning module (yes/no) was used as explanatory variables.

The PolrME models (and proportional odds logistic regression model) estimate odds ratios (ORs) for falling into a higher EEG total score category, comparing a given level of the explanatory variables to the reference group (eg, posttest vs pretest or e-learning module completed vs not completed) for a given study participant. Similarly, ePolrME models estimate ORs for higher EEG total scores. The interpretation remains consistent when the models are applied to the secondary outcome, EEG basic knowledge. To assess differences in recognizing seizure patterns, artifacts, and normal cortical activity, a binomial general linear mixed effects model with logit link function was fitted to the correct EEG rating (binary outcome, 1=correct, 0=false). The data set was prepared to contain a row for each rating of an EEG by a participant (12 EEG recordings in each test and participant). The model included a random intercept for test nested within study participant and the type of EEG (normal cortical activity, artifact, or seizure pattern) and test as explanatory variables, and estimated ORs for correct rating of EEG recordings. All analyses were performed using R (version 4.1.2; R Foundation for Statistical Computing) [29].

Ethical Considerations

This study involves human participants but does not fall within the scope of the Swiss Human Research Act, and the responsible ethics committee (Cantonal Ethics Committee, Zürich, Switzerland) exempted this study (Req-2024 - 00833). All participants were volunteers and gave electronic consent to participate in the study.

Results

Overview

In all, 102 PEM providers were invited to participate in the study, and 61 consented to take the pretest (Figure 2), including 25 SMS, 15 JMS, and 21 RNs (Table 1). A total of 53 (86.9%) participants were female, 25 (41.0%) participants were SMS, 15 (24.6%) participants were JMS, and 21 (34.4%) participants were RNs. Participants had minimal or no prior experience in implementing and interpreting EEGs.

Table . Participant characteristics.

	Overall (n=61)	Control group (n=13)	Finisher (n=29)	Nonfinisher (n=19)
Sex, n (%)				
Female	53 (86.9)	11 (84.6)	24 (82.8)	18 (94.7)
Male	8 (13.1)	2 (15.4)	5 (17.2)	1 (5.3)
Professional group, n (%)				
Senior medical staff	25 (41.0)	2 (15.4)	21 (72.4)	2 (10.5)
Junior medical staff	15 (24.6)	1 (7.7)	4 (13.8)	10 (52.6)
Registered nurses	21 (34.4)	10 (76.9)	4 (13.8)	7 (36.8)
ED[a] experience, n (%)				
0-5 y	15 (24.6)	4 (30.8)	9 (31.0)	2 (10.5)
5-10 y	11 (18.0)	4 (30.8)	5 (17.2)	2 (10.5)
10-15 y	10 (16.4)	1 (7.7)	9 (31.0)	0 (0.0)
+15 y	10 (16.4)	3 (23.1)	2 (6.9)	5 (26.3)
JMS[b]	15 (24.6)	1 (7.7)	4 (13.8)	10 (52.6)
Prior EEG[c] experience, n (%)				
No (no prior experience)	18 (29.5)	5 (38.5)	6 (20.7)	7 (36.8)
No (passive knowledge)	18 (29.5)	4 (30.8)	10 (34.5)	4 (21.1)
Yes (neonatal EEG[c] [aEEG[d]])	19 (31.1)	3 (23.1)	9 (31.0)	7 (36.8)
Yes (practical experience)	6 (9.8)	1 (7.7)	4 (13.8)	1 (5.3)
Yes (proficient)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)

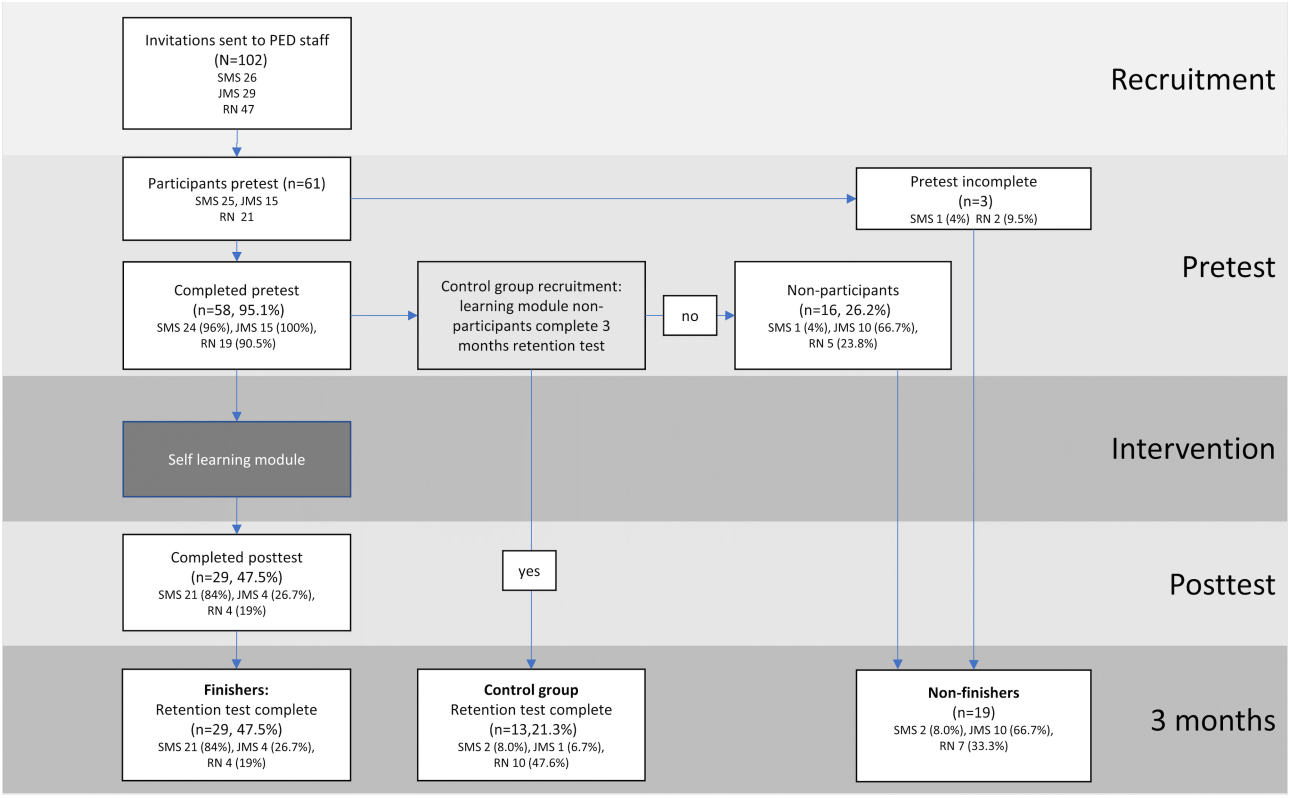
[a]ED: emergency department.

[b]JMS: junior medical staff.

[c]EEG: electroencephalogram.

[d]aEEG: amplitude integrated electroencephalogram.

Figure 2. Flowchart of the study's stages from recruitment to pretest, intervention, posttest, and study conclusion. Percentages are calculated relative to the initial participant pool who took the pretest (n=61). Subgroup percentages always refer to the respective subgroup sample. JMS: junior medical staff (residents); RN: registered nurse; SMS: senior medical staff (attendings/consultants; fellows).



Of the 61 participants who took the pretest, 58 completed it and subsequently gained access to the e-learning module. Among these, 29 completed the e-learning module, posttest, and retention test and were designated as finishers (Table 1). Participants who only completed the pretest (n=29) and did not interact with the e-learning module formed a control group and were invited to take the 3-month test, resulting in 13 controls with a pretest and a 3-month test. The group of nonfinishers (n=19) consisted of 3/61 without and 16/61 with a complete pretest. The finisher rate among SMS, the primary focus group of our e-learning module, was 84% (21/25), while it was 26.7% (4/15) in JMS, and 19.0% (4/21) in RNs. The distribution of the various participants is shown in Figure 2.

Primary Outcome EEG Total Score

The primary outcome, EEG total score, increased considerably on average following completion of the e-learning module among all finishers. However, it subsequently decreased between the posttest and the retention test, although it remained higher than at the pretest (Figures 3A and 3B, Table 2, Figure S1 in Multimedia Appendix 1).

Figure 3. (A) Line plot for the progression of EEG total scores over pretest, posttest, and 3-month retention test of finishers: top panel senior medical staff (SMS, n=21), middle panel junior medical staff (JMS, n=4), and bottom panel registered nurses (RN, n=4). (B) Barplots for number of correct answers on pretest, posttest, and 3-month retention tests and per EEG domain artifacts (Artif), normal cortical activity (Norm), and seizure patterns (Path). (C) Lineplot showing the progression of scores in basic EEG knowledge; staff groups are shown in 3 separate panels. (D) Lineplot for progression of nonguessed correct responses (EEG total score) across the 3 tests with top panel SMS, middle panel JMS, and bottom panel RNs. EEG: electroencephalogram; JMS: junior medical staff; RN: registered nurse; SMS: senior medical staff.

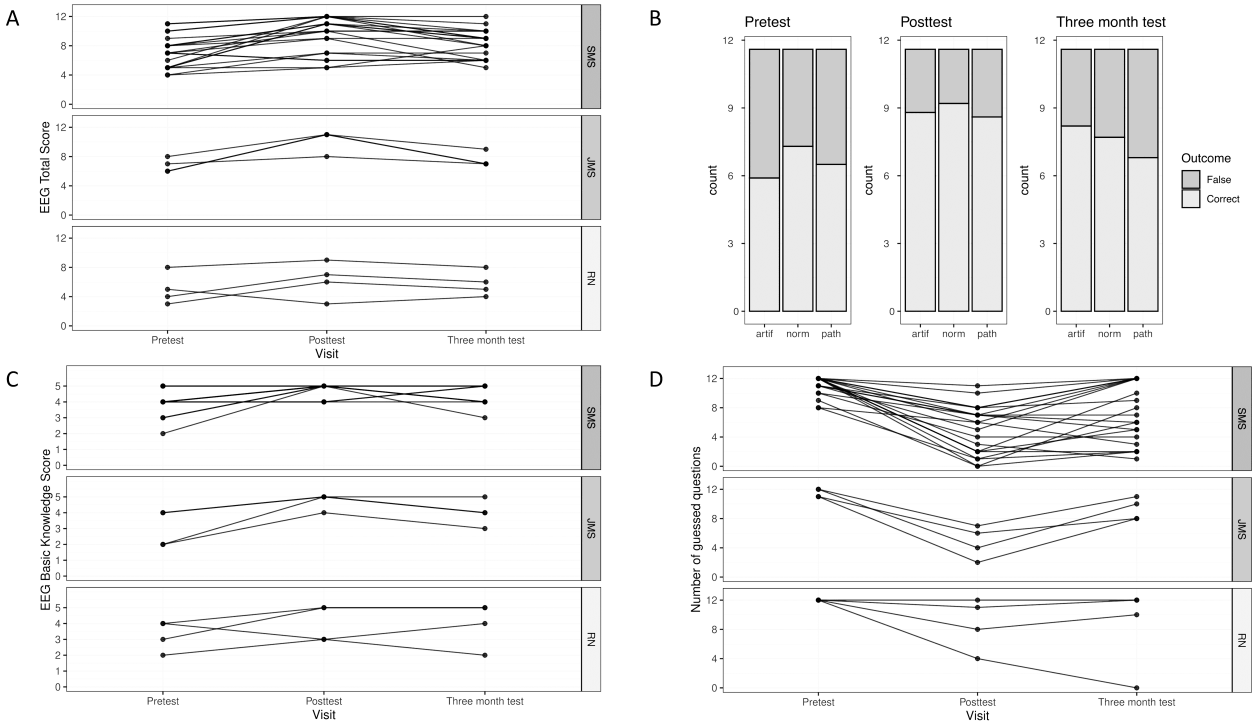


Table . Primary outcome electroencephalogram (EEG) total scores of finishers, finisher subgroups, control group, and nonfinishers at all 3 tests (maximum score=12). Results are summarized as median and IQR (1st and 3rd quartile).

	Finishers total (n=29)	Finisher SMS[a] (n=21)	Finisher JMS[b] (n=4)	Finisher RNs[c] (n=4)	Control group (n=13)	Nonfinisher (n=19)
EEG (total scores pretest), median (IQR)	7.00 (5.00-8.00)	7.00 (5.00-8.00)	6.50 (6.00-7.25)	4.50 (3.75-5.75)	5.00 (3.00-7.00)	6.00 (5.00-7.00)
EEG (total scores posttest), median (IQR)	10.00 (7.00-11.00)	10.00 (7.00-12.00)	11.00 (10.25-11.00)	6.50 (5.25-7.50)	0.00 (0.00-0.00)	0.00 (0.00-0.00)
EEG (total scores 3m[d] test), median (IQR)	8.00 (6.00-9.00)	9.00 (6.00-10.00)	7.00 (7.00-7.50)	5.50 (4.75-6.50)	4.00 (3.00-7.00)	0.00 (0.00-0.00)

[a]SMS: senior medical staff.

[b]JMS: junior medical staff.

[c]RN: registered nurse.

[d]3m: 3 months.

Results from the ePOLR model provide evidence for an increase in the odds of achieving higher EEG total scores at the posttest compared with the pretest (OR 24.18, 95% Wald CI 7.398-79.043; *P*<.001). Similarly, the odds of achieving higher EEG total scores at the retention test compared with the pretest increased by a factor of 3.7 (OR 3.7, 95% Wald CI 1.406-9.741; *P*=.008). Furthermore, differences emerged in EEG total scores among professional groups, with SMS achieving the highest scores, followed by JMS and RNs (OR compared with SMS of 0.79, 95% Wald CI 0.08-8.06; *P*=.84; and 0.03, 95% Wald CI 0.00-0.33; *P*=.005, respectively; Table 3).

Table . Odds ratio (OR) estimates with 95% Wald CIs from the enhanced proportional odds logistic regression model on the primary outcome, electroencephalogram total score. The model included 87 observations from 29 participants.

	OR (95% CI)	*P* value
JMS[a] versus SMS[c]	0.79 (0.08-8.06)	.84
RNs[b] versus SMS	0.03 (0.00-0.33)	.005
Posttest versus pretest	24.18 (7.40-79.05)	<.001
3 mo retention test versus pretest	3.70 (1.41-9.74)	.008

[a]JMS: junior medical staff.

[b]RN: registered nurse.

[c]SMS: senior medical staff.

Secondary Outcomes

Among SMS, we observed that EEG basic knowledge generally increased after the intervention, with a less pronounced decline at the retention test than at the EEG total score. This pattern was also seen in JMS and RNs, albeit with some individual variations (Figure 3C). In the ePolrME model, SMS odds for a higher score in basic EEG knowledge after the module increased significantly for the posttest (OR 32.89, 95% Wald CI 7.137-151.569; *P*<.001) and for the retention test (OR 12.96, 95% Wald CI 3.421-49.122; *P*<.001). JMS and RNs showed a decrease in odds for a higher score compared with SMS (OR 0.09 Wald CI 0.01-0.79, *P*=.03; and OR 0.06, Wald CI 0.01-0.52, *P*=.01; respectively).

In the secondary outcome analysis of EEG domains, the number of correct answers for normal cortical activity (norm), artifacts (artif), and seizure patterns (path) is shown in Figure 3B. Overall, the number of correct answers increased markedly in the posttest and remained higher in the retention test compared with baseline (Table 4). We observed minor variations among the 3 domains across these tests.

Table . Odds ratio (OR) estimates with 95% Wald CIs from the binomial generalized linear mixed-effects model on classification accuracy (binary outcome for correctness of answer). The model was fitted with a logit link function and a random intercept for visit nested within participant and included 1044 observations from 29 participants.

	OR (95% CI)	*P* value
Intercept	1.16 (0.78-1.73)	.47
Artifact versus pathological	1.15 (0.83-1.61)	.40
Normal cortical activity versus pathological	1.40 (1.00-1.96)	.049
Posttest versus pretest	2.75 (1.95-3.87)	<.001
Three-month retention test versus pretest	1.50 (1.09-2.07)	.014

To assess the participants' assurance level, we analyzed the guess rates for 12 EEG recording questions across professional groups (Figure 3D). The pretest revealed high guess rates across all groups, with SMS guessing between 8 and 12 questions, and many participants guessing all. Completion of the e-learning module was followed by a notable reduction in guess rates across all groups, with some senior staff not guessing any. However, at the 3-month retention test, guess rates increased again for all groups compared with posttest.

We evaluated the confidence levels of finishers in their interpretation of pocEEG by focusing on overall pocEEG skills, ability to detect pocEEG signals correctly, knowledge about potential artifacts, and proficiency in identifying artifacts on a Likert scale (Figures S2-S6 in Multimedia Appendix 1). During the pretest, responses regarding overall pocEEG skills varied widely across all groups, indicating a significant disparity in confidence levels. However, an improvement in confidence was observed across all occupational groups during the posttest. This increased confidence largely remained at the retention test, albeit with some variations among groups. Notably, SMS consistently exhibited higher confidence in overall skills throughout the study compared with JMS and RNs. Confidence in detecting pocEEG signals varied widely across occupational groups during the pretest but improved significantly after the completion of the e-learning module. This increased confidence remained high in the retention test, although JMS and RNs showed lower confidence levels than SMS. Likewise, confidence in identifying EEG artifacts was initially low across all occupational groups but improved notably after the e-learning module. This increased confidence generally remained in the retention test, with some variations observed among groups. Throughout the study, confidence in interpreting pocEEG recordings remained relatively low across all occupational groups, with only slight increases observed in certain subgroups and a general trend of continued uncertainty or disagreement.

Effect of the E-Learning Module: Finishers Versus Control Group

Comparing finishers with the control group revealed that the e-learning module led to increased EEG total scores at the retention test after 3 months.

The odds of achieving higher EEG total scores in the retention test were significantly increased by a factor of 9 for the finishers compared with the controls (OR 9.075, 95% CI 2.181-37.763;

P=.002). Furthermore, adjusting for initial EEG knowledge at pretest was important because participants with better pretest performance also performed better at the 3-month retention test (OR 2.053, 95% CI 1.505-2.800; *P*<.001).

Discussion

Principal Findings

This study evaluated the effectiveness of a novel e-learning EEG module in improving pocEEG interpretation skills among SMS, JMS, and RNs in the PED. Our findings indicate a notable improvement in EEG knowledge and interpretation skills across all participant groups following the module, which was particularly evident in the EEG total score, our primary outcome. These results are consistent with the literature emphasizing the importance of targeted educational interventions for nonexperts in interpreting EEGs in critical care settings [18]. To our knowledge, this study represents the first attempt to evaluate knowledge transfer of basic EEG skills to EEG nonexperts in the PED and the first investigation into training for reduced lead EEG and pocEEG skills. A recent review has also identified this topic as a research priority [17].

The primary outcome, the EEG total score, showed a substantial increase in correct classifications of EEG recordings immediately after module completion and, to a lesser extent, at the 3-month follow-up. This finding is encouraging because it demonstrates both the immediate impact and the lasting effect of the educational module, albeit with a decline over time. The increase in EEG total scores was particularly pronounced among the primary target group, SMS.

However, due to our limited sample size for JMS and RNs, we did not assess whether the increase in EEG total scores or EEG knowledge in general differed from SMS. Overall, finishers were significantly more likely to achieve higher EEG total scores at the 3-month retention test compared with the control group, with the caveat that the control group had a divergent composition (Figure 2 and Table 1).

The finisher rate among SMS was 84%, considerably higher than among JMS and RNs. This discrepancy can be attributed to several factors: JMS typically rotate through our PED for 3 to 6 months, and both JMS and most RNs lack nonclinical time in the PED and remote access to the hospital's computer system, thus preventing their participation from home. This underscores the need for tailored educational approaches that consider the specific needs, time constraints, and baseline competencies of various health care professionals.

Seizure pattern recognition appeared to be a slightly more challenging task for participants (Figure 3B) in our set of test EEGs. Extrapolating from our experience with point-of-care ultrasound, we had previously hypothesized that identifying seizure patterns would be more straightforward than discerning normal activity [30] or artifacts, but findings may differ with a larger set of EEGs.

Nevertheless, the improvement observed in identifying artifacts and normal cortical activity is encouraging and indicates a fundamental enhancement in EEG interpretation skills. The 3 domains of normal signal, artifact, and seizure pattern form the foundational pillars of continuous EEG teaching [31]. They are crucial for improving neurocritical care in the PED, where timely and accurate EEG interpretation can significantly impact patient management and outcomes.

Our additional observations, focusing on self-assessment and confidence measures, indicated a notable increase in confidence levels among participants. The initial high rate of guesses, which decreased significantly post training, highlights the initial uncertainty among staff members in interpreting EEGs. However, the resurgence in guess rates at the 3-month follow-up suggests that periodic refresher modules or ongoing training might be necessary to maintain confidence and competence levels. This retention challenge, as with any critical care skill, is particularly pronounced in ED settings [32,33].

A recent review by Taran et al [18] outlined numerous studies conducted in critical care settings involving nurses and physicians as participants and focusing on EEG educational initiatives. These studies often evaluated the use of processed EEGs, such as amplitude-integrated EEGs, with raw EEGs being less common [18]. Only one study in the adult ED context improved the abilities of emergency physicians to identify seizure patterns; that study used a learning module with full montage standard EEG examples [20]. Studies investigating long-term retention of EEG knowledge reported better scores [19], predominantly in intensive care units, where trainees had frequent exposure to continuous EEG.

One of the key strengths of our study is its high participation rate among SMS, our primary focus group. The 3-tiered longitudinal assessment helps to gauge knowledge retention over time and assess the long-term impact of the intervention. The involvement of diverse occupational groups in the study also provides a broader understanding of educational needs across different levels of PEM providers. By assessing guess rates, we also examined participants' confidence in their answers.

However, our study also has limitations. Its participants are all from a single pediatric academic tertiary center, which limits generalizability. The sample size of finishers other than SMS was small and limited the statistical analysis of differences between professional groups. In addition, the disparity in subgroup sizes contributed to substantial uncertainty in OR estimates, particularly for the EEG total score and EEG basic knowledge score. The reliance on self-assessment measures for confidence may introduce bias and may be influenced by local cultural factors. The composition of the control group was suboptimal due to the finite availability of study participants at the study site, which precluded the recruitment of matched controls. Selective participation and initial participant attrition limit the generalizability of our findings to broader interdisciplinary teams. In addition, the test instrument in this study was tailored for the local educational context and was not formally validated. It is unclear whether any participants engaged in interpreting pocEEGs on clinical shifts during the interval prior to the retention test, which could have influenced their knowledge retention. Finally, the decline in knowledge

retention over time suggests the need for ongoing training and support.

Conclusions

Our study demonstrates that a targeted educational module can improve EEG knowledge and interpretation skills among PED staff. This finding has important implications for enhancing neurocritical care in pediatric emergency settings, where rapid and accurate diagnosis is crucial for patient outcomes. Moreover, the decline in skill retention over time underscores the critical need for continuous education and ongoing training to ensure that these vital skills are maintained.

Acknowledgments

This study would not have been possible without the help of the nursing and medical staff in the Pediatric Emergency Department of the University Children's Hospital Zurich. We would like to thank Prof Geetha Chari (Department of Neurology, SUNY Downstate Medical Center, Brooklyn, NY, US) for sharing the teaching material that she used in her study [20]. Simon Milligan helped with editorial assistance. LS received grant money for investigator-initiated research from the Anna Mueller Grocholski Foundation, Zurich, Switzerland; and the Children's Research Center, University Children's Hospital Zurich, University of Zurich, Zurich, Switzerland.

Data Availability

The data sets used and analyzed in this study are in Multimedia Appendix 2.

Authors' Contributions

LS was responsible for conceptualization, methodology, software development, investigation, resource provision, data curation, drafting the original manuscript, reviewing and editing the manuscript, and creating visualizations. MHS contributed to formal analysis, visualization, and manuscript review and editing. SVF was involved in methodology, formal analysis, visualization, and manuscript review and editing. MS contributed to conceptualization, methodology, investigation, and manuscript review and editing. GR was responsible for validation and reviewing and editing the manuscript. BKB contributed to conceptualization, methodology, software development, validation, investigation, supervision, and reviewing and editing the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional figures.
[PDF File, 905 KB - mededu_v11i1e69395_app1.pdf]

Multimedia Appendix 2
Dataset.
[ZIP File, 5985 KB - mededu_v11i1e69395_app2.zip]

Checklist 1
Checklist for Reporting Results of Internet E-Surveys (CHERRIES).
[PDF File, 909 KB - mededu_v11i1e69395_app3.pdf]

References

1. Pallin DJ, Goldstein JN, Moussally JS, Pelletier AJ, Green AR, Camargo CA Jr. Seizure visits in US emergency departments: epidemiology and potential disparities in care. Int J Emerg Med 2008 Jun;1(2):97-105. [doi: 10.1007/s12245-008-0024-4] [Medline: 19384659]
2. Chavez H, Garcia CT, Sakers C, Darko R, Hannan J. Epidemiology of the critically Ill child in the resuscitation bay. Pediatr Emerg Care 2018 Jan;34(1):6-9. [doi: 10.1097/PEC.0000000000000682] [Medline: 26999585]
3. Lutz N, Vandermensbrugghe NG, Dolci M, Amiet V, Racine L, Carron PN. Pediatric emergencies admitted in the resuscitation room of a Swiss university hospital. Pediatr Emerg Care 2014 Oct;30(10):699-704. [doi: 10.1097/PEC.0000000000000231] [Medline: 25272075]
4. Simma L, Stocker M, Lehner M, Wehrli L, Righini-Grunder F. Critically Ill children in a Swiss pediatric emergency department with an interdisciplinary approach: a prospective cohort study. Front Pediatr 2021;9(1029):721646. [doi: 10.3389/fped.2021.721646] [Medline: 34708009]

5. Gaínza-Lein M, Fernández IS, Ulate-Campos A, Loddenkemper T, Ostendorf AP. Timing in the treatment of status epilepticus: from basics to the clinic. Seizure 2019 May;68:22-30. [doi: 10.1016/j.seizure.2018.05.021] [Medline: 29884518]
6. Sánchez Fernández I, Sansevere AJ, Guerriero RM, et al. Time to electroencephalography is independently associated with outcome in critically ill neonates and children. Epilepsia 2017 Mar;58(3):420-428. [doi: 10.1111/epi.13653] [Medline: 28130784]
7. Fung FW, Abend NS. EEG monitoring after convulsive status epilepticus. J Clin Neurophysiol 2020 Sep;37(5):406-410. [doi: 10.1097/WNP.0000000000000664] [Medline: 32890062]
8. Tay SKH, Hirsch LJ, Leary L, Jette N, Wittman J, Akman CI. Nonconvulsive status epilepticus in children: clinical and EEG characteristics. Epilepsia 2006 Sep;47(9):1504-1509. [doi: 10.1111/j.1528-1167.2006.00623.x] [Medline: 16981867]
9. Fernández IS, Loddenkemper T, Datta A, Kothare S, Riviello JJJ, Rotenberg A. Electroencephalography in the pediatric emergency department: when is it most useful? J Child Neurol 2014 Apr;29(4):475-482. [doi: 10.1177/0883073813483570] [Medline: 23594820]
10. Kothare SV, Khurana DS, Valencia I, Melvin JJ, Legido A. Use and value of ordering emergency electroencephalograms and videoelectroencephalographic monitoring after business hours in a children's hospital: 1-year experience. J Child Neurol 2005 May;20(5):416-419. [doi: 10.1177/08830738050200050401] [Medline: 15968926]
11. Barcia Aguilar C, Sánchez Fernández I, Loddenkemper T. Status epilepticus-work-up and management in children. Semin Neurol 2020 Dec;40(6):661-674. [doi: 10.1055/s-0040-1719076] [Medline: 33155182]
12. Stephens CM, Mathieson SR, McNamara B, et al. Electroencephalography quality and application times in a pediatric emergency department setting: a feasibility study. Pediatr Neurol 2023 Nov;148:82-85. [doi: 10.1016/j.pediatrneurol.2023.08.016]
13. Nozawa M, Terashima H, Tsuji S, Kubota M. A simplified electroencephalogram monitoring system in the emergency room. Pediatr Emerg Care 2019 Jul;35(7):487-492. [doi: 10.1097/PEC.0000000000001033] [Medline: 28072672]
14. Yamaguchi H, Nagase H, Nishiyama M, et al. Nonconvulsive seizure detection by reduced-lead electroencephalography in children with altered mental status in the emergency department. J Pediatr 2019 Apr;207:213-219. [doi: 10.1016/j.jpeds.2018.11.019] [Medline: 30528574]
15. Simma L, Bauder F, Schmitt-Mechelke T. Feasibility and usefulness of rapid 2-channel-EEG-monitoring (point-of-care EEG) for acute CNS disorders in the paediatric emergency department: an observational study. Emerg Med J 2021 Dec;38(12):919-922. [doi: 10.1136/emermed-2020-209891] [Medline: 33127740]
16. Takase R, Sasaki R, Tsuji S, Uematsu S, Kubota M, Kobayashi T. Benzodiazepine use for pediatric patients with suspected nonconvulsive status epilepticus with or without simplified electroencephalogram: a retrospective cohort study. Pediatr Emerg Care 2022 Sep 1;38(9):e1545-e1551. [doi: 10.1097/PEC.0000000000002811] [Medline: 35947072]
17. Simma L, Kammerl A, Ramantani G. Point-of-care EEG in the pediatric emergency department: a systematic review. Eur J Pediatr 2025 Mar 7;184(3):231. [doi: 10.1007/s00431-025-06059-y] [Medline: 40053132]
18. Taran S, Ahmed W, Pinto R, et al. Educational initiatives for electroencephalography in the critical care setting: a systematic review and meta-analysis. Can J Anaesth 2021 Aug;68(8):1214-1230. [doi: 10.1007/s12630-021-01962-y] [Medline: 33709264]
19. Legriel S, Jacq G, Lalloz A, et al. Teaching important basic EEG patterns of bedside electroencephalography to critical care staffs: a prospective multicenter study. Neurocrit Care 2021 Feb;34(1):144-153. [doi: 10.1007/s12028-020-01010-5] [Medline: 32495314]
20. Chari G, Yadav K, Nishijima D, Omurtag A, Zehtabchi S. Improving the ability of ED physicians to identify subclinical/electrographic seizures on EEG after a brief training module. Int J Emerg Med 2019 Mar 27;12(1):11. [doi: 10.1186/s12245-019-0228-9] [Medline: 31179946]
21. Davey Z, Gupta PB, Li DR, Nayak RU, Govindarajan P. Rapid response EEG: current state and future directions. Curr Neurol Neurosci Rep 2022 Dec;22(12):839-846. [doi: 10.1007/s11910-022-01243-1] [Medline: 36434488]
22. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208. [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
23. Hirsch LJ, Fong MWK, Leitinger M, et al. American clinical neurophysiology society's standardized critical care EEG terminology: 2021 version. J Clin Neurophysiol 2021 Jan 1;38(1):1-29. [doi: 10.1097/WNP.0000000000000806] [Medline: 33475321]
24. Simma L, Romano F, Schmidt S, Ramantani G, Bölsterli BK. Integrating neuromonitoring in pediatric emergency medicine: exploring two options for point-of-care electroencephalogram (pocEEG) via patient monitors-a technical note. J Pers Med 2023 Sep 20;13(9):1411. [doi: 10.3390/jpm13091411] [Medline: 37763178]
25. Lee HC, Jung CW. Vital Recorder—a free research tool for automatic recording of high-resolution time-synchronised physiological data from multiple anaesthesia devices. Sci Rep 2018;8(1):1527. [doi: 10.1038/s41598-018-20062-4]
26. La Barge G. Pre-and post-testing with more impact. J Ext 2007;45(6):17 [FREE Full text]
27. Eysenbach G. Improving the quality of web surveys: the checklist for reporting results of internet e-surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
28. Tamási B, Hothorn T. tramME: mixed-effects transformation models using template model builder. R J 2021;13(2):398-418 [FREE Full text]

29. R Core Team. R: A Language and Environment for Statistical Computing: R Foundation for Statistical Computing; 2021.
30. O'Brien AJ, Brady RM. Point-of-care ultrasound in paediatric emergency medicine. J Paediatr Child Health 2016 Feb;52(2):174-180. [doi: 10.1111/jpc.13098] [Medline: 27062620]
31. Fernandez A, Asoodar M, van Kranen-Mastenbroek V, Majoie M, Balmer D. What do you see? signature pedagogy in continuous electroencephalography teaching. J Clin Neurophysiol 2025 Jan 1;42(1):81-86. [doi: 10.1097/WNP.0000000000001075] [Medline: 38376951]
32. Saloum D. Critical skills: use them or lose them. Ann Emerg Med 2013 May;61(5):599. [doi: 10.1016/j.annemergmed.2012.10.044] [Medline: 23622029]
33. Green SM, Ruben J. Emergency department children are not as sick as adults: implications for critical care skills retention in an exclusively pediatric emergency medicine practice. J Emerg Med 2009 Nov;37(4):359-368. [doi: 10.1016/j.jemermed.2007.05.048] [Medline: 18022780]

Abbreviations

AMS: altered mental status
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
ED: emergency department
EEG: electroencephalogram
ePolrME: mixed-effects enhanced proportional odds logistic regression
JMS: junior medical staff
NCSE: nonconvulsive status epilepticus
OR: odds ratio
PED: pediatric emergency department
PEM: pediatric emergency medicine
pocEEG: point-of-care electroencephalogram
PolrME: mixed-effects proportional odds logistic regression
REDCap: Research Electronic Data Capture
RN: registered nurse
SE: status epilepticus
SMS: senior medical staff

Edited by B Lesselroth; submitted 28.11.24; peer-reviewed by NH Mohamad Zainal; revised version received 17.04.25; accepted 20.06.25; published 20.08.25.

Please cite as:
Simma L, Schneeberger MH, von Felten S, Seiler M, Ramantani G, Bölsterli BK
E-Learning for Pediatric Emergency Department Staff in Point-of-Care Electroencephalogram Interpretation: Prospective Cohort Study
JMIR Med Educ 2025;11:e69395
URL: https://mededu.jmir.org/2025/1/e69395
doi:10.2196/69395

Evaluation and Uptake of an Online ADHD Psychoeducation Training for Primary Care Health Care Professionals: Implementation Study

Blandine French[1], PhD; Hannah Wright[1], MSc; David Daley[2], PhD; Elvira Perez Vallejos[1], PhD; Kapil Sayal[1], PhD; Charlotte L Hall[1], PhD

[1]Institute of Mental Health, Jubilee Campus, University of Nottingham Innovation Park, Triumph Road, Nottingham, United Kingdom
[2]Applied Psychological Practice, University of Nottingham Innovation Park, Nottingham, United Kingdom

Corresponding Author:
Blandine French, PhD
Institute of Mental Health, Jubilee Campus, University of Nottingham Innovation Park, Triumph Road, Nottingham, United Kingdom

Abstract

Background: Health care professionals seldom receive training on neurodevelopmental conditions such as attention-deficit/hyperactivity disorder (ADHD). An online training was co-developed to address some of the gaps in knowledge and understanding in primary care. A randomized controlled trial demonstrated that the training increased knowledge and confidence and improved practice.

Objective: This report highlights the implementation of the training in practice and follow-up 4 years post evaluation.

Methods: The online ADHD training comprises 2 modules: "Understanding ADHD" and "The Role of the GP," each taking approximately 45 minutes to complete. The training targets general practitioners primarily but is open to other health care professionals and parents. Feedback was collected through a survey at the end of the training, and the training has been widely adopted by various organizations internationally and nationally.

Results: Between December 2019 and January 2024, the "Understanding ADHD" module was accessed more than 13,486 times, while the "Role of the GP" module was accessed 7018 times, primarily by users from the United States and the United Kingdom. Survey results from both modules showed positive feedback with high ratings for usefulness, likelihood to inform practice, and recommendation to colleagues. Some suggestions for improvement included reducing the negative focus on ADHD consequences and incorporating more positive aspects of ADHD.

Conclusions: This ADHD online training program, despite facing implementation challenges, has seen positive outcomes, including international translation and high user ratings. Suggestions for improvement were received, but some were not feasible due to regional variations in ADHD pathways. The training's impact extended beyond GPs to other health care professionals, although the COVID-19 pandemic posed obstacles to dissemination efforts. Nonetheless, ongoing plans aim to expand the training's implementation globally.

(JMIR Med Educ 2025;11:e59365) doi:10.2196/59365

KEYWORDS

ADHD; online training; general practice; psychoeducation; training; implementation report; primary care; primary care health professionals; healthcare professionals; survey; feedback; training program; neurodevelopmental conditions

Introduction

Background

Attention-Deficit Hyperactivity Disorder (ADHD) is a neurodevelopmental condition characterized by symptoms of persistent inattention or hyperactivity-impulsivity, which causes clinical impairment in academic and social functioning [1] affecting approximately 5% of children [2] and 2.5% of adults [3]. Having ADHD carries many additional risks [4], and these are worsened when unsupported and undiagnosed [5], often leading to an increase in access to primary care services [6].

In the United Kingdom, access to care for children and adults with attention-deficit hyperactivity disorder (ADHD) is complex, starting from general practitioners (GPs) in primary care or schools who refer on to specialist services for diagnosis and treatment (psychiatry, Children and Adolescent Mental Health Services, etc). Health care professionals, including GPs, receive little or no training on ADHD. This significantly impacts access to care for many children and adults as GPs are the main

gatekeepers for specialist services [7,8]. To fill this gap in ADHD training, an online training was coproduced with GPs to improve GPs' knowledge about ADHD [9]. The stepwise, coproduction approach toward developing this online ADHD training for GPs began with a preparatory workshop in order to highlight the relevant topics to be included in the intervention, from which educational videos were then developed, as well as the content and format for the training. Two workshops were then conducted with GPs, leading to further refinement of the video content and subsequently the final intervention. A pilot usability study (N=10 GPs) was conducted to assess the intervention's acceptability, feasibility, and accessibility. The online training included interactive psychoeducation elements reinforced with activities and videos lasting a total of 45 minutes. The content included enough information for GPs to identify ADHD and better understand the condition. The resulting intervention was then evaluated through a randomized controlled trial (RCT) [10] in GP practices based in England where 221 GPs took part. The evaluation of this training demonstrated that GPs' knowledge and confidence significantly improved, misconceptions decreased, and attitudes and reported practice changed [10]. The unique aspect of this training lies in the strong coproduction element, with GPs being involved throughout the development and review process and within the evaluation through an RCT, which is rarely done for education packages. The coproduction element, reviews, and evaluation aspect of the original project took over 2 years from January 2018 to March 2020. To our knowledge, no other online training package has been developed and evaluated for ADHD in primary care in the United Kingdom.

The original evaluation of this training terminated in March 2020. Since then, we have spent time implementing the training in GP practices, in alignment with the British National Institute for Healthcare Research (NIHR) priority settings in digital support for primary care [11].

This publication describes the impact, engagement, and implementation of the original ADHD training program, 4 years beyond completion of data collection.

Aims and Objectives

This project aimed to implement and evaluate the ADHD online training into practice, beyond the scope of the original project. The implementation and impact of the training are measured through website access analytics and responses to survey questions.

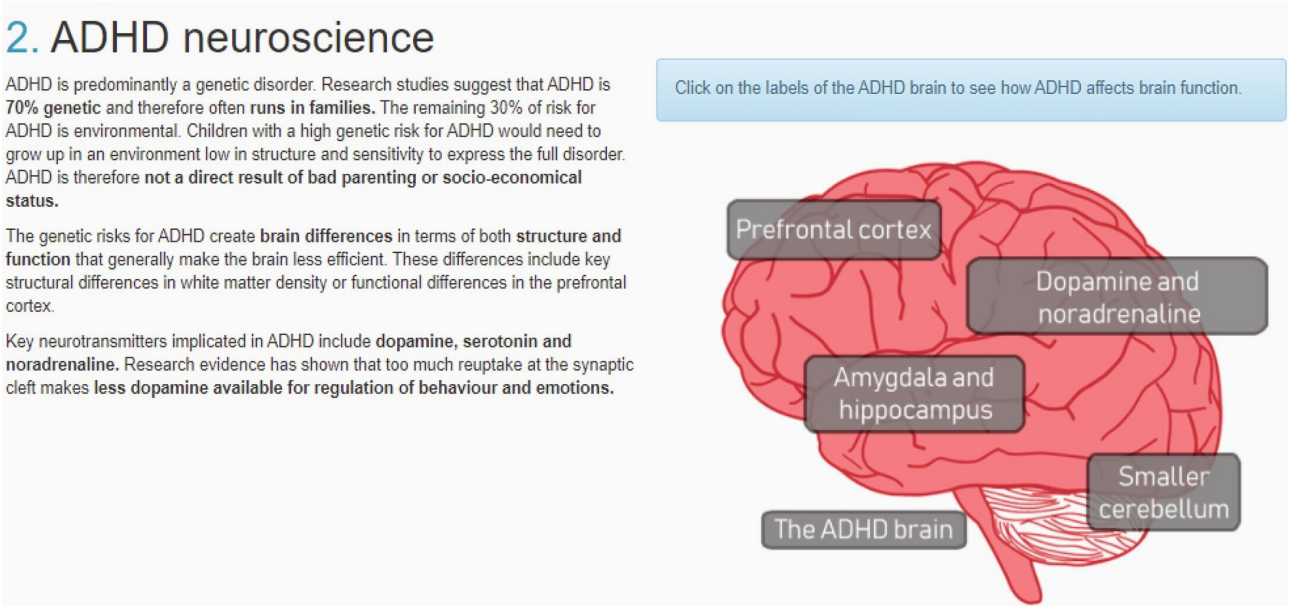
Methods

Blueprint Summary and Technical Design

The online ADHD training is a psychoeducation program consisting of 2 modules, one on "Understanding ADHD" and one on "the role of the GP in the care pathway." The training takes approximately 45 minutes to complete and can be accessed freely online (link in [12]; example of an education module page in Figure 1).

The training is hosted on a secure platform within the University of Nottingham Health E-learning and Media team (HELM). It has been developed to be easily accessible on a computer or a mobile device. Further details on the original project can be found online [13].

Figure 1. Example of educational module page from the training. ADHD: attention-deficit/hyperactivity disorder.



Target

The training is principally aimed at general practitioners or other medical doctors. However, while the module "Role of the GP" in the care pathway focuses on primary care, the module "Understanding ADHD" can be useful for anyone including other health care professionals, education professionals, and parents. Additionally, although the training has elements of the care pathway that are specific to the United Kingdom, a large proportion of the training is relevant to other countries and settings as well (psychoeducation on ADHD, effect of medication, types of interventions, etc) and therefore was found to be useful in different contexts. The original RCT had specific inclusion criteria, but for this aspect of implementation, no

restrictions were implemented on either module. The online training is freely available and hosted on a university server that is widely secure and accessible, including from health care servers (eg, SystmOne). This was essential as a lot of health care servers can block external links.

Survey

The survey was developed by the lead researchers. Following a pragmatic approach prioritizing quick time response, only 4 questions were presented to participants in the training. As the training is freely available and no longer part of a study, the questions were optional and presented upon completion of each module. Not all participants took part in both modules, and we wanted to capture the values of each. While the survey is aimed primarily at GPs, anyone could take part in the module, and we anticipated that many other health care professionals would take part in this training. Therefore, the questions were targeted to capture clinical impact on practice with health care, in line with the initial aim of the training. In addition to 3 demographic questions (age, occupation, and gender), 4 evaluation questions were presented:

1. How useful did you find the information in this program?
2. How likely is this information going to inform your practice?
3. Would you recommend this training to your colleagues?
4. Any other comments on the intervention?

Participants were asked to select a score on a scale of 1 - 10 (1: not at all and 10: extremely) to represent how much they agreed with the evaluation questions.

Data

The data and analytics generated from the survey are stored online within the HELM platform and only accessible by the HELM team. The feedback questionnaire was anonymous and voluntary for anyone taking part in the training.

Participating Entities and Dissemination

Many partner organizations have adopted, distributed, and implemented the training over the last 4 years. These include the Royal College of General Practice (RCGP), ADHD Europe, the Association for Child and Adolescent Mental Health (ACAMH), the Academic Health Science Network (AHSN), local GP training hubs, European ADHD research networks (EUNETHYDIS), the University of Montpellier, The University of Dublin and the ADHD collective.

Presentations about the training have been delivered by the lead researcher to groups locally, nationally (eg RCGP) and internationally (eg ADHD Europe). The training has been accredited by the RCGP, the leading professional organization for GP training and accreditation in the United Kingdom, as part of its Continuing Professional Development (CPD) program. Accreditation included further peer review from GPs and refinement of text in line with national and international guidelines (National Institute for Health and Care Excellence and *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*). Internationally, the training has been translated into 3 languages (German, French, and Spanish), and ongoing collaboration with leading European ADHD research networks (EUNETHYDIS) has started to develop evaluation and implementation of the translated versions.

The training development and initial RCT evaluation were funded by the Economic and Social Research Council (ESRC) through a doctorate training program. The training also received a non-profit grant from Takeda (a pharmaceutical company) to support the online development and trial.

Sustainability and Budget

The original project funds from Takeda allowed for the intervention to be developed and hosted on the free accessible HELM platform. A booster grant from the same funders also allowed for the translations into other languages to be completed, but aside from these, no other budget was available for the long-term implementation of the training. As a university employee, the lead researcher has driven the implementation in her own time by giving workshops, training, and presentations to specific groups over the last few years.

Ethical Considerations

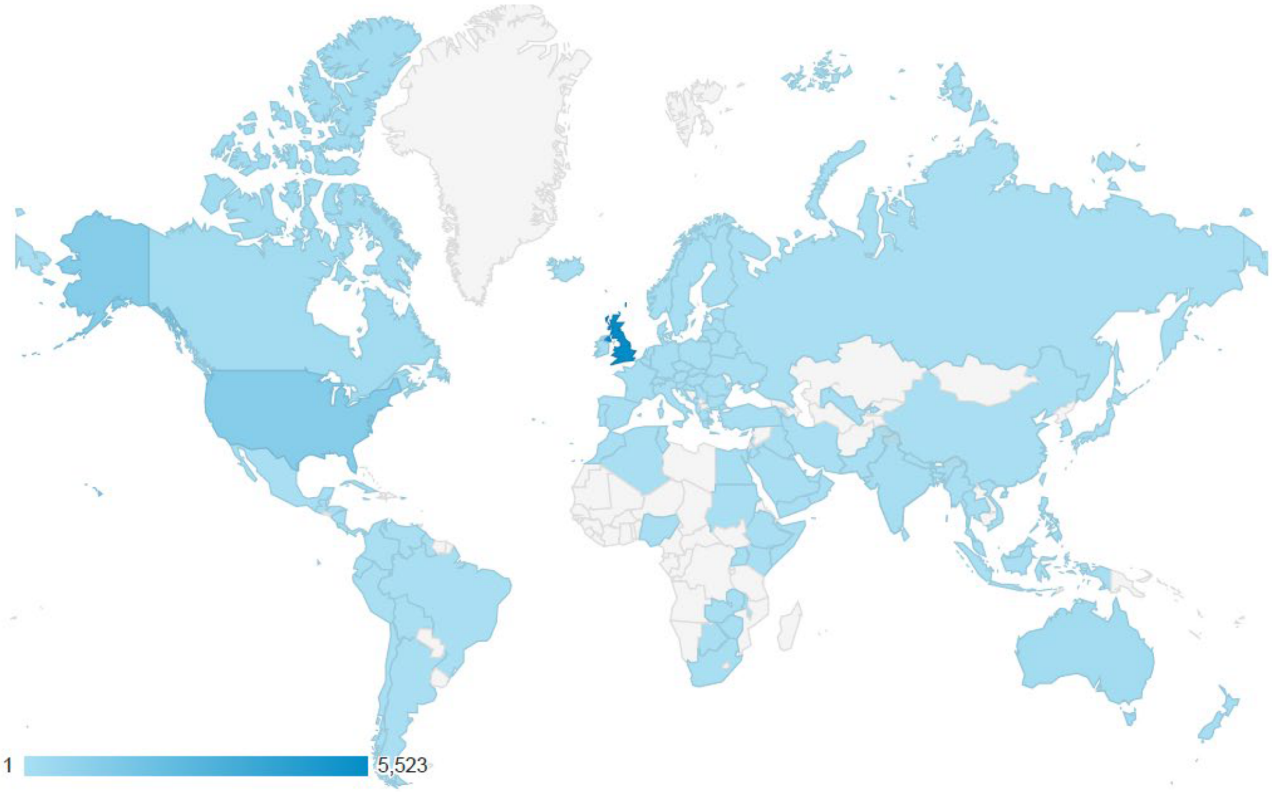
Ethics approval for the RCT and the ongoing evaluation was received from the University of Nottingham, Faculty of Medicine Research Ethics Committee (reference 19/HRA/1028) and from the Nottinghamshire Healthcare National Health Service (NHS) Foundation Trust Research and Development department (project ID 257567). Participants in the survey consented to their data being used. In accordance with ethical standards with IRB and with the Helsinki Declaration of 1975, as revised in 2000. Participants were not compensated for their time in providing feedback. All responses were anonymous and unidentifiable.

Results

Coverage

Between December 2019 and January 2024, the "Understanding ADHD" module was accessed more than 13,486 times with 1484 (11%) of users returning visitors. Many partner organizations have adopted and disseminated the training. Most users were from the United States and the United Kingdom (33% and 57%, respectively). The "role of the GP in the care pathway" module was accessed 7,018 times with 716 (10.2%) of returning visitors. Most users were from the United States and the United Kingdom (20% and 76%, respectively), but the training was also accessed by users based in another 120 countries (Figure 2).

Figure 2. Countries that have accessed the attention-deficit/hyperactivity disorder (ADHD) training from 2019 to 2024.



Outcomes: Survey Results

"Understanding ADHD" Module

A total of 648 participants responded (134 males, 496 females, and 18 unknown), with a mean age of 40.4 (SD 11.32) years (range 16 - 81 years). On average, participants rated the information in the resource 8.47 (SD 1.93) for usefulness and 8.26 (SD 2.11) for how likely this information was going to inform their practice. A total of 611 of 631 (96.8%) participants would recommend the training resource to their colleagues. Table 1 represents the demographics and responses given, separated into different occupations.

Table . Demographic data and average evaluation questionnaire responses for the "Understanding ADHD[a]" module, separated by occupation category.

Variables	Occupation				
	GP[b] (n=66)	Other health care professionals (n=285)	Mental health professionals (n=91)	Health care students (n=47)	Other occupations[c] (n=159)
Gender					
Men	22	74	13	4	21
Women	43	209	76	41	127
Unknown gender	1	2	2	2	11
Age (years), mean (SD)[d]	44.97 (11.74)	40.89 (11.30)	40.01 (11.56)	31.85 (9.20)	40.57 (12.83)
Usefulness rating[e], mean (SD)	8.36 (1.89)	8.61 (1.83)	8.69 (1.62)	8.39 (1.92)	8.13 (2.20)
Likelihood that information would inform practice[e], mean (SD)	8.30 (1.83)	8.49 (1.95)	8.64 (1.66)	8.40 (2.05)	7.56 (2.55)
% Respondents that would recommend the training resource to their colleagues	96.97	96.74	100.00	97.87	94.74

[a]ADHD: attention-deficit/hyperactivity disorder.

[b]GP: general practitioner.

[c]Other occupations include non–health care students, homemakers, project managers, teachers, and teaching assistants.

[d]Data taken only from those who responded. Overall, 13 did not respond to age.

[e]On a scale of 1 - 10, 1 being not at all, and 10 being extremely.

Positive Feedback

The free-text response box yielded 117 responses about the resource, 104 (89%) of which were positive. Participants felt that the resource was comprehensive and clear and appreciated the range of mediums used, including sound bites, videos, text, and interactive quizzes. Multiple comments acknowledged the benefit of including personal lived experiences to help solidify the information included.

Many commented that the resource will be valuable to them in their practice as a GP or healthcare professional, eg, "I am better informed and ready to be more useful to my patients," and others said it was helpful to them on a more personal level (eg, understanding their children or colleagues with ADHD).

> *This intervention is amazing and the value availed of by the public, ADHDers, teachers and medical professionals cannot be underestimated in fact I would say it is immeasurable! [Homemaker]*
>
> *Fantastic resource. Thank you so much. The simple and accessible explanations are brilliant. [Researcher]*
>
> *Easy to read, absorb and navigate. Highly informative educational resource, one of the best I have come across as an individual who suffers from ADHD, thank you! [Carer]*

Suggestions for Improvement

The most common critical feedback was that the module focused too heavily on the negative consequences and risks associated with ADHD. Participants expressed that there are positives associated with ADHD that could be mentioned and that people with ADHD can still be successful, employed, and intelligent. Some highlighted concern that the slides on the problems with ADHD "played into the stigma that often stops people from being accepted for an assessment in the NHS."

> *It would also be good to emphasise the potential strengths in individuals with ADHD for example creativity and the benefits of hyper-focus. [Clinical psychologist]*

Another issue was the lack of representation in the videos for non-White individuals, and therefore more diversity in ethnicity was requested. A few comments suggested the need for an explicit reference to the differential presentation and diagnosis rates between males and females.

"Role of the GP" Module

A total of 308 participants responded to this module feedback questionnaire (64 males, 241 females, and 3 unknown), with a mean age of 39.9 (SD 11.15) years (range 18 - 67 years). On average (mean), participant rating was 8.12 (SD 2.16) for usefulness, and 7.98 (SD 2.14) on how likely this information is going to inform your practice, demonstrating that overall participants responded positively, finding the program to be

useful for their knowledge and improving their practice. Additionally, 291 of 299 respondents (97.3%) would recommend the training resource to their colleagues.

Table 2 shows the age, gender, and answers, separated by the participant's professional group. The ratings appear to be quite similar between each occupation group. The GP group gave the lowest rating for usefulness, likelihood to inform practice, and likelihood to recommend to colleagues, whereas mental health professionals gave the highest.

Table . Demographic data and average evaluation questionnaire responses for the "Role of the GP" module, separated by occupation category.

Variables	Occupation				
	GP[a] (n=66)	Other health care professionals (n=128)	Mental health professionals (n=30)	Health care students (n=30)	Other occupations[b] (n=54)
Gender					
Men	19	30	6	4	5
Women	46	98	24	26	47
Unknown gender	1	0	0	0	2
Age (years), mean (SD)[c]	45.40 (9.59)	39.00 (10.27)	39.40 (11.15)	31.50 (12.35)	39.67 (12.42)
Usefulness rating[d], mean (SD)	7.92 (2.16)	8.02 (2.02)	8.67 (1.60)	8.40 (2.04)	8.15 (2.30)
Likelihood that information would inform practice[d], mean (SD)	7.74 (2.18)	7.93 (2.09)	8.53 (1.67)	8.50 (2.01)	7.78 (2.39)
% Respondents that would recommend the training resource to their colleagues	93.80	98.40	100.00	96.60	98.08

[a]GP: general practitioner.

[b]Other occupations include non–health care students, a local government manager, a homemaker, and a teaching assistant.

[c]Data taken only from those who responded. Overall, 4 did not respond to age.

[d]On a scale of 1 - 10, 1 being not at all, and 10 being extremely.

Positive Feedback

The free-text response box yielded 75 responses about this resource, 61 of which were positive (81%). Many comments said that the resource was "very interesting," "informative," and "useful." There was also praise regarding the clarity and presentation of the information, which facilitated easy understanding. Some respondents also commented that it was particularly insightful to include a video from the point of view of a GP discussing their own difficulties in getting a diagnosis, adding a unique and important perspective. These positive comments came from participants with a range of backgrounds demonstrating the benefits of this resource for GPs as well as other professions.

> *Clear and well presented. So many people complicate ADHD and so it becomes difficult to learn - this was a perfect way to introduce it - and the GP with ADHD was a fantastic resource. [GP]*

> *Very detailed, lots of information and suitable diagrams and videos of experiences also helped. [Student Mental Health Nurse]*

> *This is a really good teaching resource. I will be recommending that all new nursing staff also have access to this. [Clinical Specialist ADHD Nurse]*

Suggestions for Improvement

A few of the free-text responses gave some suggestions for other things to include in this module. A couple of participants asked for more information about the assessment process and specific information about the role of the GP in shared care, with the roles of nurses and support workers. There was a suggestion from a few GPs to include free resource signposting for them to give to patients and parents to help with management.

> *It could do with more detail and clinical scenarios. Also important is the association and differential diagnosis of other comorbidities, addiction/personality disorders/anxiety depression. That may indeed be beyond the scope of this website. [GP]*

As these participants acknowledged themselves, this module focused on the role of the GP in ADHD, so did not aim to include comorbid diagnoses.

Another participant gave a specific suggestion for the "What is ADHD" section to include clearer differentiation about the presentation in girls and boys or women and men and more detail on how ADHD impacts mental health when it is not diagnosed or well-treated.

A care assistant expressed their frustration about the care system in their part of the United Kingdom in referring to the Association for Child and Adolescent Mental Health or adult mental health services for ADHD, commenting that "Your information makes going to the doctors and getting a diagnosis seem easy."

Finally, a couple of other respondents stated that this resource did not add anything new to their knowledge and would not change their practice as a GP, although this may not be reflective of the quality of the resource, but of this participant's existing knowledge being sufficient or extensive.

Discussion

Principal Findings

Overall, the implementation of this funding has been very positive, particularly in the light of no secured economic or staff resources to support this. The program has been translated internationally, reaching more than 120 countries and has been translated into 3 different languages. The feedback has been predominantly positive, and the intervention has received consistently high ratings regarding its acceptability and usefulness.

Lessons Learned

Many lessons were learned from the feedback as well as from the dissemination and implementation process.

The feedback highlighted some limitations about how the training had been framed, the lack of representation, and gender differences. The coproduction aspect of this training aimed to maximize the accessibility and usability of the training by GPs. Service users also reviewed the training before its dissemination, but their input was not as significant as that from GPs. Some of these issues might have been avoided by more thorough service-user involvement. Striking the balance between the different stakeholders was complex as, in this instance, their input at times was not compatible.

The dissemination of the training in practice within the United Kingdom was also very complex. In the United Kingdom, there is no single training organization that can adopt training and trickle down to all GPs in a top-down manner. Each area of the United Kingdom has separate training programs that are decided at a regional level, and linking with all different regions is difficult. Therefore, the implementation had to be conducted in a bottom-up manner, which is time-consuming and has limited reach. We were also unable to reliably report data pertaining to user engagement, such as completion rate and number of visits. These numbers were biased by the ongoing use of the tools as a training module, and therefore the numbers do not represent realistic interactions.

It is also important to note that the dissemination of the training started in March 2020. In many countries including the United Kingdom, this was the beginning of the COVID-19 pandemic. This would have had significant impacts on the implementation of the training. First, GPs' priorities had to change very quickly, and practices adapted in a very stressful environment. Dealing with the direct consequences of the pandemic became the priority, with taking part in CPD or a better understanding of ADHD becoming less important. Although the demand for confinement significantly increased ADHD symptoms expressions and associated impairment for some [14] and an increase in referrals postpandemic [15], it was difficult for GPs to prioritize this issue, as is demonstrated by the significant decrease in referrals during lockdowns [16]. Second, the research team was not able to actively reach out, disseminate, and give presentations for a long time, which also impacted the implementation of the training in practice.

Finally, the implementation process has not been without challenges. The UK system for training health care professionals is not easy to navigate. Outreach events have been useful in engaging stakeholders and generating uptake in the training; however, this led to only pockets of health care professionals being trained and did not extend wider than the reach for the event. The aim of the evaluation was to be pragmatic so that we could evaluate the implementation in routine primary care settings; for this reason, the evaluation questions were simple, and free text comments were only provided by a minority of users.

Some suggestions for improvement have been raised, which will be considered in alterations to the training, but it is important to note that these were from a minority of users (30/9000). Additionally, some of the suggestions were not possible to implement, for example, in local pathways and national practices. The local pathways for ADHD vary widely between regions and countries. They also often change regularly, and information quickly becomes outdated [17-19]. Therefore, while the information on pathways would be very beneficial, it is impossible to capture all regional and national variations.

While the training was primarily aimed at GPs, it was always clear that it would benefit many other health care professions working alongside primary care. By gaining a better understanding of the GP's role, the roles of other professionals will become clearer. Therefore, it is very positive that so many other health care professionals accessed the training and gained a better understanding of ADHD, significantly widening the impact and reach of the resource.

In conclusion, this coproduced and evidence-based training shows ongoing benefits, acceptability, and usefulness in practice. The results from this implementation demonstrated a wider use to other health care professionals and international reach. Ongoing implementation plans aim to support further the wider implementation of this training, principally in other countries.

Conflicts of Interest

DD has in the provided educational talks for Medice and Takeda . He has also reported grants, personal fees and nonfinancial support from Takeda, Medice, ACAMH Learn, the New Forest Parenting Programme and World ADHD congress, book royalties from the sale of a self-help version of the New Forest Parenting Programme, and compensation for the provision of training and supervision in the New Forest Parenting Programme. BF reports personal fees and nonfinancial support from Takeda and Medice. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Diagnostic and Statistical Manual of Mental Disorders, 5th edition: American Psychiatric Publishing, Inc.; 2013. [doi: 10.1176/appi.books.9780890425596]
2. Sayal K, Prasad V, Daley D, Ford T, Coghill D. ADHD in children and young people: prevalence, care pathways, and service provision. Lancet Psychiatry 2018 Feb;5(2):175-186. [doi: 10.1016/S2215-0366(17)30167-0] [Medline: 29033005]
3. Song P, Zha M, Yang Q, et al. The prevalence of adult attention-deficit hyperactivity disorder: a global systematic review and meta-analysis. J Glob Health 2021 Feb 11;11:04009. [doi: 10.7189/jogh.11.04009] [Medline: 33692893]
4. French B, Nalbant G, Wright H, et al. The impacts associated with having ADHD: an umbrella review. Front Psychiatry 2024;15:1343314. [doi: 10.3389/fpsyt.2024.1343314] [Medline: 38840946]
5. French B, Daley D, Groom M, Cassidy S. Risks associated with undiagnosed ADHD and/or autism: a mixed-method systematic review. J Atten Disord 2023 Oct;27(12):1393-1410. [doi: 10.1177/10870547231176862] [Medline: 37341291]
6. Prasad V, Rezel-Potts E, White P, et al. Use of healthcare services before diagnosis of attention-deficit/hyperactivity disorder: a population-based matched case-control study. Arch Dis Child 2023 Dec 14;109(1):46-51. [doi: 10.1136/archdischild-2023-325637] [Medline: 37903632]
7. French B, Sayal K, Daley D. Barriers and facilitators to understanding of ADHD in primary care: a mixed-method systematic review. Eur Child Adolesc Psychiatry 2019 Aug;28(8):1037-1064. [doi: 10.1007/s00787-018-1256-3] [Medline: 30552584]
8. French B, Perez Vallejos E, Sayal K, Daley D. Awareness of ADHD in primary care: stakeholder perspectives. BMC Fam Pract 2020 Feb 28;21(1):45 [FREE Full text] [doi: 10.1186/s12875-020-01112-1] [Medline: 32111169]
9. French B, Daley D, Perez Vallejos E, Sayal K, Hall CL. Development and evaluation of an online education tool on attention deficit hyperactivity disorder for general practitioners: the important contribution of co-production. BMC Fam Pract 2020 Nov 1;21(1):224. [doi: 10.1186/s12875-020-01289-5] [Medline: 33131490]
10. French B, Hall C, Perez Vallejos E, Sayal K, Daley D. Assessing the efficacy of online ADHD awareness training in primary care: pilot randomised control trial evaluation with nested qualitative interviews. JMIR Med Educ 2020 Jul;6(2):e19871 [FREE Full text]
11. Leach B, Parkinson S, Gkousis E, et al. Digital facilitation to support patient access to web-based primary care services: scoping literature review. J Med Internet Res 2022 Jul 14;24(7):e33911. [doi: 10.2196/33911] [Medline: 35834301]
12. ADHD GP training. ADHD Project. URL: https://www.nottingham.ac.uk/helmopen/rlos/practice-learning/mental-health/adhd/ [accessed 2025-06-30]
13. GP ADHD training. NDLAB. URL: https://ndlab.org.uk/research/gp-adhd-training/ [accessed 2025-06-22]
14. Rogers MA, MacLean J. ADHD symptoms increased during the Covid-19 pandemic: a meta-analysis. J Atten Disord 2023 Jun;27(8):800-811. [doi: 10.1177/10870547231158750] [Medline: 36879524]
15. Smith MCF, Mukherjee RAS, Müller-Sedgwick U, Hank D, Carpenter P, Adamou M. UK adult ADHD services in crisis. BJPsych Bull 2024 Feb;48(1):1-5. [doi: 10.1192/bjb.2023.88] [Medline: 38058161]
16. FitzPatrick P, Antczak K, Lynch F, Lynch S, McNicholas F. General practitioner referrals to child and adolescent mental health services: did they differ during Covid-19? Ir J Psychol Med 2023 Sep;40(3):457-459. [doi: 10.1017/ipm.2023.13] [Medline: 36855804]
17. Khandaker GM, Gandamaneni PK, Dibben CRM, Cherukuru S, Cairns P, Ray MK. Evaluating care pathways for community psychiatry in England: a qualitative study. J Eval Clin Pract 2013 Apr;19(2):298-303. [doi: 10.1111/j.1365-2753.2012.01822.x] [Medline: 22360292]
18. Crocker T, Johnson O, King S. The suitability of care pathways for integrating processes and information systems in healthcare. Transforming Government 2009 Jul 31;3(3):289-301. [doi: 10.1108/17506160910979379]
19. Jones A, Kamath PD. Issues for the development of care pathways in mental health services. J Nurs Manag 1998 Mar;6(2):87-95. [doi: 10.1046/j.1365-2834.1998.00042.x] [Medline: 9582782]

Abbreviations

ADHD: attention-deficit/hyperactivity disorder
CPD: Continuing Professional Development
GP: general practitioners
RCT: randomized controlled trial

Edited by D Chartash; submitted 10.04.24; peer-reviewed by K McClatchey, MB Ellur, TN Willig; revised version received 03.04.25; accepted 03.04.25; published 11.07.25.

Please cite as:
French B, Wright H, Daley D, Perez Vallejos E, Sayal K, Hall CL
Evaluation and Uptake of an Online ADHD Psychoeducation Training for Primary Care Health Care Professionals: Implementation Study
JMIR Med Educ 2025;11:e59365
URL: https://mededu.jmir.org/2025/1/e59365
doi:10.2196/59365

Making Medical Education Courses Visible: Theory-Based Development of a National Database

Andi Gashi[1,2], MMed; Monika Brodmann Maeder[2,3], PD, Dr med, MME; Eva K. Hennel[2], PhD, Dr med, MME

[1]Medizinische Fakultät, University of Bern, Bern, Switzerland

[2]Forschung und Entwicklung, Schweizerisches Institut für ärztliche Weiter- und Fortbildung SIWF/ISFM, Elfenstrasse 18, Bern, Switzerland

[3]Universitätsklinik für Notfallmedizin, Inselspital Bern, Universitätsspital Bern, University of Bern, Bern, Switzerland

Corresponding Author:
Andi Gashi, MMed
Medizinische Fakultät, University of Bern, Bern, Switzerland

Abstract

Background: Medical education has undergone professionalization during the last decades, and internationally, educators are trained in specific medical education courses also known as "train the trainer" courses. As these courses have developed organically based on local needs, the lack of a general structure and terminology can confuse and hinder educators' information and development. The first aim of this study was to conduct a national search, analyze the findings, and provide a presentation of medical education courses based on international theoretical frameworks to support Swiss course providers and educators searching for courses. The second aim was to provide a blueprint for such a procedure to be used by the international audience.

Objective: In this study, we devised a scholarly approach to sorting and presenting medical education courses to make their content accessible to medical educators. This approach is presented in detailed steps and our openly available exemplary database to make it serve as a blueprint for other settings.

Methods: Following our constructivist paradigm, we examined content from medical education courses using a theory-informed inductive data approach. Switzerland served as an example, covering 4 languages and different approaches to medical education. Data were gathered through an online search and a nationwide survey with course providers. The acquired data and a concurrently developed keyword system to standardize course terminology are presented using Obsidian, a software that shows data networks.

Results: Our iterative search included several strategies (web search, survey, provider enquiry, and snowballing) and yielded 69 courses in 4 languages, with varying terminology, target audiences, and providers. The database of courses is interactive and openly accessible. An open-access template database structure is also available.

Conclusions: This study proposes a novel method for sorting and visualizing medical education courses and the competencies they cover to provide an easy-to-use database, helping medical educators' practical and scholarly development. Notably, our analysis identified a specific emphasis on undergraduate teaching settings, potentially indicating a gap in postgraduate educational offerings. This aspect could be pivotal for future curriculum development and resource allocation. Our method might guide other countries and health care professions, offering a straightforward means of cataloging and making information about medical education courses widely available and promotable.

(JMIR Med Educ 2025;11:e62838) doi:10.2196/62838

KEYWORDS

curriculum mapping; faculty development; competencies; database; medical education

Introduction

Ensuring high-quality health care necessitates the presence of well-trained medical educators [1]. Internationally, this has led to the development of frameworks that define their role and the creation of educational pathways for certification in many predominantly Anglophone countries. The terms "medical education" and "medical educator" are used with varying definitions in various settings. For this study, we define "medical educators" as the diverse group of health care professionals who teach, but regarding their learners, we focused on undergraduate medical students and physicians only. Hence, we include medical educators who are active in multiple settings: from clinical supervision and classroom instruction to the development and implementation of curricula. Their roles span the continuum of medical education from undergraduate through postgraduate training and continuing professional development.

Recent studies have identified significant tensions faced by medical educators, including lack of defined career structures, insufficient recognition of teaching roles, and challenges in developing educational identities [2,3]. While formal training programs alone cannot fully address these structural issues, they

serve as a crucial stepping stone in professionalizing medical education.

Globally, this coincides with the effort to provide these educators with a defined career path and support for their teaching activities. Notably, the Royal College of Physicians and Surgeons in Canada has introduced a "Clinician Educator Diploma," rooted in the findings of studies by Sherbino et al [4,5]. Similarly, the University of Michigan Medical School in the United States offers a Master of Health Professions Education [6-8], and in the United Kingdom, a variety of programs align with the "Professional Standards for medical, dental, and veterinary educators" established by the Academy of Medical Educators [9]. These initiatives exemplify the efforts to emphasize the value of quality medical education and facilitate access to quality training for educators.

Drawing on Wenger's Communities of Practice theory [10], these structured pathways and frameworks serve not only as career development tools but also as boundary objects that connect different medical education communities and facilitate knowledge sharing. When educators can easily identify and access training opportunities, they can more effectively participate in and contribute to their professional community. The visibility and accessibility of educational resources play a crucial role in fostering community development and knowledge exchange across different regions and language groups, ultimately supporting the growth of medical education as a professional field.

Lack of Career Pathways and an Opaque Landscape

Nevertheless, many regions internationally do not provide a career pathway for medical educators, nor do they use specific frameworks yet [3,11,12]. While some countries offer incentives such as continuous medical education (CME) credits for educational training, this is not standardized. Similarly in Switzerland, formal requirements for teaching qualifications are not yet standardized; there are some existing incentives for medical educators to pursue didactic training. Many didactic courses, particularly those offered or recognized by the Swiss Institute for Postgraduate Medical Education (SIWF), provide CME credits, independent from whether they concern undergraduate, postgraduate, or CME. Additionally, some specialty training programs, such as Internal Medicine, have incorporated mandatory didactic courses into their curriculum requirements [13]. However, these incentives remain fragmented and vary across specialties and institutions.

Still, in everyday reality, senior staff members are expected to educate while juggling everyday clinical tasks with providing education for students and junior doctors, often lacking dedicated time for teaching activities, as described in literature from Australia, Canada, and the United Kingdom [14-17] and in a 2008 Association for Medical Education in Europe guide [18]. Similarly, in Switzerland, where this study took place, there are few examples of clinics hiring specific teaching staff or at least dedicating specific hours in employment contracts to teaching. This practice is still not the norm and is rarely seen outside of specific pilot programs [19]. All of this leads to a diverse, somewhat undefined, and opaque medical education landscape and hinders high-quality medical education.

Aims and Research Questions

The aims of this study are twofold. The first aim of this study was to conduct a national search, analyze the findings, and provide a presentation of medical education courses based on international theoretical frameworks to support mainly Swiss course providers and educators searching for courses.

The second aim was to provide a blueprint for such a procedure to be used by the international audience.

Our research questions are (1) How do we conduct the search for courses and their content? and (2) How can we present the courses in an accessible way that is translatable to other regions?

Methods

Research Paradigm and Use of Theory

Guided by a constructivist research paradigm, which implies that no objective reality exists but that knowledge is created by social interactions [20], this study adopted a subjectivist inductive approach, using a theory-informed inductive data analysis method, described by Varpio et al [21].

This approach allowed a dynamic interplay between intermediate results and potential theoretical frameworks during data collection until we found the most fitting theoretical lens. This reflective and adaptive approach also ensured the relevance of the study results to the roles and competencies of medical educators as they were reflected in the Swiss data.

After a literature review of possible frameworks [4,5,7,9,22-29] and thorough deliberation by the research team, we selected the framework by Sidhu et al [28] as the best fit to support data analysis. Our decision to adopt this framework was strategic, not only for its clear enough lens through which to categorize and analyze course content, due to its comprehensive integration of 67 texts on educator competency domains, but also because it encompassed multiple health care professions, not just physicians. It comprehensively synthesizes educator competencies and identifies 6 distinct domains: "Teaching and Facilitating Learning," "Designing and Planning Learning," "Assessment of Learning," "Educational Research and Scholarship," "Educational Leadership and Management," and "Educational Environment, Quality, and Safety." This approach provides a robust, inclusive structure for understanding and evaluating educator competencies across different health professions.

This study lays the groundwork for a national strategy in Switzerland to enhance the quality of medical education. The first step was establishing a comprehensive database of existing educational offerings. This database facilitates an iterative and reciprocal examination of the course landscape, identifying gaps in the current educational landscape and recognizing requirements that may diverge from international frameworks. This approach ensures that the subsequent framework development and the certification of educators and courses are tailored to the unique needs and priorities of the Swiss medical education system. Additionally, it also allows us to identify and highlight areas lacking coverage by comparing with existing provisions.

Selection and Eligibility Criteria for Course Inclusion and Exclusion

Our study employed a systematic approach to searching for courses, with carefully defined inclusion and exclusion criteria.

The inclusion criteria encompassed the geographical scope, that is courses offered in Switzerland or online courses offered by Swiss institutions; the target group that is courses aimed at educators teaching physicians or medical students; and the content focus, that is courses with a primary focus on teaching skills.

The study's exclusion criterion was individual (1-to-1) offerings.

The limitation of the geographical scope ensured feasibility and was chosen with regard to the first study aim of building a national database. Our target group of educators who teach medical students and physicians, rather than including educators for learners in all health care professions, was driven by several methodological and practical considerations. First, our research team's expertise lies specifically in physician education, allowing us to conduct more nuanced and informed analysis within this domain. Second, our primary data collection method through the Joint Commission of Swiss Medical Schools (SMIFK/CIMS) naturally oriented our study toward medical education targeting physician settings.

With the formulation of the content focus in teaching skills, we deliberately excluded courses with generalized skills that could apply to any professional setting. These "soft skills" courses—such as generic presentation techniques, leadership pitching, or broadly applicable communication strategies—were set aside to concentrate on educational content specifically tailored to medical teaching contexts.

These criteria helped to refine the search and survey strategies, allowing for a targeted collection of data that is pertinent to the goals of this study.

Data Collection

First Phase: Online Search

We initiated our investigation with an extensive online search (May to July 2023), employing various search strategies to explore clinic, faculty, university, and specialists' association websites to find information on courses. We approached the search from an estimated end-user perspective, simulating how a medical professional seeking to expand their teaching skills might investigate educational opportunities. This meant using primarily search engines and direct institutional websites (including every medical faculty n=11, university hospital n=9, cantonal and larger regional hospital websites n=18, medical associations n=45, and the official SIWF website). Our search extended across Switzerland's linguistic diversity, using keywords in the country's 3 official languages (German, French, and Italian) and English to ensure no regional offerings were overlooked. Exemplary search strings, detailed in Multimedia Appendix 1, were designed to capture the multifaceted nature of medical education across different linguistic and regional contexts. Additionally, we used a pragmatic snowball-like method where we let discovered sources lead to additional course offerings. While not claiming absolute comprehensiveness, this method aimed to provide an overview of medical education courses in the Swiss context.

Second Phase: Widening the Search

Subsequently (July to December 2023), we sent out an e-mail targeting representatives within the Joint Commission of the Swiss Medical Schools (SMIFK/CIMS), a commission that includes deans and other stakeholders from all Swiss medical faculties. (see Multimedia Appendix 1 for survey questions). The survey did not systematically track response rates or potential selection biases. To widen our scope and better represent non-university medical institutions, we distributed a modified version of the survey to Human Resources departments of 25 larger hospitals in Switzerland in December 2023.

Through this dual approach of detailed online searches and surveys, we sought to capture a comprehensive snapshot of medical education offerings in Switzerland. It allowed us to gather insights not only from academic institutions but also from larger health care providers, thus offering a comprehensive view of the medical education landscape in Switzerland.

Course data were collected using a spreadsheet attached to the surveys, asking the respondents to fill in information about their courses and refer us to additional information if applicable. The specific metadata sought is presented in Table 1. Missing data were filled out by the authors, either by referring to online information or by consulting with the respective course contact persons directly. This step was important to enrich and verify the information found via the online search.

Table . Course metadata Items.

Item	Description
Title	Course title
Course ID	ID number applied to the course by the researchers
Provider	Institution providing the course
Contact information	Contact information for further inquiries (contact person and, if available, e-mail)
URL	Link to the course website (if available)
Location	Onsite or blended or hybrid or online
Duration	Duration of course as specified by the provider
Language	Course language (German, French, Italian, and English)
Costs	Costs of the course in Swiss Francs (CHF)
Certification	If the course yields a certification of attendance, title, or similar
Accreditation	Number of CME (continuous medical education)—credits or ECTS—credits for a study program
Target audience	For example, Attending Physicians in student care, all health care staff, bedside tutors, etc

Data Analysis

In the data collection phase, we observed an interplay between the course offerings and our theoretical framework, Sidhu et al [28]. It is important to note, however, that despite our openness to various health care professions at the start of the project, we ultimately focused our analysis exclusively on courses aimed at educators teaching physicians or medical students. This decision was made to maintain the scope of our study within manageable bounds, ensuring our research remained focused and relevant to our primary audience.

Our intention to mirror the actual landscape of Swiss medical education in its present state guided the database construction. Therefore, the selection of keywords for sorting courses and the strategy for their visualization were meticulously developed to emphasize prevalent topics and reveal the intricacies of course content, ensuring our methodology resonates with the real-world context of medical education. We also aimed to maintain the original terminology and language of course descriptions where possible to respect Switzerland's multilingual landscape and to accurately reflect the providers' offerings.

Building the Course Database

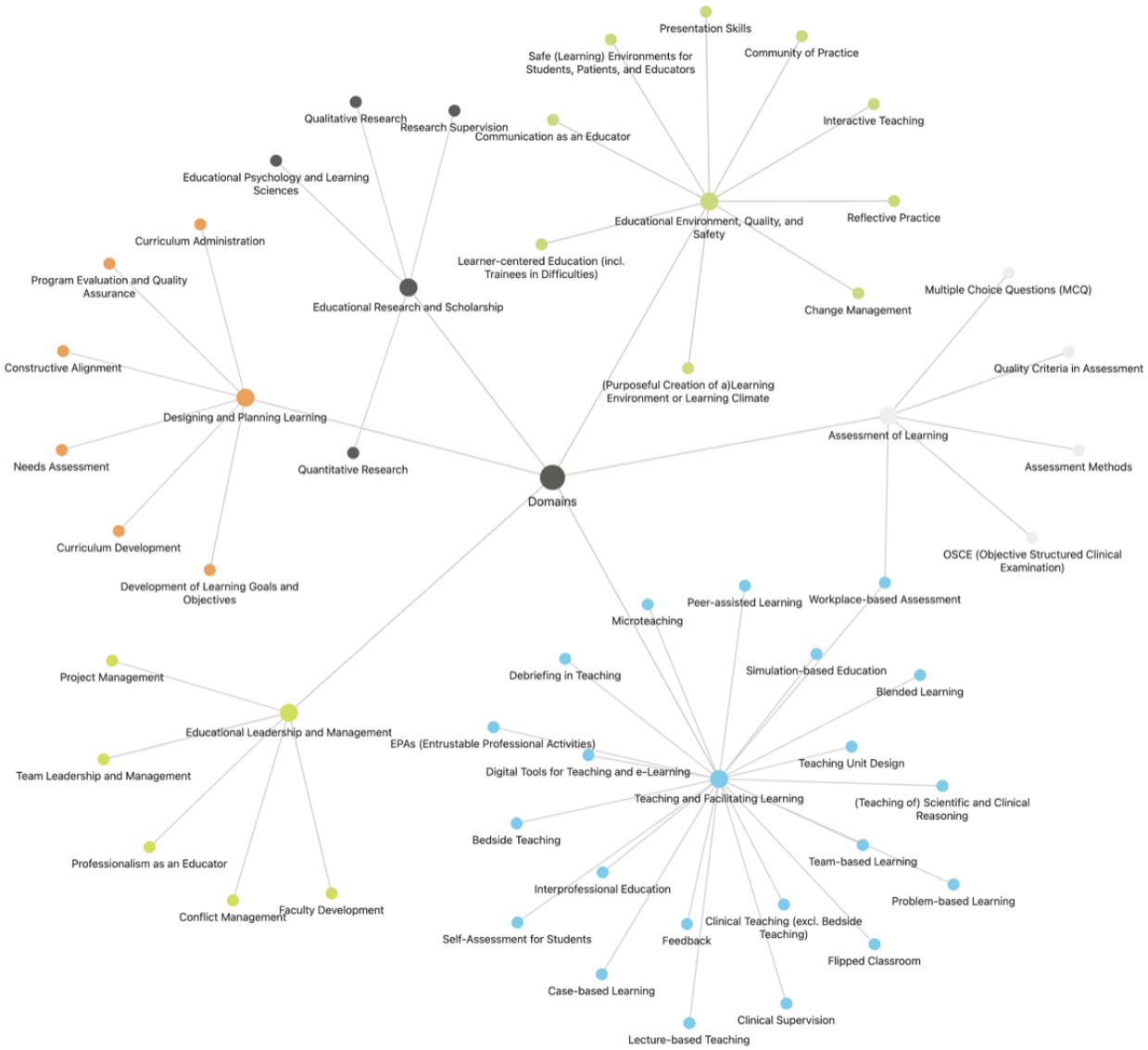
We chose Obsidian [30] as our platform for several practical reasons: its intuitive interface and its free access model for educational use made it accessible for users without extensive technical expertise, aligning with our goal for easy usability and adaptability. Obsidian allows for direct, interactive engagement with the data; this feature not only facilitated a more nuanced understanding of the data but also enabled us to refine and validate our keyword mapping and course organization strategies. Furthermore, Obsidian offered an integrated solution for both analyzing and publishing the data online, using Obsidian Publish [30], simplifying the transition from data collection to dissemination. This integration was particularly advantageous for providing an interactive, accessible resource for medical educators and ensuring the database's utility extended beyond our research team.

The spreadsheet data were then transferred by hand into an Obsidian database, with each course receiving its own Markdown file. Each file included a metadata section on the course and expanded sections on course descriptions, learning objectives and target audience, if available. All course data is provided in the original language and, if not already in English, was translated into English using DeepL, a free online translation tool [31].

Course Keywording, Standardization, and Refinement of Terminology

AG initially reviewed the compiled comprehensive course data to refine and validate our approach to organizing Swiss medical education course data, identifying preliminary keywords reflective of the content's breadth. Subsequently, AG and EH collaborated to align these keywords with the competency domains and definitions provided by Sidhu et al [28], ensuring a robust mapping grounded in established educational frameworks (see Multimedia Appendix 2 for a visual representation of the mapping process). The interplay between the domains provided by Sidhu et al. and the keywords as displayed in Obsidian.md is visualized in Figure 1.

Figure 1. Obsidian visualization of domains and keywords.



AG initially reviewed the compiled comprehensive course data to refine and validate our approach to organizing Swiss medical education course data, identifying preliminary keywords reflective of the content's breadth. Subsequently, AG and EH collaborated to align these keywords with the competency domains and definitions provided by Sidhu et al [28], ensuring a robust mapping grounded in established educational frameworks (see Multimedia Appendix 2 for a visual representation of the mapping process). The interplay between the domains provided by Sidhu et al and the keywords as displayed in Obsidian.md is visualised in Figure 1 .

This thorough process not only enhanced the accessibility and navigability of course data but also underscored our commitment to precision and educational integrity in documenting the landscape of medical education in Switzerland.

In our methodology, we iteratively developed a set of keywords to authentically represent various aspects of Switzerland's medical education landscape. Each keyword, crucial for clarity and precision, was documented and defined in separate Markdown files within Obsidian. Notably, integrating these definitions directly into Obsidian's Markdown documents enabled the keywords to serve as interactive reference points. By simply hovering over them, users can access a preview of the definition, improving the coherence and navigability of our documentation.

Database Publication

In the development of our study, we aimed to create a comprehensive and interactive database encapsulating the entirety of Switzerland's medical education landscape, specifically designed for ease of use and utility by medical educators. To make Switzerland's medical education course offerings easily navigable and useful for the academic community, we used Obsidian Publish [30] for the database's implementation. This platform was selected to ensure that the database was as interactive and accessible as possible.

To strengthen the trustworthiness of our search strategies we directly contacted each course provider to verify the representation of their course content and confirm the currency and correctness of the information.

Additionally, we prepared to release a template on GitHub to ensure our methodology was transparent and could be replicated in other contexts. This template provides the structure and template files necessary for creating a similar database, focusing on the organization and presentation of data. By offering these template files, we intended to facilitate the adoption of our approach by researchers or educators interested in developing educational landscape databases for different settings. The combination of using Obsidian Publish for the database and GitHub for sharing the template underscores our commitment to accessibility, transparency, and the potential for our work to be adapted and applied broadly.

Reflexivity

Acknowledging the importance of reflexivity for our constructivist approach, we critically examined the influence of our roles and backgrounds on this study. Each researcher brought distinct perspectives to conceptualizing a medical educator's role: AG started this project shortly after finishing his studies and provided insights on studying medicine, role definitions, and teaching cultures from a German-speaking and Italian-speaking Swiss region. MB, with a robust background as a practicing internal and emergency medicine doctor and a Master of Medical Education from the University of Bern obtained in 2006, brought insights both from her tenure as president of the SIWF and her experience in teaching emergency medicine courses. EH transitioned from clinical practice and medical education in Germany (Master of Medical Education in 2020) into medical education research in Switzerland (PhD in Medical Education), enriching our project with her comparative view and interest in theoretical concepts. To further enhance the validity of our keyword selection and study design, we engaged with 3 medical educators from Germany and Switzerland who have vast experience in curriculum development. Their feedback was instrumental in refining our course keywords, ensuring our framework resonated with the complexities and nuances of medical education across different linguistic and educational landscapes.

Study Setting

The present study was carried out in Switzerland, with a focus on identifying and analyzing medical education courses available within the country. The data collection phase spanned from May 2023 to December 2023. Subsequent data processing occurred concurrently and extended until March 2024.

Ethical Considerations

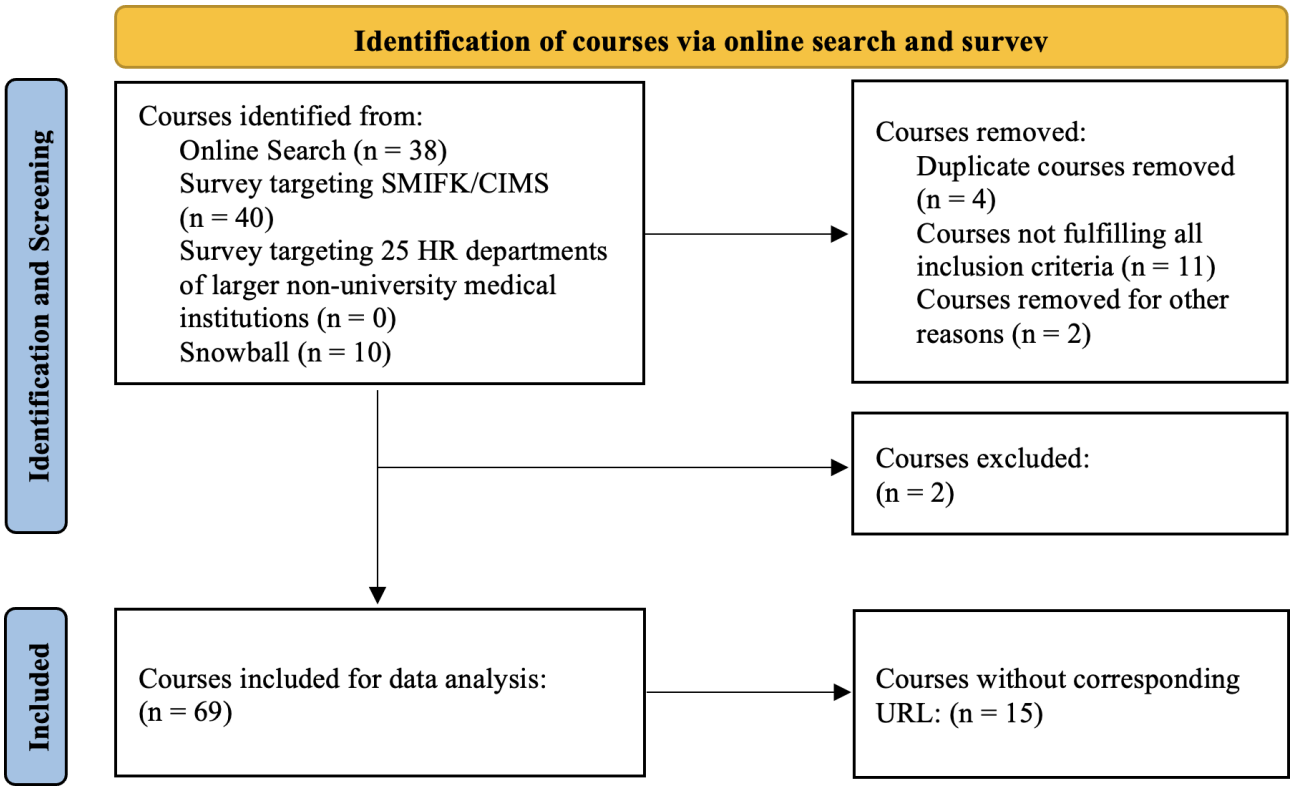
As no participants were involved, no ethical approval was needed. Everybody who supported this study by providing data, did so voluntarily without any incentives or conflicts of interest.

Results

Courses and Keywords

The initial online search, including the snowball method-like approach, resulted in 38 courses. The survey targeting the SMIFK/CIMS members resulted in 40 additional courses. The survey targeting Human Resources departments of 25 larger, non-university medical institutions in Switzerland yielded no additional new courses. In total, our search method yielded 69 eligible courses, as visible in Figure 2. Of the 69 total eligible courses identified, 36 were found through online searches alone, while 33 were discovered exclusively through survey responses.

Figure 2. Flowchart of the course identification process. One course that omitted whether physicians were part of its target audience expressed the wish upon further inquiry not to be included in this study for now but to be re-evaluated in future similar endeavors due to the current reorganization of the course offerings. Another course provider told us upon request for further information that our information on their course provided as a survey result was outdated and that they could not share further details about their course. Adapted from Page et al [32].



To ensure the reliability and currency of our findings, we checked back our data with the respective course providers. Of the 69 courses, 54 had active URLs, while 15 courses were documented without direct web links. These courses were primarily identified through institutional contacts and verified through direct communication in order to maintain the integrity of our comprehensive search strategy.

In our findings, undergraduate teaching courses outnumbered those aimed at postgraduate education (55 out of 69 courses aimed at undergraduate training), while most courses were offered by medical faculties (50 out of 69). The majority of courses were offered in French. In total, 11 courses were offered either additionally or exclusively in English. Most courses were offered onsite, meaning in an offline, face-to-face setting. The descriptive statistics derived from the course metadata, summarizing language use, provider types, course duration, and pricing, are summarized in Table 2.

Table . Aggregated course statistics.

Description	Count
Total number of courses	69
Average cost (CHF)	880
Courses offered free of charge	28
Courses with unknown costs	10
Course location	
Onsite	50
Hybrid	7
Blended	4
Online	6
Unspecified	2
Course provider	
Swiss Institute for Postgraduate Medical Education (SIWF)	8
Medical faculties	50
Specialists' associations	3
University hospitals	5
Other	3
Course language[a]	
German	36
French	40
Italian	7
English	11
Romansh	0
Target setting[b]	
Undergraduate teaching	55
Postgraduate teaching	28
Duration, h, average (range)	42 (1.5–8000)

[a]The count for languages may exceed the total number of courses because some courses are offered in more than one language.

[b]The count may exceed the total number of courses because some courses target both teaching settings.

In matching the course content across the 6 educational competency domains delineated by Sidhu et al, we identified 52 keywords. The distribution of these keywords was uneven, with 42% (22/52) of the keywords falling into the "Teaching and Facilitating Learning" domain. In contrast, the domains of "Educational Leadership and Management" and "Assessment of Learning" were represented to a lesser extent, with only 5 keywords each, while "Educational Environment, Quality, and Safety" had 10 keywords. The "Educational Research and Scholarship" domain showed the least coverage with 4 keywords. Individual courses varied in their keyword coverage, with an average of 4 keywords per course (median: 3); one comprehensive course covered 33 keywords across domains. The distribution of keywords across domains is presented in Table 3, with a graphical overview represented in Figure 1.

Table . Domains and keyword distribution among courses.

Keywords	Keywords per domain
Assessment of Learning	5
Designing and Planning Learning	6
Educational Environment, Quality, and Safety	10
Educational Leadership and Management	5
Educational Research and Scholarship	4
Teaching and Facilitating Learning	22
Keywords covered per Course (average, median, max)	4, 3, 33
Sum of keywords	52

Database and Template Publication

Using Obsidian publish, the database containing the Swiss Course Data were published online [33] (Screenshots of the landing page and an exemplary course are provided with Figures 3 and 4, please refer to Multimedia Appendix 3 for a video example of navigating the database). We also published an exemplary, ready-to-use database template with a Markdown folder and file structure as a GitHub-Repository, together with instructions explaining the structure and use of the repository to facilitate adoption by other users [34].

Figure 3. Screenshot of the publicly accessible database landing page.

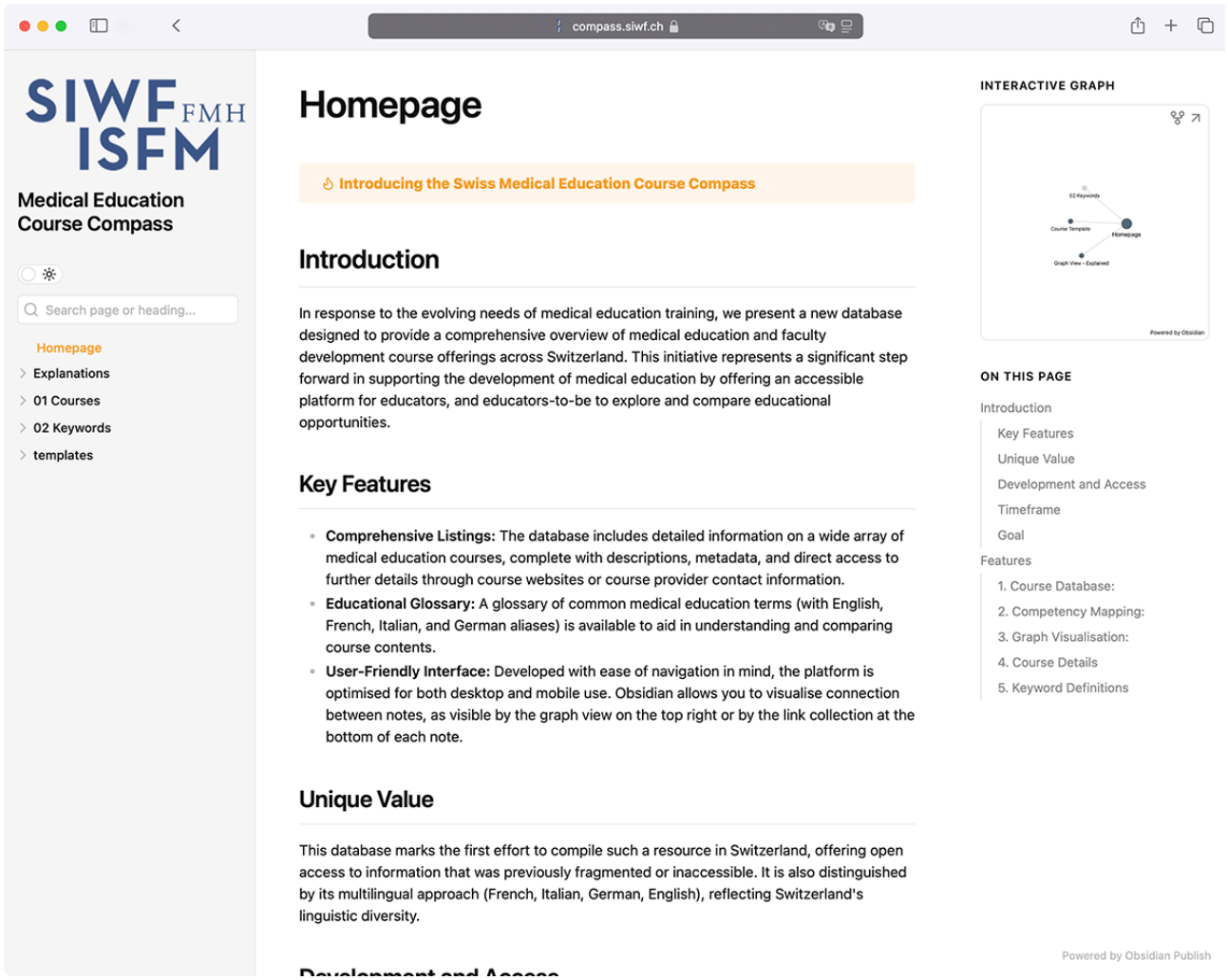
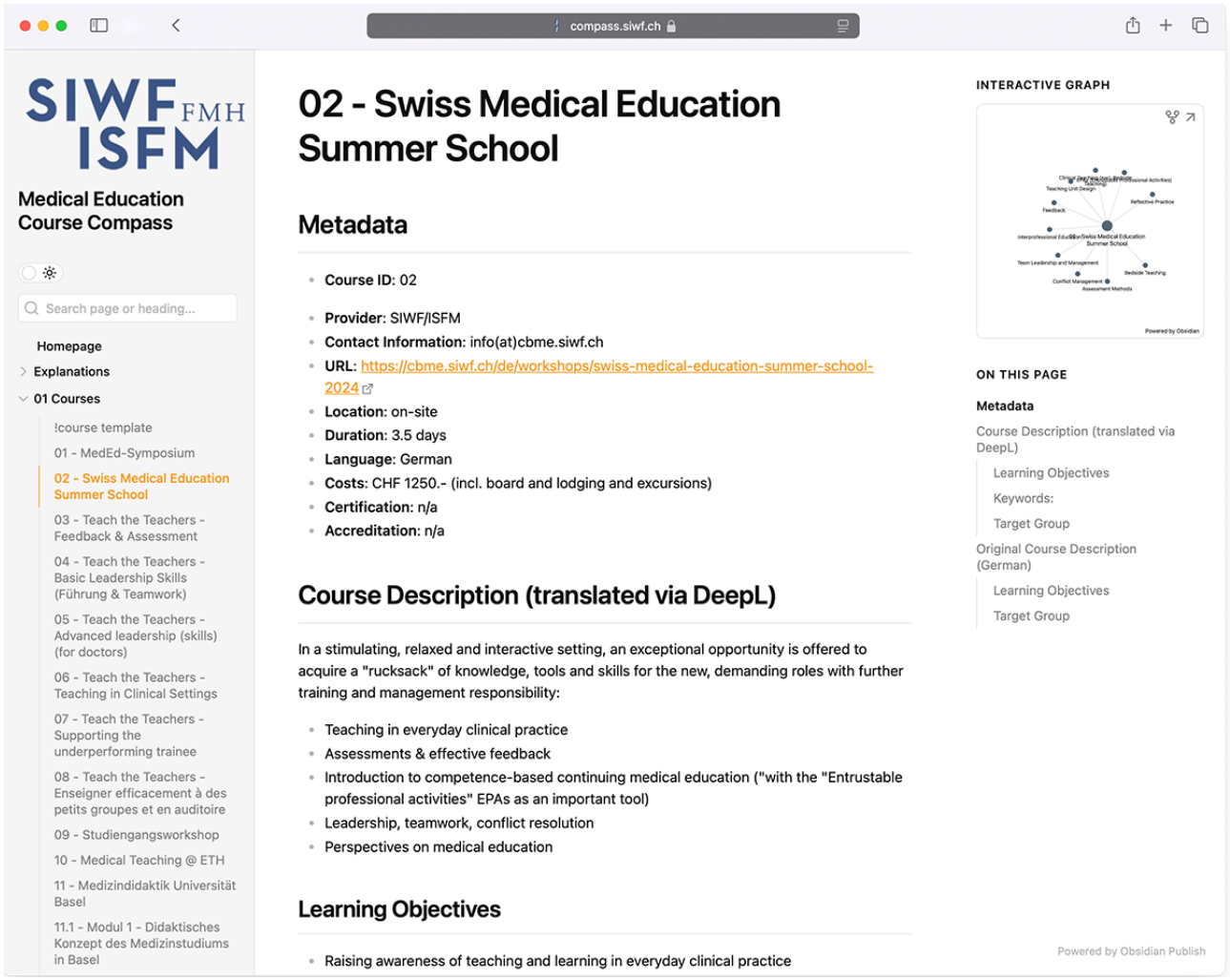


Figure 4. Exemplary screenshot of one course.



Discussion

Principal Findings

In line with the 2 aims of this study, our results cover 2 aspects. First, we searched for, analyzed, and presented the Swiss courses. We found that 69 courses are provided, of which 55 target teachers of undergraduate learners and 28 target teachers of postgraduate learners. The courses were offered in several languages (52% German, 58% French, 10% Italian, 16% English) and several formats (72% onsite, 10% in a hybrid manner, 6% as a blended-learning format, and 9% online). Course content covered mainly the “Teaching and Facilitating Learning” domain, which contained 42% (22/52) of all identified keywords and was also the most prominent in course offerings, with the keyword ’feedback’ being featured most frequently, appearing in 40.6% (28/69) of all courses.

Second, we created an internationally transferable strategy based on an international framework, using free-for-educational-use software, with published search strings to be adapted by others, and providing all metadata open access.

This study addressed the need for an approachable method of sorting and displaying courses by providing a comprehensive, accessible database of medical education courses in Switzerland. This initiative is particularly relevant in contexts where medical education training lacks clear structure and uniformity, especially across different linguistic and cultural regions.

While we acknowledge that our search strategy may not have uncovered every course, we believe this data set sufficient for our study’s objectives and that it reflects the real-world challenges faced by aspiring medical educators in navigating the educational landscape.

Challenges in Navigating the Medical Education System

Several pervasive challenges exist in medical education. These include a discernible lack of dedicated teaching time [35], insufficient, non-transparent, or misdirected funding [36], and an absence of recognition for educators within the medical field [37]. Such issues create a universal backdrop against which specific national education systems can be analyzed.

Our research provides a structured analysis of medical education courses in the Swiss system, illuminating both publicly accessible offerings and those that emerge through professional networks. Significantly, nearly half of the identified courses (33 out of 69) were discovered only through contact with course providers rather than public online searches, revealing a critical accessibility gap. This finding suggests that course discovery often depends heavily on existing professional networks and insider knowledge, potentially creating barriers for newcomers

to the field who lack such connections. This study is the first of its kind to investigate medical didactics courses in a structured manner in Switzerland, providing an overview and allowing new perspectives.

Our study uncovers a complex challenge within the Swiss medical education system: navigating the educational landscape shaped by 4 national languages—German, French, Italian, and Romansh. Among these, German, French, and Italian are official languages. In this multilingual setting, English emerges as a crucial lingua franca in international research and academia and as a common medium within Switzerland, facilitating communication across linguistic barriers. This is reflected in the educational offerings, with 11 courses being conducted partially or entirely in English. Such linguistic plurality further aggravates the existing variability in didactic methods and medical education and research terminology [38,39], presenting additional challenges to the creation of an educational framework. This situation is further complicated by issues of information accessibility—with 15 courses not having available URLs, it renders it difficult for potential participants to find accurate, up-to-date information about these educational opportunities online.

Implications for Medical Educator Career Pathways

Building on the initial findings, our work, as the first comprehensive compilation of available didactic courses within the Swiss medical education system, can potentially enable more transparency in the field. By illuminating the current educational offerings, our research could facilitate faculty development and inform career choices for medical educators, as they now have a more straightforward overview of the resources and opportunities available to them. This transparency is a critical step towards enhancing career pathways, as it allows for a strategic approach to personal and professional development within the context of medical education.

However, it is essential to note that our study revealed a predominance of courses aimed at teaching in undergraduate (medical school) settings. While this finding initially suggests a potential undersupply of educational opportunities targeting postgraduate teaching, it notably reflects Switzerland's institutional structure, where universities are primarily responsible for undergraduate medical education but not residency training or CME. This structural arrangement, where different institutions oversee different stages of medical education, may contribute to the observed imbalance in course offerings. The relative scarcity of training for postgraduate settings might indicate a need for a broader array of courses catering to the ongoing development of medical professionals beyond their initial degrees. Addressing this gap could lead to a more robust and comprehensive educational framework, ensuring that medical educators are well-equipped to foster the next generation of medical practitioners at all stages of their professional journey.

Comparison to International Models

Our comparative analysis with international models highlights notable differences in structuring medical education career pathways. Particularly in Anglophone countries, such as Australia, Canada, the United Kingdom, and the United States, well-established career pathways and educator frameworks are actively promoted and used, unlike in Switzerland, where such structured frameworks are absent [3,11,12].

When aligning our findings with the integrative framework by Sidhu et al [28], which outlines 6 educator competency domains, we observed that Swiss medical education predominantly emphasizes the domain of "Teaching and Facilitating Learning." This domain is focused on enhancing learning through suitable methods and resources, including assessment for learning. This heavy focus may inadvertently lead to a neglect of the other 5 domains ("Educational Leadership and Management," "Educational Environment, Quality, and Safety", "Designing and Planning Learning," "Assessment of Learning," and "Educational Research and Scholarship"), suggesting a possible imbalance in the educational emphasis. Our keyword map corroborates this emphasis, revealing a detailed and nuanced depiction of the "Teaching and Facilitating Learning" domain. Our mapping strategy aimed to reflect the specificity and depth of course offerings without excessive summarization, thus indicating their relative prominence by the frequency with which they appear in our data set.

The predominance of the "Teaching and Facilitating Learning" domain could stem from its direct applicability in educational settings and the relative ease of teaching and assessing these skills. This focus on tangible teaching methods and resources, which offer clear benefits and outcomes, could make it a natural focal point for educators aiming to directly impact student achievement. However, this emphasis may inadvertently lead to the undervaluation of other critical educator competencies, such as "Educational Leadership and Management" or "Educational Research and Scholarship." These domains encompass more abstract competencies that resemble attitudes or overarching professional dispositions rather than concrete skills, presenting challenges for direct instruction due to their less immediately visible impacts and harder-to-quantify qualities.

The challenge of effectively imparting the more abstract domains within medical education and the observable predilection for addressing more accessible teaching topics can be analogized to the tendency to assess readily quantifiable competencies [40]. This parallel might reflect an educational predilection for what can be straightforwardly taught and measured, perhaps at the expense of more profound, more complex competencies that are less amenable to conventional assessment methodologies. Such a tendency may not fully encapsulate the multifaceted nature of medical educator competencies, underscoring a potential disjunction between educational priorities and the comprehensive skill set required for clinical excellence.

The marked predominance of teaching methods and learning facilitation within the Swiss context, as opposed to a more balanced distribution across the 6 domains, is an intriguing phenomenon. We propose that this area warrants further inquiry. Investigating why there is such a strong focus on these teaching competencies within Switzerland, especially compared to the broader scope seen internationally, could yield insights that

inform future developments in medical education. This analysis might ultimately contribute to enhancing educator frameworks and diversifying professional development opportunities.

Future Directions and Potential Impact

The insights garnered from this study lay the groundwork for developing a unified and expansive framework for medical education in Switzerland, emphasizing the need for a formal recognition and certification process for medical education courses and medical educators. By enhancing awareness of available courses and clarifying the expectations for a medical educator, this initiative could significantly improve the quality of medical education. Increased awareness is crucial for establishing clear competencies in the first place, and it also serves to acknowledge and validate existing efforts while encouraging the development of future educational opportunities. The overview of courses also offers the basis for future discussions, which have already begun. Currently, the database does not give information on the quality of courses, as this could not reliably be assessed. For a future version of the database, methods to assess course quality and display this transparently are developed.

In addition, the overview on courses in comparison with the framework has shown some blind spots of the current training. Courses on “Assessment of Learning,” “Educational Leadership and Management,” and “Educational Research and Scholarship” seem to be rare. As a follow-up to this study, the responsible groups (SMIFK and SIWF) will discuss these gaps and take this as a base for future course offerings.

Furthermore, this focus on a structured framework could resonate on an international level, offering a model for contexts where clear career pathways for medical educators are similarly undefined, thus having the potential to inform and shape international standards in medical education.

While this study focused exclusively on courses on medical education (physician-adjacent) settings to maintain a manageable scope, we deliberately chose Sidhu et al’s framework for its interprofessional applicability across health care education. Combined with our intentionally transparent and replicable methodology, this provides a strong foundation for other health care professions to adapt our approach to mapping their own educational landscapes. The template we developed could serve as a starting point for similar analyses in other health professions’ education systems.

Limitations

Our study encountered different challenges. The research relied primarily on internet-searchable courses, potentially overlooking non-digital or unadvertised educational offerings. Additionally, our snowball search approaches introduce potential selection biases.

As we focused on assembling an overview instead of the details of courses, the depth of content validation regarding, for example, course quality, pedagogical methodologies, and educational effectiveness is limited.

Future research should focus on course content and course quality and investigate the usability and comprehensiveness of our database, including an accessibility audit of course information, a comprehensive comparison with expert-identified course offerings, and systematic verification of course details. We are planning to investigate the implications of the use of the database, such as changes in content offered and new strategies to acknowledge medical educator careers.

Conclusions

The methodology showcased in our study serves as a valuable template for international adaptation, offering insights into how diverse educational systems can be evaluated, organized, and presented effectively. By innovatively mapping the educational landscape with the example of Switzerland, we have provided a replicable approach that could aid other countries, even with similar linguistic and cultural challenges. The creation of our database amplifies the visibility and transparency of medical education courses available, facilitating better decision-making for educators. This enhanced access to information allows medical professionals to chart their career progressions more effectively, emphasizing the benefits of more transparent navigation, especially in environments without well-defined educational pathways.

Through this pioneering work, we aim to contribute to the global conversation on medical education, fostering a more interconnected and accessible educational landscape that benefits educators and students alike.

Acknowledgments

We thank Dr phil-nat. Elke Bayha, Dr rer nat Felix Joachimski, MME, and Dr med vet Melanie Simon, MME, for their critical review and helpful comments on developing the keyword map. We also thank all course providers and survey participants for their collaboration on this project and for so generously answering our questions on course details.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Document outlining our Search Strategy for the online search.
[DOCX File, 23 KB - mededu_v11i1e62838_app1.docx]

Multimedia Appendix 2
Screenshot of Miro Board used for mapping between domains and keywords which helped facilitate discussions between authors.
[PNG File, 1372 KB - mededu_v11i1e62838_app2.png]

Multimedia Appendix 3
Video example of navigating the publicly accessible online database
[MP4 File, 204005 KB - mededu_v11i1e62838_app3.mp4]

References

1. Norman G. Medical education: past, present and future. Perspect Med Educ 2012 Mar;1(1):6-14. [doi: 10.1007/s40037-012-0002-7] [Medline: 23316454]
2. Sethi A, Ajjawi R, McAleer S, Schofield S. Exploring the tensions of being and becoming a medical educator. BMC Med Educ 2017 Mar 23;17(1):62. [doi: 10.1186/s12909-017-0894-3] [Medline: 28335820]
3. Bartle E, Thistlethwaite J. Becoming a medical educator: motivation, socialisation and navigation. BMC Med Educ 2014 May 31;14(1):110. [doi: 10.1186/1472-6920-14-110] [Medline: 24885740]
4. Sherbino J, Frank JR, Snell L. Defining the key roles and competencies of the clinician–educator of the 21st century. Acad Med 2014 May;89(5):783-789. [doi: 10.1097/ACM.0000000000000217] [Medline: 24667507]
5. Sherbino J, Snell L, Dath D, Dojeiji S, Abbott C, Frank JR. A national clinician-educator program: A model of an effective community of practice. Med Educ Online 2010 Dec 6;15(1):5356. [doi: 10.3402/meo.v15i0.5356] [Medline: 21151594]
6. Fitzgerald JT, Burkhardt JC, Kasten SJ, et al. Assessment challenges in competency-based education: a case study in health professions education. Med Teach 2016 May;38(5):482-490. [doi: 10.3109/0142159X.2015.1047754] [Medline: 26052881]
7. Gruppen LD, Burkhardt JC, Fitzgerald JT, et al. Competency-based education: programme design and challenges to implementation. Med Educ 2016 May;50(5):532-539. [doi: 10.1111/medu.12977] [Medline: 27072442]
8. Master of HPE of university of michigan medical school. Learning Health Sciences. 2017 Mar 17. URL: https://medicine.umich.edu/dept/lhs/education/degree-programs/master-health-professions-education-mhpe/curriculum [accessed 2023-11-03]
9. Academy of Medical Educators. Professional Standards for Medical, Dental and Veterinary Educators, 4th edition: Academy of Medical Educators; 2021. URL: https://www.medicaleducators.org/write/MediaManager/Documents/AoME_Professional_Standards_4th_edition_1.0_(web_full_single_page_spreads).pdf [accessed 2025-04-09]
10. Wenger E. Communities of practice and social learning systems: the career of a concept. In: Blackmore C, editor. Social Learning Systems and Communities of Practice: Springer; 2010:179-198. [doi: 10.1007/978-1-84996-133-2_11]
11. Walport M. Medically-and dentally-qualified academic staff: recommendations for training the researchers and educators of the future. : Academic Careers Sub-committee of Modernising Medical Careers and the UK Clinical Research Collaboration; 2005 URL: https://www.ukcrc.org/wp-content/uploads/2014/03/Medically_and_Dentally-qualified_Academic_Staff_Report.pdf [accessed 2025-04-09]
12. Church H, Brown MEL. Rise of the Med-Ed-ists: achieving a critical mass of non-practicing clinicians within medical education. Med Educ 2022 Dec;56(12):1160-1162. [doi: 10.1111/medu.14940] [Medline: 36148497]
13. SGAIM - Schweizerische Gesellschaft für Allgemeine Innere Medizin. Fachärztin oder facharzt für allgemeine innere medizin - weiterbildungsprogramm vom. 2022 Jan 1. URL: https://www.sgaim.ch/fileadmin/user_upload/Weiterbildung/aim_wbp_d_1_.pdf [accessed 2025-01-16]
14. Jelinek GA, Weiland TJ, Mackinlay C. Supervision and feedback for junior medical staff in Australian emergency departments: findings from the emergency medicine capacity assessment study. BMC Med Educ 2010 Nov 2;10(1):74. [doi: 10.1186/1472-6920-10-74] [Medline: 21044342]
15. Steinert Y, McLeod PJ, Boillat M, Meterissian S, Elizov M, Macdonald ME. Faculty development: a “field of dreams”? Med Educ 2009 Jan;43(1):42-49. [doi: 10.1111/j.1365-2923.2008.03246.x] [Medline: 19140996]
16. Fish RM, Gawne SJ, Machin L. Balancing medical education with service in the workplace: a qualitative case study. JWL 2022 Jan 27;34(2):176-187. [doi: 10.1108/JWL-05-2021-0064]
17. Piquette D, Moulton CA, LeBlanc VR. Balancing care and teaching during clinical activities: 2 contexts, 2 strategies. J Crit Care 2015 Aug;30(4):678-684. [doi: 10.1016/j.jcrc.2015.03.002] [Medline: 25776896]
18. Ramani S, Leinster S. AMEE Guide no. 34: teaching in the clinical environment. Med Teach 2008;30(4):347-364. [doi: 10.1080/01421590802061613] [Medline: 18569655]
19. Hohl F. 42-Stunden-Woche mit grossem Plus. Schweiz Ärzteztg . [doi: 10.4414/saez.2023.1267925043]
20. Ng SL, Baker L, Cristancho S, Kennedy TJ, Lingard L. Qualitative research in medical education: methodologies and methods. In: Swanwick T, Forrest K, O’Brien BC, editors. Understanding Medical Education, 1st edition: Wiley, Vol. 2018:427-441. [doi: 10.1002/9781119373780]
21. Varpio L, Paradis E, Uijtdehaage S, Young M. The distinctions between theory, theoretical framework, and conceptual framework. Acad Med 2020 Jul;95(7):989-994. [doi: 10.1097/ACM.0000000000003075] [Medline: 31725464]
22. Crosby RJ. AMEE Guide No 20: the good teacher is more than a lecturer - the twelve roles of the teacher. Med Teach 2000 Jan;22(4):334-347. [doi: 10.1080/014215900409429]

23. Hesketh EA, Bagnall G, Buckley EG, et al. A framework for developing excellence as A clinical educator. Med Educ 2001 Jun;35(6):555-564. [doi: 10.1046/j.1365-2923.2001.00920.x] [Medline: 11380858]
24. Srinivasan M, Li STT, Meyers FJ, et al. "Teaching as a Competency": competencies for medical educators. Acad Med 2011 Oct;86(10):1211-1220. [doi: 10.1097/ACM.0b013e31822c5b9a] [Medline: 21869655]
25. Parson L, Childs B, Elzie P. Using competency-based curriculum design to create a Health Professions Education Certificate Program that Meets the needs of students, administrators, faculty, and patients. Health Professions Education 2018 Sep;4(3):207-217. [doi: 10.1016/j.hpe.2018.03.008]
26. Walsh A, Koppula S, Antao V, et al. Preparing teachers for competency-based medical education: Fundamental teaching activities. Med Teach 2018 Jan;40(1):80-85. [doi: 10.1080/0142159X.2017.1394998] [Medline: 29113520]
27. Fallis D, Irwin S, Cervero R, Durning S. Frameworks to guide faculty development for health professions education: a scoping review. J Contin Educ Health Prof 2022 Jul 1;42(3):180-189. [doi: 10.1097/CEH.0000000000000376] [Medline: 34459440]
28. Sidhu NS, Allen KJ, Civil N, et al. Competency domains of educators in medical, nursing, and health sciences education: an integrative review. Med Teach 2023 Feb;45(2):219-228. [doi: 10.1080/0142159X.2022.2126758] [Medline: 36179761]
29. Barrier J. Conférence Internationale des Doyens des Facultés de Médecine d' Expression Française. Recommandations pour la formation Pédagogique des enseignants de Médecine. 1998. URL: https://docplayer.fr/19159965-Recommandations-pour-la-formation-pedagogique-des-enseignants-de-medecine.html [accessed 2023-11-03]
30. Li S, Xu E. Dynalist Inc. Obsidian.md. URL: https://obsidian.md/ [accessed 2023-11-02]
31. DeepL SE. DeepL Translator. URL: https://www.deepl.com/translator [accessed 2024-02-04]
32. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
33. Homepage - Medical Education Course Compass. URL: https://compass.siwf.ch/Homepage [accessed 2024-05-20]
34. Gashi A. Andi-gashi/mededcoursecompass. 2024 Apr 1. URL: https://github.com/andi-gashi/mededcoursecompass [accessed 2024-05-05]
35. Aldeen AZ, Gisondi MA. Bedside teaching in the emergency department. Acad Emerg Med 2006 Aug;13(8):860-866. [doi: 10.1197/j.aem.2006.03.557] [Medline: 16766739]
36. Dacre J, Walsh K. Funding of medical education: the need for transparency. Clin Med (Northfield) 2013 Dec;13(6):573-575. [doi: 10.7861/clinmedicine.13-6-573]
37. Fantaye AW, Kitto S, Hendry P, et al. Attributes of excellent clinician teachers and barriers to recognizing and rewarding clinician teachers' performances and achievements: a narrative review. Can Med Educ J 2022 May;13(2):57-72. [doi: 10.36834/cmej.73241] [Medline: 35572019]
38. Wojtczak A. Medical education terminology. Med Teach 2002 Jul;24(4):357-357. [doi: 10.1080/01421590220145699] [Medline: 12193314]
39. Finn GM, Charmer B, Burton OE, et al. How to define clinical education research terminology: a glossary. Clin Teach 2023 Aug;20(4):e13605. [doi: 10.1111/tct.13605] [Medline: 37503773]
40. Foster N. 21st Century competencies: challenges in education and assessment Innov Assess Meas Support Complex Ski. URL: https://www.oecd.org/en/publications/innovating-assessments-to-measure-and-support-complex-skills_e5f3e341-en/full-report.html [accessed 2025-04-09]

Abbreviations

CME: continuous medical education
SIWF: Swiss Institute for Postgraduate Medical Education
SMIFK/CIMS: Joint Commission of the Swiss Medical Schools

Edited by B Lesselroth; submitted 12.06.24; peer-reviewed by AMH Chen, D Jackson; revised version received 19.01.25; accepted 25.02.25; published 16.04.25.

Please cite as:
Gashi A, Brodmann Maeder M, Hennel EK
Making Medical Education Courses Visible: Theory-Based Development of a National Database
JMIR Med Educ 2025;11:e62838
URL: https://mededu.jmir.org/2025/1/e62838
doi:10.2196/62838

Large Language Models in Biochemistry Education: Comparative Evaluation of Performance

Olena Bolgova[1], MD, PhD; Inna Shypilova[2], MD, PhD; Volodymyr Mavrych[1], MD, PhD

[1]College of Medicine, Alfaisal University, Al Takhassousi St, Riyadh, Saudi Arabia

[2]School of Medicine, St Mathews University, George Town, Cayman Islands

Corresponding Author:
Volodymyr Mavrych, MD, PhD
College of Medicine, Alfaisal University, Al Takhassousi St, Riyadh, Saudi Arabia

Abstract

Background: Recent advancements in artificial intelligence (AI), particularly in large language models (LLMs), have started a new era of innovation across various fields, with medicine at the forefront of this technological revolution. Many studies indicated that at the current level of development, LLMs can pass different board exams. However, the ability to answer specific subject-related questions requires validation.

Objective: The objective of this study was to conduct a comprehensive analysis comparing the performance of advanced LLM chatbots—Claude (Anthropic), GPT-4 (OpenAI), Gemini (Google), and Copilot (Microsoft)—against the academic results of medical students in the medical biochemistry course.

Methods: We used 200 USMLE (United States Medical Licensing Examination)–style multiple-choice questions (MCQs) selected from the course exam database. They encompassed various complexity levels and were distributed across 23 distinctive topics. The questions with tables and images were not included in the study. The results of 5 successive attempts by Claude 3.5 Sonnet, GPT-4 - 1106, Gemini 1.5 Flash, and Copilot to answer this questionnaire set were evaluated based on accuracy in August 2024. Statistica 13.5.0.17 (TIBCO Software Inc) was used to analyze the data's basic statistics. Considering the binary nature of the data, the chi-square test was used to compare results among the different chatbots, with a statistical significance level of $P<.05$.

Results: On average, the selected chatbots correctly answered 81.1% (SD 12.8%) of the questions, surpassing the students' performance by 8.3% ($P=.02$). In this study, Claude showed the best performance in biochemistry MCQs, correctly answering 92.5% (185/200) of questions, followed by GPT-4 (170/200, 85%), Gemini (157/200, 78.5%), and Copilot (128/200, 64%). The chatbots demonstrated the best results in the following 4 topics: eicosanoids (mean 100%, SD 0%), bioenergetics and electron transport chain (mean 96.4%, SD 7.2%), hexose monophosphate pathway (mean 91.7%, SD 16.7%), and ketone bodies (mean 93.8%, SD 12.5%). The Pearson chi-square test indicated a statistically significant association between the answers of all 4 chatbots ($P<.001$ to $P<.04$).

Conclusions: Our study suggests that different AI models may have unique strengths in specific medical fields, which could be leveraged for targeted support in biochemistry courses. This performance highlights the potential of AI in medical education and assessment.

(JMIR Med Educ 2025;11:e67244) doi:10.2196/67244

KEYWORDS

ChatGPT; Claude; Gemini; Copilot; biochemistry; LLM; medical education; artificial intelligence; NLP; natural language processing; machine learning; large language model; AI; ML; comprehensive analysis; medical students; GPT-4; questionnaire; medical course; bioenergetics

Introduction

Recent breakthroughs in artificial intelligence (AI), especially in large language models (LLMs), have started a new era of innovation across diverse fields, with medicine leading the charge in this technological revolution. The integration of AI into various medical disciplines such as oncology, radiology, and pathology has demonstrated its advancing clinical uses and its potential to revolutionize health care delivery [1-3]. As new LLMs continue to emerge and evolve, AI is poised to fundamentally reshape our understanding and approach to medicine, offering unprecedented opportunities for improved patient care, diagnostics, and medical education [4].

While academic interest in AI has surged in recent years, integrating AI technologies in educational settings, particularly medicine, has been uneven and fraught with challenges. Among

many AI tools available, ChatGPT has emerged as a potential game-changer in medical education [5,6]. This sophisticated language model, powered by advanced neural networks, demonstrates a remarkable ability to interpret prompts and generate human-like responses, making it difficult to distinguish from human-produced language.

LLM's underlying transformer architecture enables it to excel in natural language understanding, continuously processing and adapting to new information. This adaptability, combined with its vast knowledge base, presents promising opportunities for enhancing teaching and learning methodologies in medical education [7]. AI-powered tools such as ChatGPT may be particularly effective in addressing persistent challenges in student engagement, offering interactive and personalized learning experiences that traditional teaching methods often struggle to provide [8].

OpenAI's GPT-4 and GPT-3.5, Google's Gemini, and Anthropic's Claude have emerged as frontrunners, offering unique capabilities and potential medical education and practice applications. As of 2024, the AI landscape in health care has become increasingly diverse, with over 20 LLMs available for public use. Among them, 4 are the most promising.

Anthropic developed Claude, an AI assistant known for its strong natural language understanding and generation capabilities. It has been trained on a wide range of data and is designed to be helpful, harmless, and honest. Claude has shown particular strength in tasks requiring nuanced understanding and ethical reasoning [9].

Created by OpenAI, GPT-4 is the latest GPT series iteration. It represents a significant advancement over its predecessor, GPT-3, with improved language understanding, generation, and reasoning capabilities. GPT-4 has demonstrated impressive performance across various domains, including coding, creative writing, and analytical tasks [10].

Developed by Google AI, Gemini is a multimodal AI model capable of understanding and generating text, images, and other forms of data. It comes in different sizes and is optimized for various tasks and computational requirements. Gemini has shown strong performance in complex reasoning tasks and can understand context across different modalities [11].

Created by GitHub in collaboration with OpenAI, Copilot is an AI pair programmer designed to assist developers by suggesting code completions and entire functions. It is now an integral part of Microsoft Windows. While primarily focused on coding tasks, Copilot's underlying language model has shown capabilities in understanding and generating natural language [12].

One primary method for assessing the capabilities of LLMs in knowledge-based fields, including medicine, is their performance on multiple-choice tests [13-16]. The release of GPT-4 by OpenAI in 2023 marked a significant milestone, demonstrating impressive test-taking abilities across various domains [17]. Similarly, Claude 2 from Anthropic, released in June 2023, has gained attention for its ability to process larger input spaces (up to 100,000 tokens), potentially allowing for a more comprehensive analysis of medical texts and case studies [8].

The high accuracy demonstrated by ChatGPT-4 in answering multiple-choice questions (MCQs) compared to medical students' performance is particularly noteworthy. It suggests that AI could be an effective study aid, helping students review and reinforce their knowledge across various medical subjects. However, it is essential to view AI as a complementary tool rather than a replacement for MCQs that have transformed from their conventional use as assessment tools to become a versatile educational approach in medical curricula. MCQs stimulate students' cognitive abilities and promote active interaction with study materials. By using advanced generative AI-driven language models to address MCQs in medical physiology and other subjects, educators may provide students with an innovative and engaging learning experience, potentially enhancing their grasp of essential medical concepts, traditional teaching methods, or human expertise [18,19].

Recent studies have begun to compare the performance of different AI models in medical education contexts. For instance, Claude, an LLM developed by Anthropic, has shown promising results in solving medical MCQs. Some studies have indicated that Claude demonstrated a high frequency of right answers and explanations compared to ChatGPT-3.5 [8,20]. These comparative studies are crucial in understanding the strengths and limitations of different AI models in medical education. They help educators and researchers identify the most suitable tools for specific learning objectives and contexts within medical curricula.

Despite the promising results, it is important to note the variability in AI performance across different studies and question types. For example, while some studies reported high accuracy rates for ChatGPT in physiology tests [5,8], others found lower performance rates, particularly as the complexity and difficulty of questions increased [21,22]. This variability underscores the need for careful consideration when integrating AI tools into medical education. Educators must be aware of these tools' strengths and limitations and ensure they are used appropriately to complement, rather than replace, traditional teaching methods.

It is important for educational strategies to prioritize the integration of LLMs into the curriculum as a vital aspect of the learning process. This integration should enable students to cultivate critical thinking and analytical skills, particularly in understanding the constraints of AI. LLMs have the potential to offer students in-depth knowledge and diverse viewpoints, facilitating a more thorough comprehension of intricate medical concepts [23]. By using the output of LLMs and working alongside educators to draw upon their existing knowledge, students can actively participate in the learning process. This collaborative approach allows for the refinement of their understanding and insights. The future of medical education depends on the seamless integration of human expertise with AI-powered tools [3,19,23].

The aim of this study was to conduct a comprehensive analysis comparing the performance of advanced LLM chatbots—GPT-4, Claude, Copilot, and Gemini—against the academic results of

medical students in biochemistry. The research objectives were to evaluate the following hypotheses:

- The AI chatbots will perform similarly to medical students on factual recall and basic concept application questions in biochemistry but may show differences in performance on complex problem-solving scenarios.
- There will be significant variation in performance among the different AI models, with newer models (GPT-4 and Claude) potentially showing higher accuracy compared to earlier versions.
- The AI-driven LLMs' performance will vary across different biochemistry topics, with potentially stronger performance in areas requiring systematic pathway analysis and weaker performance in topics requiring integration of clinical context.

Methods

Study Design

This study focused on a comparative analysis of the capabilities of different AI-driven LLMs in the medical biochemistry course. The research included an examination of 4 chatbots currently available to the public: Claude (Anthropic), GPT-4 (OpenAI), Gemini (Google), and Copilot (Microsoft).

A total of 200 scenario-based MCQs with 4 options and a single correct answer were randomly chosen from the medical biochemistry course's examination database for medical students and validated by 2 independent experts. The study did not include questions with images and tables. The selected questions encompassed various levels of complexity. They were distributed across 23 distinctive categories: structural proteins and associated diseases, globular proteins and hemoglobin, red blood cells (RBCs) and anemia, structure and function of amino acids, structure and function of proteins, bioenergetics and electron transport chain, enzymes, glycolysis and gluconeogenesis, glycogen, signaling mechanisms, pyruvate dehydrogenase and Krebs cycle, cholesterol metabolism, eicosanoids, fatty acid metabolism, fructose and galactose metabolism, hexose monophosphate pathway, ketone bodies, lipoproteins, lysosomal storage diseases, amino acid metabolism, fast and fed state, heme metabolism, and nitrogen metabolism.

Data Collection

For the testing phase, each selected chatbot was required to answer a set of 200 questions, and their performance was evaluated against the responses provided by medical students for the same set of questions. Claude 3.5 Sonnet, GPT-4 - 1106, Gemini 1.5 Flash, and Copilot proficiency in responding to MCQs was assessed in the last 2 weeks of August 2024. An OpenAI paid subscription was obtained to get GPT-4 access.

Each chatbot was given the prompt "generate the list of correct answers for the following MCQs" and provided with a first set of 50 questions; following with the same prompt and 3 more sets of 50 MCQs each, totally there were 200 MCQs in the questionnaire. After that, this procedure was repeated 5 times (no time period between the attempts was assigned). The results of 5 successive attempts by each chatbot to answer this questionnaire set were meticulously recorded in a Microsoft Excel spreadsheet and evaluated based on accuracy. A total of 4000 answers from LLMs were analyzed.

Five random answers were generated and analyzed for the same MCQ set using the RAND() function in Excel (Microsoft 365) to compare chatbot results with random guessing.

Data Analysis

The answers provided by each LLM were recorded and input into the Excel spreadsheet (Microsoft 365). The data from each (1-5) attempt was matched with the answer key and compared with all previous attempts, finding the percentage of repeated and correct answers among them. After that, a detailed item analysis was performed for each chatbot concerning different question categories.

Statistica 13.5.0.17 (TIBCO Software) was used to analyze the data's basic statistics. Considering the data's binary nature, the chi-square test was used to compare results among the different chatbots.

Results

Overview

According to our data, on average, 4 selected chatbots accurately answered 81.1% (SD 12%) of 200 MCQs from the medical biochemistry course. This result was 8.3% (P=.02) above the students' average (mean 72.8%, SD 12.7%) and almost 4 times better than randomly generated responses (mean 22%, SD 2.9%) for the same questions.

There was a significant variation in correct responses among the chatbots. The best result was recorded for Claude (92.5%, SD 0%), followed by GPT-4 (mean 85.1%, SD 1%) and Gemini (mean 78.5%, SD 0%), which were better than the students' average. Copilot showed the lowest result (mean 64%, SD 0%; Figure 1).

Interestingly, all chatbots answered 104 (52%) of the 200 questions correctly in all attempts. General item analysis revealed that eicosanoids, bioenergetics and electron transport chain, hexose monophosphate pathway, and ketone bodies were the 4 best topics, with the mean (SD) results for all chatbots being 100% (0%), 96.4% (7.2%), 91.7% (16.7%), and 93.8% (12.5%), respectively.

In contrast, the lowest results were recorded for globular proteins and hemoglobin (mean 58.4%, SD 26.4%), lipoproteins (mean 64.6%, SD 20.3%), and fructose and galactose metabolism questions (mean 65.8%, SD 29.9%).

After that, each chatbot's results for all 23 topics were evaluated (Figure 2).

Figure 1. Percentile of correct answers from different chatbots and students on 200 multiple-choice questions from the medical biochemistry course.

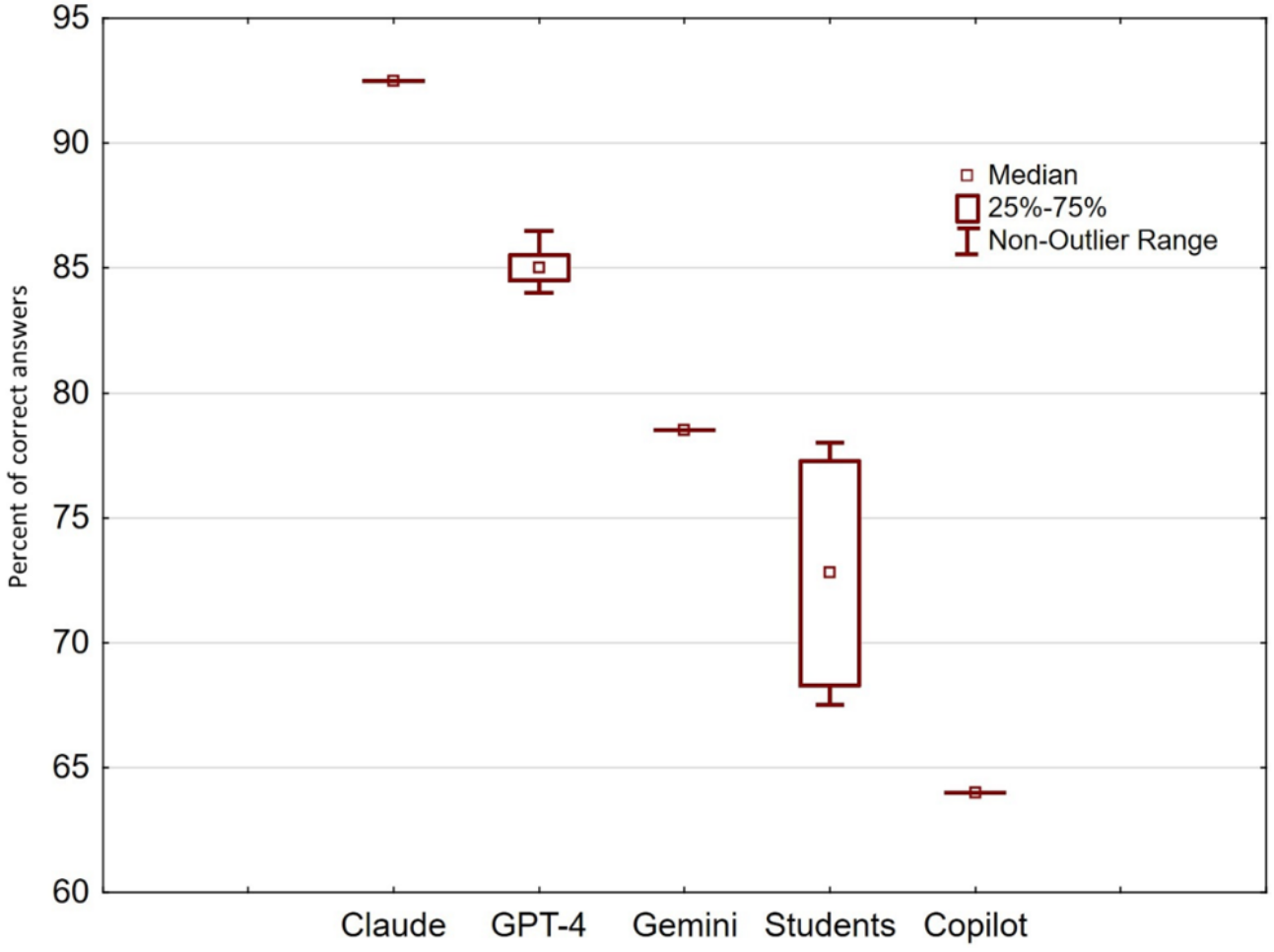
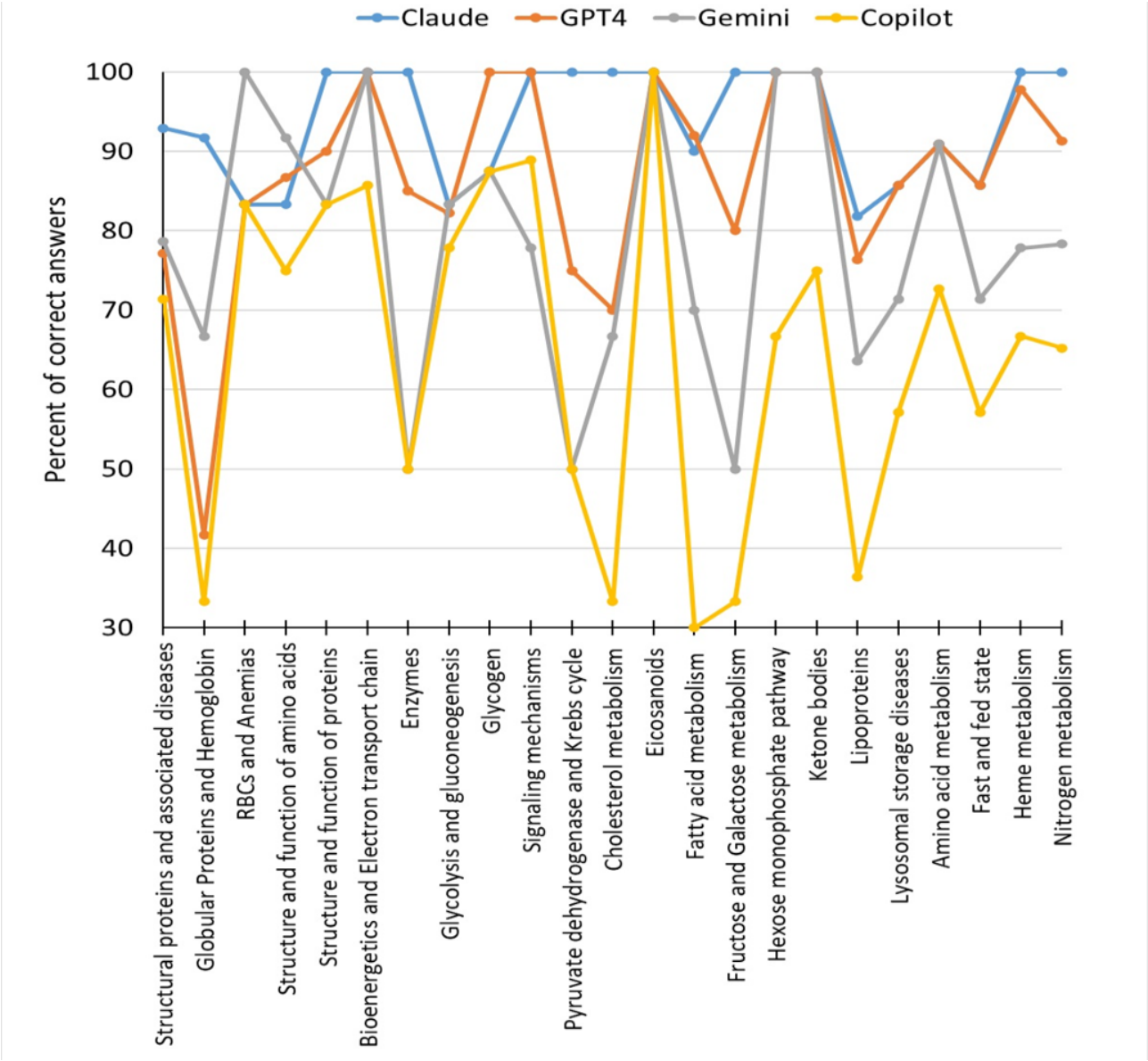


Figure 2. Evaluation of chatbot performance in different topics of the medical biochemistry course. RBC: red blood cell.



Claude

Claude, offered by Anthropic, provided 92.5% (185/200) correct answers to the set of biochemistry MCQs. The answers in this second and all subsequent attempts were identical to the first. As the chatbot claims, its knowledge base has not changed between attempts, and it applies the same reasoning to answer each question. It was the best result among the 5 chatbots, 19.7% better than the student average and 70.5% superior to random guessing. The item analysis suggested that Claude correctly answered all questions (100%) from the following 12 categories: structure and function of proteins, bioenergetics and electron transport chain, enzymes, signaling mechanisms, pyruvate dehydrogenase and Krebs cycle, cholesterol metabolism, eicosanoids, fructose and galactose metabolism, hexose monophosphate pathway, ketone bodies, heme metabolism, and nitrogen metabolism. The lowest result (81.8%) was recorded for the lipoproteins. For the rest of the topics, the percentile of correct answers was 83.3% - 91.7%.

Claude did not solve 15 (7.5%) out of 200 MCQs from the entire questionnaire set. These were comprehensive questions about RBCs, hemoglobin, enzymes, biotin deficiency, and lipoproteins.

GPT-4

The results of 5 successive ChatGPT-4 (OpenAI) attempts to answer the set of 200 biochemistry MCQs showed 85.1% (SD 1%) correct answers on average. The best result of its 5 attempts was 86.5%, 13.7% better than the average for medical students and 64.5% above random guessing. The fourth attempt was the most successful; the mean results of the other 4 attempts were close to 85% (range 84% - 85.5%). The coincidence generated by GPT-4 answers with the previous attempts was 91.5% - 94.5%, and the coincidence of correct answers among them was in the range 81% - 83.6%.

Of the 200 questions, 158 (79%) were answered correctly across all 5 attempts and considered a solid knowledge area for GPT-4. Most of these MCQs were recall questions, but some were complex and required critical thinking. The item analysis indicated that the best 6 categories with 100% correct answers

were bioenergetics and electron transport chain, glycogen, signaling mechanisms, eicosanoids, hexose monophosphate pathway, and ketone bodies. The lowest result was recorded for globular proteins and hemoglobin questions—only 41.7% of the correct answers. For the rest of the topics, the percentile of correct answers was 77.1% - 97.8%.

GPT-4 did not answer 17 (8.5%) of the 200 MCQs from the entire questionnaire set in any 1 out of all 5 attempts. These were more comprehensive questions about defective proteins, oxygen saturation, anemia, amino acids, glycogen, glycolysis and gluconeogenesis, and lipoproteins.

Gemini

Google recently introduced Gemini as a successor to Bard. The results of 5 attempts by Gemini to answer the set of 200 biochemistry MCQs showed 157 (78.5%) correct answers, 5.7% above the average for medical students and 56.5% above the random answers. Unlike Bard, 5 successive attempts from Gemini were similar; the same answers were received.

The item analysis of these 157 correct answers shows that Gemini did the best (100% accurate) for questions in the following 5 categories: RBCs and anemia, bioenergetics and electron transport chain, eicosanoids, hexose monophosphate pathway, and ketone bodies. Most of these MCQs were recall questions. The lowest 50% results were recorded for the following 3 categories: enzymes, pyruvate dehydrogenase and Krebs cycle, and fructose and galactose metabolism. Gemini's responses in other topics were in the 63.6% - 91.7% interval. Gemini did not answer 43 (21.5%) of the 200 MCQs from the entire questionnaire set, which were comprehensive questions mostly about proteins, enzymes, the Krebs cycle, fatty acids, fructose, and galactose metabolism.

Copilot

Microsoft's Copilot can accept only up to 2000 characters in the prompt, so only 2 to 7 MCQs can be answered at a time, which is inconvenient to work with. The results received on the first try were not different from 4 successive attempts, so there was zero variation among all 5 attempts. Copilot generated 128 (64%) accurate answers for the same set of 200 MCQs from the biochemistry course, 8.8% lower than the average medical student but 42% better than random guessing.

The item analysis of these 126 correct answers indicated that these MCQs were mostly recall questions. The best result was shown for the eicosanoids category (100%), and the lowest was for fatty acid metabolism (only 30% of correct answers). Copilot's responses in other topics vary from 33.3% to 88.9%. Copilot did not answer 72 (36%) of the 200 MCQs from the questionnaire set. These questions concerned proteins, hemoglobin, amino acids, enzymes, fatty acids, pyruvate dehydrogenase, Krebs cycle, and fast and fed state.

Pearson Chi-Square Test Results

Table 1 shows the results of the Pearson chi-square test, which we used due to the binary nature of the data to compare the performance of the different AI-driven chatbots against each other.

The null hypothesis was rejected because the *P* value for all chatbots was less than α (*P*=.05), so there is a statistically significant association between the answers of all 4 chatbots.

Table . Pearson chi-square test results to compare the performance of Claude, GPT-4, Gemini, and Copilot against each other.

Large language models	Chi-square (*df*)	*P* value
Claude × GPT-4	19.7 (1)	<.001
Claude × Gemini	6.1 (1)	.01
Claude × Copilot	4.1 (1)	.04
GPT-4 × Gemini	33.1 (1)	<.001
GPT-4 × Copilot	15.9 (1)	<.001
Gemini × Copilot	23.5 (1)	<.001

Discussion

Principal Findings

Medical education is rapidly evolving, with AI playing an increasingly significant role. In this context, evaluating AI efficacy and relevancy to results is crucial, particularly given the precision and depth of understanding required in medical practice. AI-driven LLMs such as ChatGPT, Claude, Copilot, and Gemini have been compared against medical students in various studies, revealing both the strengths and limitations of AI in medical education. These comparisons show how AI can enhance human learning while also highlighting areas where it may not measure up. Research into AI's role in medical training has uncovered intriguing possibilities and important constraints [1,5,7].

MCQs form a cornerstone of assessment in medical education. Analyzing these questions is vital as it allows educators to assess their effectiveness in testing higher-order thinking and clinical reasoning skills, ensuring that assessments accurately reflect the competencies required for medical practice [18]. While LLMs have demonstrated impressive capabilities in answering queries and simulating scenarios, the depth and breadth of their understanding, particularly concerning MCQs in medical exams, still requires thorough evaluation [19].

The comparative analysis of LLMs and medical students in biochemistry assessment reveals several intriguing patterns that both confirm and challenge our initial hypotheses. While we anticipated comparable performance between AI models and medical students, the results demonstrated that LLMs not only matched but significantly exceeded student performance, with

an 8.3% higher average score (P=.02) across 200 medical biochemistry questions. This finding particularly supports our hypothesis regarding factual recall and concept application, though with a more pronounced advantage for AI systems than initially predicted. The observed variation in performance among different LLM platforms—ranging from Claude's exceptional 92.5% (185/200) accuracy to Copilot's more modest 64% (128/200)—aligns with our hypothesis about performance differences between AI models, suggesting that architectural and training differences significantly impact their capabilities in specialized medical knowledge domains.

Comparison to Literature

Recent studies have shown that LLMs, specifically GPT-4, often outperform medical students on MCQ items in board and licensing exams. This finding underscores the significance of MCQs in medical licensing exams, extensively used in crucial assessments worldwide. Examples include the Peruvian National Licensing Medical Examination, the United States Medical Licensing Examination (USMLE), the United Kingdom Medical Licensing Assessment (UKMLA), and the Australian Medical Council (AMC) Exam [20,24-26]. The widespread use of MCQs is attributed to their effectiveness in evaluating higher-order skills through complex clinical scenarios, analysis, and problem-solving. These questions assess students' ability to integrate information, reflecting real-world challenges and shaping competent professionals. It is well correlated with the results of our study, which have shown that the selected 4 chatbots answered correctly to 81.1% (SD 12%) of the 200 questions from the medical biochemistry course, which is 8.3% above the students' average.

Another comprehensive study compared the results of 4 LLMs across 163 questions from sample NBME (National Board of Medical Examiners) clinical subject exams. The results were striking: GPT-4 achieved a perfect score of 100% (163/163), significantly outperforming GPT-3.5, Claude, and Bard. GPT-3.5 scored 82.2% (134/163), Claude 84.7% (138/163), and Bard 75.5% (123/163). The statistical superiority of GPT-4 was evident, with no significant differences observed among the other 3 models [27]. Interestingly, while GPT-4 excelled across all subject exams, the different models demonstrated variable strengths. GPT-3.5 performed best in family medicine and obstetrics and gynecology, Claude in surgery, and Bard in surgery and neurology. The surgery exam yielded the highest average score across all models, while family medicine had the lowest. GPT-4's exceptional performance may be attributed to its extensive training data, which exceeded 45 terabytes by September 2021, despite not being specifically fine-tuned for medical data [10].

Our data contradict this clinical study and suggest that GPT-4 did well with 85% (170/200) of correct answers but is not currently the most proficient chatbot for biochemistry questions. The best result was recorded for Claude, with an impressive 92.5% (185/200) of the correct answers. Gemini took third place with 78.5% (157/200) of correct answers, which is still above the student's average of 72.8% (SD 5.2%) for the same questions. The lowest result was recorded for Copilot (128/200, 64%).

These findings highlight the potential of LLMs in medical education and practice. Their ability to tackle complex medical questions opens doors to innovative clinical decision support, research, and education applications. However, it is worth noting that GPT-4, the only LLM in this study not available for free, could be less accessible to a broad range of students, potentially limiting its widespread use in educational settings.

Several studies have evaluated ChatGPT's performance in biochemistry. One study examined GPT-3.5's potential as a self-study adjunct for medical students in biochemistry, using 200 questions. ChatGPT provided correct answers to 58% (116/200) of the biochemistry questions. While this performance allowed it to pass the university's medical biochemistry exam, the study suggests there is room for improvement in GPT-3.5 as a comprehensive and reliable self-learning tool [28].

Another study focused on ChatGPT's ability to address higher-order questions in medical biochemistry. Using GPT-3.5, researchers conducted a web-based cross-sectional study presenting 200 randomly selected, complex reasoning questions from an institutional question bank, classified according to CBME (Competency-Based Medical Education) curriculum modules. Two expert biochemistry academicians evaluated responses on a 0 - 5 scale. The AI achieved a median score of 4 (IQR 3.5-4.5), which was comparable to a hypothetical value of 4 (P=.16) but significantly lower than the maximum of 5 (P=.001). These results suggest that GPT-3.5 shows promise as an effective tool for addressing complex questions in medical biochemistry, demonstrating its potential in handling higher-order thinking tasks in this field [29].

Our research confirms that GPT-4 has significant improvements and is superior to GPT-3.5. Our data suggest that GPT-4 responded correctly to 84% - 86.5% of MCQs, and 79% answered correctly across all 5 attempts.

Implications of Findings

The implications of AI's performance in medical education extend beyond mere test-taking abilities. LLMs can answer complex medical questions that raise important questions about the future of medical education and topics in which LLMs demonstrate proficiency, so they may be used to assist students. The detailed analysis of MCQs in our study revealed that questions from 4 topics are well answered by all chatbots: eicosanoids, bioenergetics, electron transport chain, and ketone bodies. In contrast, the lowest results were recorded for globular proteins and hemoglobin, lipoproteins, and fructose and galactose metabolism questions. However, there was a significant difference in the 4 LLMs performances. Claude showed the most impressive results and answered all questions (100%) from 12 categories: structure and function of proteins, bioenergetics and electron transport chain, enzymes, signaling mechanisms, pyruvate dehydrogenase and Krebs cycle, cholesterol metabolism, eicosanoids, fructose and galactose metabolism, hexose monophosphate pathway, ketone bodies, heme, and nitrogen metabolism.

In conclusion, the rapid advancements in AI technology, particularly in medical education, present opportunities and challenges. While LLMs have shown impressive capabilities

in answering medical exam questions, it is crucial to remember that medical education encompasses more than just knowledge acquisition. Clinical skills, empathy, ethical decision-making, and the ability to navigate complex health care systems are all integral parts of medical training that current AI models may not fully capture.

As we progress, we must continuously evaluate AI's role in medical education, ensuring that it complements rather than replaces human expertise. Our findings also have important implications for assessment strategies in medical education. The ability of LLMs to surf the net and do better than medical students on MCQ-based evaluation is an assault on the traditional ways of measuring medical performance and calls for a better understanding of how medical knowledge and skills should be assessed. While such results provide ideas on how to develop a curriculum and manage educational resources, they also highlight the need to ensure that the value of AI in measuring certain aspects of medical training, such as clinical reasoning, interaction with patients, and even decision-making ethics, is always respected. This underscores the need for medical education to continue emphasizing the development of comprehensive clinical skills beyond what can be measured through standardized testing.

Future Directions

Future research in this field should pursue several key routes to better understand and implement AI technologies in medical education. Long-term studies are needed to evaluate the impact of LLM integration on student learning outcomes, particularly focusing on how AI-assisted learning affects knowledge retention, clinical reasoning development, and overall academic performance. These studies should incorporate diverse assessment methods beyond MCQs, including case-based scenarios, open-ended questions, and practical clinical applications of biochemistry knowledge across different medical disciplines to understand whether the observed performance patterns are consistent.

Strengths and Limitations

This study represents one of the first comprehensive comparisons between multiple leading LLMs and medical students in the specific context of medical biochemistry education. The large sample size of 200 questions provided a robust dataset for analysis, covering a broad spectrum of biochemistry topics typically encountered in medical education. The inclusion of multiple LLM platforms (GPT-4, Claude, Copilot, and Gemini) allowed for a nuanced comparison of AI capabilities across different models, providing valuable insights into their relative strengths and potential applications in medical education.

Several limitations should be considered when interpreting these results. This study's findings on different chatbot proficiencies are limited to MCQs from the biochemistry course, which may not represent other medical questions or contexts. In addition, the sample size of 200 questions, excluding questions with images or tables, may not capture the full range of difficulty levels or content areas.

LLMs receive regular updates, which result from training on inputs and tuning so that they may provide different answers depending on the testing date. Another limitation is that GPT-4, which performed well, is not freely available, potentially limiting its applicability in widespread educational settings.

Conclusions

LLMs such as ChatGPT, Claude, Copilot, and Gemini have impressive capabilities in answering MCQs, often outperforming medical students. In this study, the selected chatbots outperformed students' results. These findings highlight the potential of AI in medical education and assessment. Different LLMs exhibit varying strengths in different topics of medical biochemistry courses. In this study, Claude showed the best performance, followed by GPT-4, Gemini, and Copilot. This variability suggests that different AI models may have unique strengths in specific medical fields, which could be leveraged for targeted educational support. The strong performance of LLMs in answering complex medical questions raises important considerations for the future of medical education. While AI demonstrates proficiency in knowledge-based assessments, it is crucial to remember that medical training encompasses more than just information recall. Clinical reasoning, empathy, ethical decision-making, and navigating health care systems remain essential components that current AI models may need to capture fully.

Acknowledgments

The authors thank the Dean of the College of Medicine at Alfaisal University, Prof Khaled Al-Kattan, and the Head of the Department of Anatomy and Genetics, Prof Paul Ganguly, for their support in this research.

Conflicts of Interest

None declared.

References

1. Liu PR, Lu L, Zhang JY, Huo TT, Liu SX, Ye ZW. Application of artificial intelligence in medicine: an overview. Curr Med Sci 2021 Dec;41(6):1105-1115. [doi: 10.1007/s11596-021-2474-3] [Medline: 34874486]
2. Garcia-Vidal C, Sanjuan G, Puerta-Alcalde P, Moreno-García E, Soriano A. Artificial intelligence to support clinical decision-making processes. EBioMedicine 2019 Aug;46(27-29):27-29. [doi: 10.1016/j.ebiom.2019.07.019] [Medline: 31303500]

3. Ellahham S. Artificial intelligence: the future for diabetes care. Am J Med 2020 Aug;133(8):895-900. [doi: 10.1016/j.amjmed.2020.03.033] [Medline: 32325045]
4. Singhal K, Azizi S, Tu T, et al. Publisher correction: large language models encode clinical knowledge. Nature New Biol 2023 Aug;620(7973):E19. [doi: 10.1038/s41586-023-06455-0] [Medline: 37500979]
5. Meo SA, Al-Masri AA, Alotaibi M, Meo MZS, Meo MOS. ChatGPT knowledge evaluation in basic and clinical medical sciences: multiple choice question examination-based performance. Healthcare (Basel) 2023 Jul 17;11(14):2046. [doi: 10.3390/healthcare11142046] [Medline: 37510487]
6. Bolgova O, Shypilova I, Sankova L, Mavrych V. How well did ChatGPT perform in answering questions on different topics in gross anatomy? Eur J Med Health Sci 2023;5(6):94-100. [doi: 10.24018/ejmed.2023.5.6.1989]
7. Roos J, Kasapovic A, Jansen T, Kaczmarczyk R. Artificial intelligence in medical education: comparative analysis of ChatGPT, Bing, and medical students in Germany. JMIR Med Educ 2023 Sep 4;9:e46482. [doi: 10.2196/46482] [Medline: 37665620]
8. Agarwal M, Goswami A, Sharma P. Evaluating ChatGPT-3.5 and Claude-2 in answering and explaining conceptual medical physiology multiple-choice questions. Cureus 2023 Sep;15(9):e46222. [doi: 10.7759/cureus.46222]
9. Welcome to Claude. Anthropic. URL: https://docs.anthropic.com/en/docs/welcome [accessed 2024-09-06]
10. GPT-4 turbo and GPT-4. OpenAI. URL: https://tinyurl.com/ycec959y [accessed 2024-09-06]
11. Gemini models. Google AI. URL: https://ai.google.dev/gemini-api/docs/models/gemini [accessed 2024-09-06]
12. Microsoft 365 Copilot. Microsoft. URL: https://learn.microsoft.com/en-us/office365/servicedescriptions/office-365-platform-service-description/microsoft-365-copilot [accessed 2024-09-06]
13. Mavrych V, Ganguly P, Bolgova O. Using large language models (ChatGPT, Copilot, PaLM, Bard, and Gemini) in gross anatomy course: comparative analysis. Clin Anat 2025 Mar;38(2):200-210. [doi: 10.1002/ca.24244] [Medline: 39573871]
14. Laupichler MC, Rother JF, Grunwald Kadow IC, Ahmadi S, Raupach T. Large language models in medical education: comparing ChatGPT- to human-generated exam questions. Acad Med 2024 May 1;99(5):508-512. [doi: 10.1097/ACM.0000000000005626] [Medline: 38166323]
15. Liu M, Okuhara T, Chang X, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. J Med Internet Res 2024 Jul 25;26:e60807. [doi: 10.2196/60807] [Medline: 39052324]
16. Mavrych V, Bolgova O. Evaluating AI performance in answering questions related to thoracic anatomy. MOJ Anat Physiol 2023;10(1):55-59. [doi: 10.15406/mojap.2023.10.00339]
17. Brin D, Sorin V, Vaid A, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep 2023 Oct 1;13(1):16492. [doi: 10.1038/s41598-023-43436-9] [Medline: 37779171]
18. Bharatha A, Ojeh N, Fazle Rabbi AM, et al. Comparing the performance of ChatGPT-4 and medical students on MCQs at varied levels of Bloom's taxonomy. Adv Med Educ Pract 2024;15(393-400):393-400. [doi: 10.2147/AMEP.S457408] [Medline: 38751805]
19. Goyal M, Agarwal M, Goel A. Interactive learning: online audience response system and multiple choice questions improve student participation in lectures. Cureus 2023;15(7):e42527. [doi: 10.7759/cureus.42527]
20. Torres-Zegarra BC, Rios-Garcia W, Ñaña-Cordova AM, et al. Performance of ChatGPT, Bard, Claude, and Bing on the Peruvian National Licensing Medical Examination: a cross-sectional study. J Educ Eval Health Prof 2023;20:30. [doi: 10.3352/jeehp.2023.20.30] [Medline: 37981579]
21. Gilson A, Safranek CW, Huang T, et al. Correction: How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2024 Feb 27;10:e57594. [doi: 10.2196/57594] [Medline: 38412478]
22. Friederichs H, Friederichs WJ, März M. ChatGPT in medical school: how successful is AI in progress testing? Med Educ Online 2023 Dec;28(1):2220920. [doi: 10.1080/10872981.2023.2220920] [Medline: 37307503]
23. Lin Z. Why and how to embrace AI such as ChatGPT in your academic life. R Soc Open Sci 2023 Aug;10(8):230658. [doi: 10.1098/rsos.230658] [Medline: 37621662]
24. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
25. Lai UH, Wu KS, Hsu TY, Kan JKC. Evaluating the performance of ChatGPT-4 on the United Kingdom Medical Licensing Assessment. Front Med (Lausanne) 2023;10:1240915. [doi: 10.3389/fmed.2023.1240915] [Medline: 37795422]
26. Kleinig O, Gao C, Bacchi S. This too shall pass: the performance of ChatGPT-3.5, ChatGPT-4 and New Bing in an Australian medical licensing examination. Med J Aust 2023 Sep 4;219(5):237. [doi: 10.5694/mja2.52061] [Medline: 37528548]
27. Abbas A, Rehman MS, Rehman SS. Comparing the performance of popular large language models on the National Board of Medical Examiners sample questions. Cureus 2024 Mar;16(3):e55991. [doi: 10.7759/cureus.55991] [Medline: 38606229]
28. Surapaneni KM, Rajajagadeesan A, Goudhaman L, et al. Evaluating ChatGPT as a self-learning tool in medical biochemistry: a performance assessment in undergraduate medical university examination. Biochem Mol Biol Educ 2024;52(2):237-248. [doi: 10.1002/bmb.21808] [Medline: 38112255]

29. Ghosh A, Bir A. Evaluating ChatGPT's ability to solve higher-order questions on the competency-based medical education curriculum in medical biochemistry. Cureus 2023;15(4):e37023. [doi: 10.7759/cureus.37023]

Abbreviations

AI: artificial intelligence
AMC: Australian Medical Council
CBME: Competency-Based Medical Education
LLMs: large language models
MCQ: multiple-choice question
NBME: National Board of Medical Examiners
RBC: red blood cell
UKMLA: United Kingdom Medical Licensing Assessment
USMLE: United States Medical Licensing Examination

Edited by B Lesselroth; submitted 06.10.24; peer-reviewed by M Agarwal, M Malhotra; revised version received 20.01.25; accepted 08.03.25; published 10.04.25.

Please cite as:
Bolgova O, Shypilova I, Mavrych V
Large Language Models in Biochemistry Education: Comparative Evaluation of Performance
JMIR Med Educ 2025;11:e67244
URL: https://mededu.jmir.org/2025/1/e67244
doi:10.2196/67244

Medical Students' Acceptance of Tailored e–Mental Health Apps to Foster Their Mental Health: Cross-Sectional Study

Catharina Grüneberg[1]; Alexander Bäuerle[1,2], PhD; Sophia Karunakaran[1], MSc; Dogus Darici[3], Dr med; Nora Dörrie[1,2], Dr med; Martin Teufel[1,2], Prof Dr med; Sven Benson[2,4], Prof Dr; Anita Robitzsch[1,2], Dr med

[1]Clinic for Psychosomatic Medicine and Psychotherapy, LVR-University Hospital, University of Duisburg-Essen, Virchowstraße 174, Essen, Germany

[2]Center for Translational Neuro- and Behavioral Sciences, University of Duisburg-Essen, Essen, Germany

[3]Institute of Anatomy and Neurobiology, University of Münster, Münster, Germany

[4]Institute for Medical Education, University Hospital Essen, University of Duisburg-Essen, Essen, Germany

Corresponding Author:
Anita Robitzsch, Dr med
Clinic for Psychosomatic Medicine and Psychotherapy, LVR-University Hospital, University of Duisburg-Essen, Virchowstraße 174, Essen, Germany

Abstract

Background: Despite the high prevalence of mental health problems among medical students and physicians, help-seeking remains low. Digital mental health approaches offer beneficial opportunities to increase well-being, for example, via mobile apps.

Objective: This study aimed to assess the acceptance, and its underlying predictors, of tailored e–mental health apps among medical students by focusing on stress management and the promotion of personal skills.

Methods: From November 2022 to July 2023, a cross-sectional study was conducted with 245 medical students at the University of Duisburg-Essen, Germany. Sociodemographic, mental health, and eHealth-related data were assessed. The Unified Theory of Acceptance and Use of Technology (UTAUT) was applied. Differences in acceptance were examined and a multiple hierarchical regression analysis was conducted.

Results: The general acceptance of tailored e–mental health apps among medical students was high (mean 3.72, SD 0.92). Students with a job besides medical school reported higher acceptance ($t_{107.3}$=–2.16; P=.03; P_{adj}=.027; Cohen d=4.13) as well as students with higher loads of anxiety symptoms ($t_{92.4}$=2.36; P=.02; P_{adj}=.03; Cohen d=0.35). The t values were estimated using a 2-tailed t test. Regression analysis revealed that acceptance was significantly predicted by anxiety symptoms (β=.11; P=.045), depressive symptoms (β=–.11; P=.05), internet anxiety (β=–.12; P=.01), digital overload (β=.1; P=.03), and the 3 UTAUT core predictors—performance expectancy (β=.24; P<.001), effort expectancy (β=.26; P<.001), and social influence (β=.43; P<.001).

Conclusions: The high acceptance of e–mental health apps among medical students and its predictors lay a valuable basis for the development and implementation of tailored e–mental health apps within medical education to foster their mental health. More research using validated measures is needed to replicate our findings and to further investigate medical students' specific needs and demands regarding the framework of tailored e–mental health apps.

(JMIR Med Educ 2025;11:e58183) doi:10.2196/58183

KEYWORDS

eHealth; medical education; medical students; tailored interventions; UTAUT; intention to use; e–mental health apps; app; foster; cross-sectional study; mental health problems; physician; well-being; mobile apps; acceptance; assessment; mental health apps

Introduction

Background

Medical students have a heightened incidence of mental health problems, namely anxiety [1,2] and depression [3-5], and are confronted with stressful situations throughout their careers [6,7]. Elevated levels of depression and anxiety among medical students and physicians exert considerable influence on personal well-being and patient safety [8], emphasizing the urgent need for targeted preventive and support programs [7,9-11]. The necessity for assessable and easily accessible interventions to foster mental health and well-being is of utmost importance in the medical student population [12,13].

In recent years, the surge in the significance of digitalization within health care and medical education has been noteworthy [14-16]. This trend has been particularly pronounced during the COVID-19 pandemic and persists afterward [17]. A report disseminated by a German health insurance entity in 2023

scrutinized students' health, with a specific focus on postpandemic developments and the pivotal role of digital education and instruction [18]. The report underscored the critical importance of stress prevention and mental health initiatives [18]. Digital mental health approaches present promising avenues for surmounting barriers and enhancing the use of mental health support, for example, through mobile apps [13,19,20].

Analyzing factors influencing the acceptance of a mobile app is essential, and further research on actual uptake, adoption, and adherence is needed [21-26]. Incorporating future users directly into the development process is crucial for optimizing the adherence of new technologies and should be focused within research [27,28].

Few studies have delved into e–mental health promotion and the prevention of psychological distress among medical students and have shown that uptake of mental health support remains low due to barriers such as mental health stigma or data safety [6,29-31]. To date and to the best of our knowledge, no study has examined the acceptance of tailored e–mental health apps among medical students using a validated model. For this reason, the Unified Theory of Acceptance and Use of Technology (UTAUT) was applied in this study to lay the foundation for the development of an application especially tailored to the students' needs and demands to foster mental health by focusing on stress management and promotion of personal skills at University Duisburg-Essen. The UTAUT evaluates the acceptance of technological systems consisting of 4 primary predictors—performance expectancy (PE), effort expectancy (EE), social influence (SI), and facilitating conditions (FC)—and has been adjusted to investigate the acceptance of eHealth interventions along with their underlying factors [23-25]. Numerous studies have used the UTAUT framework in the context of eHealth interventions among different samples [32-35].

Objectives

Due to the evident progression of digitalization and its concomitant potential to enhance mental health while simultaneously acknowledging the existing impediments to leveraging these opportunities, this study is specifically oriented toward investigating the acceptance of tailored e–mental health apps and their foundational predictors among medical students, using the validated UTAUT model as the analytical framework.

While prior research has underscored the significance of promoting mental health among medical students [7,9,12,31], limited attention has been given to evaluate e–mental health approaches focusing on the promotion and the prevention of psychological distress among medical students using validated measures, such as the UTAUT model, and tailored approaches [28,36,37].

This study will address the following research questions: (1) What is the extent of acceptance of e–mental health apps among medical students? (2) Are there differences in acceptance among medical students based on sociodemographic and mental health data? (3) What factors predict acceptance among medical students?

Methods

Study Design and Participants

A cross-sectional study was conducted to assess acceptance and to analyze drivers and barriers of tailored e–mental health apps among medical students. The study was presented to medical students in the 5th year at the Medical Faculty of University Duisburg-Essen, North-Rhine-Westphalia, Germany, during the course of psychosomatic medicine. Following the course, students were given the opportunity to participate voluntarily. The participants of the study were recruited from November 2022 to July 2023. Of the 305 students attending the course, 245 (80.3%) students gave their informed consent to participate in the study. Of the 245 participants, 16 (6.5%) participants were eliminated from the sample because of missing data. In total, 229 (93.5%) students were included in the final data analysis. We applied no inclusion or exclusion criteria. Medical students were invited to participate in the study through direct contact in the context of psychosomatic medicine courses. All participants were aged 18 years or above.

Ethical Considerations

The study was conducted in accordance with the Declaration of Helsinki and has been approved by the ethics committee of the Medical Faculty of the University of Duisburg-Essen (21 - 10196-BO). Participation was anonymous, voluntary, and without any compensation. Prior to the start of the questionnaire, written informed consent was obtained and the students received background information on the purpose of the study.

Assessment Instruments

The survey consisted of a paper-pencil questionnaire with self-developed items. Additionally, validated scales were used. The measures encompassed sociodemographic, eHealth-related, and mental health data. The primary outcome was the acceptance of an e–mental health app by using the conceptual framework of the UTAUT model's theory.

Sociodemographic Data

Sociodemographic data contained age, gender, marital status, employment besides medical school (occupational status), and working hours per week (0 - 5, 5 - 10, 10 - 15, and >15 hours).

Mental Health Data

To obtain mental health data, the validated PHQ-4 (Patient Health Questionnaire-4) measure consisting of two 2-item measures—PHQ-2 (symptoms of depression, Patient Health Questionnaire-2) and GAD-2 (symptoms of general anxiety disorder, Generalized Anxiety Disorder-2)—were used [38,39]. Answers were given on a 4-point Likert scale (0="never" to 3="nearly every day"). A cutoff score of 3 or more is described to be an indicator of depression (PHQ-2) [38] or general anxiety (GAD-2) [40]. Internal consistencies measured by the Cronbach α were sufficient with α=0.82 (95% CI 0.76 - 0.87) for GAD-2 and α=0.81 (95% CI 0.73 - 0.86) for PHQ-2. Self-generated questions were used to assess life quality (0="very low" to 10="very good"), mental health (0="very low" to 10="very good"), physical health (0="very low" to 10="very good"), and

importance of promoting mental well-being (0="not important" to 10="very important") on numerical rating scales.

eHealth-Related Data

eHealth-related data were assessed by measuring digital overload, internet anxiety, and digital competence. Internet anxiety and digital overload were both measured on a 5-point Likert scale (1="strongly disagree" to 5="strongly agree"). Internal consistency measured by the Cronbach α was low to sufficient with α=0.68 (95% CI 0.6 - 0.75) for the digital overload scale and sufficient with α=0.81 (95% CI 0.72 - 0.87) for the internet anxiety scale. These scales were previously published and established [34,35,41,42]. Digital competence was measured with a numerical rating scale (0="low" to 10="high").

Acceptance and UTAUT Predictors

To assess medical students' acceptance of using tailored e–mental health apps, a modified UTAUT questionnaire [24] was applied. The adapted UTAUT model consisted of 14 items and measured items on a 5-point Likert scale (1="strongly disagree" to 5="strongly agree"). Acceptance, operationalized as behavioral intention (BI) to use technology, is forecasted by PE, EE, and SI [25]. PE reflects the individual's belief in the benefits they will derive from using the technology. EE signifies the perceived ease of use. SI gauges the extent to which an individual believes that their relatives or friends would endorse the use of the technology. Four items were used to assess BI and PE. Acceptance, operationalized as BI, represented the dependent variable. Two predictors of acceptance—EE and SI—were measured with 3 items each. Internal consistency (Cronbach α) was excellent for BI (α=0.91, 95% CI 0.89 - 0.93) and PE (α=0.92, 95% CI 0.89 - 0.94), sufficient for SI (α=0.83, 95% CI 0.77 - 0.87), and low to sufficient for EE (α=.67, 95% CI 0.57 - 0.75).

Statistical Analysis

For data and statistical analysis, SPSS Statistics version 26 (IBM Corp) and R through RStudio version 4.3.1 (The R Foundation for Statistical Computing; Posit Software) were used. The raw data were collected from the survey, extracted, and processed. Relevant assumptions and prerequisites were tested prior to any statistical test [43-46]. The level of significance was set at α=0.05 for all tests. To minimize α error inflation for multiple comparisons Bonferroni correction was used and *P* values were adjusted. Sum scores (PHQ-4 scale, PHQ-2 scale, and GAD-2 scale) and mean scores (internet anxiety and digital overload) were computed. Mean scores for the UTAUT model were computed: BI, PE, EE, and SI. Consistent with previous research, acceptance scores, operationalized as BI, were categorized as "low acceptance" from 1 to 2.34, "moderate acceptance" from 2.35 to 3.67, and "high acceptance" from 3.68 to 5 [33,41,47]. Descriptive statistics (percentage and absolute count, mean scores, distributions, and standard deviations) of scales, items, and acceptance categories were performed. Additionally, explorative data analysis was conducted. Internal consistencies such as the Cronbach α and item-total correlation were calculated for scales. The normal distribution of the dependent variable (acceptance) was tested graphically and by the Kolmogorov-Smirnov test. Although violations against normal distribution were detected, parametric tests could be used according to the central limit theorem (n>30) and the robustness of the *t* test and Welch-ANOVA against normal distribution violations [44]. Means of acceptance were compared between groups using the *t* test (occupational status, PHQ-2, and GAD-2) and Welch-ANOVA (gender and marital status). The predictive model of acceptance was tested using multiple hierarchical regression analyses. The following predictors were included stepwise: sociodemographic data, mental health data (PHQ-2 and GAD-2), eHealth-related data, and the UTAUT core predictors (EE, SI, and PE). Linearity could be assumed and was analyzed using a scatter plot of the residuals against fitted values. Multicollinearity was not detected because all values of the variance inflation factor were <5. The normality of residuals could be assumed due to the central limit theorem. Homoscedasticity was proven and analyzed using a scatter plot of the standardized residuals and adjusted predicted values. According to Cohen *d*, effect sizes were reported and interpreted as small (0.2), medium (0.5), and large (0.8) [48].

Results

Study Population

In this sample, participants' age ranged from 20 to 37 years (mean 25.05, SD 2.82 years). Medical students experienced low digital overload (mean 2.85, SD 0.92) and low internet anxiety (mean 1.72, SD 0.79). Digital competence was high among medical students (mean 6.97, SD 1.72; range 0 - 10). For detailed characteristics, see Table 1.

Table . Sociodemographic and mental health data of participants (n=229).

Variable		N (%)	Mean (SD)	Acceptance, n (%)		
				Low[a]	Moderate[b]	High[c]
Gender						
	Woman	157 (68.6)	—[d]	13 (8.3)	42 (26.8)	102 (65)
	Man	70 (30.6)	—	9 (12.9)	21 (30)	40 (57.1)
	Nonbinary	2 (0.9)	—	0 (0)	2 (100)	0 (0)
Marital status (n=228)						
	Single, divorced, or separated	139 (61)	—	14 (10.1)	45 (32.4)	80 (57.6)
	Married or in a relationship	89 (39)	—	8 (9)	20 (22.5)	61 (68.5)
Job						
	Yes	165 (72.1)	—	14 (8.5)	43 (26.1)	108 (65.5)
	No	64 (28)	—	8 (12.5)	22 (34.4)	34 (53.1)
Working hours per week (n=166)						
	0 - 5	37 (22.3)	—	2 (5.4)	9 (24.3)	26 (70.3)
	5 - 10	84 (50.6)	—	8 (3.8)	21 (25)	55 (65.5)
	10 - 15	25 (15.1)	—	2 (8)	7 (28)	16 (64)
	>15	20 (12.1)	—	2 (10)	7 (35)	11 (55)
Mental health[e]		—	6.82 (1.72)	7.4 (2.2)	6.9 (2.3)	6.7 (2.4)
Physical health[e]		—	8.00 (1.84)	8.5 (1.6)	8 (1.8)	7.9 (1.9)
Life quality[e]		—	7.92 (1.67)	8.4 (1.1)	7.9 (1.7)	7.9 (1.7)
Promotion of mental well-being[e]		—	8.68 (1.80)	7.9 (2.3)	7.9 (1.7)	9 (1.7)
PHQ-2[f] score (range 0 - 6)		—	1.26 (1.42)	—	—	—
	Low (≤2)	201 (87.8)	0.84 (0.84)	21 (10.5)	57 (28.4)	123 (61.2)
	High (≥3)	28 (12.2)	4.25 (1.11)	1 (1.7)	8 (28.6)	19 (67.9)
GAD-2[g] score (range 0 - 6)		—	1.85 (1.51)	—	—	—
	Low (≤2)	178 (77.7)	1.19 (0.76)	20 (11.2)	52 (29.2)	106 (59.6)
	High (≥3)	51 (22.3)	4.16 (1.17)	2 (3.9)	13 (25.5)	36 (70.6)

[a]Low acceptance, with scores ranging from 1 to 2.34.

[b]Moderate acceptance, with scores ranging from 2.35 to 3.67.

[c]High acceptance, with scores ranging from 3.68 to 5.

[d]Not applicable.

[e]Higher scores indicate higher levels of mental health, physical health, life quality, or importance of promoting mental well-being (range 0 - 10).

[f]PHQ-2: Patient Health Questionnaire-2.

[g]GAD-2: Generalized Anxiety Disorder-2.

Acceptance of Tailored e–Mental Health Apps

The general acceptance of tailored e–mental health apps among medical students was high (mean 3.72, SD 0.92). Dividing the acceptance categories from low to high, 62% (142/229) participants showed high acceptance (mean 4.31, SD 0.45), 28.4% (65/229) showed moderate acceptance (mean 3.11, SD 0.28), and 9.6% (22/229) showed low acceptance (mean 1.76, SD 0.42).

Between groups, significant differences in acceptance were identified between occupational status ($t_{107.3}$=–2.16; P=.03; P_{adj}=.03; Cohen d=4.13) and GAD-2 groups ($t_{92.4}$=2.36; P=.02; P_{adj}=.03; Cohen d=0.35) using a 2-tailed t test. Students with a job besides medical school reported higher acceptance of tailored e–mental health apps than students without a job. Medical students with high GAD-2 levels (high load of anxiety symptoms) showed higher acceptance than students with low GAD-2 levels (low load of anxiety symptoms). No significant

differences between acceptance were found regarding PHQ-2 groups (low and high), gender (female, male, and divers), and marital status via ANOVA and *t* test (P_{adj}>.5).

Hierarchical Linear Regression Analysis and Predictors of Acceptance

A hierarchical linear regression analysis was conducted to evaluate predictors of acceptance among medical students regarding tailored e–mental health apps.

Sociodemographic data were included in the first step, explaining 3.6% of the variance in acceptance (R^2=0.036; R^2_{adj}=0.022; $F_{3,222}$=2.72; *P*=.045). Occupational status emerged as a significant positive predictor (β=.31; *P*=.03).

In the second step, mental health data were added to the analysis, increasing the explained variance to 6.4% (R^2=0.064; R^2_{adj}=0.042; $F_{5,220}$=2.99; *P*=.01). GAD-2 was identified as a significant predictor (β=.12; *P*=.03) of acceptance.

In the third step, eHealth-related data were added to the model, which further explained 8.2% of the variance in acceptance (R^2=0.082; R^2_{adj}=0.048; $F_{8,217}$=2.14; *P*=.02).

In the fourth and final step, the UTAUT predictors (EE, PE, and SI) were added (overall model), resulting in a comprehensive model that explained 65.8% of the variance in acceptance (R^2=0.658; R^2_{adj}=0.647; $F_{11,214}$=37.47; *P*<.001). The following variables (UTAUT core predictors) showed a significant positive prediction: UTAUT PE (β=.22, *P*<.001), UTAUT EE (β=.32, *P*<.001), and UTAUT SI (β=.44; *P*<.001).

To sum up, within the overall model, the UTAUT predictors, PHQ-2 and GAD-2 sum scores, internet anxiety, and digital overload were associated with the acceptance of tailored e–mental health apps among medical students. For a detailed overview of the hierarchical regression model of acceptance, see Table 2.

Table . Hierarchical regression model of acceptance (the extended Unified Theory of Acceptance and Use of Technology model; n=226).

Predictors	β[a]	β[b]	*t* [c]	R^2 [d]	ΔR^2[e]	*P* value
Intercept	–.22	–.00	–0.46	—[f]	—	.64
Step 1: Sociodemographic data	—	—	—	0.036	0.036	—
Gender	.04	.02	0.53	—	—	.59
Age	.01	.03	0.80	—	—	.43
Occupational status	.16	.08	1.78	—	—	.08
Step 2[g]: Mental health data	—	—	—	0.064	0.028	—
PHQ-2[h], sum score	–.07	–.11	–1.98	—	—	.05
GAD-2[i], sum score	.07	.11	2.02	—	—	.04
Step 3[g]: eHealth-related data	—	—	—	0.082	0.018	—
Digital overload	.10	.10	2.18	—	—	.03
Internet anxiety	–.14	–.12	–2.49	—	—	.01
Digital competence	.01	.03	0.47	—	—	.64
Step 4[g]: UTAUT[j] core predictors	—	—	—	0.658	0.576	—
Social influence	.44	.43	7.57	—	—	<.001
Performance expectancy	.22	.24	4.31	—	—	<.001
Effort expectancy	.32	.26	5.24	—	—	<.001

[a]Unstandardized coefficient beta.

[b]Standardized coefficient beta.

[c]Test statistics were estimated using a 2-tailed *t* test.

[d]Multiple R^2 reported, determination coefficient.

[e]Changes in R^2.

[f]Not applicable.

[g]In steps 2, 3, and 4, only the newly included variables are presented.

[h]PHQ-2: Patient Health Questionnaire-2.

[i]GAD-2: Generalized Anxiety Disorder-2.

[j]UTAUT: Unified Theory of Acceptance and Use of Technology.

Discussion

Principal Findings

This study focused on examining the acceptance of tailored e–mental health apps and the factors influencing their use to promote medical students' mental health.

The general acceptance was high. Students with a job besides medical school reported higher acceptance as well as students with higher loads of anxiety symptoms. Acceptance was significantly predicted by occupational status, anxiety symptoms, depressive symptoms, internet anxiety, digital overload, and the 3 UTAUT core predictors—PE, EE, and SI.

The participants in this sample reported higher overall acceptance compared to previous research involving different target groups [23,33,42]. A qualitative study conducted by Dederichs et al [12] corroborates our findings, elucidating universally positive perspectives among medical students regarding internet- and mobile-based interventions. Preceding investigations have posited that augmented levels of educational attainment are concomitant with elevated acceptance scores [32,49], concurrently accentuating the advantages of e–mental health methodologies, including their low-threshold nature, temporal flexibility, and provision of anonymous support [12].

Among our cohort, self-rated promotion of mental well-being was highly valued, indicating general interest in mental health promotion as an important prerequisite and determinant of increasing acceptance.

The UTAUT core predictors elucidated the majority of the variance in acceptance, substantiating the model's efficacy in appraising e–mental health acceptance among medical students

and aligning with antecedent research [25,33,42]. Despite prior investigations indicating age [25,32,42] and gender [25,49] as salient determinants influencing acceptance within heterogeneous populations, these variables did not achieve statistical significance in this study. This lack of significance may be attributed to the existence of comparable stress factors affecting all participants uniformly.

A notable proportion, 12.2% (28/229), displayed indicators suggestive of depressive symptoms (PHQ-2), while 22.3% (51/229) exhibited symptoms indicative of a general anxiety disorder (GAD-2). These findings are consistent with extant research documenting the psychological vulnerability of medical students, illustrating elevated levels of anxiety and depression [1,6,50]. This underscores the imperative for psychological support interventions [2,3,10]. Our analysis revealed that mental health data concerning anxiety symptoms positively predicted acceptance within our model, aligning with prior research [51]. In contrast to that, depressive symptoms were associated with lower acceptance within our model. The acceptability may be decreased among students with higher depressive symptoms due to fear of additional loads. Furthermore, barriers, such as mental health stigma or data safety, were described as known challenges within previous research focusing on help-seeking behavior [30,36]. Additional information and educative programs or interventions may have beneficial effects to increase help-seeking and decrease stigma [29,52-55], but their impact needs to be investigated further.

Students concurrently managing part-time employment and medical school responsibilities demonstrated higher acceptance scores. Research specifically focusing on the mental health of working medical students is scarce [9,56]. Based on the findings, we would suggest that the additional load due to a part-time job results in higher acceptance levels of mental health support programs but this needs to be investigated further.

A study by Joiner et al [57] found that individuals born after 1993 exhibited lower internet anxiety and higher internet identification, reinforcing our findings. In our sample, most of the participants were born in the 1990s and 2000s. Internet anxiety and digital overload were observed at low levels and significant predictors of acceptance in the overall regression model. Aligning with previous research [23], high levels of internet anxiety were associated with decreased acceptance.

Digital competence was high within our sample. High internet identification and regular use of digital media might have influenced digital competence within our sample. Information on digital skills [58], preventive strategies, and digitalization need to be integrated further within medical education [15].

While acceptance and potential usage constitute crucial prerequisites for the implementation of digital approaches [23,59], it is imperative to acknowledge additional factors, including barriers and risks associated with the promotion of such approaches. Notably, skepticism and a lack of knowledge regarding e–mental health apps among medical students underscore the necessity for augmented information dissemination and increased personal experience with digital health approaches [22,36]. Attention must be directed toward addressing stigma and concerns related to data security [30,36]. Comprehensive assessments of additional barriers influencing actual usage and dropout rates are warranted in the implementation of e–mental health approaches [19,60].

The outcomes of this study establish a foundational framework for subsequent research endeavors and the implementation of e–mental health apps within the realm of medical education. The imperative for further implementation and rigorous evaluation of digital interventions for medical students is underscored.

Limitations

This study has limitations that should be considered when interpreting the presented results. It should be noted that studies assessing medical students' acceptance with validated instruments are still scarce and comparability is limited. The cross-sectional design does not allow causal inferences. Overall, overrepresentation may diminish representativeness, generalizability, and external validity, which is a common bias in research. In the context of a tailored design approach, additional stakeholders should be integrated into future studies [61]. The intention-behavior gap should be considered, as our study assessed theoretical willingness rather than actual usage. Within this study, the Cronbach α, a conservative measure assessing reliability, was used, and it should be noted that the Cronbach α of the EE scale and digital overload scale were lower compared to those observed in previous studies [33,35,41,42]. One possible explanation may be inconsistent response patterns; therefore, the interpretation should be done with caution. According to previous studies [21-28], adherence, actual usage, and dropout rates of e–mental health approaches should be investigated further. While the 3 fundamental predictors of the UTAUT model—EE, PE, and SI—remain crucial, additional factors should be focused on to comprehensively grasp and optimize acceptance levels further.

Conclusions

In this investigation, the focus was on evaluating the acceptance of tailored e–mental health apps and its influencing factors in promoting medical students' mental health. The overall acceptance was found to be high, with students having part-time jobs alongside medical school and students with elevated anxiety levels reporting even higher levels of acceptance. Besides the 3 UTAUT core predictors (PE, EE, and SI), additional significant predictors influence acceptance among medical students including occupational status, anxiety symptoms, depressive symptoms, internet anxiety, and digital overload. As digitalization transforms the medical sector, integrating supportive digital tools into medical education requires a focus on promoting a healthy learning environment and well-being among future physicians. Preventive strategies, including addressing barriers like stigma, are crucial. This study contributes valuable insights in order to develop and implement a digital application to foster medical students' mental health focusing on stress management and promotion of personal skills at Medical University Duisburg-Essen, Germany.

Acknowledgments

We gratefully acknowledge the support of the Open Access Publication Fund at the University of Duisburg-Essen.

Data Availability

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

Authors' Contributions

Conceptualization: AR, MT, AB
Data curation: AR, AB, CG
Formal analysis: CG, SK
Investigation: AR, MT, AB, SB
Methodology: AR, MT, AB, SB
Project administration: AR, AB, SB
Supervision: AR, MT, AB, SB
Writing – original draft: CG
Writing – review & editing: SB, DD, ND

Conflicts of Interest

None declared.

References

1. Quek TTC, Tam WWS, Tran BX, et al. The global prevalence of anxiety among medical students: a meta-analysis. Int J Environ Res Public Health 2019 Jul 31;16(15):2735. [doi: 10.3390/ijerph16152735] [Medline: 31370266]
2. Thapa B, Sapkota S, Khanal A, Aryal BK, Hu Y. Predictors of depression and anxiety among medical students. J Nepal Health Res Counc 2023 Sep 8;21(1):63-70. [doi: 10.33314/jnhrc.v21i1.4514] [Medline: 37742151]
3. Rotenstein LS, Ramos MA, Torre M, et al. Prevalence of depression, depressive symptoms, and suicidal ideation among medical students: a systematic review and meta-analysis. JAMA 2016 Dec 6;316(21):2214-2236. [doi: 10.1001/jama.2016.17324] [Medline: 27923088]
4. Santabárbara J, Olaya B, Bueno-Notivol J, et al. Prevalence of depression among medical students during the COVID-19 pandemic: a systematic review and meta-analysis. Rev Med Chil 2021 Nov;149(11):1579-1588. [doi: 10.4067/S0034-98872021001101579] [Medline: 35735320]
5. Cohen AM, Braun K, Hübner N, Scherner PV, Jurkat HB. Influencing factors on stress management in medical students—with special consideration of depression [In German]. Nervenarzt 2022 May;93(5):468-475. [doi: 10.1007/s00115-021-01183-0] [Medline: 34487197]
6. de Sá e Camargo ML, Torres RV, Cotta KCG, da Silva Ezequiel O, Lucchetti G, Lucchetti ALG. Mental health throughout the medical career: a comparison of depression, anxiety, and stress levels among medical students, residents, and physicians. Int J Soc Psychiatry 2023 Aug;69(5):1260-1267. [doi: 10.1177/00207640231157258]
7. Voltmer E, Köslich-Strumann S, Voltmer JB, Kötter T. Stress and behavior patterns throughout medical education—a six year longitudinal study. BMC Med Educ 2021 Aug 28;21(1):454. [doi: 10.1186/s12909-021-02862-x] [Medline: 34454487]
8. Melnyk BM, Kelly SA, Stephens J, et al. Interventions to improve mental health, well-being, physical health, and lifestyle behaviors in physicians and nurses: a systematic review. Am J Health Promot 2020 Nov;34(8):929-941. [doi: 10.1177/0890117120920451] [Medline: 32338522]
9. Wege N, Muth T, Li J, Angerer P. Mental health among currently enrolled medical students in Germany. Pub Health (Fairfax) 2016 Mar;132:92-100. [doi: 10.1016/j.puhe.2015.12.014] [Medline: 26880490]
10. Pelzer A, Sapalidis A, Rabkow N, Pukas L, Günther N, Watzke S. Does medical school cause depression or do medical students already begin their studies depressed? A longitudinal study over the first semester about depression and influencing factors. GMS J Med Educ 2022;39(5):Doc58. [doi: 10.3205/zma001579] [Medline: 36540560]
11. Krümmel A, Laiker I, Wrona KJ, Aschentrup L, Dockweiler C. Acceptance and use of digital interventions for distress prevention among students: results from a qualitative interview study using the Unified Theory of Acceptance and Use of Technology 2 (UTAUT2) [In German]. Präv Gesundhförder 2023;18(4):508-516. [doi: 10.1007/s11553-022-00985-7]
12. Dederichs M, Weber J, Pischke CR, Angerer P, Apolinário-Hagen J. Exploring medical students' views on digital mental health interventions: a qualitative study. Internet Interv 2021 Sep;25:100398. [doi: 10.1016/j.invent.2021.100398] [Medline: 34026567]
13. D'Adamo L, Paraboschi L, Grammer AC, et al. Reach and uptake of digital mental health interventions based on cognitive-behavioral therapy for college students: a systematic review. J Behav Cogn Ther 2023 Jun;33(2):97-117. [doi: 10.1016/j.jbct.2023.05.002] [Medline: 37724304]

14. Apolinário-Hagen J. Current perspectives on e-mental-health self-help treatments: exploring the “black box” of public views, perceptions, and attitudes toward the digitalization of mental health care. In: Menvielle L, Audrain-Pontevia AF, Menvielle W, editors. The Digitization of Healthcare: New Challenges and Opportunities: Palgrave Macmillan; 2017:205-223. [doi: 10.1057/978-1-349-95173-4_12]
15. Kuhn S, Frankenhauser S, Tolks D. Digital learning and teaching in medical education: already there or still at the beginning? Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2018 Feb;61(2):201-209. [doi: 10.1007/s00103-017-2673-z] [Medline: 29234823]
16. Aulenkamp J, Mikuteit M, Löffler T, Schmidt J. Overview of digital health teaching courses in medical education in Germany in 2020. GMS J Med Educ 2021;38(4):Doc80. [doi: 10.3205/zma001476] [Medline: 34056069]
17. Smith KA, Blease C, Faurholt-Jepsen M, et al. Digital mental health: challenges and next steps. BMJ Ment Health 2023 Feb;26(1):e300670. [doi: 10.1136/bmjment-2023-300670]
18. Gesundheitsreport 2023 – wie geht’s Deutschlands Studierenden? Techniker Krankenkasse. 2023. URL: https://www.tk.de/resource/blob/2149886/e5bb2564c786aedb3979588fe64a8f39/2023-tk-gesundheitsreport-data.pdf [accessed 2025-01-08]
19. Paganin G, Apolinário-Hagen J, Simbula S. Introducing mobile apps to promote the well-being of German and Italian university students: a cross-national application of the technology acceptance model. Curr Psychol 2022 Oct 27:1-12. [doi: 10.1007/s12144-022-03856-8] [Medline: 36320558]
20. Harrer M, Apolinário-Hagen J, Fritsche L, et al. Effect of an internet- and app-based stress intervention compared to online psychoeducation in university students with depressive symptoms: results of a randomized controlled trial. Internet Interv 2021 Apr;24:100374. [doi: 10.1016/j.invent.2021.100374] [Medline: 33718001]
21. Apolinário-Hagen J, Hennemann S, Fritsche L, Drüge M, Breil B. Determinant factors of public acceptance of stress management apps: survey study. JMIR Ment Health 2019 Nov 7;6(11):e15373. [doi: 10.2196/15373] [Medline: 31697243]
22. Apolinário-Hagen J, Hennemann S, Kück C, et al. Exploring user-related drivers of the early acceptance of certified digital stress prevention programs in Germany. Health Serv Insights 2020;13:1178632920911061. [doi: 10.1177/1178632920911061] [Medline: 32206013]
23. Philippi P, Baumeister H, Apolinário-Hagen J, et al. Acceptance towards digital health interventions: model validation and further development of the Unified Theory of Acceptance and Use of Technology. Internet Interv 2021 Dec;26:100459. [doi: 10.1016/j.invent.2021.100459] [Medline: 34603973]
24. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep;13(3):319-340. [doi: 10.2307/249008]
25. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q 2003;27(3):425-478. [doi: 10.2307/30036540]
26. Dwivedi YK, Rana NP, Tamilmani K, Raman R. A meta-analysis based modified Unified Theory of Acceptance and Use of Technology (meta-UTAUT): a review of emerging literature. Curr Opin Psychol 2020 Dec;36:13-18. [doi: 10.1016/j.copsyc.2020.03.008] [Medline: 32339928]
27. Oti O, Pitt I. Online mental health interventions designed for students in higher education: a user-centered perspective. Internet Interv 2021 Dec;26:100468. [doi: 10.1016/j.invent.2021.100468] [Medline: 34703772]
28. Dederichs M, Nitsch FJ, Apolinário-Hagen J. Piloting an innovative concept of e-mental health and mHealth workshops with medical students using a participatory co-design approach and app prototyping: case study. JMIR Med Educ 2022 Jan 10;8(1):e32017. [doi: 10.2196/32017] [Medline: 35006085]
29. Chew-Graham CA, Rogers A, Yassin N. “I wouldn’t want it on my CV or their records”: medical students’ experiences of help-seeking for mental health problems. Med Educ 2003 Oct;37(10):873-880. [doi: 10.1046/j.1365-2923.2003.01627.x] [Medline: 12974841]
30. Berliant M, Rahman N, Mattice C, Bhatt C, Haykal KA. Barriers faced by medical students in seeking mental healthcare: a scoping review. MedEdPublish 2016;12:70. [doi: 10.12688/mep.19115.1]
31. Michaeli D, Keough G, Perez-Dominguez F, et al. Medical education and mental health during COVID-19: a survey across 9 countries. Int J Med Educ 2022 Feb 26;13:35-46. [doi: 10.5116/ijme.6209.10d6] [Medline: 35226614]
32. Hennemann S, Beutel ME, Zwerenz R. Drivers and barriers to acceptance of web-based aftercare of patients in inpatient routine care: a cross-sectional survey. J Med Internet Res 2016 Dec 23;18(12):e337. [doi: 10.2196/jmir.6003] [Medline: 28011445]
33. Damerau M, Teufel M, Musche V, et al. Determining acceptance of e-mental health interventions in digital psychodiabetology using a quantitative web-based survey: cross-sectional study. JMIR Form Res 2021 Jul 30;5(7):e27436. [doi: 10.2196/27436] [Medline: 34328429]
34. Bäuerle A, Frewer AL, Rentrop V, et al. Determinants of acceptance of weight management applications in overweight and obese individuals: using an extended Unified Theory of Acceptance and Use of Technology model. Nutrients 2022 May 8;14(9):1968. [doi: 10.3390/nu14091968] [Medline: 35565935]
35. Bäuerle A, Mallien C, Rassaf T, et al. Determining the acceptance of digital cardiac rehabilitation and its influencing factors among patients affected by cardiac diseases. J Cardiovasc Dev Dis 2023 Apr 17;10(4):174. [doi: 10.3390/jcdd10040174] [Medline: 37103053]

36. Mayer G, Gronewold N, Alvarez S, Bruns B, Hilbel T, Schultz JH. Acceptance and expectations of medical experts, students, and patients toward electronic mental health apps: cross-sectional quantitative and qualitative survey study. JMIR Ment Health 2019 Nov 25;6(11):e14018. [doi: 10.2196/14018] [Medline: 31763990]
37. Ungar P, Schindler AK, Polujanski S, Rotthoff T. Online programs to strengthen the mental health of medical students: a systematic review of the literature. Med Educ Online 2022 Dec;27(1):2082909. [doi: 10.1080/10872981.2022.2082909] [Medline: 35642839]
38. Kroenke K, Spitzer RL, Williams JBW. The Patient Health Questionnaire-2: validity of a two-item depression screener. Med Care 2003 Nov;41(11):1284-1292. [doi: 10.1097/01.MLR.0000093487.78664.3C] [Medline: 14583691]
39. Kroenke K, Spitzer RL, Williams JBW, Monahan PO, Löwe B. Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. Ann Intern Med 2007 Mar 6;146(5):317-325. [doi: 10.7326/0003-4819-146-5-200703060-00004] [Medline: 17339617]
40. Plummer F, Manea L, Trepel D, McMillan D. Screening for anxiety disorders with the GAD-7 and GAD-2: a systematic review and diagnostic metaanalysis. Gen Hosp Psychiatry 2016;39:24-31. [doi: 10.1016/j.genhosppsych.2015.11.005] [Medline: 26719105]
41. Stoppok P, Teufel M, Jahre L, et al. Determining the influencing factors on acceptance of eHealth pain management interventions among patients with chronic pain using the Unified Theory of Acceptance and Use of Technology: cross-sectional study. JMIR Form Res 2022 Aug 17;6(8):e37682. [doi: 10.2196/37682] [Medline: 35976199]
42. Rentrop V, Damerau M, Schweda A, et al. Predicting acceptance of e–mental health interventions in patients with obesity by using an extended unified theory of acceptance model: cross-sectional study. JMIR Form Res 2022 Mar 17;6(3):e31229. [doi: 10.2196/31229] [Medline: 35297769]
43. Pfeifer A. Datenanalyse Mit SPSS Für Windows [Book in German]: De Gruyter; 1996.
44. Gollwitzer M, Eid M, Schmitt M. Statistik und Forschungsmethoden [Book in German]: Beltz Verlagsgruppe; 2017.
45. Sedlmeier P, Burkhardt M. Datenanalyse mit R: Beschreiben, Explorieren, Schätzen und Testen [Book in German]: Pearson Deutschland; 2021.
46. Wollschläger D. Grundlagen der Datenanalyse mit R : Eine anwendungsorientierte Einführung. In: Ruhmann I, editor. Eine Anwendungsorientierte Einführung: Springer Spektrum; 2020. [doi: 10.1007/978-3-662-53670-4]
47. Esber A, Teufel M, Jahre L, In der Schmitten J, Skoda EM, Bäuerle A. Predictors of patients' acceptance of video consultation in general practice during the coronavirus disease 2019 pandemic applying the Unified Theory of Acceptance and Use of Technology model. D Health 2023;9:20552076221149317. [doi: 10.1177/20552076221149317] [Medline: 36815005]
48. Cohen J. Statistical Power Analysis for the Behavioral Sciences: Routledge; 1977. [doi: 10.4324/9780203771587]
49. Crisp DA, Griffiths KM. Participating in online mental health interventions: who is most likely to sign up and why? Depress Res Treat 2014;2014:790457. [doi: 10.1155/2014/790457] [Medline: 24804089]
50. Nikolic A, Bukurov B, Kocic I, et al. Smartphone addiction, sleep quality, depression, anxiety, and stress among medical students. Front Public Health 2023;11:1252371. [doi: 10.3389/fpubh.2023.1252371] [Medline: 37744504]
51. Lin J, Faust B, Ebert DD, Krämer L, Baumeister H. A web-based acceptance-facilitating intervention for identifying patients' acceptance, uptake, and adherence of internet- and mobile-based pain interventions: randomized controlled trial. J Med Internet Res 2018 Aug 21;20(8):e244. [doi: 10.2196/jmir.9925] [Medline: 30131313]
52. Wimsatt LA, Schwenk TL, Sen A. Predictors of depression stigma in medical students: potential targets for prevention and education. Am J Prev Med 2015 Nov;49(5):703-714. [doi: 10.1016/j.amepre.2015.03.021] [Medline: 26141915]
53. Apolinário-Hagen J, Fritsche L, Bierhals C, Salewski C. Improving attitudes toward e-mental health services in the general population via psychoeducational information material: a randomized controlled trial. Internet Interv 2018 Jun;12:141-149. [doi: 10.1016/j.invent.2017.12.002] [Medline: 30135778]
54. Kirschner B, Goetzl M, Curtin L. Mental health stigma among college students: test of an interactive online intervention. J Am Coll Health 2022;70(6):1831-1838. [doi: 10.1080/07448481.2020.1826492] [Medline: 33048656]
55. Nazari A, Garmaroudi G, Foroushani AR, Hosseinnia M. The effect of web-based educational interventions on mental health literacy, stigma and help-seeking intentions/attitudes in young people: systematic review and meta-analysis. BMC Psychiatry 2023 Sep 4;23(1):647. [doi: 10.1186/s12888-023-05143-7] [Medline: 37667229]
56. Shao R, He P, Ling B, et al. Prevalence of depression and anxiety and correlations between depression, anxiety, family functioning, social support and coping styles among Chinese medical students. BMC Psychol 2020 Apr 22;8(1):38. [doi: 10.1186/s40359-020-00402-8] [Medline: 32321593]
57. Joiner R, Gavin J, Brosnan M, et al. Comparing first and second generation digital natives' internet use, internet anxiety, and internet identification. Cyberpsychol Behav Soc Netw 2013 Jul;16(7):549-552. [doi: 10.1089/cyber.2012.0526] [Medline: 23675995]
58. Burzyńska J, Bartosiewicz A, Januszewicz P. Dr. Google: Physicians—the web—patients triangle: Digital skills and attitudes towards e-health solutions among physicians in south eastern Poland—a cross-sectional study in a pre-COVID-19 era. Int J Environ Res Public Health 2023 Jan 5;20(2):978. [doi: 10.3390/ijerph20020978] [Medline: 36673740]
59. Bautista J, Schueller SM. Understanding the adoption and use of digital mental health apps among college students: secondary analysis of a national survey. JMIR Ment Health 2023 Mar 22;10:e43942. [doi: 10.2196/43942] [Medline: 36947115]

60. Apolinario-Hagen J, Harrer M, Salewski C, Lehr D, Ebert DD. Acceptance and use of e-mental health services among university students: secondary analysis of an experiment. Prav Und Gesundhford 2023;18(2):196-203. [doi: 10.1007/s11553-022-00945-1]
61. Irish M, Kuso S, Simek M, et al. Online prevention programmes for university students: stakeholder perspectives from six European countries. Eur J Public Health 2021 Jul 7;31(31 Suppl 1):i64-i70. [doi: 10.1093/eurpub/ckab040] [Medline: 34240152]

Abbreviations

BI: behavioral intention
EE: effort expectancy
FC: facilitating conditions
GAD-2: Generalized Anxiety Disorder-2
PE: performance expectancy
PHQ-2: Patient Health Questionnaire-2
PHQ-4: Patient Health Questionnaire-4
SI: social influence
UTAUT: Unified Theory of Acceptance and Use of Technology

Edited by B Lesselroth; submitted 08.03.24; peer-reviewed by C Papan, K Koelkebeck, M Marendic; revised version received 13.09.24; accepted 24.09.24; published 24.01.25.

Please cite as:
Grüneberg C, Bäuerle A, Karunakaran S, Darici D, Dörrie N, Teufel M, Benson S, Robitzsch A
Medical Students' Acceptance of Tailored e–Mental Health Apps to Foster Their Mental Health: Cross-Sectional Study
JMIR Med Educ 2025;11:e58183
URL: https://mededu.jmir.org/2025/1/e58183
doi:10.2196/58183

Integration of an Audiovisual Learning Resource in a Podiatric Medical Infectious Disease Course: Multiple Cohort Pilot Study

Garrik Hoyt[1*], BTech; Chandra Shekhar Bakshi[2], PhD; Paramita Basu[1,3*], PhD

[1]Touro University, New York, NY, United States

[2]New York Medical College, New York, NY, United States

[3]New York College of Podiatric Medicine, 53 E 124th St, New York, NY, United States

[*]these authors contributed equally

Corresponding Author:
Paramita Basu, PhD
Touro University, New York, NY, United States

Abstract

Background: Improved long-term learning retention leads to higher exam scores and overall course grades, which is crucial for success in preclinical coursework in any podiatric medicine curriculum. Audiovisual mnemonics, in conjunction with text-based materials and an interactive user interface, have been shown to increase memory retention and higher order thinking.

Objective: This pilot study aims to evaluate the effectiveness of integrating web-based multimedia learning resources for improving student engagement and increasing learning retention.

Methods: A quasi-experimental study was conducted with 2 cohorts totaling 158 second-year podiatric medical students. The treatment group had access to Picmonic's audiovisual resources, while the control group followed traditional instruction methods. Exam scores, final course grades, and user interactions with Picmonic were analyzed. Logistic regression and correlation analyses were conducted to examine the relationships between Picmonic access, performance outcomes, and student engagement.

Results: The treatment group (n=91) had significantly higher average exam scores (P<.001) and final course grades (P<.001) than the control group (n=67). Effect size for the average final grades (d=0.96) indicated the practical significance of these differences. Logistic regression analysis revealed a positive association between Picmonic access with an odds ratio of 2.72 with a 95% confidence interval, indicating that it is positively associated with the likelihood of achieving high final grades. Correlation analysis revealed a positive relationship (r=0.25, P=.02) between the number of in-video questions answered and students' final grades. Survey responses reflected increased student engagement, comprehension, and higher user satisfaction (3.71 out of 5 average rating) with the multimedia-based resources compared to traditional instructional resources.

Conclusions: This pilot study underscores the positive impact of animation-supported web-based instruction on preclinical medical education. The treatment group, equipped with Picmonic, exhibited improved learning outcomes, enhanced engagement, and high satisfaction. These results contribute to the discourse on innovative educational methods and highlight the potential of multimedia-based learning resources to enrich medical curricula. Despite certain limitations, this research suggests that animation-supported audiovisual instruction offers a valuable avenue for enhancing student learning experiences in medical education.

(JMIR Med Educ 2025;11:e55206) doi:10.2196/55206

KEYWORDS

learning retention; preclinical education; podiatric medical education; audiovisual learning resources; multimedia-based learning resource; animation-supported learning tools; mnemonics; spaced repetition

Introduction

Long-term retention is crucial for higher exam scores and overall course grades in preclinical coursework. A recent examination of popular board preparatory resources has provided insight into the different trends that students have experienced in their self-directed learning [1]. Incorporating digital resources appeared to be as effective, if not more so, than regular text-based learning [2]. With greater interest shown in digital learning, curiosity has arisen regarding students' sentimental value of animation-styled instruction. One study has shown students' fondness for learning increased with animation instruction as an exciting new way to learn, further increasing permanent learning [3].

Various methods of increasing learning efficiency in medical education have been explored, such as digital recordings, visual mnemonics, and flashcard systems. Many students find the aforementioned methods to be an excellent supplement to the

usual textbook-based learning, resulting in higher test scores, particularly within the medical field [4-7]. Enhanced use of digital teaching tools is effective in providing students with basic science information and has shown to be useful in improving their preparation for clerkship [8].

Mnemonics are a commonly used memory technique in medical school. A mnemonic links to well-known knowledge, sometimes invoking humor or emotions [9]. Web-based learning positively impacts information retention and learning efficiency [10,11]. Audiovisual (AV) mnemonics, in conjunction with text-based materials and an interactive user interface, have been shown to increase memory retention and higher order thinking [12,13].

Picmonic [14], a web-based AV learning resource, uses immersive videos, clinical case questions, flashcards, and high-yield notes, as well as picture mnemonics to cover various aspects of the preclinical and clinical practice curriculum [15-17]. A study by Yang et al [12] observed improved student performance in free-recall and paired-matching tests when using Picmonic. Another study by Abdalla et al [13] underscores how important memory and knowledge retention are to a medical student's grades. Adding AV modalities increases a student's ability to remember information over an extended period of time. Students who had undergone AV sessions had higher marks on response answer questions, shorter time spent answering questions, and a higher memory consolidation after specific time benchmarks. Further studies of using mnemonics, particularly food eponyms in pathology-related education, have shown relevance in learning and retaining pathology knowledge in addition to being useful for United States Medical Licensing Examination boards preparation, clinical clerkship preparation, and future practice [18].

Other studies have examined the usefulness of incorporating AV instructional tools in various levels of education [19-21], including medical education [1,12,22,23], and found them helpful for improving student engagement and learning experiences [12,22,23]. However, there seems to be a dearth of studies exploring the usage of multimedia web-based learning resources in podiatric medical education. Our goal is to evaluate how integrating a multimedia web-based learning resource affects student engagement and learning retention in a preclinical course. Though there are many online learning resources available for medical students to bolster their learning, we selected the commercially available web-based platform Picmonic due to the shorter length of the videos. Since this resource was integrated into the course in the form of low-stakes assignments with the purpose of serving as a supplementary resource, in addition to the textbook and the instructor-provided materials, it was important to ensure that the videos did not take up too much time.

This pilot study aims to offer insight into the use of tools like Picmonic that uses AV media and mnemonics to supplement traditional learning resources in podiatric medical education. To achieve this goal, we have tried to determine in second-year students in a podiatric medicine program (P) if students who have access to Picmonic, an interactive video-based learning system with mnemonics as an additional supplementary resource (I), show higher course performance and experience better learning retention and engagement with the learning material (O), compared to those with access to textbooks and other instructor-provided course materials only (C), when enrolled in a preclinical infectious disease course in their third semester (T). Course performance and knowledge retention were determined by comparing average final exam scores between a treatment group that had access to Picmonic in addition to textbooks and other instructor-provided material and a control group that relied only on the same textbooks and instructor provided materials. Students' perception of the usefulness of this platform as a learning resource and engagement with course materials were assessed using a survey instrument and analysis of correlation between the number of in-video questions answered, the number of times the video was watched, and the accuracy of video-embedded quiz attempts.

Methods

Study Design

A sample of 158 second-year podiatric medical students enrolled in the Infectious Disease course in the third semester at New York College of Podiatric Medicine (NYCPM) were observed in 2 consecutive cohorts. The cohorts consisted of a control group of 67 students taking the course in 2021 and a treatment group of 91 students taking the course in 2022. Participants in the treatment group used the multimedia web-based learning tool Picmonic as a learning resource, while participants in the control group did not. All students were given the same didactic instruction, textbooks, and other traditional learning resources. The study was conducted as a posttest-only, nonequivalent group, quasi-experimental design [24,25]. Although sample selection was nonrandom, it is assumed that the 2 sample groups are similar in their baseline characteristics as they were both in the same curricular level within the program at the time of taking the course. In addition, the initial knowledge and skill level of the students in the 2 cohorts were determined to be equivalent based on their average cumulative grade point average (GPA) data from the earlier semesters in the program and the average incoming Medical College Admission Test (MCAT) scores and undergraduate GPA. Both cohorts started the course and third semester with similar average standardized test scores, similar mean incoming cumulative GPAs, and were given similar course content and assessments. Other confounders like educator quality and digitalization were addressed by using the same instructors and learning management system for the delivery of course content to both cohorts. There were also no changes made in course instruction, course content, syllabus, grading, or objectives between the 2 cohorts. It was also ensured that contextual confounders such as new academic initiatives or changes in course leadership, program objectives, and fallout from the pandemic did not occur during the period of the study.

In the treatment group, students were given 5% participation credit, which was awarded on the completion of the video-based assignments. A customized playlist of assigned videos (aligned to the lecture topics) curated from the Picmonic video database was created by the instructors to be watched by the students on their own time and answer the embedded quiz questions shown in Multimedia Appendix 1. Each set of assigned videos and

quizzes had to be completed before the scheduled in-class lecture on that topic to get credit. Data on the number and frequency of videos watched, multiple attempts at answering video-embedded questions, and quiz accuracy was recorded and monitored using the instructor's dashboard provided to faculty in the Picmonic platform.

Similarly, in the control group, a 5% participation grade was awarded for active participation in the live discussions held during class time based on prior review of the posted instructional materials and assigned readings from the course textbook to be completed before lecture sessions. The contribution of all other course assessments was weighed identically in both control and treatment cohorts. The instructors, textbooks, lectures, instructor-provided materials, and exams used were kept the same between the 2 cohorts.

Exam scores for the treatment group were collected as posttest observations over the course of the semester. The control group underwent comparable nontreated observations [15,16]. Feedback about user experience was gathered from students enrolled in the treatment group at the end of the course through an electronic web-based semi-structured survey questionnaire modified from Haleem et al [17] consisting of 7 required questions included in the survey instrument as shown in Multimedia Appendix 2. NYCPM's Institutional Review Board (IRB) granted ethical approval for this study. The students enrolled in the treatment group were sent the survey link, which the student participants voluntarily filled out. The responses to the 7 questions listed under 4 items in the survey instrument were collected, then analyzed and reported as detailed in the Checklist for Reporting Results of Internet E-Surveys (CHERRIES) adapted from [26] and included as Multimedia Appendix 3. Standard strategies appropriate for this type of quasi-experimental study design, consisting of nonrandomized sampling and posttest-only nonequivalent groups, were used for analysis of the data collected [24,25,27].

Data Analysis

Statistical analysis, data processing, and model fitting were performed in Jupyter Notebook, MATLAB, and Excel. Descriptive statistics, including mean, standard deviation, minimum, maximum, and quartile ranges, were calculated for both the treatment and control group's exam scores and final grade. Boxplots and histograms display score distributions, central tendency, spread, and outliers. Statistical analysis encompassed Levene tests, Welch *t* tests, and the calculation of Cohen *d* as the effect size [28].

Logistic regression analysis enables us to explore the association between access to Picmonic and receiving a score of 90% or higher through a logistic regression model built using Python's Statsmodels library for statistical analysis [29]. To fit the model, we used access to Picmonic as a binary predictor (1=access, 0=no access) to predict whether a student achieved a final grade of 90% or higher in the course (1=final grade ≥0.9; 0=final grade <0.9). We obtained the odds ratio from the model by exponentiating the coefficient for the independent variable with the base of the natural logarithm [30]. The log-likelihood ratio was used to evaluate if access to the resource is a relevant predictor of high final grades.

Correlation analysis was performed to determine the strength of the relationship between usage metrics—number of questions answered, videos played, and quiz accuracy —and students' final grades [30]. We calculated Pearson *r* using a dataset of user interactions with the assigned videos and embedded quizzes on the platform [31].

Survey analysis was conducted using user experience data gathered from students enrolled in the treatment group at the end of the course through an electronic web-based questionnaire sent out by email. Students first answered 4 questions about Picmonic, focusing on information retention, concept understanding, higher test scores, and its usefulness as a learning supplement. Next, students were asked to answer 3 questions regarding their level of satisfaction, frequency of use, and favorite features of the platform—as shown in the Student Experience Survey Instrument in Multimedia Appendix 1. Researchers manually categorized answers to questions regarding their favorite features in Excel. Accordingly, summary statistics were calculated and compared using the data collected from student survey responses.

Ethical Considerations

This study was approved by NYCPM's IRB (23575) in May 2022. Informed consent was waived off by the IRB since students agree to the use of unidentifiable education data for research purposes at registration.

As per institutional policy, the IRB approval for this study is a blanket approval provided for all curriculum-related studies, which are undertaken at the college using deidentified aggregate course data rather than individual scores. The original consent or blanket IRB approval covers secondary analysis without additional consent since all incoming new students are required to sign a consent form agreeing to the use of unidentifiable course and education-related data for research purposes at the beginning and is applicable throughout their enrollment in the program.

All students enrolled in the courses that were included in this study were informed about the research and were made aware that the deidentified aggregate course performance data and their feedback would be used to gather data for this pilot study. This information was also reiterated when they were given the survey instruments to record their feedback which was optional for them to fill out.

All data used in this study are course-level aggregate data calculated from score-related data that are anonymized or deidentified.

No compensation was provided for participation in the research study as the courses used are required as part of the podiatric medical curriculum. The students were made aware their feedback would be collected in the form of responses to a survey questionnaire which was optional to complete. Transparency and fairness were ensured by clarifying that the survey instrument was not mandatory and without any consequences for participants who opted out of responding to the questions included in the survey.

Results

Statistical Analysis

The difference in distribution of exam scores and final grades among the treatment and control groups was visualized using bar graphs (Figure 1A), and box plots (Figure 1B). The summary statistics (Table 1) show the central tendencies using mean exam scores and mean final grade, the spread of the scores using standard deviation, and the shape of the score distributions within each group (Figure 1B), which were used to identify potential differences between the groups. The treatment group had significantly higher average exam scores for most of the course exams and had higher final grades compared to the control group (Figure 1A). The difference in the average scores of the first 2 exams (P<.001), the third exam (P=.04), and the final course grades (grand total) (P<.001) between the 2 groups with and without access to Picmonic was significant. There were also significant differences in variance for exam 1, exam 2, exam 3, and the final grade between the treatment and control groups (Table 2).

Figure 1. Comparison of student performance in course assessment between treatment group (T) and control group (C) based on (**a**) scaled mean test scores and final course grades from treatment group (having access to Picmonic or AV instruction) and control group (without access to Picmonic or audiovisual instruction) and (**b**) score distributions within each group.

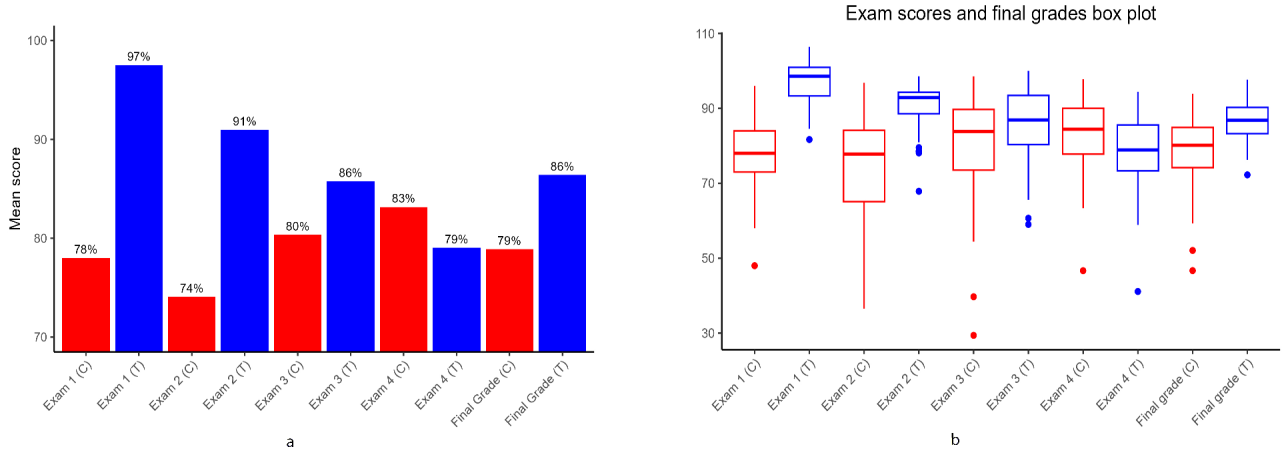


Table . Statistical analysis of student performance in course assessments for treatment and control groups.

Exam	Control, mean (SD)	Treatment, mean (SD)	Mean difference	Welch t test P value	Pooled SD	Cohen d
Exam 1	0.780 (0.098)	0.975 (0.052)	0.195	<.001	0.075	2.398
Exam 2	0.741 (0.143)	0.910 (0.052)	0.169	<.001	0.101	1.486
Exam 3	0.803 (0.133)	0.858 (0.089)	0.054	.005	0.110	0.463
Exam 4	0.832 (0.091)	0.790 (0.092)	−0.041	.006	0.092	−0.449
Final grade	0.789 (0.094)	0.864 (0.049)	0.075	<.001	0.072	0.957

Table . Analysis of variance of exam scores and final grades between treatment and control groups.

Assessment	Levene test statistic	P value	
Exam 1	19.442	<.001[a]	Significant compared to control
Exam 2	52.632	<.001[b]	Significant
Exam 3	4.354	.04	Significant
Exam 4	0.127	.72	Not significant
Final grade	14.652	<.001[c]	Significant

[a](1.92×10^{-5})
[b](1.77×10^{-11})
[c](1.87×10^{-4})

Levene test statistic values for potential differences in variance of exam scores and final grades between the 2 groups revealed significant differences in variance for exam 1, exam 2, exam 3, and the final grade (Table 2). The results of Welch t test used to compare the average final exam scores and final grades indicated statistically significant differences between the 2 groups' first 3 exams and final grades for the course (P<.01) (Table 1). Cohen d values calculated to quantify the observed

differences between the treatment and control groups across all exams and the final grade revealed a large effect size for exam 1 (d=2.397), exam 2 (d=1.486), and the final grade (d=0.957) (Table 1).

Logistic Regression

A logistic regression analysis between access to Picmonic and the likelihood of achieving a high final course grade of 90 out of 100 (90%) or above resulted in an odds ratio which indicates that, assuming all other factors are constant, students in our study with access to Picmonic were 2.72 times more likely to have received a final grade of 90% or higher and a letter grade of A in the course. The log-likelihood ratio P value is .02 (P<.05); therefore, we reject the null hypothesis that the base model with only the intercept is better than the model with access to Picmonic used as the predictor (Table 3).

Table . Regression analysis of the association between access to Picmonic and receiving a high final grade.

Predictor	Coefficient	SE	z value	P value[a]	Lower CI	Upper CI
Intercept	−1.998	0.377	−5.303	<.001	−2.737	−1.260
Picmonic access	0.971	0.446	2.180	.03	0.098	1.845

[a]Model log-likelihood ratio P value =.02.

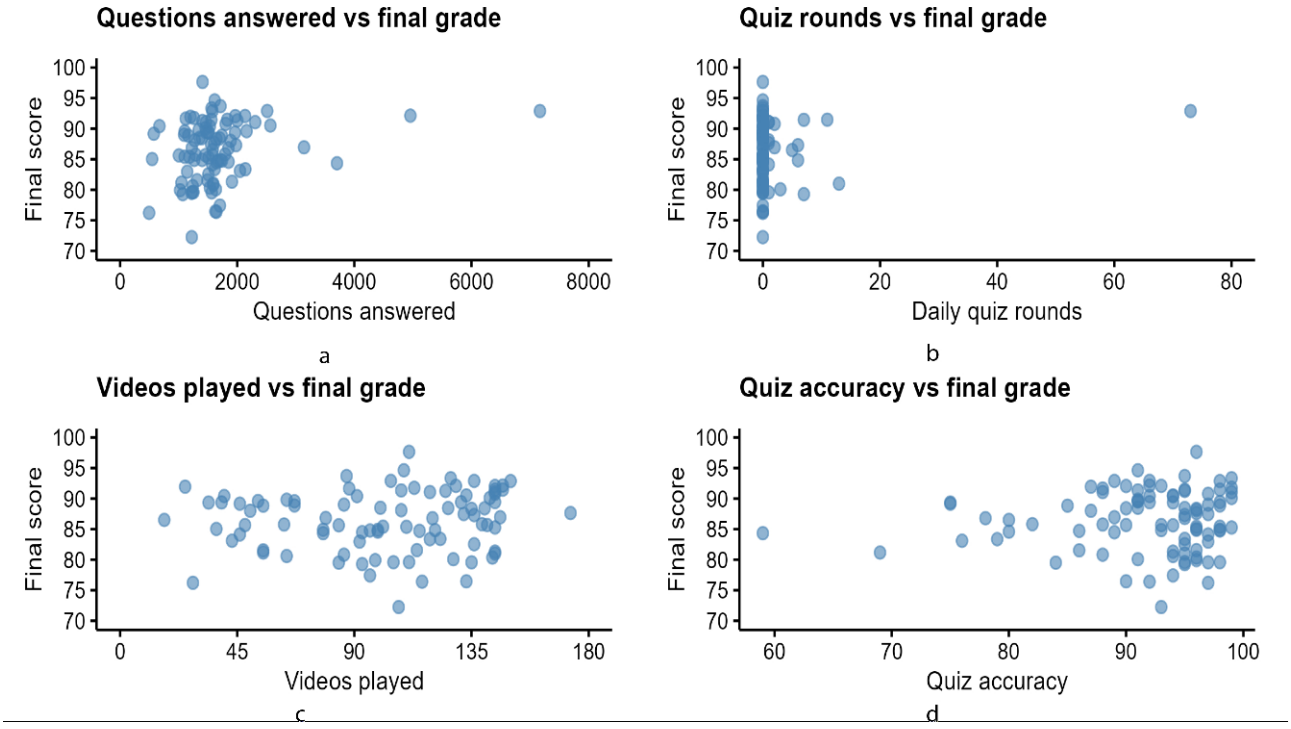
Correlation Analysis

Pearson correlation coefficient calculated to explore the relationship between final course grade and various usage statistics (number of videos played, questions answered, and quiz accuracy) show that students in the treatment group who answered more questions on the platform tended to get a higher final score (r=0.25, P=.02) (Table 4 and Figure 2).

Table . Analysis of correlation between final grades and platform usage metrics.

Final grade with:	Pearson r	P value
Questions answered	0.247965674	.02
Daily quiz rounds	0.124515676	.24
Videos played	0.09980258	.35
Quiz accuracy	0.074970912	.48

Figure 2. Correlation analysis between final course grades and (**a**) the number of in-video questions answered, (**b**) number of daily quiz round attempts, (**c**) number of times videos watched, and (**d**) overall accuracy of embedded quizzes.



Survey Analysis

The response rate for item 1 of the survey instrument, which consisted of 4 questions about the students' perceived usefulness of Picmonic, was 73% (66 out of 91). In this item, students indicated strong agreement regarding Picmonic's positive impact on information retention, concept understanding, higher test scores, and usefulness as a supplementary learning tool (Figure 3A). In item number 4, students also reported an average satisfaction rating of 3.71 out of 5 (Figure 3B). Out of 53 students who responded to item number 2 in the survey questionnaire, 36 accessed Picmonic at least once a week—predominantly 1‐2 times per week (Figure 3C). In item number 3, 50 out of 91 students responded to the open-ended questions regarding user experience or preferences about their favorite feature of Picmonic with some choosing more than 1 feature. The number of times each feature was reported as preferred is listed and compared in Table 5. Noteworthy features of Picmonic highlighted by students included videos, quizzes and questions, mnemonics, and content (Table 5).

Figure 3. Analysis of student experience feedback data. Student experience survey insights and summary of qualitative open-ended student feedback data showing (**a**) perceived effectiveness of Picmonic on various learning outcomes and retention, (**b**) user satisfaction level, and (**c**) frequency of use.

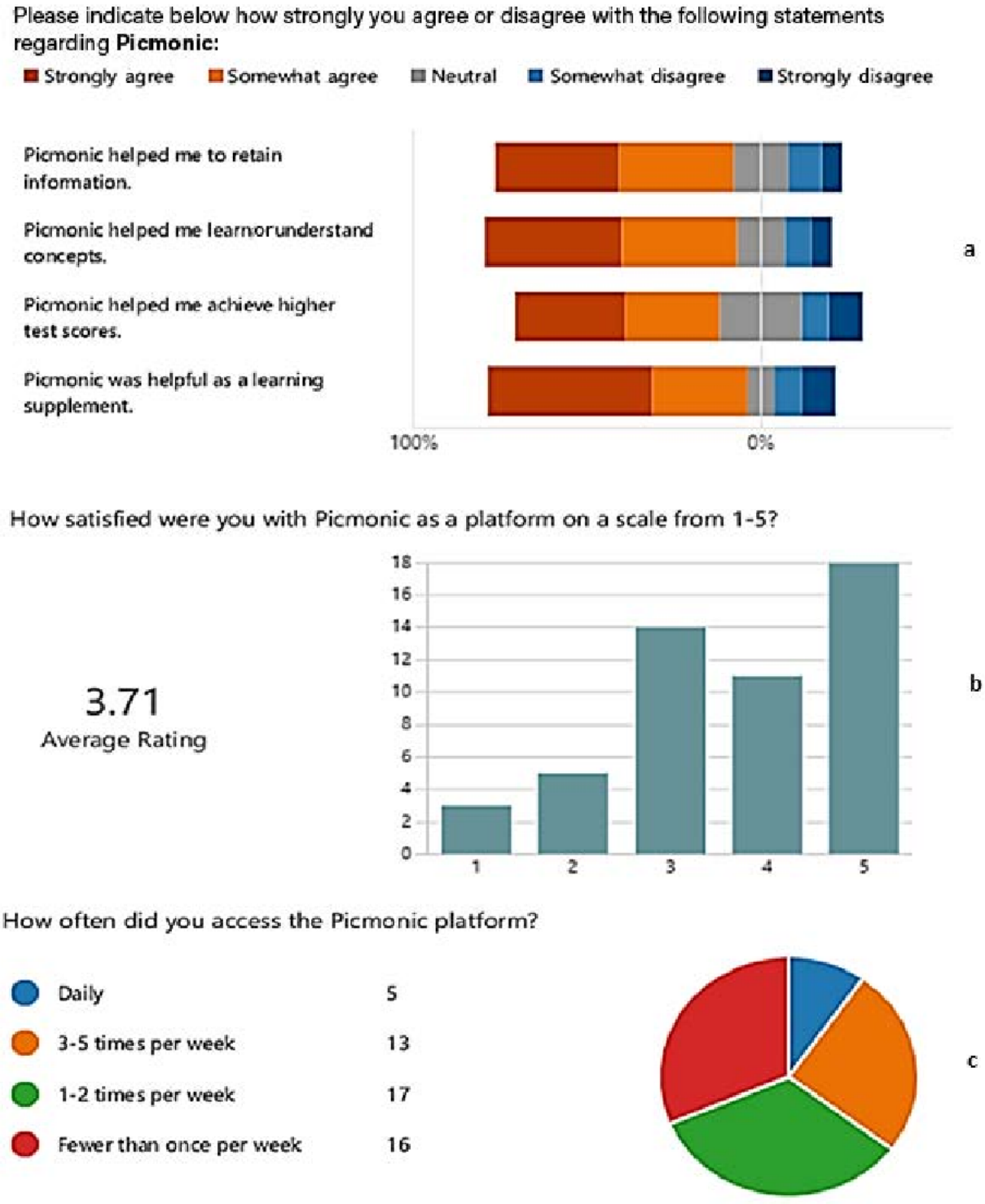


Table . Comparative analysis of student feedback performed on open-ended questions regarding user experience or preferences.

	Feedback (N=50), n (%)
Videos	15 (30)
Quizzes and questions	9 (18)
Mnemonics	9 (18)
Content	9 (18)
Images	6 (12)
User interface	4 (8)
General	2 (4)

Discussion

Principal Findings

In our study, we observed that students with access to the multimedia and mnemonic-based AV learning resource scored higher on most exams, had higher final grades, and were more likely to receive a final grade of 90% or higher. The resulting effect sizes for the treatment group are large enough to be meaningful in the real world; however, other factors may have contributed to the observed effect size. Students in the treatment group who answered more questions on the platform tended to get a higher final score. Most students used Picmonic at least once a week to learn the course content. The lower response rate on this item asking about the frequency of accessing Picmonic in a week could be due to the inability of participants to remember their frequency of access at the time of completing the survey several weeks after the completion of the course and semester. Students had positive opinions regarding Picmonic's platform user experience and its effectiveness in helping them retain information, learn and understand concepts, and achieve higher test scores. In this course, the students reported the videos to be their most preferred feature, followed by mnemonics and self-assessment quiz questions associated with the videos, which allowed them to test their knowledge both during and after watching the assigned videos on their own time with unlimited attempts. In this context, it is important to note that though students generally prefer to watch videos on their own rather than attending class, the 2 groups were treated equally since both the reading assignments (for the control group) and the Picmonic-based video assignments (for the treatment group) were assigned to be completed outside of class on their own time. Additionally, the students in the treatment group were assigned to watch the videos and complete the video-embedded quizzes within a specific time frame to mimic the reading assignments given to the control group. To receive the 5% participation grade, the treatment group had to complete the Picmonic video assignments within the instructor-provided deadline corresponding to the weekly topical schedule, which is similar to that of the control group, rather than watching the assigned videos at their own pace throughout the semester.

Comparison With Previous Research

Our experimental results indicate that students with access to the multimedia-based AV mnemonic learning resource scored higher on most exams, had higher final grades, and were more likely to receive a final grade of 90% or higher. The difference between the grades achieved by the 2 groups is large enough to be meaningful in the real world. These findings are consistent with previous studies that have demonstrated the effectiveness of AV mnemonics and web-based learning tools in enhancing memory retention and learning outcomes in medical education [12,16].

Yang et al [12] observed improved student performance in free-recall and paired-matching tests when using an earlier version of the multimedia-based learning platform that we have used here that was released almost 10 years ago, while the current version that we have used has newer, redesigned, more impactful, and shorter videos, though still based on the same type of picture mnemonics and principles of spaced repetition and visual learning. The currently available version that was used in this study also has improved dashboard features and assessment capabilities compared to the older version. Abdalla et al [13] also found that students who had undergone AV sessions had higher marks on response answer questions, shorter time spent answering questions, and higher memory consolidation after specific time benchmarks. These studies underscore the importance of memory and knowledge retention in medical students' academic performance.

Results examining user interaction with the resource showed that the more questions a student answered on a multimedia-based AV learning platform (like Picmonic) using spaced repetition and mnemonics, the higher their grade tended to be. This finding aligns with research highlighting the benefits of interactive user interfaces and spaced repetition in increasing memory retention and higher order thinking [12].

Survey responses indicated that students found the resource useful for learning concepts, retaining information, and achieving higher test scores. This is consistent with previous research demonstrating increased student engagement, comprehension, and satisfaction with multimedia-based resources compared to traditional instructional methods [15,18-20]. Studies by Tackett et al [23] examined student engagement with commercially produced medical education videos incorporated into a preclinical course and also found the videos to be helpful for student learning and improved students' experiences.

Overall, our findings contribute to the growing body of evidence supporting the integration of multimedia-based learning resources and AV mnemonics in medical education curricula

to enhance student learning experiences and outcomes [7,8,12,16,17].

Limitations

While this investigation showed promising initial findings, we recognize that the study design limits its ability to make unequivocal causal inferences about the impact of the multimedia tool alone on the outcomes. The sample size is relatively small, and the lack of randomization in the study design may limit the generalizability of the findings. The quasi-experimental design with nonrandom assignment to treatment and control groups restricts the establishment of a cause-and-effect relationship [24], which could potentially affect the internal validity of the study and its ability to accurately infer whether the change in outcomes was caused by the intervention. Without randomized selection of the 2 groups, we cannot rule out potential unmeasured differences due to systematic differences in sample selection [23]. The study only examined the effect of the implementation of the intervention in 1 preclinical infectious disease course, limiting assessment of the tool's effectiveness. Additionally, the 5% participation credit component was implemented differently for the control (class participation) versus treatment (video watching) groups, which could impact effort levels.

The 2 most likely influential unmeasured confounding variables in our study are potential differences in the overall academic aptitude of the students in each sample [13], as well as potential differences in the amount of time spent with study materials between the 2 groups due to the spaced repetition provided in the Picmonic platform, which could impact the results. The potential difference in student aptitude between the 2 groups should be somewhat mitigated by the fact that both groups were 2nd-year medical students at the same college when being tested. Additionally, the 2 groups had similar average standardized test scores and mean incoming GPAs at the beginning of the semester. The absence of pretest observations comparing the treatment group to the control group makes it difficult to know whether any differences between the 2 groups could potentially be attributed to pre-existing factors rather than the multimedia learning tool alone [14].

Since this study was done on a very limited number of students and involved only 1 course focused on the topic of preclinical medical microbiology and infectious diseases, this effect may not be generalized for all topics or types of courses. This course is heavily dependent on memorization and recall due to the nature of the topics covered, which may also contribute to the impact of the integration of visual learning with mnemonics and spaced repetition and therefore may not be equally applicable in another course that does not require extensive memorization.

Conclusions

Our study shows strong agreement amongst students that Picmonic helped them achieve key learning outcomes. Usage data revealed a positive relationship between final grades and students' usage of platform features; the number of in-video questions answered had a stronger correlation than the number of times videos were watched or the accuracy of topic-associated quiz attempts. Students that were given access to Picmonic did better on exams; however, it must be noted that due to the nonrandomized sampling process, posttest-only study design, limited implementation, and small sample size, it is difficult to conclude whether the difference was due solely to the integration of the new tool. The improved course grades and test scores in the treatment group may have been due to the inherent confounding factors like comprehension and retention skills or increased contact time with the course topics provided by the platform features. Our pilot study focused on the integration of Picmonic, a multimedia-based learning resource, in only 1 course, but implementing it across more courses over multiple semesters would strengthen the assessment of the tool's effectiveness. Despite its limitations, this study provides insight into the potential benefits of integrating multimedia learning resources in podiatric medical education. However, a larger study that implements this type of learning resource on a larger scale, and in more preclinical and clinical courses throughout the curriculum, is needed to further analyze its effectiveness in the podiatric medical curriculum.

Acknowledgments

This work originated at New York College of Podiatric Medicine; the authors express their deep gratitude for the support received. The authors of this paper would like to thank Rhonda Altonen, MLS, MS (Director of Library Services, Sheldon L. Sirota Memorial Library, Touro College of Osteopathic Medicine and College of Pharmacy, Touro University) for her insights about the specific features of Picmonic, the multimedia-based learning resource used in this study, as well Adrian Rice, Associate Registrar, New York College of Osteopathic Medicine, for providing summarized cumulative grade point average data of the control and treatment cohorts. The authors of this paper would also like to thank Picmonic for providing additional data on usage of platform features for further analysis which provided detailed insights.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Example of the playlist of Picmonic videos and embedded quiz questions on Gram positive bacilli assigned to the treatment group students.

[PNG File, 417 KB - mededu_v11i1e55206_app1.png]

Multimedia Appendix 2
Student user experience survey instrument.
[PNG File, 186 KB - mededu_v11i1e55206_app2.png]

Multimedia Appendix 3
Checklist for Reporting Results of Internet E-Surveys (CHERRIES).
[PDF File, 104 KB - mededu_v11i1e55206_app3.pdf]

References

1. O'Hanlon R, Laynor G. Responding to a new generation of proprietary study resources in medical education. J Med Libr Assoc 2019 Apr;107(2):251-257. [doi: 10.5195/jmla.2019.619] [Medline: 31019395]
2. Kyaw BM, Posadzki P, Paddock S, Car J, Campbell J, Tudor Car L. Effectiveness of digital education on communication skills among medical students: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019 Aug 27;21(8):e12967. [doi: 10.2196/12967] [Medline: 31456579]
3. Nur Yasak Cevi˙zci˙ R. An investigation into the student opinions and observations towards animation-supported instruction in social sciences lesson. Int J Turkish Lit Cult Educ 2023 Jan 1;12(1):374-391. [doi: 10.7884/teke.5619]
4. Harris DM, Chiang M. An analysis of Anki usage and strategy of first-year medical students in a structure and function course. Cureus 2022 Mar;14(3):e23530. [doi: 10.7759/cureus.23530] [Medline: 35494926]
5. Wechsler B. Improvement of medical terminology education with interactive learning module. S D Med 2022 Oct;75(10):459. [Medline: 36889270]
6. Brahler CJ, Walker D. Learning scientific and medical terminology with a mnemonic strategy using an illogical association technique. Adv Physiol Educ 2008 Sep;32(3):219-224. [doi: 10.1152/advan.00083.2007]
7. Maag M. The effectiveness of an interactive multimedia learning tool on nursing students' math knowledge and self-efficacy. Comput Inform Nurs 2004;22(1):26-33. [doi: 10.1097/00024665-200401000-00007] [Medline: 15069846]
8. Myers SR. The changing face of medical education. J Med Educ Curric Dev 2014 Jan;1:S17270. [doi: 10.4137/JMECD.S17270]
9. West N. Mnemonics are useful memory tools in modern medicine. Ugeskr Laeger 2014 Dec 8;176(50):V66204. [Medline: 25498182]
10. Noor U, Younas M, Saleh Aldayel H, Menhas R, Qingyu X. Learning behavior, digital platforms for learning and its impact on university student's motivations and knowledge development. Front Psychol 2022;13:933974. [doi: 10.3389/fpsyg.2022.933974] [Medline: 36506979]
11. Abarghouie MHG, Omid A, Ghadami A. Effects of virtual and lecture-based instruction on learning, content retention, and satisfaction from these instruction methods among surgical technology students: a comparative study. J Educ Health Promot 2020;9:296. [doi: 10.4103/jehp.jehp_634_19] [Medline: 33426100]
12. Yang A, Goel H, Bryan M, et al. The Picmonic(®) Learning System: enhancing memory retention of medical sciences, using an audiovisual mnemonic Web-based learning platform. Adv Med Educ Pract 2014;5:125-132. [doi: 10.2147/AMEP.S61875] [Medline: 24868180]
13. Abdalla MMI, Azzani M, Rajendren R, et al. Effect of story-based audiovisual mnemonics in comparison with text-reading method on memory consolidation among medical students: a randomized controlled trial. Am J Med Sci 2021 Dec;362(6):612-618. [doi: 10.1016/j.amjms.2021.07.015] [Medline: 34606752]
14. Picmonic. URL: https://www.picmonic.com/ [accessed 2025-01-24]
15. Bhangu S, Provost F, Caduff C. Introduction to qualitative research methods - part I. Perspect Clin Res 2023;14(1):39-42. [doi: 10.4103/picr.picr_253_22] [Medline: 36909216]
16. Krishnan P. A review of the non-equivalent control group post-test-only design. Nurse Res 2019 Sep 21;26(2):37-40. [doi: 10.7748/nr.2018.e1582] [Medline: 30226337]
17. Haleem A, Javaid M, Qadri MA, Suman R. Understanding the role of digital technologies in education: a review. Sust Oper Comput 2022;3:275-285. [doi: 10.1016/j.susoc.2022.05.004]
18. Seto C, Zayat V. A spoonful of eponyms helps the pathology go down: using food eponyms and visual mnemonics in preclinical pathology education. Med Sci Educ 2022 Feb;32(1):131-140. [doi: 10.1007/s40670-021-01474-w] [Medline: 35154897]
19. Harris DA, Krousgrill C. Distance education: new technologies and new directions. Proc IEEE 2008 May 20;96(6):917-930. [doi: 10.1109/JPROC.2008.921612]
20. Giovannella C. Effect induced by the Covid-19 pandemic on students' perception about technologies and distance learning. Ludic, co-design and tools supporting smart learning ecosystems and smart education. Smart Innov Syst Technol 2020;197:105-116. [doi: 10.1007/978-981-15-7383-5_9]

21. Sosa Neira EA, Salinas J, De Benito B. Emerging technologies (ETs) in education: a systematic review of the literature published between 2006 and 2016. Int J Emerg Technol Learn 2017;12(5):128. [doi: 10.3991/ijet.v12i05.6939]
22. Grainger R, Liu Q, Geertshuis S. Learning technologies: a medium for the transformation of medical education? Med Educ (Chicago Ill) 2021 Jan;55(1):23-29. [doi: 10.1111/medu.14261]
23. Tackett S, Green D, Dyal M, et al. Use of commercially produced medical education videos in a cardiovascular curriculum: multiple cohort study. JMIR Med Educ 2021 Oct 7;7(4):e27441. [doi: 10.2196/27441] [Medline: 34617911]
24. Harris AD, McGregor JC, Perencevich EN, et al. The use and interpretation of quasi-experimental studies in medical informatics. J Am Med Inform Assoc 2006;13(1):16-23. [doi: 10.1197/jamia.M1749] [Medline: 16221933]
25. Shadish WR, Cook TD, Campbell DT. Experimental and Quasi-Experimental Designs for Generalized Causal Inference: Houghton Mifflin; 2002.
26. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004;6(3):e34. [doi: 10.2196/jmir.6.3.e34]
27. Maxwell SE, Delaney HD. Designing Experiments and Analyzing Data: A Model Comparison Approach, 2nd edition: Lawrence Erlbaum Associates; 2004. [doi: 10.4324/9781410609243]
28. Cohen J. Statistical Power Analysis for the Behavioral Sciences, 2nd edition: Routledge; 1988.
29. Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with Python. Presented at: Python in Science Conference; Jun 28 to Jul 3, 2010; Austin, Texas. [doi: 10.25080/Majora-92bf1922-011]
30. Hosmer DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression, 3rd edition: Wiley; 2013. [doi: 10.1002/9781118548387]
31. Witte R, Witte J. Statistics, 9th edition: Wiley; 2010.

Abbreviations

AV: audiovisual
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
GPA: grade point average
IRB: Institutional Review Board
MCAT: Medical College Admission Test
NYCPM: New York College of Podiatric Medicine

Edited by B Lesselroth; submitted 05.12.23; peer-reviewed by K Malale, S Mukhida; revised version received 15.10.24; accepted 03.12.24; published 11.02.25.

Please cite as:
Hoyt G, Bakshi CS, Basu P
Integration of an Audiovisual Learning Resource in a Podiatric Medical Infectious Disease Course: Multiple Cohort Pilot Study
JMIR Med Educ 2025;11:e55206
URL: https://mededu.jmir.org/2025/1/e55206
doi:10.2196/55206

Performance of ChatGPT-3.5 and ChatGPT-4 in the Taiwan National Pharmacist Licensing Examination: Comparative Evaluation Study

Ying-Mei Wang[1,2,3,4], MBA; Hung-Wei Shen[1,2,4], MBA; Tzeng-Ji Chen[5,6,7], Dr Med; Shu-Chiung Chiang[1,8], PhD; Ting-Guan Lin[2,4], BS

[1]Department of Medical Education and Research, Taipei Veterans General Hospital Hsinchu Branch, 81, Section 1, Zhongfeng Road, Zhudong, Hsinchu, Taiwan

[2]Department of Pharmacy, Taipei Veterans General Hospital Hsinchu Branch, Hsinchu, Taiwan

[3]School of Medicine, National Tsing Hua University, Hsinchu, Taiwan

[4]Hsinchu County Pharmacists Association, Hsinchu, Taiwan

[5]Department of Family Medicine, Taipei Veterans General Hospital Hsinchu Branch, Hsinchu, Taiwan

[6]Department of Family Medicine, Taipei Veterans General Hospital, Taipei, Taiwan

[7]Department of Post-Baccalaureate Medicine, National Chung Hsing University, Taichung, Taiwan

[8]Institute of Hospital and Health Care Administration, School of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan

Corresponding Author:
Ying-Mei Wang, MBA
Department of Medical Education and Research, Taipei Veterans General Hospital Hsinchu Branch, 81, Section 1, Zhongfeng Road, Zhudong, Hsinchu, Taiwan

Abstract

Background: OpenAI released versions ChatGPT-3.5 and GPT-4 between 2022 and 2023. GPT-3.5 has demonstrated proficiency in various examinations, particularly the United States Medical Licensing Examination. However, GPT-4 has more advanced capabilities.

Objective: This study aims to examine the efficacy of GPT-3.5 and GPT-4 within the Taiwan National Pharmacist Licensing Examination and to ascertain their utility and potential application in clinical pharmacy and education.

Methods: The pharmacist examination in Taiwan consists of 2 stages: basic subjects and clinical subjects. In this study, exam questions were manually fed into the GPT-3.5 and GPT-4 models, and their responses were recorded; graphic-based questions were excluded. This study encompassed three steps: (1) determining the answering accuracy of GPT-3.5 and GPT-4, (2) categorizing question types and observing differences in model performance across these categories, and (3) comparing model performance on calculation and situational questions. Microsoft Excel and R software were used for statistical analyses.

Results: GPT-4 achieved an accuracy rate of 72.9%, overshadowing GPT-3.5, which achieved 59.1% (P<.001). In the basic subjects category, GPT-4 significantly outperformed GPT-3.5 (73.4% vs 53.2%; P<.001). However, in clinical subjects, only minor differences in accuracy were observed. Specifically, GPT-4 outperformed GPT-3.5 in the calculation and situational questions.

Conclusions: This study demonstrates that GPT-4 outperforms GPT-3.5 in the Taiwan National Pharmacist Licensing Examination, particularly in basic subjects. While GPT-4 shows potential for use in clinical practice and pharmacy education, its limitations warrant caution. Future research should focus on refining prompts, improving model stability, integrating medical databases, and designing questions that better assess student competence and minimize guessing.

(JMIR Med Educ 2025;11:e56850) doi:10.2196/56850

KEYWORDS

artificial intelligence; ChatGPT; chat generative pre-trained transformer; GPT-4; medical education; educational measurement; pharmacy licensure; Taiwan; Taiwan national pharmacist licensing examination; learning model; AI; Chatbot; pharmacist; evaluation and comparison study; pharmacy; statistical analyses; medical databases; medical decision-making; generative AI; machine learning

Introduction

Background

With the advent of the artificial intelligence (AI) era, applications of AI in the medical field have increased with ChatGPT (OpenAI) being the most notable examples. ChatGPT is a large language model based on a generative pretrained transformer developed by OpenAI. ChatGPT-3.5 (GPT-3.5) was the first publicly accessible version, while ChatGPT-4 (GPT-4) was the subscription version. GPT-4 surpasses GPT-3.5 in advanced reasoning, almost nearing human-level performance in professional and academic examinations [1,2]. For instance, GPT-4 ranked in the top 10% of scores on a law examination, whereas GPT-3.5 ranked in the bottom 10% [3]. Additionally, GPT-3.5 resolved 90% of false-belief tasks, achieving the level of a 7-year-old child, whereas GPT-4 resolved 95% of these tasks [4]. Following its launch, ChatGPT has been extensively studied and discussed in both the medical and educational fields [5]. The most widely recognized performance of GPT-3.5 has been on the United States Medical Licensing Examination (USMLE) [6,7]; however, GPT-3.5's performance did not meet expectations in other examinations [8-11]. Gradually, Nori et al [12]observed that the accuracy of GPT-4 was higher than that of the GPT-3.5 on the USMLE, and further studies confirmed that GPT-4 outperforms GPT-3.5 [13-16]. However, there has been limited research on its performance in pharmacy examinations.

In the field of pharmacy, GPT-3.5 has exhibited commendable performance in clinical toxicology and pharmacology [17,18], although it has not passed the National Pharmacist Licensing Examination (NPLE) in Taiwan [19]. However, GPT-4 has outperformed GPT-3.5 in drug information [20] and China's Pharmacist Licensing Examination [21]. Generative AI models, a large language model, has been applied in drug development and novel drug design [22-24], pharmacovigilance [25,26], pharmacokinetic model development [27], pharmacy education, and research writing [28,29].

Goal of the Study

According to previous studies, GPT-3.5 failed to pass the NPLE, indicating its limitations in pharmacy education. Based on these findings, we hypothesized that GPT-4 would outperform GPT-3.5 in this context, demonstrating greater proficiency. To test this hypothesis, this study compared the performance of GPT-3.5 and GPT-4 on Taiwan's NPLE. Additionally, we conducted a comprehensive assessment of their performance across various question types, with a focus on pharmacy-related tasks such as pharmacokinetic calculation and clinical decision-making scenarios. This analysis aims to determine the practical applications of GPT-4 in pharmacy education and establish guidelines for its optimal use in this field.

Methods

Background

The NPLE in Taiwan is divided into 2 stages. The first stage focuses on 3 basic subjects: pharmacology and pharmaceutical chemistry, pharmaceutical analysis and pharmacognosy (including traditional Chinese medicine), and pharmaceutics and biopharmaceutics. The second stage focuses on 3 clinical subjects: dispensing and clinical pharmacy, pharmacotherapy, and pharmacy administration and pharmacy law. The first and second stages of the examination have 240 and 210 multiple-choice questions, respectively. Pharmacy students typically complete the first-stage exam after completing their third year of university coursework. They become eligible for the second-stage exam only after passing the first examination, completing their internships and obtaining their graduation certificates. After passing the second-stage examination, candidates receive their pharmacist certificate, allowing them to practice as a pharmacist legally.

Data Source

This study used the 2-stage NPLE questions released by the Ministry of Examination in February 2023, with each subject exam lasting for 1 hour. The version of NPLE used in this study was the most recent available at the time of research. We used both GPT-3.5 (free version) and GPT-4 (licensed version). No temperature settings were applied. Examination questions were manually fed into GPT-4 and GPT-3.5 sequentially. To simulate student responses, complete questions were entered into the models without tailored prompts. One question was input at a time, and the responses were recorded for analysis. Since GPT-3.5 cannot process images and image functionality of GPT-4 was unavailable during the analysis, only text-based questions were used. Questions containing graphics, such as chemical structures, tables, symbols, and formulas were excluded. Both models were presented with the same set of questions under identical conditions. Due to the limitations on the number of times the model could be used and required cooling time between queries, all questions were answered sequentially and not timed to avoid any potential bias introduced by time constraints.

Study Design

The study was divided into 3 parts; the first part compared the accuracy of GPT-4 and GPT-3.5, as well as in different subjects. The second part compared the accuracy of GPT-4 and GPT-3.5 across different question types. These questions were categorized into 4 types: memory-based questions (1 correct word answer out of 4 options, low-level thinking; Figure 1), judgment questions (1 correct statement out of 4, medium-level thinking; Figure 2), reverse questions (1 incorrect statement out of 4, medium to high-level thinking; Figure 3), and comprehension questions (multiple-choice or matching types, high-level thinking; Figure 4). One pharmacist classified the questions according to these established categories and the second pharmacist reviewed the classifications. In the event of disagreement, a third pharmacist was consulted for the final decision. All pharmacists had over 10 years of experience in medical center hospitals or community teaching hospitals. The third part compared the accuracy of GPT-4 and GPT-3.5 for calculation-based and case scenario questions (Figure 5). Model testing for this study was conducted from May 10 to July 20, 2023.

Figure 1. Template of a memory-based question (choose 1 correct word from 4 options, requiring low-level thinking).

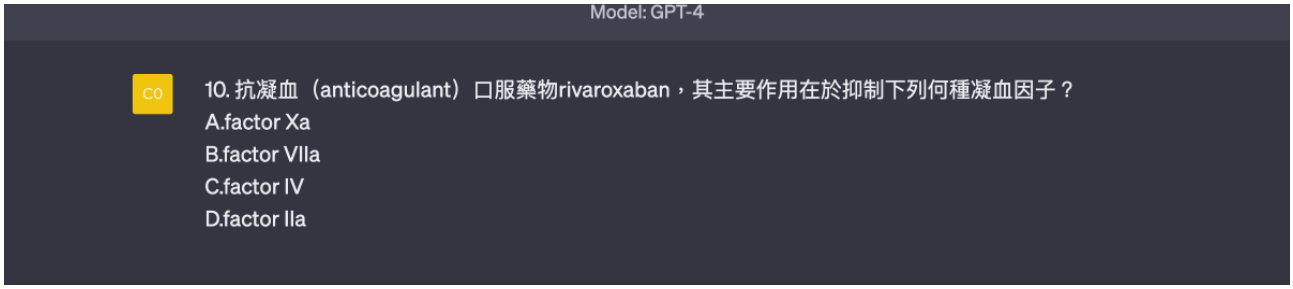


Figure 2. Template of a judgment question (choose 1 correct statement from 4 options, requiring medium-level thinking).

Figure 3. Template of a reverse question (choose 1 incorrect statement from 4 options, requiring medium- to high- level thinking).

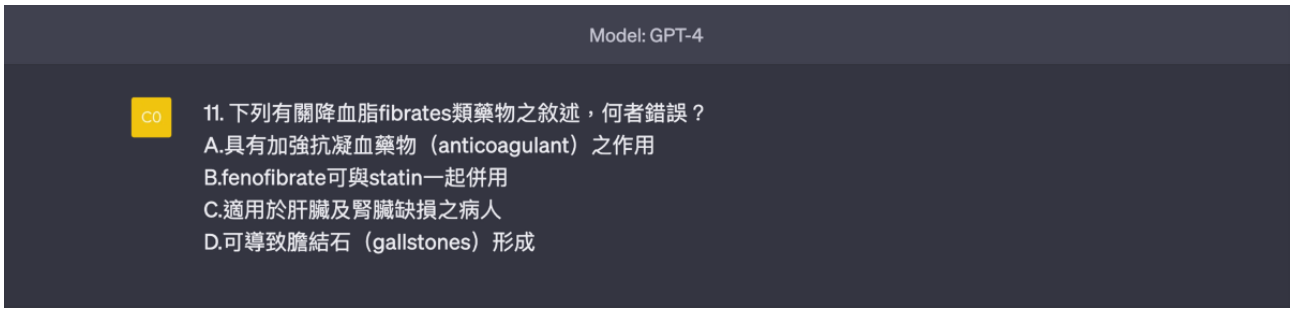
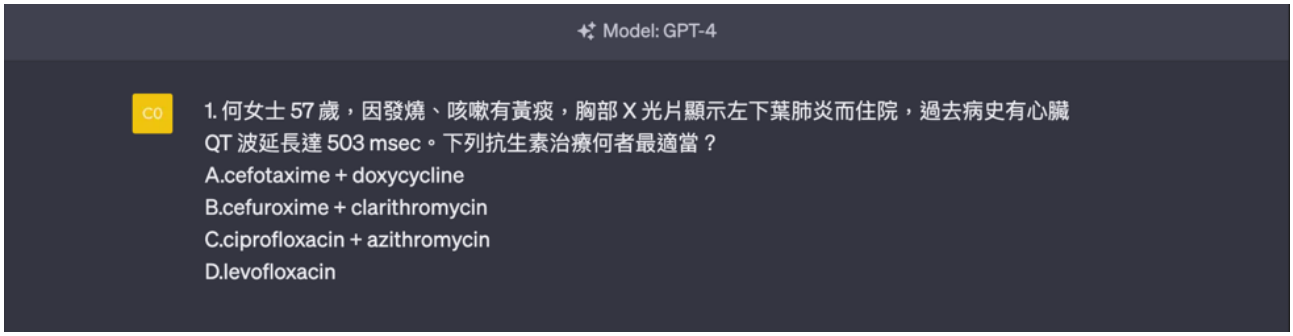


Figure 4. Template of a comprehension questions (multiple-choice or matching types, requiring high- level thinking).

Figure 5. Template of a case scenario question.



Statistical Analysis

Microsoft Excel 2019 was used to compare the accuracy rates of the 2 models. χ^2 tests were used to compare the overall accuracy rates of answers obtained using GPT-3.5 and GPT-4. McNemar tests were used to compare the consistency in answers between GPT-3.5 and GPT-4, and for the calculation-based and situational question types using R software (version 4.2.2; R Foundation for Statistical Computing).

Ethical Considerations

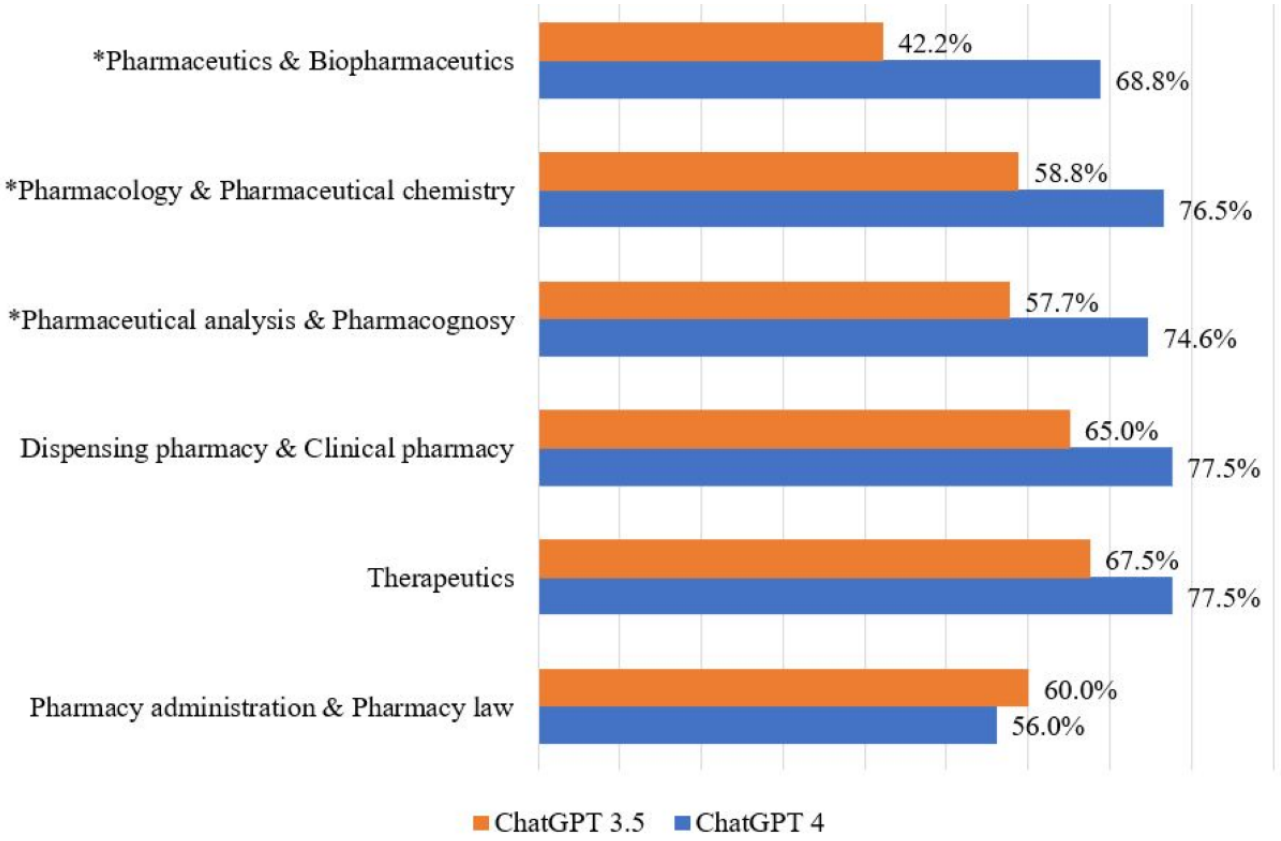
This study involved comparing the performance of ChatGPT-4 and ChatGPT-3.5 in the pharmacist licensing examination. It did not involve human participants. As per the guidelines of the 'Human Research Cases Exempted from Ethics Review Board' issued by the Ministry of Health and Welfare, Taiwan, this study was exempted from Ethics Review Board analysis.

Results

Accuracy in Different Subjects

In total, 203 and 210 questions were included for analysis from the first- and second-stage examinations, respectively, after excluding 37 questions containing graphical elements (N=413) (Figure 6). GPT-4 had an overall accuracy of 72.9% (301/413), easily passing the test (60% threshold) and outperforming GPT-3.5 which achieved an accuracy of 59.1% (244/413; $P<.001$). In terms of accuracy by stage, GPT-4's overall accuracy was significantly higher than that of GPT-3.5 (73.4% vs 53.2% or 149/203 vs 108/203; $P<.001$) in basic subjects of the first stage. GPT-4 also significantly outperformed GPT-3.5 in each of the 3 basic subjects. In the clinical subjects of the second stage, GPT-4's accuracy was higher but not statistically significant than that of GPT-3.5 (72.4% vs 64.8% or 152/210 vs 136/210; $P=.096$). In pharmacy administration and pharmacy law, GPT-4's accuracy was lower than that of GPT-3.5 (56% vs 60% or 28/50 vs 30/50; $P=.96$). Among individual subjects, significant differences were observed in pharmacology and pharmaceutical chemistry ($P=.02$), pharmaceutical analysis and pharmacognosy ($P=.02$), and pharmaceutics and biopharmaceutics ($P=.002$). No significant differences were noted in dispensing pharmacy and clinical pharmacy ($P=.07$), pharmacotherapeutics ($P=.10$), and pharmacy administration and pharmacy law ($P=.48$).

Figure 6. Accuracy comparison of ChatGPT-3.5 and ChatGPT-4 across different subjects. *P<.05.



The overall consistency among answers significantly differed between the 2 models (68%, P<.001), with GPT-4 showing consistent correct answers in 49.4% (n=204) of cases and consistent incorrect answers in 18.6% (n=77) of cases (Table 1).

Table . Performance comparison of consistency between ChatGPT-3.5 and ChatGPT-4.

ChatGPT-3.5 responses	GPT-4	
	Correct answers, n (%)	Incorrect answers, n (%)
Correct answer	204 (49.4)	38 (9.2)
Incorrect answer	94 (22.8)	77 (18.6)

Accuracy in Different Question Types

Among the 413 examination questions analyzed, memory-based questions were the most common (n=254, 61.5%), followed by judgment questions (n=82, 19.9%), reverse questions (n=46, 11.1%), and comprehension questions (n=31, 7.5%). GPT-4 and GPT-3.5 did not differ significantly in terms of accuracy of answers between question types (P=.461 vs P=.18; Table 2). GPT-4 is significantly better than GPT-3.5 in memory-based questions (P<.001) and comprehension-based questions(P=.03).

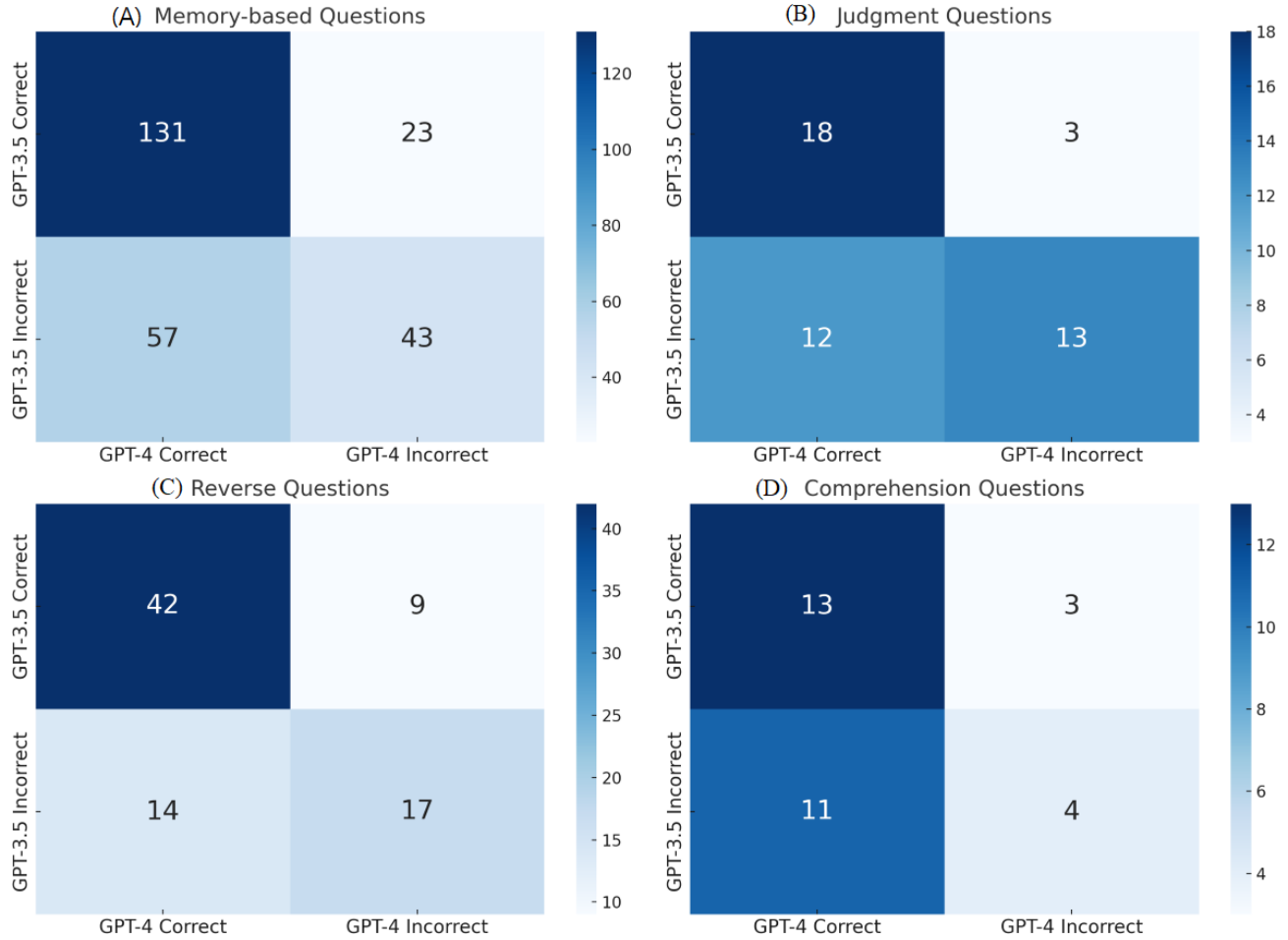
Table . Accuracy comparison of ChatGPT-3.5 and ChatGPT-4 by question type.

Question type	GPT-3.5 Correct answers, n (%)	GPT-4 Correct answers, n (%)	Total, n (%)	P value
Memory-based questions	155 (61)	188 (74)	254 (61.5)	<.001[a]
Judgment questions	21 (45.7)	30 (65.2)	46 (11.1)	.06
Reverse questions	51 (62.6)	56 (68.3)	82 (19.9)	.41
Comprehension questions	16 (51.6)	24 (77.4)	31 (7.5)	.03[a]

[a]P<.05.

Figure 7 shows the performance comparison of GPT-3.5 and GPT-4 across question types. The data provided insights into the relative strengths and weaknesses of each model.

Figure 7. Performance comparison of GPT-3.5 and GPT-4 across question types (A) memory-based, (B) judgement, (C) reverse , and (D) comprehension. The heatmaps display the number of answers, with darker shades indicating higher counts of correct responses and highlighting model performance.



Further analysis of the discrepancies between the models revealed no significant difference in questions answered incorrectly by GPT-3.5 but correctly by GPT-4 (n=94) and vice versa (n=38) across the 4 question types (P=.27 vs P=.95).

For calculation-based questions, GPT-4 showed higher accuracy than that of GPT-3.5 (80% vs 40%, P=.03), with the most pronounced difference in pharmaceutics and biopharmaceutics subjects. In scenario-based questions, GPT-4 also outperformed GPT-3.5 in terms of accuracy (63% vs 44.4%, P=.41), though the difference was nonsignificant.

Discussion

Principal Findings

This study demonstrates that GPT-4 significantly outperformed GPT-3.5 in the Taiwan NPLE, surpassing the passing threshold, especially in basic pharmacy subjects. These subjects, which have only a 13.82% passing rate among human students, are particularly challenging. GPT-4 excelled in areas such as pharmacology, pharmaceutical chemistry, pharmaceutical analysis, and pharmaceutics, consistently providing correct answers and comprehensive explanations. Although GPT-4 also performed better than GPT-3.5 in clinical subjects such as dispensing pharmacy and therapeutics, the performance gap was narrower in these areas.

In specific subjects like pharmacodynamics, pharmacokinetics, and drug-related topics in the autonomic nervous system, GPT-4 consistently provided accurate responses, where GPT-3.5 often faltered. Additionally, GPT-4 exhibited superior accuracy in bioavailability, dosing, and pharmacokinetic calculations. However, GPT-4's accuracy dropped in topics like herbal medicines and pharmacy law, emphasizing the need for further model refinement in these areas [30].

Comparison with Literature

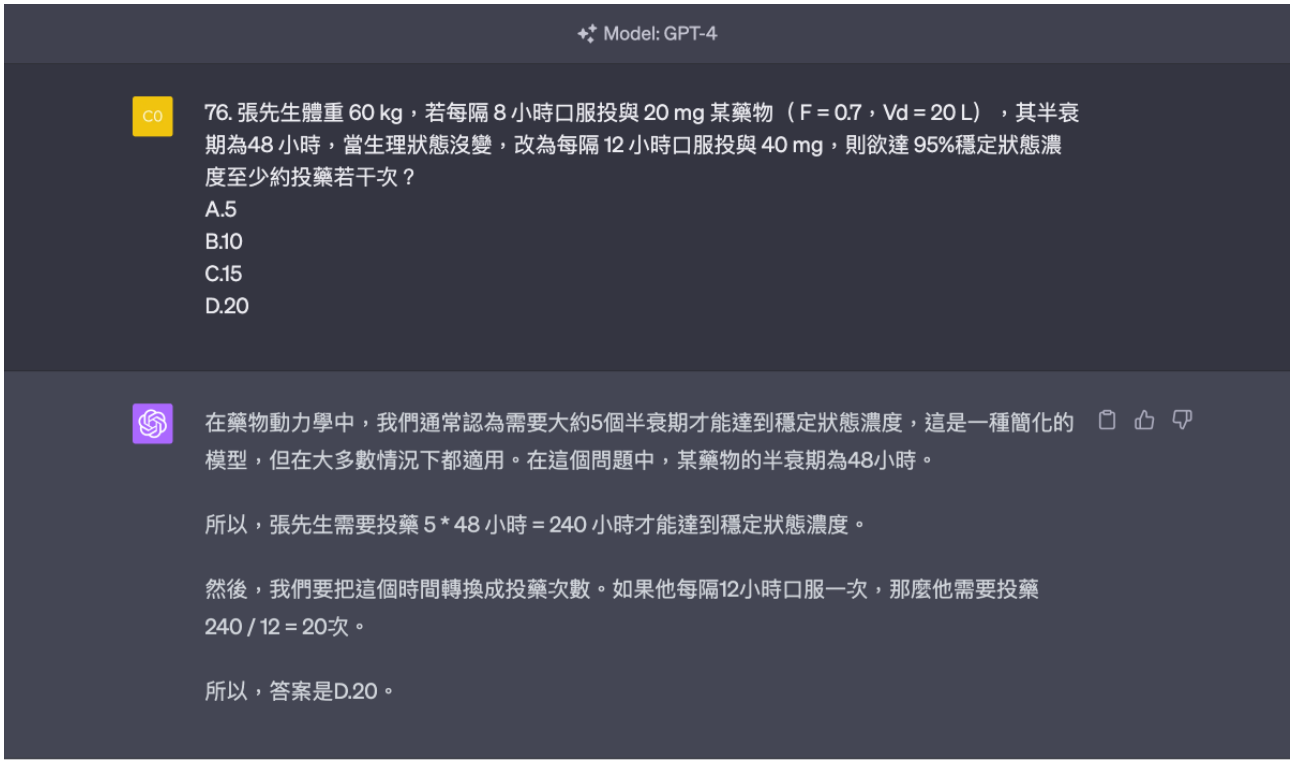
Previous studies have established that GPT-4 consistently outperforms GPT-3.5 in various medical exams, including the Australian Medical Licensing Examination [31], Canadian Radiology Examination [15], Turkish Medical Examination [32], and Japanese Medical Licensing Examination [33]. In many of these examinations, GPT-4 consistently achieved scores above 70% [34-36]. This study aligns with those findings, showing GPT-4's superior performance in the Taiwan NPLE. Unlike prior research that focused on real-world clinical applications [37-43], this study comprehensively assessed the models across various pharmacy domains.

A study by Choi [44] reported that GPT-3.5 performed well on memory-based questions but struggled with problem-solving, whereas GPT-4 demonstrated better performance in comprehension and judgment tasks. Similarly, a radiology study suggested that GPT-4 outperformed GPT-3.5 on higher-order thinking questions but not on lower-order questions [15]. These

findings slightly differ from the results of our study, where GPT-3.5 exhibited higher accuracy in both memory-based (low-level thinking) and reverse (mid-level thinking) questions. However, GPT-4 surpassed GPT-3.5 across all question types, particularly in comprehension (high-level thinking) and memory-based (low-level thinking) questions. In judgment, reverse, and comprehension questions—tasks that demand more advanced reasoning—GPT-4 demonstrated superior accuracy with fewer errors compared to GPT-3.5. Additionally, GPT-4's ability to correct errors made by GPT-3.5 reinforces its potential as a more reliable model for pharmacy-related assessments.

Further, GPT-4 significantly outperformed GPT-3.5 in calculation questions. While GPT-3.5 provided step-by-step explanations but often guessed the final answer—a phenomenon known as hallucination' due to insufficient training—GPT-4 exhibited stronger logical reasoning (Figure 8) with over 80% accuracy. However, it still made errors in 20% of cases, indicating the need for needed during its use [21,45]. In clinical applications, modifying prompts has been shown to improve GPT's accuracy [46]. For integrated analysis questions, GPT-4's performance was slightly better than GPT-3.5, consistent with findings from a nursing licensure examination in Japan [14].

Figure 8. Template of the questions that GPT-4 exhibited stronger logical reasoning.



Implications for Education

The study highlights GPT-4's potential as an educational tool, particularly in pharmacy education. GPT-4 can offer extensive practice opportunities for pharmacy students across both basic and clinical subjects, providing both correct answers and detailed explanations [18,47] to enhance understanding. Given the lower passing rates among pharmacy students in basic subjects among that were challenging, GPT-4 could assist in individualized learning. Its strength in comprehension and integrated analysis questions makes it a valuable resource for fostering critical thinking skills.

Despite its advancements over GPT-3.5, GPT-4's occasional inconsistencies suggest that model stability is not yet perfect. Questions correctly answered by GPT-3.5 were not always consistently answered by GPT-4. Nevertheless, GPT-4's accuracy, approaching 80% suggests that it can serve as an effective learning supplement, provided educators guide students in minimizing potential errors. For instance, specifying clearer prompts, such as "Please do not add your own opinions", may help mitigate hallucinations and enhance its use in educational settings.

In addition, educators should consider adjusting the format of examinations by replacing memory-based questions with comprehension questions, which can reduce the chances of guessing and better assess students' true intelligence.

Limitations

The primary limitation of this study is the time frame during which the models were tested (ie, from May 10 to July 20, 2023), which may affect the reproducibility of the results if retested in the future. Additionally, both GPT-3.5 and GPT-4 struggled with recognizing structural diagrams, limiting their performance in areas such as pharmaceutical chemistry and pharmacognosy. These limitations, consistent with previous research, highlight the need for cautious application of GPT models in fields that require visual recognition [11,48,49]. Additionally, the models showed poorer performance in subjects with less available training data and specific medical knowledge such as pharmacy law and traditional medicine, indicating potential biases in the models' training. We suggest that future

efforts in model development should focus on incorporating more diverse and comprehensive data to reduce such biases.

Conclusions

This study demonstrates that GPT-4 outperforms GPT-3.5 in the Taiwan NPLE, particularly in pharmacy expertise, calculation ability, and situational case studies, with a notable advantage in basic subjects. It is recommended that GPT-4 be applied in clinical pharmacy practice (ie, patient education, drug consultation) and pharmacy education, particularly to support self-directed learning. However, given its limitations, caution is advised when integrating GPT-4 into clinical settings and educational programs. Future research should focus on refining prompts, improving model stability, integrating medical databases, and enhancing comprehensive questions to evaluate student competence more effectively while minimizing the chance of guessing correct answers.

Acknowledgments

This work was supported by Taipei Veterans General Hospital Hsinchu Branch (2024-VHCT-P-0008) and the authors would like to thank Wallace Academic Editing (https://www.editing.tw/) for English language editing.

Conflicts of Interest

None declared.

References

1. ChatGPT: optimizing language models for dialogue. OpenAI. URL: https://chatgpt.r4wand.eu.org/ [accessed 2023-03-03]
2. Research index. OpenAI. URL: https://openai.com/research/gpt-4 [accessed 2023-08-03]
3. OpenAI. GPT-4 technical report. arXiv. Preprint posted online on Mar 4, 2024. [doi: 10.48550/arXiv.2303.08774]
4. Kosinski M. Evaluating large language models in theory of mind tasks. Proc Natl Acad Sci U S A 2024 Nov 5;121(45):e2405460121. [doi: 10.1073/pnas.2405460121] [Medline: 39471222]
5. Wang YM, Chen TJ. ChatGPT surges ahead: GPT-4 has arrived in the arena of medical research. J Chin Med Assoc 2023 Sep 1;86(9):784-785. [doi: 10.1097/JCMA.0000000000000955] [Medline: 37406215]
6. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
7. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Dig Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
8. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination? A descriptive study. J Educ Eval Health Prof 2023;20:1. [doi: 10.3352/jeehp.2023.20.1] [Medline: 36627845]
9. Fijačko N, Gosak L, Štiglic G, Picard CT, John Douma M. Can ChatGPT pass the life support exams without entering the American Heart Association course? Resuscitation 2023 Apr;185:109732. [doi: 10.1016/j.resuscitation.2023.109732] [Medline: 36775020]
10. Weng TL, Wang YM, Chang S, Chen TJ, Hwang SJ. ChatGPT failed Taiwan's Family Medicine Board Exam. J Chin Med Assoc 2023 Aug 1;86(8):762-766. [doi: 10.1097/JCMA.0000000000000946] [Medline: 37294147]
11. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. Ophthalmol Sci 2023 Dec;3(4):100324. [doi: 10.1016/j.xops.2023.100324] [Medline: 37334036]
12. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv. Preprint posted online on Apr 12, 2023. [doi: 10.48550/arXiv.2303.13375]
13. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. Sci Rep 2023 Nov 22;13(1):20512. [doi: 10.1038/s41598-023-46995-z] [Medline: 37993519]
14. Kaneda Y, Takahashi R, Kaneda U, et al. Assessing the performance of GPT-3.5 and GPT-4 on the 2023 Japanese Nursing Examination. Cureus 2023 Aug;15(8):e42924. [doi: 10.7759/cureus.42924] [Medline: 37667724]
15. Bhayana R, Bleakney RR, Krishna S. GPT-4 in radiology: improvements in advanced reasoning. Radiology 2023 Jun;307(5):e230987. [doi: 10.1148/radiol.230987] [Medline: 37191491]
16. Oh N, Choi GS, Lee WY. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. Ann Surg Treat Res 2023 May;104(5):269-273. [doi: 10.4174/astr.2023.104.5.269] [Medline: 37179699]
17. Sabry Abdel-Messih M, Kamel Boulos MN. ChatGPT in clinical toxicology. JMIR Med Educ 2023 Mar 8;9:e46876. [doi: 10.2196/46876] [Medline: 36867743]

18. Nisar S, Aslam MS. Is ChatGPT a good tool for T&CM students in studying pharmacology? SSRN. Preprint posted online on Jan 17, 2023. [doi: 10.2139/ssrn.4324310]
19. Wang YM, Shen HW, Chen TJ. Performance of ChatGPT on the Pharmacist Licensing Examination in Taiwan. J Chin Med Assoc 2023 Jul 1;86(7):653-658. [doi: 10.1097/JCMA.0000000000000942] [Medline: 37227901]
20. He N, Yan Y, Wu Z, et al. Chat GPT-4 significantly surpasses GPT-3.5 in drug information queries. J Telemed Telecare 2023 Jun 22:1357633X231181922. [doi: 10.1177/1357633X231181922] [Medline: 37350055]
21. Li D, Yu J, Hu B, Xu Z, Zhang M. ExplainCPE: A free-text explanation benchmark of Chinese Pharmacist Examination. arXiv. Preprint posted online on Oct 26, 2023. [doi: 10.48550/arXiv.2305.12945]
22. Vert JP. How will generative AI disrupt data science in drug discovery? Nat Biotechnol 2023 Jun;41(6):750-751. [doi: 10.1038/s41587-023-01789-6] [Medline: 37156917]
23. Blanco-González A, Cabezón A, Seco-González A, et al. The role of AI in drug discovery: challenges, opportunities, and strategies. Pharmaceuticals (Basel) 2023 Jun 18;16(6):891. [doi: 10.3390/ph16060891] [Medline: 37375838]
24. Savage N. Drug discovery companies are customizing ChatGPT: here's how. Nat Biotechnol 2023 May;41(5):585-586. [doi: 10.1038/s41587-023-01788-7] [Medline: 37095351]
25. Wang H, Ding YJ, Luo Y. Future of ChatGPT in pharmacovigilance. Drug Saf 2023 Aug;46(8):711-713. [doi: 10.1007/s40264-023-01315-2] [Medline: 37306853]
26. Carpenter KA, Altman RB. Using GPT-3 to build a lexicon of drugs of abuse synonyms for social media pharmacovigilance. Biomolecules 2023 Feb 18;13(2):387. [doi: 10.3390/biom13020387] [Medline: 36830756]
27. Cloesmeijer ME, Janssen A, Koopman SF, Cnossen MH, Mathôt RAA, consortium S. ChatGPT in pharmacometrics? Potential opportunities and limitations. Br J Clin Pharmacol 2024 Jan;90(1):360-365. [doi: 10.1111/bcp.15895] [Medline: 37621112]
28. Sallam M, Salim N, Barakat M, Al-Tammemi A. ChatGPT applications in medical, dental, pharmacy, and public health education: a descriptive study highlighting the advantages and limitations. Narra J 2023 Mar 29;3(1):e103. [doi: 10.52225/narra.v3i1.103]
29. Zhu Y, Han D, Chen S, Zeng F, Wang C. How can ChatGPT benefit pharmacy: a case report on review writing. Preprints. Preprint posted online on Feb 20, 2023. [doi: 10.20944/preprints202302.0324.v1]
30. Hsu HY, Hsu KC, Hou SY, Wu CL, Hsieh YW, Cheng YD. Examining real-world medication consultations and drug-herb interactions: ChatGPT performance evaluation. JMIR Med Educ 2023 Aug 21;9:e48433. [doi: 10.2196/48433] [Medline: 37561097]
31. Kleinig O, Gao C, Bacchi S. This too shall pass: the performance of ChatGPT-3.5, ChatGPT-4 and New Bing in an Australian Medical Licensing Examination. Med J Aust 2023 Sep 4;219(5):237. [doi: 10.5694/mja2.52061] [Medline: 37528548]
32. Kılıç ME. AI in medical education: A comparative analysis of GPT-4 and GPT-3.5 on turkish medical specialization exam performance. medRxiv. Preprint posted online on Jul 12, 2023. [doi: 10.1101/2023.07.12.23292564]
33. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: comparison study. JMIR Med Educ 2023 Jun 29;9:e48002. [doi: 10.2196/48002] [Medline: 37384388]
34. Guerra GA, Hofmann H, Sobhani S, et al. GPT-4 artificial intelligence model outperforms ChatGPT, medical students, and neurosurgery residents on neurosurgery written board-like questions. World Neurosurg 2023 Nov;179:e160-e165. [doi: 10.1016/j.wneu.2023.08.042] [Medline: 37597659]
35. Lewandowski M, Łukowicz P, Świetlik D, Barańska-Rybak W. ChatGPT-3.5 and ChatGPT-4 dermatological knowledge level based on the Specialty Certificate Examination in dermatology. Clin Exp Dermatol 2024 Jun 25;49(7):686-691. [doi: 10.1093/ced/llad255] [Medline: 37540015]
36. Khorshidi H, Mohammadi A, Yousem DM, et al. Application of ChatGPT in multilingual medical education: How does ChatGPT fare in 2023's Iranian Residency Entrance Examination. Inform Med Unlocked 2023;41:101314. [doi: 10.1016/j.imu.2023.101314]
37. Huang X, Estau D, Liu X, Yu Y, Qin J, Li Z. Evaluating the performance of ChatGPT in clinical pharmacy: a comparative study of ChatGPT and clinical pharmacists. Br J Clin Pharmacol 2024 Jan;90(1):232-238. [doi: 10.1111/bcp.15896] [Medline: 37626010]
38. Jairoun AA, Al-Hemyari SS, Shahwan M, Humaid Alnuaimi GR, Zyoud SH, Jairoun M. ChatGPT: threat or boon to the future of pharmacy practice? Res Social Adm Pharm 2023 Jul;19(7):975-976. [doi: 10.1016/j.sapharm.2023.03.012] [Medline: 37061346]
39. Juhi A, Pipil N, Santra S, Mondal S, Behera JK, Mondal H. The capability of ChatGPT in predicting and explaining common drug-drug interactions. Cureus 2023 Mar;15(3):e36272. [doi: 10.7759/cureus.36272] [Medline: 37073184]
40. Davies NM. Adapting artificial intelligence into the evolution of pharmaceutical sciences and publishing: technological darwinism. J Pharm Pharm Sci 2023;26:11349. [doi: 10.3389/jpps.2023.11349] [Medline: 37034476]
41. Kleebayoon A, Wiwanitkit V. Performance and risks of ChatGPT used in drug information: comment. Eur J Hosp Pharm 2023 Dec 27;31(1):85-86. [doi: 10.1136/ejhpharm-2023-003864] [Medline: 37339863]
42. Mohammed M, Kumar N, Zawiah M, et al. Psychometric properties and assessment of knowledge, attitude, and practice towards ChatGPT in pharmacy practice and education: a study protocol. J Racial Ethn Health Disparities 2024 Aug;11(4):2284-2293. [doi: 10.1007/s40615-023-01696-1] [Medline: 37428357]

43. Abu-Farha R, Fino L, Al-Ashwal FY, et al. Evaluation of community pharmacists' perceptions and willingness to integrate ChatGPT into their pharmacy practice: a study from Jordan. J Am Pharm Assoc 2023 Nov;63(6):1761-1767. [doi: 10.1016/j.japh.2023.08.020]
44. Choi W. Assessment of the capacity of chatgpt as a self-learning tool in medical pharmacology: a study using mcqs. BMC Med Educ 2023 Nov 13;23(1):864. [doi: 10.1186/s12909-023-04832-x] [Medline: 37957666]
45. Snoswell CL, Falconer N, Snoswell AJ. Pharmacist vs machine: pharmacy services in the age of large language models. Res Social Adm Pharm 2023 Jun;19(6):843-844. [doi: 10.1016/j.sapharm.2023.03.006] [Medline: 36907776]
46. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res 2023 Oct 4;25:e50638. [doi: 10.2196/50638] [Medline: 37792434]
47. Krumborg JR, Mikkelsen N, Damkier P, et al. ChatGPT: first glance from a perspective of clinical pharmacology. Basic Clin Pharma Tox 2023 Jul;133(1):3-5. [doi: 10.1111/bcpt.13879]
48. Fergus S, Botha M, Ostovar M. Evaluating academic answers generated using ChatGPT. J Chem Educ 2023 Apr 11;100(4):1672-1675. [doi: 10.1021/acs.jchemed.3c00087]
49. Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT-3.5, ChatGPT-4, and orthopaedic resident performance on orthopaedic assessment examinations. J Am Acad Orthop Surg 2023 Dec 1;31(23):1173-1179. [doi: 10.5435/JAAOS-D-23-00396] [Medline: 37671415]

Abbreviations

GPT-3.5: ChatGPT-3.5
GPT-4: ChatGPT-4
NPLE: National Pharmacist Licensing Examination
USMLE: United States Medical Licensing Examination

Edited by B Lesselroth; submitted 28.01.24; peer-reviewed by G Farid, S Zhai, WY Jen; revised version received 26.09.24; accepted 17.12.24; published 17.01.25.

Please cite as:
Wang YM, Shen HW, Chen TJ, Chiang SC, Lin TG
Performance of ChatGPT-3.5 and ChatGPT-4 in the Taiwan National Pharmacist Licensing Examination: Comparative Evaluation Study
JMIR Med Educ 2025;11:e56850
URL: https://mededu.jmir.org/2025/1/e56850
doi:10.2196/56850

Original Paper

Guidelines for Patient-Centered Documentation in the Era of Open Notes: Qualitative Study

Anita Vanka[1,2], MD; Katherine T Johnston[2,3], MD; Tom Delbanco[1,2], MD; Catherine M DesRoches[1], DrPH; Annalays Garcia[1], MD; Liz Salmi[1], AS; Charlotte Blease[4], PhD

[1]Division of General Medicine, Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA, United States

[2]Harvard Medical School, Boston, MA, United States

[3]Department of Medicine, Massachusetts General Hospital, Boston, MA, United States

[4]Department of Women's and Children's Health, Uppsala University, Uppsala, Sweden

Corresponding Author:
Anita Vanka, MD
Division of General Medicine
Department of Medicine
Beth Israel Deaconess Medical Center
330 Brookline Avenue
Deaconess 301
Boston, MA, 02215-5400
United States
Phone: 1 617 632 8350
Fax: 1 617 632 8261
Email: avanka@bidmc.harvard.edu

Abstract

Background: Patients in the United States have recently gained federally mandated, free, and ready electronic access to clinicians' computerized notes in their medical records ("open notes"). This change from longstanding practice can benefit patients in clinically important ways, but studies show some patients feel judged or stigmatized by words or phrases embedded in their records. Therefore, it is imperative that clinicians adopt documentation techniques that help both to empower patients and minimize potential harms.

Objective: At a time when open and transparent communication among patients, families, and clinicians can spread more easily throughout medical practice, this inquiry aims to develop informed guidelines for documentation in medical records.

Methods: Through a series of focus groups, preliminary guidelines for documentation language in medical records were developed by health professionals and patients. Using a structured focus group decision guide, we conducted 4 group meetings with different sets of 27 participants: physicians experienced with writing open notes (n=5), patients accustomed to reviewing their notes (n=8), medical student educators (n=7), and resident physicians (n=7). To generate themes, we used an iterative coding process. First-order codes were grouped into second-order themes based on the commonality of meanings.

Results: The participants identified 10 important guidelines as a preliminary framework for developing notes sensitive to patients' needs.

Conclusions: The process identified 10 discrete themes that can help clinicians use and spread patient-centered documentation.

(JMIR Med Educ 2025;11:e59301) doi:10.2196/59301

KEYWORDS

open notes; patient-centered documentation skills; medical student education; 21st Century Cures Act

Introduction

Reflecting long-standing tradition, medical record notes documenting clinical encounters have primarily served the doctors or other health professionals preparing them. Their diverse functions include accurate documentation of a patient's unique circumstance, refreshing clinicians' memories, documenting diagnostic reasoning, communicating cogently with colleagues, justifying charges for encounters, and serving as material for assaying quality of care.

Until about 15 years ago, clinicians prepared such notes rarely with patients or their families envisioned as potential recipients [1]. Since the turn of this century, however, the movement toward more open and transparent communication with patients has grown, and since April 2021, federal rules in the United States now mandate that all patients (with very limited permitted exceptions) are offered online and rapid access to their clinical records, including the notes written by clinicians ("open notes") [2,3].

Whether one note can serve diverse recipients remains an open question, however, studies suggest room for optimism. A large majority of clinicians experienced with open notes favor their continuation, and few report "dumbing down" what they write [4]. Moreover, extensive survey and qualitative research demonstrate that patients who review their records and read their notes feel more involved in and knowledgeable about their care, report being better prepared for visits, and indicate they are more likely to follow their clinicians' advice [5-11]. However, words matter, and studies also show that patients can feel judged, stigmatized, or offended by their notes, with potentially adverse effects on the clinician-patient relationship [12-15]. For example, a recent study at 3 diverse health systems found that 1 in 10 patients reported feeling judged or offended by an outpatient note, reflecting the perception that the note contained errors, surprises, inappropriate labeling, or evidence of disrespect [16].

To date, little empirically informed counsel about best practices with respect to patient-centered documentation has been published [17]. Although studies have demonstrated that certain language in notes influences clinician attitudes toward patients and that specific words used can negatively impact the clinician-patient relationship, what language providers should use or avoid has not been clearly described [11,15]. Within the current medical literature, some recommendations have been offered on how clinicians might better prepare for the era of open notes [18]. However, there are no evidence-informed guidelines describing concrete approaches to patient-centered language in clinical notes, and such guidance may be important for developing mindful practices.

In preparation for an educational intervention with medical students and practitioners supervising their work [19], we aimed to address this gap in knowledge and practice by drawing on the perspectives of patients, physicians facile with open notes, medical educators responsible for teaching clinical skills to early learners, and medical residents, who often teach students directly. Following dedicated discussions with these 4 groups and subsequent thematic analysis of their perspectives, we developed a set of guidelines as a preliminary framework for future initiatives aimed at teaching patient-centered documentation skills to medical students, their preceptors, and a broad range of clinicians.

Methods

We conducted 4 focus groups with discrete groups of individuals to gain an understanding of their perspectives and experiences with written medical documentation. Our goal was to develop guiding principles for best practices in patient-centered documentation skills.

Design

Our focus groups addressed experiences with written medical documentation (patient notes), with particular attention to the language used in notes. We sought a range of perspectives and structured the groups based on the type of participant. Within their respective groups, the aim was to create opportunities for interaction and comparison of responses among participants [20,21]. This methodology can generate a large volume of responses, and we anticipated a robust discussion of experiences among participants. Although there is no consensus about the ideal number of participants in focus groups, our goal was to recruit between 6 and 10 participants per group [22]. All study procedures received ethical approval in March 2022 and met exempt status both from the Beth Israel Deaconess Medical Center institutional review board (reference number 2022P000079) and the Mass General Brigham institutional review board (reference number 2022P000635).

Recruitment and Participants

We convened four groups of participants: (1) patients familiar with open notes, (2) practicing physicians facile with open notes, (3) medical student educators involved with teaching clinical skills, and (4) resident physicians working closely with medical students. Between March and April 2022, we used various forms of outreach to recruit participants: For the patient focus group, LS identified, recruited, and contacted patients through email with a flyer attachment. For the group of practicing physicians, TD identified and AG recruited individuals having active experiences with open notes at academic institutions across the country. For the group of medical student teachers, AV and KJ recruited physician educators through the national medical educator list, which included Clerkship Directors in Internal Medicine and Directors of Clinical Skills. For the focus group of residents, AV and KJ identified and contacted through email residents from Beth Israel Deaconess Medical Center and Massachusetts General Hospital who had expressed interest in medical student education. For their time, patient participants were offered a US $100 honorarium, and physicians a US $50 honorarium.

Between April and May 2022, we convened four 90-minute focus groups, each composed of specific types of participants as described above. Focus groups were conducted through Zoom (Zoom Video Communications) and recorded for transcription.

Format of Focus Groups

Each group followed the same script, with the meetings facilitated by 2 of the study leaders (AV and KJ). Drawing on a variety of stakeholder perspectives, the research team collectively devised the structured interview script. The team included medical educators, practicing physicians, a health services researcher, a patient advocacy researcher, a health services researcher, and a medical ethicist. The script was developed based on the group's experiential, ethical, and practical experience. To assess face validity, the questions were further honed, refined, and pretested with 4 doctors and 3 patients, leading to further refinements in the wording of

questions and prompts (Textbox 1). Groups opened with a standard anonymity disclosure, along with a description of the reason behind the inquiries. We offered participants the opportunity to opt out at any time during the process. Participants were encouraged to set their screen name, however, they felt comfortable and to keep their video off if so desired. To preserve anonymity, we asked participants not to mention each other's names during the session.

Textbox 1. Focus group questions for patient and physician groups.

(10 minutes per question)

Question 1: What do you recall from any previous learning experiences that focused on writing notes that patients will read, or reading notes from the perspective of a patient?

- For the patient group: "What do you recall from your experience of reading your clinical notes?"
- Follow-up question: What are some concrete examples of this?

Question 2: What should early medical students know about writing notes that will be useful to patients and encourage partnering or engagement in their care?

- Follow-up question: What are some concrete examples of this?

Question 3: What should early medical students know about writing notes with words or phrases that could be harmful to patients or their relationship with their physician?

- Follow-up question: What are some concrete examples of this?

Question 4: What should early medical students know regarding how medical vernacular or acronyms may be perceived by patients?

- For patient group: "What has been your experience when reading notes containing medical vernacular or acronyms?"
- Follow-up questions:
- What are examples of vernacular that should be avoided due to possible patient harm or creating unwanted bias?
- Are there common acronyms that should be avoided?

Question 5: What should early medical students know about including the patient's voice in the notes?

- For patient group: Describe some ways in which a physician could write your words, or ensure that the writing of your medical concerns represents your lived experience.
- Follow-up question:
- What documentation approaches could a student follow to help the patient feel authentically seen and heard when reading their clinical notes?

Question 6: What should early medical students know about how words or phrases in clinical notes may convey bias?

- For patient group: "What has been your experience in reading notes with words or phrases that you feel convey bias?"
- Follow-up question: What are some concrete examples of this?

Question 7: What key content areas are more sensitive for patient readers and should require students to receive specific guidance on documentation?

- Follow-up question:
- What topics or themes should receive special attention and teaching?
- Are there any key areas that should be omitted or avoided in notes unless discussed and reviewed with faculty?
- Prompt examples: race, obesity, firearm ownership, gender identity and health, sexual identity and health, substance use, and so on.

(5 minutes for wrap-up and debrief and closing thoughts)

Analysis

We designed and conducted an inductive thematic content analysis of the transcribed focus groups [23]. This approach was used because it is particularly appropriate for analyzing textual data by identifying patterns, themes, or categories that emerge from the data itself, rather than imposing predetermined categories or codes. This approach facilitates the identification of nuanced themes arising directly from the data, making it particularly useful when exploring new or less understood phenomena [24]. Responses were analyzed by 2 members of the research team (AV and KJ). Both are medical educators and general internists in the United States with experience in sharing online access to patients' health records (AV is an inpatient physician and KJ is a primary care physician). The research team was diverse in age and background: the lead author and one of the coders (AV) identified as of Indian background, another author (AG) identified as having Cuban descent, and

the patient-researcher author (LS) identified as having both physical and cognitive disabilities.

First, AV and KJ read the transcripts to familiarize themselves with the responses. Second, AV and KJ independently created codes through the selection of excerpts from each of the 4 focus groups. The codes were then reviewed jointly by AV and KJ to come to a consensus. Using the common list, new excerpts from each of the 4 transcripts were coded by AV and KJ, and the codes were further refined until a consensus was reached. Subsequently, first-order codes were grouped into second-order themes based on commonality of meaning. Representative comments for each theme were identified by authors AV and AG.

Ethical Considerations

All study procedures received ethical approval in March 2022 and met exempt status both from the Beth Israel Deaconess Medical Center institutional review board (reference number 2022P000079) and the Mass General Brigham institutional review board (reference number 2022P000635). This study did not meet criteria for human subjects research at either institution and thereby was deemed exempt.

Results

Overview

A total of 27 individuals participated in the focus groups: 8 patients (5 women, 2 men, and 1 individual identifying as gender nonbinary), 5 physicians with experience preparing open notes (2 women and 3 men), 7 medical student educators (6 women and 1 man), and 7 resident physicians (6 women and 1 man). The participants from the patient, physician, and medical educator groups were from diverse geographical regions of the United States. Based at large academic centers, physician participants represented different clinical specialties including internal medicine, pediatrics, and behavioral health. Educator participants were all involved with either leadership or teaching foundational clinical skills at various medical schools in the country. The resident physician participants were recruited from 2 large academic health centers in the greater Boston area.

Using the iterative coding process [25], we identified 10 major themes that could serve as guidelines for patient-centered documentation (Textbox 2). We discuss these in greater detail below, with representative comments illustrative for each theme in Table 1.

Textbox 2. Checklist of guidelines for patient-centered documentation: major themes.

Themes

- Use person-first language.
- Refer to your patients as how they want to be identified.
- Avoid abbreviations and acronyms, especially if not officially approved by the practice.
- Say what you write and write what you say.
- Verify past history information before recording it in the note.
- Avoid words that may convey bias or judgment.
- Keep descriptions of physical examinations objective.
- Empower your patients with encouraging words and clear next steps.
- Pay close attention to sensitive topics, including but not limited to sexual history, trauma history, substance use history, mental health, or illness.
- Write from your perspective.

Table 1. Examples and suggestions for each identified best practice for patient-centered documentation.

Theme	Examples and suggestions
Use person-first language	• I don't see necessarily a boundary between the language you use to describe a patient out loud and what you write in your note. It's all part of a single way that you frame patients. It's not a diabetic patient. It's a patient with diabetes. I really don't think about the way that I document my notes differently than the way that I just think about my patients, which is to think about them in a person-centered way. [#2, Medical Educator] • Even though I've lost weight, the word morbidly obese is in every note, every note I see. That doesn't bring me joy or comfort me or get me to want to interact in a positive manner. My issues have nothing to do with my weight and never did. I developed a MRSA staph infection; it was a healthcare-acquired infection.There are certain things where weight doesn't always factor in, and you don't always need weight. [#1, Patient]
Refer to your patients as how they want to be identified	• I have found the doctors that I have currently are really good about using he/him pronouns for me and referring to me as using masculine identifying language. I have had providers who are not. Never knowing what I'm going to run into in those notes is very anxiety-inducing on the gender part alone, right? Let alone, does this person think that my gender is influenced by or influencing any of my other health issues? [#5, Patient] • I tend to write and call the patient whatever they like to go by, so I'll say like, Bill is a 75-year-old man, instead of Mr. Smith, for instance. For nonbinary or transgender patients, I try to ask them what gender they would like me to document rather than just assuming it's transgender male. Then in the social history and medical history, perhaps elaborating on it more, but trying to give them voice in what I'm documenting. I also think that mentioning race without having any clear connection to anything should be discouraged completely. [#1, Resident]
Avoid abbreviations and acronyms, especially if not officially approved by the practice	• I was struck by the idea that it almost felt, as a medical student, as if you were being inducted in a secret society. It was this secret language that you now all understood, and that's what unified everyone who was part of that. It really is just so unnecessary at this point. I think what's already come out here is that there clearly are regional differences. I've never heard of "MOP" ["mother of patient"], but "ISO"["in the setting of"] is rampant around here, so clearly, there are differences in terms of some of the abbreviations that are used in specific regions and areas, and we can't then even understand each other's language. [#2, Physician] • Just the other day, we were talking about PSA and how it's prostate-specific antigen, but also pseudomonas, and also pseudoaneurysm. If a patient reads PSA and looks that up for themselves on Google, they're going to find a million things that could be. If you're going to use an acronym, maybe don't use ones that have multiple different meanings even in the medical world, and the most important parts of their diagnoses should not be abbreviated. [#1, Resident] • Medical jargon and all this evolved as a way of having more succinct communication, being able to communicate among clinical teams. I think there is some value to that, but I think what's important for a lot of educators to keep in mind is, this is an area where the students are actually—while they're learning, they also should, in some ways, be teachers for all of us who have been doing this for some time, right, and how they can take their lack of being indoctrinated by the system and bring that from the bottom up. [#3, Physician] • Then some acronyms. Point out we say SOB for shortness of breath, and, obviously, has other meanings. [#2, Resident]
Say what you write and write what you say	• There probably shouldn't be anything in the note that hasn't been discussed. If you're telling someone a diagnosis, I don't think you should go back in the note and say, "And this diagnosis is terminal" if you haven't discussed it with the patient. Knowing that the patient is likely going to go back and read the note, I don't think that it should be put in the note. Or, if the physician feels like it's important that it's put in the note, make sure it's discussed with the patient as well. There's no point in writing a good note if none of it was even said to the patient. [#7, Patient] • When we have a discussion and he puts it right in my note right then and there, those are kinds of things that help me. Today I had a call from the pharmacist about a med. They would have given me the wrong med if I wouldn't have stood up for myself and wouldn't have known what I was supposed to have, and wouldn't have had my notes there, because basically I read it right from my note to him. I think the notes become more and more valuable. [#1, Patient]

Theme	Examples and suggestions
Verify past history information before recording it in the note	• We've evolved as people from year 1 to year 5 over the course of a relationship of knowing a doctor, and our social histories often don't reflect that change. What could have been pertinent 5 years ago for a patient might not be anymore, and depending on what they were experiencing at that point in time, could be biasing. [#1, Resident] • So now it's on my to-do list when I go to appointments to make sure that's changed. Really, that shouldn't be a priority of mine when we go into medical appointments to say, "Hey, I read the note, and this needs to be changed, 'cause this doesn't represent who I am.' I think everyone has assumed the person before them has done it, and it's accurate. It takes a lot of time and a lot of effort and a lot of energy for us to be able to correct mistakes that are in our patient portals, our notes. I think it's 1 of those things where I just want to remind medical students, you can do 2 things to check with patients. Like, "Hey, I have all this down. Is that right?" Two, "If you see anything in here that's wrong, send me a MyChart message or EHR or whatever message, so I can fix it. [#5, Patient] • Some things that I notice that get copy-forwarded a lot in notes that create bias in the reader, 1 would be substance use disorders, and not really specifying or clarifying it, right? It will just say, "Polysubstance use disorder." Things like that, I think definitely have the ability to bias. [#1, Resident]
Avoid words that may convey bias or judgment	• You don't need to use the word complain. It's a pretty negative connotation. Patient "reports", or just "is having". I think avoiding the word complaint is pretty important." [#2, Resident] • I think, if you're ever thinking about using a term like that, it should be more about, what are the barriers this patient has to achieving care, and talking more about that and not just using 1 word. Using more language is actually helpful, and med students often have more time to actually ask patients about these things and have a conversation. I think that could be a really empowering point for medical students to learn more about the determinants of health that are affecting the patient more so than being like, "This patient doesn't take their medications. [#5, Resident] • I also think that mentioning race without having any clear connection to anything is—should be discouraged completely. It's such a part of vernacular, especially more physicians who are trained in a different era, so we've been actively encouraging students not to include that unless it has a very clear reason. [#2, Medical Educator] • I could imagine, sometimes, quotation marks are used to just directly capture what the patient said. I think it's about striking that balance of bringing the patient language in and having some of it as verbatim as directly as possible so that the patient can see what they said was actually recorded, but then a key part of what we're doing in notes is really clinical translation. It's about going from verbatim conversation to clinical processing and then recording our clinically processed thoughts. [#3, Physician]
Keep physical examination descriptions objective	• General appearance is very tricky, and I think we all have to ask ourselves, when is that relevant and why, and then really distill it down to what's clinically meaningful in the least judgmental way possible. Is it important that the patient is disheveled or older than stated age or has poor hygiene? It might be, but then you have to figure out, how do you relay that in the most respectful manner possible? [#5, Medical Educator]
Empower your patients with encouraging words and clear next steps	• There's a book called "The 15 Minute Hour" that's geared toward primary care providers, and 1 of my favorite takeaways from it that has always stuck with me over the years is the idea of using the word "yet" at the end of a sentence or a phrase. It's supposed to emphasize that you're not at a dead end. You can still do this. The patient has not been able to quit smoking yet. The patient has not lost weight yet. [#5, Physician] • I somehow think that I would like medical notes to be imbued with a different perspective, that we're trying to teach patients to be friends with their disease…[…]…disease is not the enemy. [#4, Patient] • I have heard from some patients that we had engaged in providing feedback on our student notes that they would really like, in reading the assessment and plan, to have some understanding of what they can do next and what the evaluation, diagnostic evaluation plan means for them. Maybe some additional content around health coaching and steps that they can take in their lives to support their health too. It could be a great way to extend that partnership beyond the visit. [#5, Medical Educator]
Pay close attention to sensitive topics, including but not limited to sexual history, trauma history, substance history, mental health or illness	

Theme	Examples and suggestions
	• Some advice I've gotten from the psychiatry consulting team at my primary care practice is […] to just state the type of trauma it was, the years that it happened, the relation to the abuser, and what the patient went through….[…]…There's no need to go into extreme detail and quotations about what the patient confides with you in clinic about, that you can still have a therapeutic bond…[…]…instead aim to strike a balance of what's useful for other providers and not needed for the patient to be reading about themselves. [#7, Resident] • There are certainly very specific topics around mental health, around sexually transmitted illnesses, reproductive health, substance use, et cetera, that definitely should be addressed in a particular manner that respects the patient's privacy… […]…early learners could certainly benefit from understanding what the implications of access to that information could be. [#2, Physician] • […] documenting your full differential diagnosis, where you might have a patient who is coming in for what seems like a respiratory illness or pneumonia, but maybe under differential, you also have a rare lung disorder or lung cancer. It's trying to strike that balance…[…]…trainees, as part of their training, are taught to document the full differential …[…]…this is an area where we're all still trying to figure out the best practices because you do want to indicate to some degree that you are thinking about these other diagnoses, but in the real world […]you might not start with an initial visit by sharing that this may be cancer, but it may be something you bring up on a follow-up visit when your initial working diagnosis doesn't seem to be correct and it seems more serious than that. I know the official recommendation has been to just document what you discuss…[…]…I think it's more complicated than that. [#3, Physician]
Write from your perspective	• Something I've been experimenting with a lot is using the first person in the assessment and plan, which I don't know if I was ever explicitly told not to in medical school, but I feel like I thought I wasn't supposed to. Over the past year, I've started to come around to this idea of, the objective is the objective, but the assessment and plan is my medical opinion. A lot of times, I'll say, "I'm worried about", or, "I think this might be at play. I think it's unlikely, but maybe cancer. I think anxiety might be driving some of this." I feel like that really couches it as, this is my opinion of what's going on. It doesn't negate what the patient thinks. [#5, Physician]

Use Person-First Language

Across all 4 focus groups, participants agreed that patients should be referred to as individuals with certain clinical conditions, rather than identifying them primarily by their medical condition (Table 1). Some patients identified a preference for the use of discrete numbers when addressing a person's weight, rather than describing the individual as "obese" or even "with obesity," and to be mindful of whether weight needs to be in the note if not relevant to issues being addressed (Table 1). All participants from the patient group described the importance of being mindful of phrases and words in notes that can trigger trauma for patients. Finally, participants in all groups identified descriptions of words and phrases in the medical record considered potentially depersonalizing or dehumanizing.

Refer to Your Patients as How They Want to Be Identified

The participants recommended that patients be identified according to their own expressed preferences. Participants from the resident physicians' group suggested that when first introducing oneself to a patient, one should request and document how the patient wishes to be identified in the record. Patient participants noted that honorifics and gender identity should never be presumed (Table 1). In addition, patient participants explained that patients may want to be identified by their life roles, such as their profession, or background, in addition to honorifics and names. Medical educators recommended that introductory "History of Present Illness" sentences should mention factors important to understanding and planning the care of that person at that point in time, with the understanding that these factors are dynamic and change over time. Physicians familiar with open notes suggested it is important to be thoughtful and deliberate about whether to include demographic factors, epidemiologic factors, and medical and social history in the opening sentences of a note, given the effect such information may have on framing how a patient, family member, or other clinician reads the note.

Avoid Abbreviations and Acronyms, Especially if not Officially Approved by the Practice

Participants in all groups felt that medical shorthand can cause confusion and misinterpretation by patients. Physician, resident, and educator participants noted that while some abbreviations are widely understood within health care, others are interpreted differently within and across a given specialty. They suggested that this, in turn, could lead to clinician confusion, thereby further amplifying the risk of patient confusion (Table 1). Physician participants noted that learning to use medical shorthand in notes was akin to being inducted into a "secret society" (Table 1). Illustrative examples suggested by the 4 focus groups included: "SOB" (shortness of breath), "F/U" (follow-up), "ISO" (in the setting of), "NAEON" (no acute events overnight), and "MOP" (mother of the patient). In addition, physician participants pointed out that shorthand for certain medical diagnoses was problematic, given the frequent lack of clarity. They cited "HFrEF" (heart failure with reduced ejection fraction); "HFpEF" (heart failure with preserved ejection fraction); and "AKI" (acute kidney injury). Several residents discussed the risk that abbreviations and acronyms may perpetuate discriminatory biases within notes. Medical educator participants noted that this was an area in which practicing clinicians can learn from students, given that many have not yet been inducted into contemporary systems of

medical vernacular. Finally, the medical educators noted that even if abbreviations were minimized, notes still may not be completely understandable, underscoring the need for open communication with patients.

Say What You Write and Write What You Say

Patients universally stressed the importance of notes accurately reflecting what was done and discussed during a visit. Residents and physicians emphasized that clinicians should be mindful about not documenting issues, such as possible diagnostics considerations, that were not discussed directly with the patient, and that this principle should be used consistently for all documentation (Table 1). Medical educators identified the importance of guiding learners carefully in interactions between documenting explicit clinical reasoning with robust documentation of differential diagnoses and setting expectations for patients. They suggested that clinicians point out to their patients that some notes are intended to be comprehensive and may include diagnostic possibilities that are unlikely, but nevertheless important to record. In contrast, patients suggested there are situations in which clinicians should not write too much about a potentially sensitive topic that may trigger a harmful response in patients. Instead, patients proposed clinicians might record that a discussion of a "challenging topic" took place during the visit; although details might be excluded, such reference would remind the patient of the conversation. On the other hand, several patient participants noted that accurate and detailed representations of a visit may help empower patients with managing their care, strengthen their agency, and enhance their understanding of medical recommendations.

Verify Past History Information Before Recording it in the Note

Another theme was the importance of verifying patients' past information in the note to avoid "note bloat" (ie, copy and paste from previous notes). Participants from all groups identified "cutting and pasting" as increasing the risk of mistakes. Notably, all patients agreed that the "note bloat" phenomenon can have a negative impact on how patients view their notes, and possibly on their relationship with clinicians, resulting in their feeling the need to advocate for themselves to ensure accurate representation of their stories (Table 1). Several patient participants suggested that copying over all aspects of the social history when not relevant to a specific visit might trigger trauma since this section of the note can replicate and reiterate sensitive information. Two residents noted the issue of "copy-forward" (copying previous information from the record into new notes), thereby propagating bias and stigma, and potentially risking mistrust of the clinician. Medical educators noted that untruthful documentation, at times propagated by copy-and-paste templates, can lead to mistrust in the patient-physician relationship and adversely impact care.

Avoid Words That May Convey Bias or Judgment

With striking overlap across all 4 groups, participants offered many examples of words and phrases that may reflect bias or unfair judgment of patients. Suggestions about common phrases and words to avoid included clinical vernacular such as "complains," "denies," "claims," "refuses," and "endorses." Participants recommended neutral and judgment-free words such as "says," "reports," "does not report," "concerns," and "did not tolerate." Several medical educators suggested renaming "chief complaint" as "chief concern." Participants also discussed problems with labels, such as "poor historian," "noncompliant," "difficult patient," "nonadherent," "uncooperative," "having poor health literacy," and "left AMA." Patient participants agreed on the importance of describing explanations for observed behaviors, rather than using judgmental language, such as, "Mr. X is unable to take his insulin most days of the week due to inability to refrigerate the medication at his place of work." Consideration about how to document race was another common concern. Given that race is without clear relevance to most medical concerns, many believed it should not be included in the written documentation. Finally, the pros and cons of using quotation marks and the patients' own language were discussed. Participants noted that doing so would accurately capture what was said. However, in certain contexts, this might be interpreted as sarcastic, questioning, or making light of what the patient communicated, and participants recommended that quotation marks should be used carefully.

Keep Descriptions of Physical Examinations Objective

Another emergent theme was ensuring that descriptions of physical examinations avoided phrases that could be perceived as offensive. Examples included: "disheveled," "older than stated age," "obese"; and judgmental descriptors, such as "argumentative." Participants recommended omitting descriptors such as "pleasant" or "delightful," in part because the absence of such terminology might inadvertently suggest to the reader that the patient is "unpleasant." Several participants urged considering whether language pertaining to appearance is helpful. All recommended avoiding the words "obese" or "morbidly obese" and to first consider whether the information is relevant to an issue at hand. If so, they recommended quantitative descriptors, for example, BMI or the actual weight.

Empower Patients With Encouraging Words and Clear Next Steps

Participants suggested patients often benefit from reading encouraging words and the next steps regarding their conditions. Physicians noted that empowering language, particularly in the "Assessment and Plan" section of a note, could stimulate patients to engage in their care more actively. One physician suggested adding the word "yet" at the end of the sentence or phrase, signifying progress with a specific goal (Table 1). Patients agreed that empowering notes could facilitate partnership with the medical team. Medical educators suggested adapting notes to ensure they capture expert or emerging clinical reasoning, while still engaging patients in the next steps of a plan. Relatedly, several participants suggested using notes to embed reminders to patients, their care partners, and other members of their care team by stating explicitly what was recommended during the visit, especially with respect to specific next steps and goals.

Pay Close Attention to Sensitive Topics, Including but not Limited to Sexual History, Trauma History, Substance Use History, Mental Health, or Illness

Another emergent theme was mindfulness in documenting topics such as substance use history, mental health and illness, gender identity, sexual identity and health, history of trauma, disagreements between patients and clinicians, significant illnesses, and comprehensive diagnostic reasoning. Medical educators noted the challenges of teaching early learners how best to document such information in the social history section. Physicians noted the importance of documenting any disagreements with patients in an objective and respectful manner.

Write From Your Perspective

Medical notes, in particular the "Assessment and Plan" portion of the note, reflect the perspective of the health care professional at a given point in time. Participants uniformly noted that the patient's perspective should also be included, and that descriptions should be factual and based on direct observations. Physician and resident participants alike suggested the use of the first-person pronoun "I" when writing the "Assessment and Plan," as well as using the phrase "at this point in time." Physicians advised that this wording signals that diagnostic considerations may change as a situation evolves.

Discussion

Principal Results

At a time of increasingly open and transparent communication between patients and clinicians, a growing body of research demonstrates the importance of words and phrases used in clinical notes, and the risk of bias and offense to patients. Patients who read their notes feel more involved in their care, better prepared for visits, and are more likely to follow their clinician's recommendations [4-10]. Just as importantly, patients can experience negative effects from their notes due to language perceived as judgmental, offensive, or stigmatizing [11-14]. Language in notes can also negatively influence other clinicians reading the note, leading to bias and impact on the patient's care [11]. To our knowledge, this is the first study attempting to define concrete guiding principles on best practices in patient-centered documentation. Informed by a qualitative analysis of active and interactive discussion among patients and clinicians, this inquiry, as described in detail above and in Textbox 2 and Table 1, identified 10 discrete guidelines for patient-centered documentation.

This inquiry was stimulated by our interest in facilitating open and transparent communication among health professionals and patients from the very beginning of a clinician's career. Medical school curricula and residency programs are just beginning to introduce the concept of patient-centered documentation, with few providing specific guidance [26-28]. Concrete advice is especially important for early medical learners. Moreover, many faculty and residents who teach medical students are also relatively new to the practice of open notes. Developing and describing guidelines, accompanied by clear and detailed examples, can help faculty both learn what we hope will evolve as best practices and teach students these skills more easily, creating an opportunity for standardized assessment of notes based on a checklist reflecting the guidelines. In addition to providing direct teaching and feedback, the presence of concrete guiding principles will allow faculty to role model these skills in patient interactions for their learners. Furthermore, given the importance and impact of language in documentation, we believe the guidelines identified in this research will serve all clinicians well.

The checklist this study generated has already been implemented within the first-year foundational clinical skills course students undertake at our medical school [19]. Students are introduced to the background and value of open notes, followed by the checklist of guidelines. Over the timeframe of this course and beyond into their core clinical training, students are expected to write patient notes using these guides and are subsequently assessed on the quality of their notes through a rubric based on the checklist. Faculty preceptors in this first-year course are also encouraged to use this checklist, with the aim of reviewing student notes and providing feedback based on these 10 themes.

Limitations

Our study has several limitations. First, owing to time constraints on the duration of the study caused in part by the COVID-19 pandemic, we did not conduct focus groups until data saturation was reached. Second, due to the need to develop this preliminary framework before the delivery of a patient-centered documentation curriculum for students, the data analysis was restricted to 2 reviewers only, limiting the validation of the data. Third, the small number of participants likely influenced the results. Fourth, and relatedly, a fuller spectrum of diversity including geography, clinical background, and health background, could produce more varied views and depth of experience with open notes. Fifth, studies have demonstrated that fewer patients with low income, limited education, or poor English proficiency are active users of the electronic health record, limiting our participants' understanding of the note-reading experience of these populations [29-31]. Our preliminary guidelines may therefore be limited when it comes to generalizing to wider patient populations, and future guidelines should seek to address this concern. Finally, the development of new charting practices and auditing may come with considerable resource demands. Typically, guidelines that assess resource-intensive interventions are more likely to include an assessment of implementation costs, real and intangible. It would also be useful to discuss the potential risks of open notes and any context-specific adjustments for special populations or clinics. This is, however, beyond the scope of this study.

Conclusions

The guidelines identified are preliminary. As our understanding grows, we expect clinicians will learn how to write notes in ways that patients find increasingly useful. In addition, there are nuanced areas to consider in various specialties and different patient populations that may require further adaptation of these guidelines. To this end, clinicians and patient advocates partnering to co-design medical education will be important, as it was in this study. Ideally, in the future, we will learn to teach practitioners effective documentation through a common

language and shared set of standard expectations. Although we expect norms and practice techniques to evolve, our study presents an initial attempt to develop practical and respectful guidelines for patient-centered documentation.

Acknowledgments

We thank the patients, physicians, and medical educators who took the time to share their experiences and perspectives, which helped our work. The authors also thank Alex Duncan who coordinated the logistics of the focus groups. This study is part of a project developing a patient-centered documentation curriculum for medical students and their prime clinical faculty. It is funded by an educational grant from the Josiah Macy Jr Foundation and a generous gift from Kate and Arnold Schmeidler. These funds supported honoraria for the focus group participants, of which there were a total of 27.

Authors' Contributions

AV, CB, KJ, TD, and CD performed conceptualization and visualization. CB and TD handled supervision. AV and KJ conducted data analysis. AV, KJ, and CB performed investigations. AV, KJ, AG, and LS handled project administration. AV and CB conducted writing–original draft preparation. AV, CB, KJ, CD, AG, TD, and LS performed writing–review and editing.

Conflicts of Interest

None declared.

References

1. Delbanco T, Walker J, Darer JD, Elmore JG, Feldman HJ, Leveille SG, et al. Open notes: doctors and patients signing on. Ann Intern Med 2010;153(2):121-125 [FREE Full text] [doi: 10.7326/0003-4819-153-2-201007200-00008] [Medline: 20643992]
2. Hudson KL, Collins FS. The 21st Century Cures Act - a view from the NIH. N Engl J Med 2017;376(2):111-113 [FREE Full text] [doi: 10.1056/NEJMp1615745] [Medline: 27959585]
3. Salmi L, Blease C, Hägglund M, Walker J, DesRoches CM. US policy requires immediate release of records to patients. BMJ 2021;372:n426. [doi: 10.1136/bmj.n426] [Medline: 33602667]
4. DesRoches CM, Leveille S, Bell SK, Dong ZJ, Elmore JG, Fernandez L, et al. The views and experiences of clinicians sharing medical record notes with patients. JAMA Netw Open 2020;3(3):e201753. [doi: 10.1001/jamanetworkopen.2020.1753] [Medline: 32219406]
5. Gerard M, Fossa A, Folcarelli PH, Walker J, Bell SK. What patients value about reading visit notes: a qualitative inquiry of patient experiences with their health information. J Med Internet Res 2017;19(7):e237 [FREE Full text] [doi: 10.2196/jmir.7212] [Medline: 28710055]
6. Gerard M, Chimowitz H, Fossa A, Bourgeois F, Fernandez L, Bell SK. The importance of visit notes on patient portals for engaging less educated or nonwhite patients: survey study. J Med Internet Res 2018;20(5):e191 [FREE Full text] [doi: 10.2196/jmir.9196] [Medline: 29793900]
7. Bell SK, Folcarelli P, Fossa A, Gerard M, Harper M, Leveille S, et al. Tackling ambulatory safety risks through patient engagement: what 10,000 patients and families say about safety-related knowledge, behaviors, and attitudes after reading visit notes. J Patient Saf 2021;17(8):e791-e799. [doi: 10.1097/PTS.0000000000000494] [Medline: 29781979]
8. Walker J, Leveille S, Bell S, Chimowitz H, Dong Z, Elmore JG, et al. OpenNotes after 7 years: patient experiences with ongoing access to their clinicians' outpatient visit notes. J Med Internet Res 2019;21(5):e13876 [FREE Full text] [doi: 10.2196/13876] [Medline: 31066717]
9. Mishra VK, Hoyt RE, Wolver SE, Yoshihashi A, Banas C. Qualitative and quantitative analysis of patients' perceptions of the patient portal experience with openNotes. Appl Clin Inform 2019;10(1):10-18 [FREE Full text] [doi: 10.1055/s-0038-1676588] [Medline: 30602196]
10. Nazi KM, Turvey CL, Klein DM, Hogan TP, Woods SS. VA openNotes: exploring the experiences of early patient adopters with access to clinical notes. J Am Med Inform Assoc 2015;22(2):380-389. [doi: 10.1136/amiajnl-2014-003144] [Medline: 25352570]
11. Wolff JL, Darer JD, Berger A, Clarke D, Green JA, Stametz RA, et al. Inviting patients and care partners to read doctors' notes: OpenNotes and shared access to electronic medical records. J Am Med Inform Assoc 2017;24(e1):e166-e172 [FREE Full text] [doi: 10.1093/jamia/ocw108] [Medline: 27497795]
12. P Goddu A, O'Conor KJ, Lanzkron S, Saheed MO, Saha S, Peek ME, et al. Do words matter? Stigmatizing language and the transmission of bias in the medical record. J Gen Intern Med 2018;33(5):685-691 [FREE Full text] [doi: 10.1007/s11606-017-4289-2] [Medline: 29374357]
13. Gudzune KA, Bennett WL, Cooper LA, Bleich SN. Patients who feel judged about their weight have lower trust in their primary care providers. Patient Educ Couns 2014;97(1):128-131 [FREE Full text] [doi: 10.1016/j.pec.2014.06.019] [Medline: 25049164]

14. Stanford FC, Kyle TK. Respectful language and care in childhood obesity. JAMA Pediatr 2018;172(11):1001-1002 [FREE Full text] [doi: 10.1001/jamapediatrics.2018.1912] [Medline: 30193352]
15. Bell SK, Mejilla R, Anselmo M, Darer JD, Elmore JG, Leveille S, et al. When doctors share visit notes with patients: a study of patient and doctor perceptions of documentation errors, safety opportunities and the patient-doctor relationship. BMJ Qual Saf 2017;26(4):262-270 [FREE Full text] [doi: 10.1136/bmjqs-2015-004697] [Medline: 27193032]
16. Fernández L, Fossa A, Dong Z, Delbanco T, Elmore J, Fitzgerald P, et al. Words matter: what do patients find judgmental or offensive in outpatient notes? J Gen Intern Med 2021;36(9):2571-2578 [FREE Full text] [doi: 10.1007/s11606-020-06432-7] [Medline: 33528782]
17. Beauchamp T, Childress J. Principles of biomedical ethics: marking its fortieth anniversary. Am J Bioeth 2019;19(11):9-12. [doi: 10.1080/15265161.2019.1665402] [Medline: 31647760]
18. Blease C, McMillan B, Salmi L, Davidge G, Delbanco T. Adapting to transparent medical records: international experience with "open notes". BMJ 2022;379:e069861. [doi: 10.1136/bmj-2021-069861] [Medline: 36410770]
19. Eng K, Johnston K, Cerda I, Kadakia K, Mosier-Mills A, Vanka A. A patient-centered documentation skills curriculum for preclerkship medical students in an open notes era. MedEdPORTAL 2024;20:11392 [FREE Full text] [doi: 10.15766/mep_2374-8265.11392] [Medline: 38533390]
20. McLafferty I. Focus group interviews as a data collecting strategy. J Adv Nurs 2004;48(2):187-194. [doi: 10.1111/j.1365-2648.2004.03186.x] [Medline: 15369499]
21. Morgan DL, Core SRM. Focus Groups as Qualitative Research. Thousand Oaks, CA: Sage Publications; 1997.
22. Carlsen B, Glenton C. What about N? A methodological study of sample-size reporting in focus group studies. BMC Med Res Methodol 2011;11:26 [FREE Full text] [doi: 10.1186/1471-2288-11-26] [Medline: 21396104]
23. Mayring P. Qualitative Content Analysis: Theoretical Foundation, Basic Procedures and Software Solution. Klagenfurth, Austria: Beltz; 2014.
24. Marks D, Yardley L. Research Methods for Clinical and Health Psychology. Thousand Oaks, CA: Sage; 2004.
25. Lester JN, Cho Y, Lochmiller CR. Learning to do qualitative data analysis: a starting point. Hum Resour Dev Rev 2020;19(1):94-106 [FREE Full text] [doi: 10.1177/1534484320903890]
26. Crotty BH, Anselmo M, Clarke D, Elmore JG, Famiglio LM, Fossa A, et al. Open notes in teaching clinics: a multisite survey of residents to identify anticipated attitudes and guidance for programs. J Grad Med Educ 2018;10(3):292-300 [FREE Full text] [doi: 10.4300/JGME-D-17-00486.1] [Medline: 29946386]
27. Niedermier VE. Teaching electronic health record documentation to medical students. J Grad Med Educ 2017;9(1):135 [FREE Full text] [doi: 10.4300/JGME-D-16-00628.1] [Medline: 28261413]
28. Gagliardi JP, Turner DA. The electronic health record and education: rethinking optimization. J Grad Med Educ 2016 Jul;8(3):325-327 [FREE Full text] [doi: 10.4300/JGME-D-15-00275.1] [Medline: 27413432]
29. Lyles C, Schillinger D, Sarkar U. Connecting the dots: health information technology expansion and health disparities. PLoS Med 2015;12(7):e1001852 [FREE Full text] [doi: 10.1371/journal.pmed.1001852] [Medline: 26172977]
30. Lyles CR, Sarkar U. Health literacy, vulnerable patients, and health information technology use: where do we go from here? J Gen Intern Med 2015;30(3):271-272 [FREE Full text] [doi: 10.1007/s11606-014-3166-5] [Medline: 25588688]
31. Tieu L, Schillinger D, Sarkar U, Hoskote M, Hahn KJ, Ratanawongsa N, et al. Online patient websites for electronic health record access among vulnerable populations: portals to nowhere? J Am Med Inform Assoc 2017;24(e1):e47-e54 [FREE Full text] [doi: 10.1093/jamia/ocw098] [Medline: 27402138]

Abbreviations

AKI: acute kidney injury
HFpEF: heart failure with preserved ejection fraction
HFrEF: heart failure with reduced ejection fraction

Edited by B Lesselroth; submitted 08.04.24; peer-reviewed by K Gray, L Shapiro, M Figueroa Gray, S Walker; comments to author 25.07.24; revised version received 11.09.24; accepted 23.11.24; published 20.01.25.

Please cite as:
Vanka A, Johnston KT, Delbanco T, DesRoches CM, Garcia A, Salmi L, Blease C
Guidelines for Patient-Centered Documentation in the Era of Open Notes: Qualitative Study
JMIR Med Educ 2025;11:e59301
URL: https://mededu.jmir.org/2025/1/e59301
doi:10.2196/59301
PMID:

Original Paper

Implementing the H&P 360 in Three Medical Institutions: Usability Study

Rupinder Hayer[1*], MPH; Joyce Tang[2*], MPH, MD; Julia Bisschops[3*], MD; Gregory W Schneider[4*], MD; Kate Kirley[1*], MS, MD; Tamkeen Khan[1*], MA, PhD; Erin Rieger[5*], MD; Eric Walford[6*], MD; Irsk Anderson[2*], MD; Valerie Press[2*], MPH, MD; Brent Williams[6*], MD

[1]American Medical Association, Improving Health Outcomes, Chicago, IL, United States

[2]Section of Hospital Medicine, University of Chicago Pritzker School of Medicine, Chicago, IL, United States

[3]Department of Humanities, Health and Society, Herbert Wertheim College of Medicine, Florida International University., Miami, FL, United States

[4]Florida International University, Miami, FL, United States

[5]Columbia University Medical Center, New York, NY, United States

[6]University of Michigan, Ann Arbor, MI, United States

[*]all authors contributed equally

Corresponding Author:
Rupinder Hayer, MPH
American Medical Association
Improving Health Outcomes
330 North Wabash Avenue
Chicago, IL, 60611
United States
Phone: 1 6308499232
Email: rupinder.hayer@ama-assn.org

Abstract

Background: The traditional history and physical (H&P) provides the basis for physicians' data gathering, problem formulation, and care planning, yet it can miss relevant behavioral or social risk factors. The American Medical Association's "H&P 360," a modified H&P, has been shown to foster information gathering and patient rapport in inpatient settings and objective structured clinical examinations. It prompts students to explore 7 domains, as appropriate to the clinical context: biomedical problems, psychosocial problems, patients' priorities and goals, behavioral history, relationships, living environment and resources, and functional status.

Objective: This study aims to examine the perceived usability of the H&P 360 outside standardized patient settings.

Methods: The H&P 360 was implemented in various clinical settings across 3 institutions. Of the 207 student participants, 18 were preclerkship, 126 were clerkship, and 63 were postclerkship; 3-8 months after implementation, we administered a student survey consisting of 14 Likert-type items (1=strongly disagree to 5=strongly agree) and 3 free-text response items to assess usability.

Results: Of the 207 students, 61 responded to the survey (response rate was 29.5%). Among all students, mean ratings on the 3 usability survey items ranged from 4.03 to 4.24. The 5 items assessing the impact on patient care had mean ratings ranging from 3.88 to 4.24. The mean ratings for the 2 student learning items were 4.10 and 4.16. Students' open-ended comments were generally positive, expressing a perceived value in obtaining a more complete contextual picture of patients' conditions and supporting the usability of the H&P 360. Survey response patterns varied across institutions and learner levels.

Conclusions: Our findings suggest that using the H&P 360 may enhance information gathering critical for chronic disease management, particularly regarding social drivers of health. As a potential new standard, the H&P 360 may have clinical usability for identifying and addressing health inequities. Future work should assess its effects on patient care and outcomes.

(JMIR Med Educ 2025;11:e66221) doi:10.2196/66221

KEYWORDS

history and physical; medical education; social drivers; social determinants of health

Introduction

The traditional history and physical (H&P) structure is central to the patient-physician interaction and remains a foundational element of medical education. Through medical history, physicians elicit 60%-80% of the information relevant to diagnosis and treatment [1]. Medical students are typically required to master the skill of gathering, synthesizing, and documenting patient information early in their training. The traditional H&P, used in most medical education settings and routine clinical practice, is primarily structured to diagnose acute medical conditions and has not evolved for generations—despite the growing prevalence of chronic diseases and the increasing influence of social and behavioral drivers of health [2,3].

The social determinants or social drivers of health (SDOH) heavily influence the health of patients and populations [4]. The World Health Organization (WHO), in its conceptual framework for action on SDOH, defines "social determinants of health" as the full set of social conditions in which people live and work [4,5]. We will use the phrases social determinants and social drivers interchangeably. We will refer to social risk factors, meanwhile, as individual-level adverse SDOH, such as housing instability or low education level, and social needs as social factors that take into account people's individual preferences and priorities in identifying and guiding social interventions [6,7].

Health systems and providers are increasingly exploring ways to better integrate health care delivery with reforms aimed at addressing the SDOH, identifying patients' social risk factors, and meeting patients' social needs [8,9]. At the upstream level, laws, policies, and regulations can be used to create community conditions that foster health. At the midstream level, providers and health systems can include screening questions to identify social risk factors and offer services that connect patients to resources to meet their social needs. At the downstream level, clinicians can tailor medical interventions to acknowledge individual social conditions [6,8,9]. To significantly improve the health of all, it is critical to emphasize addressing the broader SDOH inequities—those created and sustained by structural racism and the marginalization of specific groups, including women, Black, Hispanic/Latino, LGBTQ+ (lesbian, gay, bisexual, transgender, queer or questioning, and other diverse sexual orientations and gender identities) individuals, people living with disabilities, and other populations [4,5,10-12].

In addition to increased attention to the SDOH, both globally and particularly in the United States, chronic diseases are the largest contributors to disease burden, accounting for 90% of health care costs in the United States [13]. Health systems are increasingly addressing upstream factors, such as behavioral health, and social and environmental circumstances, to prevent and manage chronic diseases and their consequences [14,15]. Against this multidimensional backdrop, individual clinicians, residents, and medical students typically rely on the H&P examination as their principal method of information gathering. However, when using the traditional H&P to frame and organize this process, learners at all levels are not prompted to collect relevant biopsychosocial data, including social needs and health behaviors, which are key to preventing and managing chronic diseases. Recent research has shown that inaccurate and incomplete patient histories are among the leading causes of diagnostic errors [16].

The H&P 360, a modified version of the traditional H&P, was developed by the American Medical Association's (AMA) Chronic Disease Prevention and Management interest group in May 2017, building on earlier work by medical educators at the University of Michigan (UM) [12,17]. This new approach was designed to more explicitly acknowledge the SDOH, the prevalence of chronic diseases, and the importance of patients' preferences and priorities in clinical decision-making. It intentionally incorporates the WHO conceptual framework for addressing the SDOH at the micro-level of individual interaction [4,5], the Centers for Disease Control and Prevention (CDC) framework for addressing chronic diseases at the health system and community-clinic levels [12,18], and contemporary models of shared decision-making [1,19].

The H&P 360 is grounded in the idea that the central, standardized written template in medical education (ie, the traditional H&P) plays a significant role in both enhancing and constraining information-seeking related to medical decision-making. This understanding of information gathering aligns with Structuration Theory and Cognitive Load Theory. Structuration Theory posits that social practice both shapes and is shaped by the structures, such as learning templates, within which it occurs [20]. Cognitive Load Theory, meanwhile, asserts that cognitive capacity is limited and that learning is enhanced when key information is presented in manageable blocks, such as the 7 domains [21].

At a deeper level, the H&P may play an important role in shaping physicians' professional identity and role expectations, as suggested by Social Learning Theory [22,23], which posits that individuals' agency and role identities are critically influenced by the social and institutional contexts in which they develop. The H&P 360 prompts students to collect relevant biopsychosocial information, particularly social risk factors and needs, using a systematic yet flexible framework. While it retains the basic structure of the traditional H&P for eliciting biomedical information, the H&P 360 also includes general prompts for 6 additional domains: patients' priorities and goals, psychosocial problems, behavioral history, relationships, living environment and resources, and functional status. The 6 nonbiomedical domains were identified through a literature review as those consistently represented in comprehensive clinical assessment settings, including geriatrics, and care for homeless and chronically mentally ill persons, as well as in the categories of the Diagnostic and Statistical Manual of Mental Disorders (4th edition) [24], and have been applied in numerous clinical and teaching settings since 2010 [17]. See Multimedia Appendix 1 for the H&P 360 template.

When using the H&P 360, students are encouraged to ask a few questions from each of the domains as part of the standard history. These additional questions help students gain a more comprehensive understanding of a patient's biopsychosocial condition and support the development of an appropriate

treatment and management plan. In follow-up encounters, continued exploration of the 7 H&P 360 domains can foster a deeper understanding of the patient, informing chronic disease management. A previous randomized trial conducted at 4 medical schools found that medical students using the H&P 360 in a standardized patient setting collected significantly more biopsychosocial information compared with students using the traditional H&P [3]. Another study found that students who applied the H&P 360 using templated notes in the electronic health record reported improved elicitation of patient goals and perspectives, as well as identification of contextual factors and patient needs critical to preventing rehospitalization [25]. In addition to enhanced data gathering, the H&P 360 has been shown to encourage multidisciplinary team care planning [17] and to improve patient rapport (unpublished data).

The goal of this study was to examine the perceived usability of the H&P 360 by both faculty and students across a variety of clinical settings and learner levels in routine clinical teaching contexts. To assess usability in different clinical teaching environments, the AMA launched a grant program for institutions willing to implement the H&P 360 in student clinical encounters and administer a standardized postintervention survey to faculty and students across sites. We hypothesized that students and faculty across sites would appreciate the usability of the approach, but that barriers to engagement would vary by site and learner level.

Methods

Site Selection

The AMA offered funding for projects to implement the H&P 360 within clinical settings at academic institutions. Priority was given to projects aimed at developing additional supporting materials and gathering student and faculty feedback during the implementation phase. Following a call for proposals, 4 academic institutions received grants from the AMA to implement the H&P 360 across a diverse range of clinical settings and undergraduate medical education learner levels. The grant period began in January 2020 and ended in June 2021. Because of the pandemic, only 3 of the institutions were able to implement their grants. These institutions were the UM School of Medicine, the University of Chicago Pritzker School of Medicine (UC), and the Herbert Wertheim College of Medicine at Florida International University (FIU). The fourth institution was unable to implement its grant project but still incorporated the H&P 360 with its students.

The 3 grant-funded institutions implemented the H&P 360 across a variety of clinical settings and learner levels, described in detail in Table 1. Clinical settings included inpatient, outpatient, virtual, community-based clinics, and longitudinal outpatient clinics. Learner levels ranged from preclerkship to clerkship and postclerkship. The approaches each school used to introduce students and faculty to the H&P 360 are also detailed in Table 1. At all 3 sites, students were introduced to the H&P 360 through a nonstandardized 1- to 2-hour seminar. One site (UC) implemented standardized note templates within the electronic health record to facilitate documentation of the H&P 360. Faculty orientation to the H&P 360 varied across sites, ranging from emailed communications with attached introductory materials and a teaching guide (UM and UC) to virtual orientation sessions over Zoom (Zoom Communications, Inc) and some in-person sessions (FIU and UC).

Table 1. Learner level, clinical setting, and teaching context by institution.

Approaches to introducing H&P[a] 360	Learner level			Students, n	Setting	Duration of use	How the H&P 360 was introduced to students and topics covered	How the H&P 360 was introduced to faculty and topics covered
	Preclerkship	Clerkship	Postclerkship					
The University of Chicago Pritzker School of Medicine	✓	N/A[b]	N/A	18	• Preclerkship students were encouraged to utilize the H&P 360 in a longitudinal patient-partnered clinical experience (about 6 face-to-face and 2 virtual clinical sessions).	9 months	Preclerkship students attended a presentation on the format and components and received training materials including examples and the interview guide.	Faculty precepting preclerkship students attended a presentation or received an email.
The University of Chicago Pritzker School of Medicine	N/A	✓	N/A	8	• Clerkship students were encouraged to utilize the H&P 360 during COVID-19 follow-up virtual visits.	1 month	Clerkship students attended a 1-hour virtual training session with background about the H&P 360 and details on using a COVID-19–specific note template. The supporting interview guide and pocket card were leveraged as needed.	Faculty attended 2 or more presentations on the H&P 360.
The University of Chicago Pritzker School of Medicine	N/A	N/A	✓	24	• Postclerkship students were encouraged to utilize the H&P 360 in an internal medicine subinternship for 1 admission per call cycle.	1 month	Postclerkship students received an email from their course director with background about the H&P 360 and instructions to access H&P 360 templates. The supporting interview guide and pocket card were leveraged as needed.	Faculty supervising postclerkship students received an email.
University of Michigan School of Medicine	N/A	N/A	✓	39	• All students utilized the H&P 360 in an outpatient setting. • Of the 39 students, 8 used it in a community-based elective and 31 used it in a longitudinal clinic setting. • Students were encouraged to apply the H&P 360 in every encounter.	Elective (1 month) and longitudinal clinics (9 months)	A 2-hour interactive in-person seminar with case examples. The supporting interview guide and pocket card were leveraged as needed.	Email introduction and follow-up, which included teaching tips, pocket cards, profiles, and cases.

Approaches to introducing H&P[a] 360	Learner level			Students, n	Setting	Duration of use	How the H&P 360 was introduced to students and topics covered	How the H&P 360 was introduced to faculty and topics covered
	Preclerkship	Clerkship	Postclerkship					
Herbert Wertheim College of Medicine at Florida International University	N/A	✓	N/A	118	• All students utilized the H&P 360 in a virtual and longitudinal, interprofessional, home-based service-learning program.	12 months/1 academic year as part of a longitudinal program.	Introduced during an interactive didactic session on chronic disease management; the self-directed video was also available for students.	One in-person faculty development session before COVID-19; a self-directed video for faculty; and a conference presentation by Dr. Brent Williams from the University of Michigan.

[a]H&P: history and physical.

[b]N/A: not applicable.

Settings

At FIU, the investigators had to alter their initial implementation strategy due to the COVID-19 pandemic. The project team completed 1 in-person faculty orientation; subsequent sessions were delivered virtually as an online module tailored to both faculty and students. For students, the team relied solely on the online module, as the planned in-person session was canceled due to the pandemic.

At UC, initial implementation plans were also disrupted by the COVID-19 pandemic, which led to medical students being removed from traditional clinical settings. During this period, an innovative program was developed in which clerkship students conducted phone outreach to patients newly testing positive for COVID-19. Taking advantage of this novel opportunity, the H&P 360 was used to help structure these outreach calls. Because of the small number of students and faculty involved, an in-depth, interactive training program for both faculty and students was offered via Zoom meetings. In the later implementation of the H&P 360 for preclerkship students, a virtual orientation over Zoom was incorporated into their Clinical Skills course. Because of the large number of faculty serving as preceptors for this course, only new faculty preceptors—who were required to attend a mandatory orientation session—received virtual training on the H&P 360 framework. Preceptors who were not new to the program and for whom orientation attendance was not required received emailed communication about the H&P 360. Implementation for postclerkship students was further modified to email communication only. These students had rolling start times each month, and there was no formal orientation session during the clerkship to integrate separate training. The number of clinical faculty preceptors for postclerkship students was quite large, also with rolling start times every 2 weeks, making email communication regarding the H&P 360 the most feasible approach.

At UM, the H&P 360 was implemented during the postclerkship period in 2 settings: a 1-month clinical elective focused on underserved populations before the pandemic (8 students over 2 months), taught by 1 of the authors (BW), and longitudinal weekly clinics in primary care settings over 9 months during the pandemic (31 students). The longitudinal clinic rotation was chosen due to the faculty coordinator's interest in implementing the H&P 360 and its suitability for continuity settings, where the domains can be explored with patients over time. For the longitudinal rotation, students received a 2-hour introduction to the H&P 360, including case examples. Precepting faculty were sent introductory materials, teaching tips, and written case examples via email both at the start and several months into the longitudinal clinics. Many longitudinal clinics transitioned to telemedicine visits during the pandemic. In both rotations, students were encouraged—but not required—to use the H&P 360, or portions of it, in every encounter.

Survey Structure

Data collection consisted of a student survey on using the H&P 360 in undergraduate medical education settings. As the survey focused specifically on the use of the H&P 360, previously published surveys were not applicable. Theoretical frameworks guiding survey development included Bloom's taxonomy of learning objectives [26], which emphasizes synthesis and application of knowledge rather than factual recall, and the Expectancy-Value Theory of Motivation [27], which posits that learner motivation is influenced by the perceived value of new information.

The survey consisted of an initial section asking for examples of a question relevant to each of the 5 domains from the H&P 360, followed by 14 Likert-type items (response scale: 1=strongly disagree to 5=strongly agree) and 3 open-ended questions. The Likert-type items were developed using a "blueprint" of 7 potential impact areas of the H&P 360, designed by the authors. Source items were either adapted from a 10-item version used in a previous study [17] or newly created. To minimize the response burden, the survey was limited to 15

items or fewer. Items were reviewed for sensibility by small groups of medical students and residents not involved in the study, resulting in minor modifications. The final instrument included 14 Likert-type items. Two items were modified or omitted at some sites and thus were not included in the analyses. The analyzed items are shown in Tables 2 and 3. The 7 areas from the "blueprint" and their corresponding item numbers were perceived usability of the H&P 360 (items 1, 2, and 3); impact on history-taking (item 4); perceived clinical value added (items 5 and 6); promotion of understanding patients' goals (item 7); enhancement of patient-provider relationships (item 8); facilitation of care planning (item 9); and promotion of inclusion of other health professionals (item 10). Two additional items were included as global measures of educational and clinical value, respectively (items 11 and 12). By covering a broad range of topics, results from individual items could be used independently by educators to inform a wide spectrum of educational and research activities.

The 3 open-ended questions were designed to elicit specific feedback about the H&P 360: "Name two (or more) aspects of the H&P 360 you found helpful"; "Name two (or more) aspects of the H&P 360 you found challenging"; and "What changes would you recommend for the H&P 360?" A systematic review of the open-ended comments is not included in this paper. Instead, a subset of comments reflecting students' perceived value and limitations of the H&P 360 is provided in Multimedia Appendix 2.

Table 2. Mean student survey scores (Likert scale 1-5)[a] by school.

Student survey scores by school	All students (N=49) (N=61)[b], mean (SD)	All FIU[c] students (N=17), mean (SD)	All UM[d] students (N=13), mean (SD)	All UC[e] students (N=19) (N=31)[b], mean (SD)
Usability				
1. The H&P[f] 360 was easy to use	4.12 (0.78)	4.18 (0.64)	4.15 (0.69)	4.05 (0.97)
2. Elements of the H&P 360 are potentially useful in all patient interactions	4.24 (0.90)	4.35 (0.61)	4.62 (0.51)	3.89 (1.20)
3. I plan to use the H&P 360 during other rotations[b]	4.03 (0.84)	3.59 (0.94)	4.31 (0.75)	4.16 (0.73)
Impact on patient care				
4. The H&P 360 changed some of the questions I ask patients during the encounter	4.08 (0.76)	3.59 (0.87)	4.46 (0.52)	4.26 (0.56)
5. The H&P 360 helped create a more comprehensive problem list	3.88 (0.95)	4.06 (0.90)	4.15 (0.69)	3.53 (1.07)
6. The H&P 360 added valuable information that I would not otherwise know about the patient[b]	4.18 (0.79)	3.82 (1.01)	4.46 (0.78)	4.26 (0.58)
7. The H&P 360 helped me better understand patients' goals[b]	4.15 (0.68)	4.12 (0.78)	4.23 (0.44)	4.13 (0.72)
8. Using the H&P 360 facilitated a stronger provider-patient relationship[b]	4.24 (0.67)	4.12 (0.70)	4.08 (0.64)	4.35 (0.66)
9. I was able to develop management plans that incorporated information from the H&P 360	3.88 (0.83)	4.00 (0.87)	4.00 (0.41)	3.68 (1.00)
Overall impact on student learning				
10. The H&P 360 helped me learn to be a better clinician[b]	4.16 (0.64)	4.06 (0.66)	4.31 (0.63)	4.16 (0.64)
11. The H&P 360 helped improve the care I provided to my patients	4.10 (0.71)	4.06 (0.75)	4.31 (0.63)	4.00 (0.75)

[a]1=strongly disagree to 5=strongly agree.

[b]These are items with a greater number of respondents because an abbreviated version of the survey was completed by preclinical students at UC.

[c]FIU: Florida International University.

[d]UM: University of Michigan.

[e]UC: University of Chicago.

[f]H&P: history and physical.

Table 3. Mean student survey scores (Likert scale 1-5)[a] by clerkship status.

Mean student survey scores	Preclerkship students (N=15), mean (SD)	Clerkship students (N=25), mean (SD)	Postclerkship students (N=24), mean (SD)
Usability			
1. The H&P[b] 360 was easy to use	N/A[c]	4.32 (0.63)	3.92 (0.88)
2. Elements of the H&P 360 are potentially useful in all patient interactions	N/A	4.40 (0.76)	4.08 (1.02)
3. I plan to use the H&P 360 during other rotations	4.33 (0.49)	3.84 (0.90)	4.08 (0.88)
Impact on patient care			
4. The H&P 360 changed some of the questions I ask patients during the encounter	N/A	3.92 (0.91)	4.25 (0.53)
5. The H&P 360 helped create a more comprehensive problem list	N/A	4.08 (0.86)	3.67 (1.21)
6. The H&P 360 added valuable information that I would not otherwise know about the patient	4.25 (0.45)	4.04 (0.93)	4.29 (0.75)
7. The H&P 360 helped me better understand patients' goals	4.25 (0.45)	4.12 (0.83)	4.13 (0.61)
8. Using the H&P 360 facilitated a stronger provider-patient relationship	4.17 (0.58)	4.36 (0.70)	4.13 (0.68)
9. I was able to develop management plans that incorporated information from the H&P 360	N/A	4.08 (0.81)	3.67 (0.82)
10. The H&P 360 facilitated care planning that included other health professionals	3.75 (0.97)	4.08 (0.86)	4.00 (1.02)
Overall impact on student learning			
11. The H&P 360 helped me learn to be a better clinician	4.25 (0.45)	4.20 (0.65)	4.08 (0.12)
12. The H&P 360 helped improve the care I provided to my patients	N/A	4.20 (0.71)	4.00 (0.72)

[a]1=strongly disagree to 5=strongly agree.

[b]H&P: history and physical.

[c]N/A: not applicable.

The survey was administered online by all 3 sites approximately 3-8 months after implementation. Six items that presumed experience in clinical care were not administered to preclinical students participating in this study; this omission applied only to a subset of students at 1 site. Data were aggregated across all sites to calculate mean scores and SDs for each survey item, allowing comparisons by institution and by clerkship status. Because of the small number of respondents in each subgroup, we were limited to analyzing descriptive statistics and were unable to conduct psychometric analyses or hypothesis testing to statistically compare subgroups. However, the descriptive analysis was still useful for aggregating data across multiple sites and generating hypotheses. Data analysis was conducted using STATA version 13.0 (StataCorp). See Multimedia Appendix 3 for the supporting CHERRIES (Checklist for Reporting Results of Internet E-Surveys) document.

Ethics Considerations

The UM received exempt institutional review board status from the Institutional Review Boards of the UM Medical campus. FIU received exempt institutional review board status from The FIU Office of Research Integrity. The UC received exempt institutional review board status from the BSD/UCMC Institutional Review Boards at the UC. Lastly, the AMA confirmed that this study was not deemed to be research by the University of Illinois Chicago Institutional Review Board. All 4 institutions confirmed that all methods were carried out in accordance with relevant guidelines and regulations.

Results

Summary of the Survey Findings

The Likert-type survey items were organized by consensus among the authors into 3 sets to identify patterns and facilitate discussion: Usability (3 items); Impact on Patient Care (7 items); and Overall Impact on Student Learning (2 items). Results are presented for all student respondents by institutional site in Table 2 and by learner-level subgroups in Table 3. Of the 207 students, 61 (29.5%) responded to the survey. Institutional response rates were as follows: FIU, 17 out of 118 (14.4%) students; UM, 13 out of 39 (33.3%) students; and UC, 31 out of 50 (62.0%) students.

Among all students, mean ratings on the 3 survey items related to usability (ease of use, use in all encounters, and intention to use in other rotations) were high, with mean (SD) scores ranging from 4.03 (0.84) to 4.24 (0.90) (Table 2). Some students' comments suggested that efficiently using the H&P 360 requires practice. One postclerkship student commented: "(The H&P 360)...is quite long so it was challenging to hit aspects of each domain while attempting to time manage. However, hitting 1-item from each domain, chosen on a case-by-case basis, seems

quite doable." Several students raised concerns about the awkwardness of asking some questions, particularly during virtual outreach calls to patients who had newly tested positive for COVID-19. One clerkship student commented: "It did not always feel natural to fit into the conversation with every patient. Some were not very open to conversation, which is understandable since we were strangers calling them out of the blue."

Students also found that the H&P 360 positively affected patient care by expanding the range of information available for clinical decision-making and promoting stronger patient-clinician relationships. Mean ratings across the 5 related items ranged from 3.88 (SD 0.95) to 4.24 (SD 0.67; see Table 3 for details). Student feedback on clinical impact emphasized the benefits of the H&P 360 in building rapport. One clerkship student commented: "It helped me build rapport with my patient and have a better [understanding] of their life and how it affects their health." Others mentioned that the H&P 360 helped build trust and identify high-risk situations. See Multimedia Appendix 2 for additional relevant student comments.

Students also found that the H&P 360 facilitated their learning and development as clinicians, with mean ratings of 4.10 (SD 0.71) and 4.16 (SD 0.64) for the items "the H&P 360 helped me...improve the care I provided to my patients" and "…be a better clinician," respectively. One postclerkship student commented: "The H&P 360 was helpful in...[r]eturning the humanity to medicine: patients are people first—Helping to understand some of the barriers to health and disease prevention that might not otherwise be apparent."

Site-Specific Survey Findings

Some variation in student survey responses was observed across institutions (Table 2). For 2 items related to using the H&P 360 to develop problem lists and management plans, student ratings at UC were lower than those at UM or FIU. For 3 items—related to using the H&P 360 in other rotations and its role in changing some questions and adding valuable information—student ratings at FIU were lower compared with UC and UM.

Survey Findings by Learner Levels

Across learner levels, some variation in student survey responses was noted for a minority of items (Table 3). For example, preclinical students gave relatively low ratings on the item related to facilitating care planning that included other health professionals compared with their responses on other items. While clerkship students valued the H&P 360 in all patient interactions and for facilitating stronger patient relationships, their ratings were relatively low for items related to the H&P 360 changing the questions they asked and their plans to use it in future rotations. Postclerkship students gave high ratings for 9 of the 12 questions. Lower ratings were observed for items related to ease of use, creating a more comprehensive problem list, and the ability to develop management plans incorporating information from the H&P 360. The phrasing of the item on time burden evolved over time and was therefore not administered consistently across or within institutions.

Discussion

Principal Findings

Previous work has documented the advantages of the H&P 360 over the traditional H&P during single inpatient encounters [17] and with standardized patients [3]. This study examined the use of the H&P 360 across a broad range of routine, longitudinal clinical teaching settings. Medical students at 3 institutions, spanning different levels of training and diverse ambulatory, inpatient, community, and virtual settings, found the H&P 360 useful and reported a positive impact on patient care and their own learning. The perceived benefits of the H&P 360 include helping students gather relevant information on patients' goals and circumstances, as well as potential barriers and facilitators of health. It also enhances patient-provider relationships and encourages interprofessional care planning. Compared with the traditional H&P, student feedback suggests that the H&P 360 made them better clinicians. We can further speculate that by using the H&P 360, students develop a more complete picture of the patient—not just signs, symptoms, and diagnoses—but also the social and human narrative context that critically influences the presentation, management, and ultimately the outcomes of disease conditions. We suspect that gathering this more complete picture of patients' lives is one factor contributing to students' perception that the provider-patient relationship was enhanced by using the H&P 360. An important area for future investigation is the mechanism behind this enhanced relationship. Perhaps it is this more complete understanding of the patient, combined with improved patient rapport, that prompted the student comment that the H&P 360 "...return(ed) the humanity to medicine."

Implication of Findings

The observed variation across institutions and learner levels—though limited by small sample sizes and collinearity between institutions and learner levels—may offer insights into factors influencing medical students' perceived value of the H&P 360. Here, we present 3 speculative observations based on these data to encourage future research and application of the H&P 360. First, the teaching setting for the FIU students included in this study was a community-based, longitudinal, interprofessional environment with an established strong emphasis on comprehensive assessment and interprofessional care. As such, FIU students may have been less likely to perceive that the H&P 360 changed questions asked or added valuable information beyond their prior practice. Additionally, during the study period, FIU students conducted visits virtually due to COVID-19. Second, UC provided a relatively unique perspective on implementation efforts by including preclinical students and applying the H&P 360 in telehealth settings. Although the very high perceived value of the H&P 360 among both preclinical students and clinical-year students in telehealth settings is striking and promising, further investigation in other settings and institutions is needed to contextualize these findings. Finally, variation among institutions may reflect differences in overall emphasis on SDOH; the role of faculty in promoting or minimizing the study findings and application of the H&P 360; or differences in its use across inpatient, outpatient, and virtual settings.

Synthesizing evidence on the H&P 360 from this and previous studies, along with input from faculty and students and our own clinical teaching experience, we suggest that incorporating the H&P 360 into routine clinical practice involves at least 4 dimensions of learning. The general patterns and variations observed across different learner levels and clinical settings in this study support and shed light on each of these dimensions.

First, learners and educators using the H&P 360 will need to *integrate domain-based thinking alongside checklist-based approaches in data gathering.* Currently, early medical students are taught to take a history by following a memorized list of specific questions. Over time, an implicit process develops, where clinicians tailor questions, diagnoses, and management plans based on patient-specific information [28]. The direction a clinician takes for follow-up inquiry is likely influenced by many factors, including training experiences, knowledge and clinical skills in managing a wide range of issues (eg, emotional well-being, food insecurity, or safe housing), and local practice norms. Consequently, this approach is likely to vary widely among clinicians. Some naturally explore psychosocial dimensions, while others remain more narrowly focused on biomedical factors. To reduce this variation and better address the role of psychological and social factors in patients' health, the H&P 360 provides uniform, systematic prompts that help clinicians recognize social and psychological determinants of health. The H&P 360 represents a fundamental shift in learning to gather patient information by introducing 6 domains as general reference points alongside the traditional checklist focused on biomedical information.

For early learners, balancing domain-based thinking with a checklist approach can be disconcerting as they decide which specific content to include or exclude within the nonbiomedical domains. This challenge aligns with student feedback that the H&P 360 initially feels long and overwhelming when seen as a checklist of individual items, but becomes manageable and useful when viewed as a set of domain-based prompts that can be selectively explored—or revisited over multiple patient encounters. This is also consistent with findings that senior medical students using the H&P 360 identify and apply significantly more psychosocial information in their care planning than those using the traditional H&P [17]. Additionally, our finding that preclinical students found the H&P 360 added valuable information, helped them understand patients' goals, and facilitated stronger patient-provider relationships suggests that early learners can successfully incorporate domain-based thinking into routine data gathering. The interview guide that accompanies the H&P 360 can be a valuable resource in this regard. It helps students decide which domain to focus on and also supports faculty in navigating these domains during classroom teaching. See Multimedia Appendix 4 for the H&P 360 interview guide. Further exploration of domain-based thinking among medical learners is warranted to identify the best ways to provide a data-gathering framework that is both accessible in the early stages and comprehensive in the later stages of learning. We are particularly interested in methods for—and the implications of—incorporating patients' values, priorities, and goals into every clinical encounter [29,30].

Second, once familiar with domain-based thinking, medical students need to *develop skills in deciding which specific information within a domain is most relevant to a given patient encounter.* The finding that clerkship and postclerkship students from both inpatient and outpatient settings found the H&P 360 added valuable information and enhanced patient care suggests that students perceive tailoring domain-based questions to individual patients and clinical contexts as useful and facilitative for patient care.

Importantly, many students' comments revealed the emergence of skills in "modularizing" components of the H&P 360—using only those most relevant to a particular clinical context without feeling compelled to cover every domain.

Third, students need to *manage the emergent information through further inquiry or redirection.* As reflected in some student comments, medical students can feel compelled to fully elucidate or address the complex behavioral or social drivers of patients' health once identified. They also recognized that some behavioral or social needs uncovered during this process are important but do not require immediate action. Students should then redirect the interview to address matters of immediate concern (eg, potentially serious symptoms or a plan for initial hospital treatment), while simultaneously developing a plan to address longer-term issues. This process of identifying, prioritizing, and guiding the interview to optimize both disease-specific and contextual information has been demonstrated in the area of diagnostic reasoning, where a clinician listens and generates hypotheses, gathers data to test these hypotheses, and, depending on the results, offers treatment or pursues further diagnostic action [31]. We suggest that the domain-based framework of the H&P 360 facilitates the application of advanced interview skills not only to diagnostic assessment but also to management and care planning that better account for patients' psychosocial and environmental realities.

Finally, the *new information elicited with the H&P 360 must be applied to clinical management planning.* Our data suggest that, while learners generally found that information from the H&P 360 enhanced care planning, ratings in this area were lower than for other measures and lower than those observed in previous studies [3,17], particularly among early learners. We believe these findings highlight the complexity of incorporating social and behavioral information into care planning—for example, in discharging a patient facing homelessness or supporting medication adherence limited by insurance, income, transportation, or behavioral factors. By bringing these "background" issues to the forefront early in training, students can develop skills to mobilize interprofessional teams and utilize local resources as part of routine patient care. We also anticipate that by directly addressing SDOH—rooted in racial, ethnic, and socioeconomic inequities—learners will be better equipped to recognize and manage systemic and personal implicit biases that negatively impact care.

Although not emphasized in the student-oriented results presented here, our study suggested that faculty play an important role in promoting the effective use of the H&P 360. Faculty development and feedback methods varied across participating sites, ranging from interactive seminars to entirely

email-based communication, and faculty "buy-in" likely varied both within and across sites. At all sites, faculty were provided with information on the purpose, content, and suggested best practices for using the H&P 360. Anecdotally, however, learners reported little awareness or receptivity among teaching faculty toward using the H&P 360 domain framework in teaching and clinical management—except at 1 site where in-person faculty development was conducted before the COVID-19 pandemic. As professional development is influenced by cues from influential social sources, as well as practice resources and norms [22], effective application of the H&P 360 will likely require its incorporation into local teaching and clinical practices. Faculty development and the use of the 7-domain framework in teaching and clinical practice represent important areas for future investigation.

Limitations

Our study was limited by small sample sizes, which prevented more rigorous statistical analyses of the Likert-type scale data and further qualitative analysis. However, both in this work and in previous studies, student responses to the H&P 360 have been primarily positive. Additionally, it is important to acknowledge that some students experienced difficulties implementing the H&P 360 during virtual interactions and in specific clinical encounters. More implementation training on how to utilize the H&P 360 in different scenarios might be helpful for students. The survey also had a low response rate at some institutions, which may be attributable to several factors. The survey was optional and not required at all 3 sites. In some cases, faculty were unable to administer the survey immediately after course completion due to time constraints.

Lastly, this project took place at the beginning of the pandemic, which introduced many competing priorities and adjustments to the overall learning environment. The overall impact of response rates on the results is difficult to estimate, as rates varied by institution and were likely influenced by additional local factors. Information on factors that could promote or limit the effective application of the H&P 360 was not explored beyond the data collected from the student surveys. For example, curricular content encountered by students before the H&P 360, as well as organizational culture, could influence its application at each institution. Exploration of these factors was outside the scope of this study. Survey data were limited by too few observations in relevant substrata (eg, inpatient vs outpatient; longitudinal vs short-term; virtual vs face-to-face; and institutional vs community-based clinical settings) to permit meaningful subgroup analyses exploring additional variables that may impact the implementation of the H&P 360 in different settings.

Conclusions

The H&P 360 provides an enhanced template for data gathering that includes general prompts addressing key dimensions of human health not captured by the traditional H&P, such as patients' values, priorities, and goals. Our findings support the usability of the H&P 360 as a more comprehensive approach for medical students to gather patient information. Among early learners, it may be best to include a few specific illustrative items under each domain to familiarize students with the domains without requiring higher-order clinical knowledge or skills. Among later learners, the now-familiar domains can be used to promote more complete data gathering and to develop skills in integrating patients' goals, psychosocial and behavioral factors, and interprofessional teams into care planning. The H&P 360 may be particularly useful for making health inequities and their root causes more visible in routine clinical encounters, while guiding management planning to address them. Future work should measure its effects on patient care and outcomes.

Relevant topics for future investigation related to the H&P 360 include influences on students' use of the H&P 360 at different developmental stages; its use to identify and address SDOH; and methods and outcomes of faculty development to promote routine incorporation of domain-based thinking into clinical teaching and practice. To facilitate further investigation and implementation of the H&P 360 among medical schools, a set of tools and resources is available on the AMA website or authors may be contacted directly for further information.

Acknowledgments

The project was performed with financial support from the Medical Education Department within the American Medical Association. The authors thank their partners at the University of Michigan (contract ID AMA 36456), the University of Chicago (contract ID AMA 36457), and Florida International University (contract ID AMA 36572). The authors express gratitude for the planning and logistical support provided by several team members at each institution who led and coordinated the on-site study activities. They also deeply appreciate the contributions of numerous current AMA team members, notably Kevin Heckman, Kathryn Pajak, and Kimberly Loomis. The collaborating group authors include Jonathan Lio, MD; Vineet Arora, MD; Rachel Clarke, PhD, Onelia Lage, MD; and Ebony Whisenant, MD.

Authors' Contributions

RH and KK were the project leads for the AMA. JB was the lead investigator at FIU. GS was a coinvestigator at FIU. TK was the lead on the planning and conduct of analyses. VP led data analyses, and BW was the lead investigator at UM and developed the first draft of the H&P 360. All authors provided substantial contributions to the conception and design; acquisition of data; analysis and interpretation of data; drafting of the article and revising it critically for important intellectual content; final approval of the version to be published; and agreement to be accountable for all aspects of the work, ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The H&P 360 template.
[DOCX File , 570 KB - mededu_v11i1e66221_app1.docx]

Multimedia Appendix 2
Student comments.
[DOCX File , 15 KB - mededu_v11i1e66221_app2.docx]

Multimedia Appendix 3
The CHERRIES (Checklist for Reporting Results of Internet E-Surveys) checklist.
[PDF File (Adobe PDF File), 283 KB - mededu_v11i1e66221_app3.pdf]

Multimedia Appendix 4
Interview guide to support the H&P 360.
[DOCX File , 37 KB - mededu_v11i1e66221_app4.docx]

References

1. Keifenheim KE, Teufel M, Ip J, Speiser N, Leehr EJ, Zipfel S, et al. Teaching history taking to medical students: a systematic review. BMC Med Educ 2015 Sep 28;15(1):159 [FREE Full text] [doi: 10.1186/s12909-015-0443-x] [Medline: 26415941]
2. H. Kenneth W. Clinical methods: the history, physical, and laboratory examinations. Ann Intern Med 1990 Oct 01;113(7):563-563. [doi: 10.7326/0003-4819-113-7-563_2]
3. Kirley K, Hayer R, Khan T, Johnson E, Sanchez ES, Kosowicz L, et al. Expanding the traditional history and physical examination to address chronic diseases and social needs: a multisite randomized control trial of 4 medical schools. Acad Med 2020 Dec;95(11S Association of American Medical Colleges Learn Serve Lead: Proceedings of the 59th Annual Research in Medical Education Presentations):S44-S50. [doi: 10.1097/ACM.0000000000003640] [Medline: 32769457]
4. Preda A, Voigt K. The social determinants of health: why should we care? Am J Bioeth 2015 Mar 18;15(3):25-36. [doi: 10.1080/15265161.2014.998374] [Medline: 25786009]
5. World Health Organization (WHO). A Conceptual Framework for Action on the Social Determinants of Health. In: WHO. Geneva, Switzerland: WHO; 2010:1-76.
6. Thornton RLJ, Glover CM, Cené CW, Glik DC, Henderson JA, Williams DR. Evaluating strategies for reducing health disparities by addressing the social determinants of health. Health Aff (Millwood) 2016 Aug 01;35(8):1416-1423 [FREE Full text] [doi: 10.1377/hlthaff.2015.1357] [Medline: 27503966]
7. Alderwick H, Gottlieb LM. Meanings and misunderstandings: a social determinants of health lexicon for health care systems. Milbank Q 2019 Jul 08;97(2):407-419 [FREE Full text] [doi: 10.1111/1468-0009.12390] [Medline: 31069864]
8. Berwick DM. Getting serious about producing health. JAMA 2022 May 17;327(19):1865-1866. [doi: 10.1001/jama.2022.6921] [Medline: 35482356]
9. Castrucci B, Auerbach J. Meeting individual social needs falls short of addressing social determinants of health. Health Affairs. 2019 Jan 16. URL: https://www.healthaffairs.org/content/forefront/meeting-individual-social-needs-falls-short-addressing-social-determinants-health [accessed 2025-01-01]
10. World Health Organization (WHO). The Asia-Pacific Perspective: Redefining Obesity and Its Treatment. Geneva, Switzerland: WHO; 2000:1-55.
11. Maybank A, De Maio F, Lemos D, Derige DN. Embedding racial justice and advancing health equity at the American Medical Association. Am J Med 2022 Jul;135(7):803-805. [doi: 10.1016/j.amjmed.2022.01.058] [Medline: 35245496]
12. Penman-Aguilar A, Talih M, Huang D, Moonesinghe R, Bouye K, Beckles G. Measurement of health disparities, health inequities, and social determinants of health to support the advancement of health equity. Journal of Public Health Management and Practice 2016;22:S33-S42. [doi: 10.1097/phh.0000000000000373]
13. Benavidez GA, Zahnd WE, Hung P, Eberth JM. Chronic disease prevalence in the US: sociodemographic and geographic variations by zip code tabulation area. Prev Chronic Dis 2024 Mar 29;21:E14-E10. [doi: 10.5888/pcd21.230267] [Medline: 38426538]
14. Berg S. How these health systems are transforming chronic disease care. American Medical Association. 2024. URL: https://www.ama-assn.org/delivering-care/diabetes/how-these-health-systems-are-transforming-chronic-disease-care [accessed 2025-01-01]
15. Willis CD, Riley BL, Herbert CP, Best A. Networks to strengthen health systems for chronic disease prevention. Am J Public Health 2013 Nov;103(11):e39-e48. [doi: 10.2105/ajph.2013.301249]

16. Faustinella F, Jacobs RJ. The decline of clinical skills: a challenge for medical schools. Int J Med Educ 2018 Jul 13;9:195-197 [FREE Full text] [doi: 10.5116/ijme.5b3f.9fb3] [Medline: 30007951]
17. Williams BC, Ward DA, Chick DA, Johnson EL, Ross PT. Using a six-domain framework to include biopsychosocial information in the standard medical history. Teach Learn Med 2019 Sep 14;31(1):87-98. [doi: 10.1080/10401334.2018.1480958] [Medline: 30216097]
18. Bauer UE, Briss PA, Goodman RA, Bowman BA. Prevention of chronic disease in the 21st century: elimination of the leading preventable causes of premature death and disability in the USA. The Lancet 2014 Jul;384(9937):45-52. [doi: 10.1016/s0140-6736(14)60648-6]
19. Elwyn G, Frosch D, Thomson R, Joseph-Williams N, Lloyd A, Kinnersley P, et al. Shared decision making: a model for clinical practice. J Gen Intern Med 2012 Oct 23;27(10):1361-1367. [doi: 10.1007/s11606-012-2077-6] [Medline: 22618581]
20. Hardcastle MR, Usher KJ, Holmes CA. An overview of structuration theory and its usefulness for nursing research. Nursing Philosophy 2005 Aug 30;6(4):223-234. [doi: 10.1111/j.1466-769x.2005.00230.x]
21. de Jong T. Cognitive load theory, educational research, and instructional design: some food for thought. Instr Sci 2009 Aug 27;38(2):105-134. [doi: 10.1007/s11251-009-9110-0]
22. Bandura A. Social Learning Theory. Englewood Cliffs, NJ: Prentice Hall; 1977.
23. McLeod S. Albert Bandura's social learning theory. In: Simply Psychology. London, UK: Prentice Hall; 2011.
24. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM). Washington, DC: American Psychiatric Association; 2000.
25. Rieger EY, Anderson IJ, Press VG, Cui MX, Arora VM, Williams BC, et al. Implementation of a biopsychosocial history and physical exam template in the electronic health record: mixed methods study. JMIR Med Educ 2023 Feb 21;9:e42364 [FREE Full text] [doi: 10.2196/42364] [Medline: 36802337]
26. Adams NE. Bloom's taxonomy of cognitive learning objectives. J Med Libr Assoc 2015 Jul;103(3):152-153 [FREE Full text] [doi: 10.3163/1536-5050.103.3.010] [Medline: 26213509]
27. Wigfield A, Eccles JS. Expectancy-value theory of achievement motivation. Contemp Educ Psychol 2000 Jan;25(1):68-81. [doi: 10.1006/ceps.1999.1015] [Medline: 10620382]
28. Carraccio CL, Benson BJ, Nixon LJ, Derstine PL. From the educational bench to the clinical bedside: translating the Dreyfus developmental model to the learning of clinical skills. Academic Medicine 2008;83(8):761-767. [doi: 10.1097/acm.0b013e31817eb632]
29. Vermunt NPCA, Harmsen M, Westert GP, Olde Rikkert MGM, Faber MJ. Collaborative goal setting with elderly patients with chronic disease or multimorbidity: a systematic review. BMC Geriatr 2017 Jul 31;17(1):167-112 [FREE Full text] [doi: 10.1186/s12877-017-0534-0] [Medline: 28760149]
30. McEwan D, Harden SM, Zumbo BD, Sylvester BD, Kaulius M, Ruissen GR, et al. The effectiveness of multi-component goal setting interventions for changing physical activity behaviour: a systematic review and meta-analysis. Health Psychol Rev 2016 Nov 13;10(1):67-88. [doi: 10.1080/17437199.2015.1104258] [Medline: 26445201]
31. Sox HC, Higgins MC, Owens DK, Schmidler GS. Medical Decision Making. Hoboken, NJ: Wiley Blackwell; 2024.

Abbreviations

AMA: American Medical Association
CDC: Centers for Disease Control and Prevention
FIU: Florida International University
H&P: history and physical
LGBTQ+: lesbian, gay, bisexual, transgender, queer or questioning, and other diverse sexual orientations and gender identities
SDOH: social determinants or social drivers of health
UC: University of Chicago
UM: University of Michigan
WHO: World Health Organization

Edited by B Lesselroth; submitted 16.09.24; peer-reviewed by F Yang, LE Eberman; comments to author 03.11.24; revised version received 10.12.24; accepted 06.03.25; published 05.06.25.

Please cite as:
Hayer R, Tang J, Bisschops J, Schneider GW, Kirley K, Khan T, Rieger E, Walford E, Anderson I, Press V, Williams B
Implementing the H&P 360 in Three Medical Institutions: Usability Study
JMIR Med Educ 2025;11:e66221
URL: https://mededu.jmir.org/2025/1/e66221
doi:10.2196/66221
PMID:40471655

Comparing the Perceived Realism and Adequacy of Venipuncture Training on an in-House Developed 3D-Printed Arm With a Commercially Available Arm: Randomized, Single-Blind, Cross-Over Study

Susan Gijsbertje Brouwer de Koning[1], PhD; Amy Hofman[2], PhD; Sonja Gerber[3]; Vera Lagerburg[4,5], PhD; Michelle van den Boorn[1], MSc

[1]3D Lab, Department of Computerization, Automation and Medical Technology (iMED), OLVG Hospital, 9 Oosterpark, Amsterdam, The Netherlands

[2]Department of Research and Epidemiology, OLVG Hospital, Amsterdam, The Netherlands

[3]Skillslab, OLVG Hospital, Amsterdam, The Netherlands

[4]Department of Medical Physics, OLVG Hospital, Amsterdam, The Netherlands

[5]Department of Medical Physics and Instrumentation, St. Antonius Ziekenhuis, Nieuwegein, The Netherlands

Corresponding Author:
Susan Gijsbertje Brouwer de Koning, PhD
3D Lab, Department of Computerization, Automation and Medical Technology (iMED), OLVG Hospital, 9 Oosterpark, Amsterdam, The Netherlands

Related Article:

This is a corrected version. See correction statement: https://mededu.jmir.org/2025/1/e89670

Abstract

Background: The venipuncture is one of the most frequently performed procedures in health care. Arm phantoms are available for training, because the procedure itself can be challenging. These phantom arms do not represent a realistic setting and do not offer opportunities to train challenging scenarios.

Objective: This randomized, single-blind study aimed to train health care workers on both a commercially available injection arm and an in-house developed 3D-printed arm, to evaluate the perceived realism and adequacy of training on both arms.

Methods: Participants were trained on both the commercially available arm (arm A) and the 3D-printed arm (arm B). Participants were randomized and blinded from knowing which arm they started training on. A questionnaire was filled in on, among others, the perceived realism of the arm (0 for not realistic, 100 for realistic) and adequacy of the training (inadequate, moderate, or adequate).

Results: A total of 68 participants evaluated the perceived realism of arm A and B, which were scored on average 62.97 (SD 21.47) and 63.79 (SD 17.45), respectively. The difference in perceived realism of the two arms was not statistically significant (based on the paired *t* test, mean difference=−0.82, *P*=.78). Training on arm A was reported inadequate by 7% (5/68 participants), moderately adequate by 31% (21/68), and adequate by 62% (42/68). This was not significantly different from arm B (marginal homogeneity test, *P*=.74), with 4% (3/68), 38% (26/68), and 57% (39/68), respectively, reporting that the training was inadequate, moderately adequate, and adequate.

Conclusions: The 3D-printed arm is as realistic and provides an equally adequate training compared to the commercially available arm. The 3D-printed arm offers the additional possibility to design different models representing several levels of difficulty for vascular morphology. This potentially lowers the number of venipuncture failures by preparing health care workers on challenging scenarios.

(JMIR Med Educ 2025;11:e71139) doi:10.2196/71139

KEYWORDS

venipuncture; phlebotomy; blood draw; 3D printing; vascular phantom

Introduction

The venipuncture (or phlebotomy, blood draw) is one of the most frequently performed procedures in hospitals and is predominantly carried out for diagnostic purposes [1] Although the venipuncture is recognized as a safe and relatively easy procedure, it can be a challenging procedure especially in certain groups, for example, in patients with obesity, patients with a dark skin, or patients undergoing dialysis or chemotherapy. Minor complications include bruising and hematoma and occur in about 12% of venipunctures, while serious complications such as diaphoresis with hypotension or syncope (vasovagal) occur in 3% of patients [2]. There is a minor chance of nerve injury (1/67,000 venipunctures) and also phlebitis is rarely reported [3,4].

The risk of complications is enhanced in case of failure of the venipuncture, that is, a failed needle insertion where no blood samples can be collected, and a second attempt of venipuncture is required. It has been shown that in order to achieve a rate of successful venipunctures close to 99% at the first attempt, one year of practice is necessary [5]. When two attempts both fail for blood sampling, the protocol of most hospitals recommends contacting the anesthesiology department to perform the venipuncture procedure instead. In 2023, this happened 12.233 times at the Onze Lieve Vrouwe Gasthuis (OLVG), a large city hospital in Amsterdam (unpublished data, Business Intelligence Department, OLVG, 2023).

In order to lower this number, proper training and supportive supervision of health care professionals learning the venipuncture is fundamental and therefore recommended by the World Health Organization guidelines on drawing blood [6]. Medical education has evolved from the traditional apprenticeship training to more sophisticated training methods, such as training on simulators (venipuncture arm phantoms) and the use of e-learning modules for theoretical education and electronic simulation [7]. Simulation training can help to improve the venipuncture technique [8]. In addition, training can also improve self-confidence, which also might decrease the number of venipuncture failures [9].

To learn the venipuncture procedure, the current practice is a training on a commercially available arm phantom. This is a phantom of an arm with vessels filled with a liquid that simulates blood. The vessels are embedded in a firm material to enable localization of the vessels by palpation through a layer of skin. Although this arm is a major advancement to practice the venipuncture routine in a safe learning environment yet, there is room to improve training possibilities of the injection arm. Disadvantages of the current training arm are both the realism of the arm and some practical issues.

3D-printing gives the possibility to design and print tailor-made arms. This offers the opportunity to design different versions or models representing several levels of difficulty for vascular morphology to train both novice and experienced health care workers. As a first step towards these advanced tailor-made arms, a regular 3D arm phantom. The aim of this study was to compare the perceived realism and the reported adequacy of the training on a commercially available injection arm to a 3D printed arm.

Methods

Study Design

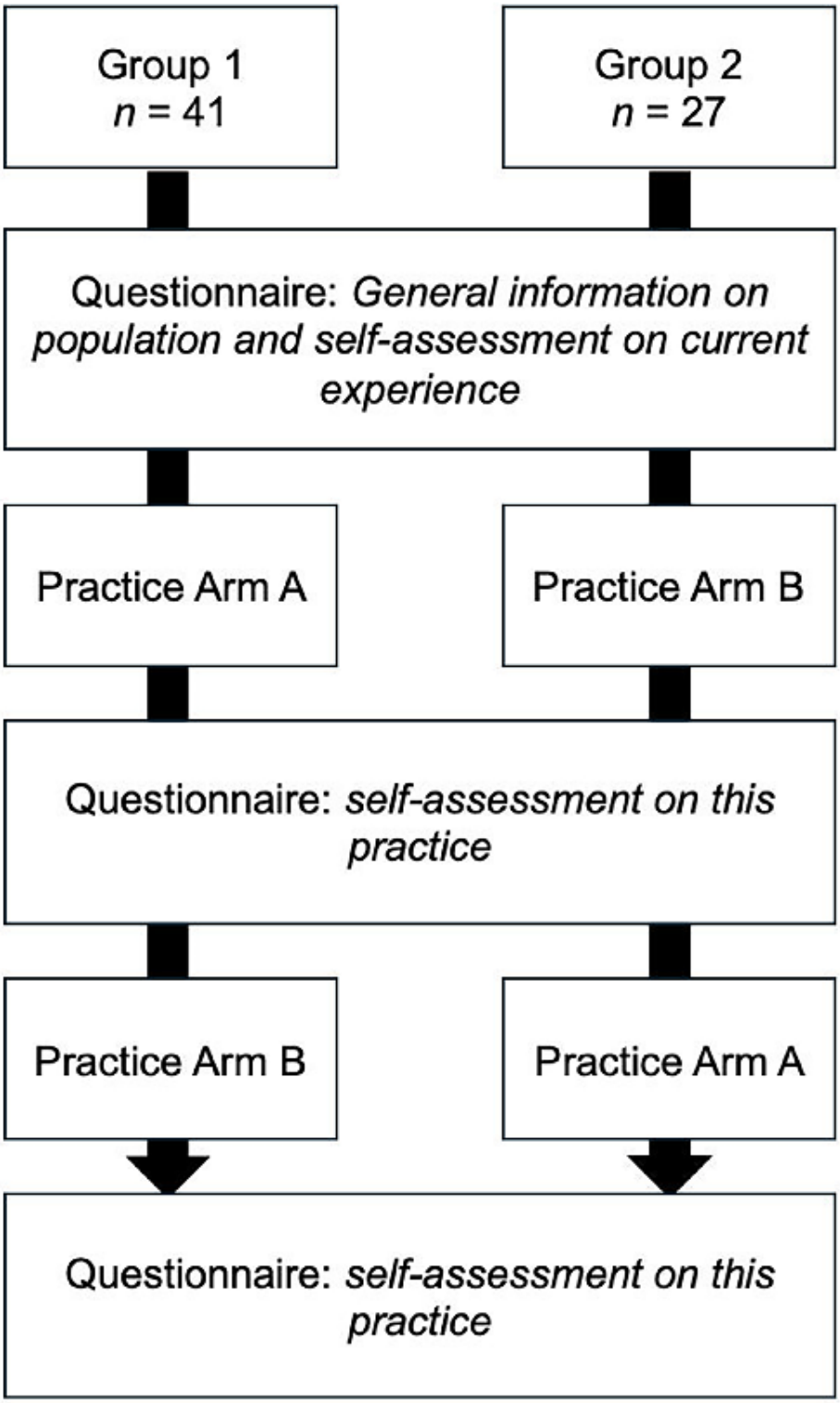
In this randomized, single-blind cross-over comparison study, participants were recruited among the hospital staff of OLVG, a large city hospital in Amsterdam. Information about the study was given via internal communication channels, that is, the intranet and digital newsletter. Participants could apply for the study voluntarily (voluntary response sampling). Requirements for the application were an age of ≥18 years and a medical license to perform peripheral intravenous injections.

All training sessions were supervised by the same trainer. The training started with a short introduction to assess the participant's initial experience and knowledge. A short explanation of the procedure was given, and participants were asked for their learning question. After this, the procedure was demonstrated by the trainer on a training arm. This arm was not part of this study. Participants could practice as often as they needed; the trainer observed and provided feedback on their actions where needed.

Participants were randomized into two groups: group 1 started the training on the commercially available injection arm (arm A) and continued with training on the 3D-printed arm (arm B), while group 2 started the training on arm B and continued training on arm A (Figure 1). The participants received a closed envelope containing the information on which arm to start the training on (A or B). Participants were blinded from knowing which arm was the commercially available arm, and which arm was the 3D-printed arm, as the same sleeves were used for arm A and arm B. Before and after training on each arm, participants were asked to fill in a questionnaire (Figure 2).

Figure 1. Commercially available venipuncture arm phantom (arm A) and the 3D-printed arm (arm B) during the training.

Figure 2. Sequence of training and questionnaires.



Ethical Considerations

Participants had to give informed consent to participate in the study. Participants were able to opt out. No compensation was provided to any participant. The study was approved by the local ethics committee of the OLVG, the Advisory Committee on Scientific Research (approval: WO19.065), and data were collected anonymously.

Measurements

The primary outcome measures were assessed using a questionnaire after each training round: perceived realism of the arm (rated on a scale from 0 [not realistic] to 100 [realistic]) and adequacy/quality of the training (three response categories: inadequate, moderate, or adequate). In addition, information about the participants (eg, role, department, or experience with venipuncture) was collected using these questionnaires (Multimedia Appendix 1).

Sample Size

No published data were available on ratings of realism or training adequacy for phantom arms. Therefore, assumptions were made based on expert and user opinions. For the commercially available arm, a mean realism score of 70 (on a 0 - 100 scale) was assumed, together with a conservatively assumed standard deviation of 20. To detect a mean difference of 10 points with 80% power at a two-sided alpha level of .05, a sample size of 65 participants was required.

Arm Phantoms

The commercially available arm phantom is designed to train venipuncture and intravenous infusion (NASCO Healthcare). Veins are accessible at the antecubital fossa, along the forearm, and at the back of the hand. There are visible veins, as well as invisible veins that can only be determined by palpation. Artificial blood can be drawn from the vessels (NASCO Healthcare, Life/form®). The arm is designed with a skin that has a soft touch when palpating the vessels.

The commercially available arm phantom for venipuncture was the basis of the design of the tailor-made 3D-printed arm: a digital 3D model of the commercially available arm was acquired through a 3D surface scan (Artec EVA 3D scanner, Artec Studio 15 Software). The digital 3D model was edited in Meshmixer (Autodesk, Inc.) and prepared for printing in the program Simplify3D. The arm was printed from glycol modified polyethylene terephthalate using the Ultimaker printer (Ultimaker BV). Silicone tubes were used to simulate the vessels (Dispomed Ltd; Figure 3). The skin of the arm was similar to the skin of the commercially available skin. This was the first time a 3D-printed arm was produced by our laboratory. There was no pilot performed on this arm in advance of using the arm for the training with participants.

Figure 3. The 3D printed arm.

Statistical Analysis

The study population was characterized using descriptive statistics. Continuous variables were presented as mean (SD) or median (IQR), depending on normality (assessed by visual inspection of histograms and Q-Q plots). Categorical variables were presented as frequencies and percentages (n, %).

As the primary outcomes were rated twice by the same participants (cross-over design), paired tests were applied. The difference in perceived realism of the arm was evaluated using a paired samples *t* test. The difference in adequacy ratings of the training was evaluated using a marginal homogeneity test for related samples.

To explore potential sequence effects (order of training, determined by randomization), results were visualized per randomization group, and perceived realism scores were compared between groups using independent sample *t* tests.

All tests were two-sided; *P* values below .05 were considered statistically significant. Analyses were performed using SPSS version 27 (IBM).

Results

Study Population

Characteristics of group 1 and group 2 are presented in Table 1. Most participants in both groups were nurses or physician assistants (Group 1 30/41, 73%; Group 2 15/27, 56%)). The median number of years of experience with the venipuncture was 1 year for both groups (IQR 0 - 4 and 0 - 3, for groups 1 and 2, respectively). Group 1 rated their competence in venipuncture on a 0 - 100 scale with a median of 50 (IQR 18 - 65), while group 2 rated their competence with a median of 40 (IQR 20 - 70). The self-rate of successful venipunctures at first attempt was approximately equal between the groups, with a median of 40 on a 0 - 100 scale. Half of the total population (34/68, 50%) performed a venipuncture daily over the last 6 months. Almost all participants (57/68, 83.8%) had a venipuncture training before, mostly on phantom arms.

Table . Characteristics and self-reported experience of study population.

Characteristic	Total population (N=68)	Group 1 (N=41)	Group 2 (N=27)
Role, n (%)			
Nurse or nurse specialist	45 (66.2)	30 (73.2)	15 (55.6)
Nurse in training	5 (7.4)	3 (7.3)	2 (7.4)
Medical doctor	3 (4.4)	1 (2.4)	2 (7.4)
Medical intern	3 (4.4)	1 (2.4)	2 (7.4)
Other[a]	12 (17.6)	6 (14.6)	6 (22.2)
Department, n (%)			
Surgery	6 (8.8)	6 (14.6)	0 (0.0)
Geriatrics	12 (17.6)	5 (12.2)	7 (25.9)
Gynecology	14 (20.6)	7 (17.1)	7 (25.9)
Intensive Care	5 (7.4)	4 (9.8)	1 (3.7)
Lung	2 (2.9)	2 (4.9)	0 (0.0)
Gastroenterology	3 (4.4)	2 (4.9)	1 (3.7)
Neurology	4 (5.9)	2 (4.9)	2 (7.4)
Oncology	3 (4.4)	2 (4.9)	1 (3.7)
Orthopedics	5 (7.4)	3 (7.3)	2 (7.4)
Emergency room	2 (2.9)	2 (4.9)	0 (0.0)
Other[b]	12 (17.6)	6 (14.6)	6 (22.2)
Mean number of years of experience with venipuncture, median (IQR)	1 (0 - 4)	1 (0 - 4)	1 (0 - 3)
Self-rating competence in venipuncturing (0 - 100), median (IQR)	50 (20 - 70)	50 (18 - 65)	40 (20 - 70)
Self-rating of successful venipunctures at first attempt (0 - 100), median (IQR)	40 (0 - 70)	40 (1 - 70)	40 (0 - 60)
Number of venipunctures within the last six months, n (%)			
Once a month or less	15 (22.1)	12 (29.3)	3 (11.1)
A few times a month	7 (10.3)	5 (12.2)	2 (7.4)
A few times a week	12 (17.6)	6 (14.6)	6 (22.2)
On a daily basis	34 (50.0)	18 (43.9)	16 (59.3)
Number of failed attempts and getting help over the last month, n (%)		0 (0 - 2)	1 (0 - 2)
I have had a venipuncture training before	57 (83.8)	33 (80.5)	24 (88.9)
I have had a training on a commercial venipuncture arm phantoms before	39 (57.4)	23 (56.1)	16 (59.3)

[a]Other: students medicine/technical medicine, obstetrician, dialysis assistant, radiologist.

[b]Other: radiology, ophthalmology, medical technology, internal medicine, otorhinolaryngology, psychiatry.

Perceived Realism of the Commercial Venipuncture Arm Phantom and the 3D-Printed Arm Phantom

The perceived realism of arms A and B in the total population were on average 62.97 (SD 21.47) and 63.79 (SD 17.45), respectively. There was no statistically significant difference between the reported perceived realism of arm A and B (mean difference −0.82, 95% CI −6.80 to 5.16, *P*=.78).

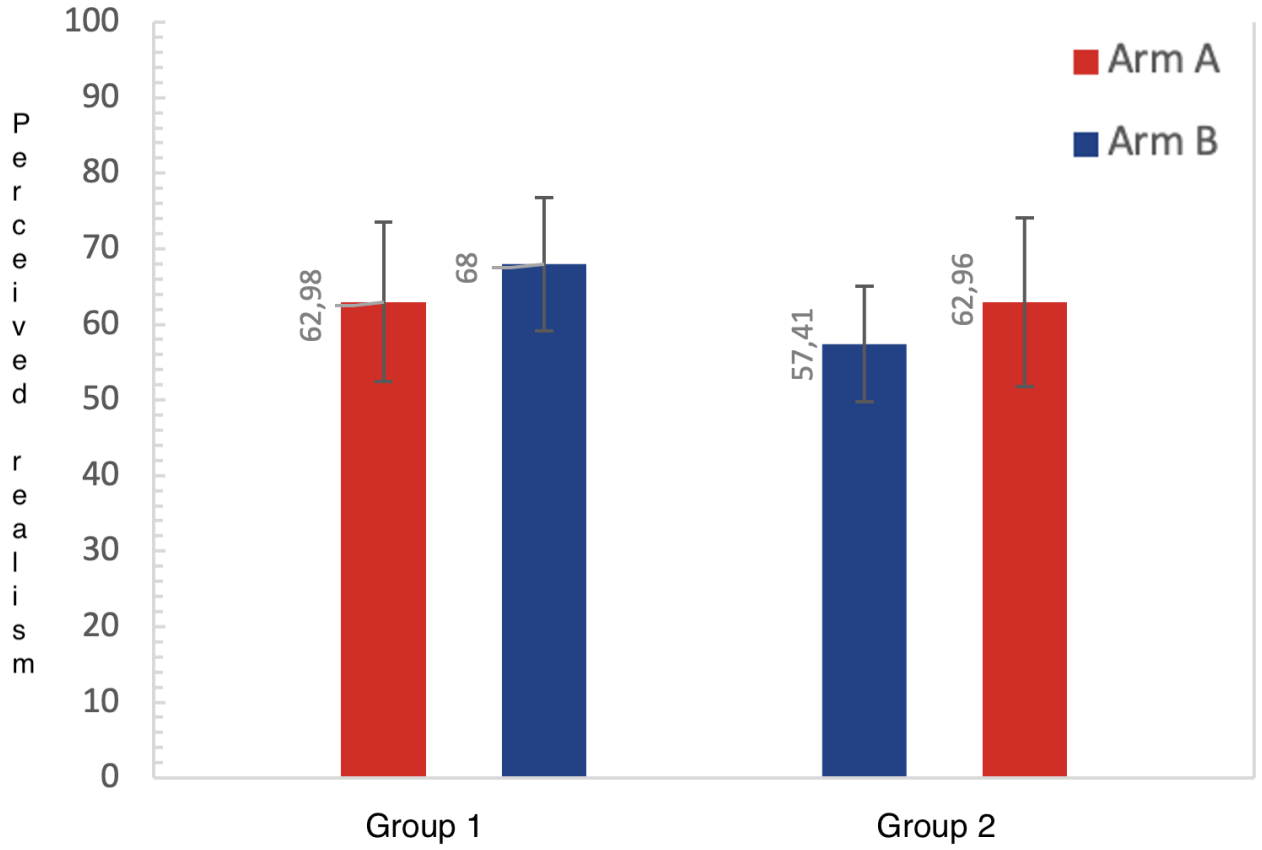
Adequacy of the Training With Commercial Venipuncture Arm Phantom and the 3D-Printed Arm Phantom

Training on arm A was reported inadequate by 5/68 (7%) of the participants, moderately adequate by 21/68 (31%), and adequate by 42/68 (62%). This was not significantly different from arm B (*P*=.74), with 4% (3 out of 68), 38% (26 out of 69) and 57% (39 out of 68), respectively, reporting that the training was inadequate, moderately adequate, and adequate.

Perceived Realism Per Randomized Group

To explore whether training sequence influenced the reporting of perceived realism, Figure 4 shows perceived realism scores stratified by randomized groups. For the commercially available arm (arm A), perceived realism did not differ between groups (mean difference 0.01, 95% CI −10.69 to 10.71, *P*=.998). In contrast, the perceived realism of the 3D-printed arm phantom (arm B) was reported to be significantly lower in group 2, who started training on this arm (mean difference 10.59, 95% CI 2.29 to 18.89, *P*=.013).

Figure 4. The perceived realism of arm A (commercially available venipuncture arm phantom) and B (3D-printed arm phantom) (0 for not realistic, 100 for realistic).



Discussion

Principal Findings

The objectives of this study were to train health care workers on both a commercially available injection arm and an in-house developed 3D-printed arm and to evaluate the perceived realism and adequacy of training on both arm phantoms. Our results indicate that the 3D-printed arm phantom is as realistic and provides an equally adequate training compared to the commercially available arm phantom. The perceived realism of the commercially available arm phantom and the 3D-printed arm phantom were on average 62.97 (out of 100, SD 21.47) and 63.79 (SD 17.45), respectively (*P*=.784). Training on the commercially available arm was reported inadequate by 5/68 (7%) of the participants, moderately adequate by 21/68 (31%), and adequate by 42/68 (62%). This was not significantly different from the 3D-printed arm phantom (*P*=.739), with 4% (3/68), 38% (26/69) and 57% (39/68), respectively, reporting that the training was inadequate, moderately adequate, and adequate.

Comparison to Prior Work

Our results confirm the results of recently published studies [10,11] indicating the perceived realism of 3D-printed arm phantoms for practicing puncturing techniques is sufficient. Hyndman et al [10] printed a 3D-model of the upper limb vasculature to practice the Seldinger technique and assessed student satisfaction on 31 medical students. The use of the 3D-printed model improved anatomic understanding and application of the Seldinger technique of the students. Raffaela et al [11] developed a 3D-printed pediatric phantom to train ultrasound-guided placement of peripheral central venous catheters in children. The model was rated as highly realistic in terms of morphology and functionality for the overall simulation, by 20 expert specialists.

Strengths and Limitations

One of the strengths of this study is that the characteristics of the two groups were comparable. Second, the questionnaire provided very valid suggestions for the improvement of adequacy of training on the next version of the 3D-printed arm, for example, a skin that endures more injections and could be easily replaced, a less heavy phantom to enable easy positioning of the arm, a skin that is thinner to resemble a real skin, and different stages of difficulty in the arm phantom, like thinner, smaller and rolling vessels. A third strength of this study is the fact that a 3D-printed arm phantom can lower the costs of training significantly and that the design can be accessible worldwide. The costs of the 3D-printed arm phantom (approximately US $80) are substantially lower compared with the commercially available arm phantom (approximately US $1100). A digital model for 3D-printing can be shared for free, like the 3D-simulator for arthroscopy given by Ferras et al [12] can be downloaded and printed for free in any 3D printer. Sritharan et al [13] are currently performing a review on how the use of open-source databases for the distribution of simulator designs used for 3D-printing can promote credible solutions for health care training while minimizing the risks of commercialization of designs for profit.

One of the limitations of the study is the potential training effect, as both randomized groups reported slightly higher perceived realism of the second training arm (although not statistically significant). Second, the limited sample size of 68 participants may reduce generalizability. A third limitation is the variability in participant experience, which could potentially influence perceptions of realism. Furthermore, the timeline of distribution into the two groups was not comparable (at the start, most participants started in group 1). As a result, the skin of particularly the 3D printed arm showed multiple injection sites and there was leakage of blood through these injection sites. This might be a reason why participants rated the arms as less realistic. As a result of these limitations, further research is recommended to validate our findings in a broader population. Ultimately, additional outcome measures, including success rates or skill retention, could be added.

Future Directions

Our study shows that the 3D-printed arm phantom is comparable to the commercially available arm phantom in terms of perceived realism and in the possibility to offer an adequate training of the venipuncture. The advantage of 3D-printing is the possibility that it offers to design and print tailor-made arms. In a 3D-printed arm, different vascular morphologies can be designed like vascular morphologies that are palpable and superficially located or deep vascular morphologies that need ultrasound guidance to puncture instead. Moreover, vascular access with aneurysms can be designed to train adequate cannulation in these challenging scenarios. A discipline in which these challenging scenarios are very common is kidney dialysis. The Kidney Disease Outcomes Quality Initiative 2019 strongly recommends training and retraining hemodialysis staff to ensure the maintenance of competency of cannulation skills [14]. They encourage future research to evaluate whether training on cannulation simulation models improves cannulation competency and reduces cannulation complications. A 3D-printed tailor-made arm would offer a simulation model adjusted for this application, in line with the strong recommendation of the Kidney Disease Outcomes Quality Initiative.

An additional advantage of the 3D-printed arm over the commercially available arm is the possible improvements on the skin. The skin can be designed in such a way that it can be easily removed for proper cleaning of blood leakage from the punctured vessels. This was a time-consuming task with the commercially available arm. Also, a different material can be chosen that does not show the location of the previous punctures and thus forces the trainee to palpate the localization of the vessels.

With implementation of the 3-D printed arm phantom for venipuncture training, substantial amounts of health care cost can be saved; when offering adequate training opportunities for venipuncture, the number of contacting the anesthesiology department to perform the venipuncture procedure instead can be lowered.

Conclusions

The 3D-printed arm phantom is as realistic and provides an equally adequate training compared to the commercially available arm phantom. However, the 3D-printed arm offers the possibility to design different models representing several levels of difficulty for vascular morphology. This potentially lowers the number of venipuncture failures by preparing health care workers on challenging scenarios.

Acknowledgments

No AI was used in the manuscript generation.

Data Availability

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

Authors' Contributions

SBDK: Conceptualization; data curation; formal analysis; investigation; methodology; resources; validation; visualization; writing—original draft;
AH: Investigation, methodology, formal analysis, writing—review and editing

SG: Investigation; resources; data curation, project administration,
VL: Investigation, methodology, formal analysis, supervision; writing—original draft, resources writing—review and editing
MB: Investigation, project administration, resources, writing—review and editing

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questionnaire
[DOCX File, 16 KB - mededu_v11i1e71139_app1.docx]

References

1. Cheung E, Baerlocher MO, Asch M, Myers A. Venous access: a practical review for 2009. Can Fam Physician 2009 May;55(5):494-496. [Medline: 19439704]
2. Galena HJ. Complications occurring from diagnostic venipuncture. J Fam Pract 1992 May;34(5):582-584. [Medline: 1578208]
3. Tsukuda Y, Funakoshi T, Nasuhara Y, Nagano Y, Shimizu C, Iwasaki N. Venipuncture nerve injuries in the upper extremity from more than 1 million procedures. J Patient Saf 2019 Dec;15(4):299-301. [doi: 10.1097/PTS.0000000000000264] [Medline: 27314202]
4. Lavery I, Ingram P. Venepuncture: best practice. Nurs Stand 2005;19(49):55-65. [doi: 10.7748/ns2005.08.19.49.55.c3936] [Medline: 16134421]
5. Vuk T, Cipek V, Jukić I. Blood collection staff education in the prevention of venepuncture failures and donor adverse reactions: from inexperienced to skilful staff. Blood Transfus 2015 Apr;13(2):338-339. [doi: 10.2450/2014.0216-14] [Medline: 25545873]
6. Geneva: world health organization. WHO guidelines on drawing blood: best practices in phlebotomy. 2010. URL: https://www.who.int/publications/i/item/9789241599221 [accessed 2025-10-20]
7. Maran NJ, Glavin RJ. Low- to high-fidelity simulation - a continuum of medical education? Med Educ 2003 Nov;37 Suppl 1(37):22-28. [doi: 10.1046/j.1365-2923.37.s1.9.x] [Medline: 14641635]
8. Fujii C. Comparison of skill in novice nurses before and after venipuncture simulation practice. JNEP 2014;4(5). [doi: 10.5430/jnep.v4n5p16]
9. Fincher RME, Lewis LA. Learning, experience, and self-assessment of competence of third-year medical students in performing bedside procedures. Acad Med 1994 Apr;69(4):291-295. [doi: 10.1097/00001888-199404000-00012] [Medline: 8155237]
10. Hyndman D, McHugh D. Simulation-based medical education: 3D printing and the Seldinger technique. IME 2024;3(3):180-189. [doi: 10.3390/ime3030016]
11. Raffaele A, Mauri V, Negrini M, et al. Elaboration and development of a realistic 3D printed model for training in ultrasound-guided placement of peripheral central venous catheter in children. J Vasc Access 2024 Nov;25(6):1767-1774. [doi: 10.1177/11297298231187005] [Medline: 37434535]
12. Ferràs-Tarragó J, Jover-Jorge N, Miranda-Gómez I. A novel arthroscopy training program based on a 3D printed simulator. J Orthop 2022;32:43-51. [doi: 10.1016/j.jor.2022.04.006] [Medline: 35601206]
13. Sritharan M, Siraj S, Brunton G, Dubrowski A. Exploring the distribution of 3D-printed simulator designs using open-source databases to facilitate simulation-based learning through a university and nonprofit collaboration: protocol for a scoping review. JMIR Res Protoc 2024 May 27;13:e53167. [doi: 10.2196/53167] [Medline: 38801764]
14. Lok CE, Huber TS, Lee T, et al. KDOQI clinical practice guideline for vascular access: 2019 update. Am J Kidney Dis 2020 Apr;75(4 Suppl 2):S1-S164. [doi: 10.1053/j.ajkd.2019.12.001] [Medline: 32778223]

Edited by T Leung; submitted 10.01.25; peer-reviewed by D McHugh, SJ Foley; revised version received 29.08.25; accepted 29.08.25; published 04.11.25.

Please cite as:
Brouwer de Koning SG, Hofman A, Gerber S, Lagerburg V, van den Boorn M
Comparing the Perceived Realism and Adequacy of Venipuncture Training on an in-House Developed 3D-Printed Arm With a Commercially Available Arm: Randomized, Single-Blind, Cross-Over Study
JMIR Med Educ 2025;11:e71139
URL: https://mededu.jmir.org/2025/1/e71139
doi:10.2196/71139

Resilience Training Web App for National Health Service Keyworkers: Pilot Usability Study

Joanna Burrell[1*], BA, PGCert, DClinPsy; Felicity Baker[2*], BSc, MPhil, DClinPsy; Matthew Russell Bennion[3,4*], BEng, MSc, MBA, PhD

[1]Department of Psychology, University of Sheffield, Cathedral Court, Sheffield, United Kingdom

[2]Ultimate Resilience Ltd, Nottingham, United Kingdom

[3]Department of Computer Science, The University of Sheffield, Sheffield, United Kingdom

[4]Digital Innovation Unit, NHS Midlands and Lancashire Commissioning Support Unit, Stoke on Trent, United Kingdom

[*]all authors contributed equally

Corresponding Author:
Joanna Burrell, BA, PGCert, DClinPsy
Department of Psychology, University of Sheffield, Cathedral Court, Sheffield, United Kingdom

Abstract

Background: It is well established that frontline health care staff are particularly at risk of stress. Resilience is important to help staff to manage daily challenges and to protect against burnout.

Objective: This study aimed to assess the usability and user perceptions of a resilience training web app developed to support health care keyworkers in understanding their own stress response and to help them put into place strategies to manage stress and to build resilience.

Methods: Nurses (n=7) and other keyworkers (n=1), the target users for the resilience training web app, participated in the usability evaluation. Participants completed a pretraining questionnaire capturing basic demographic information and then used the training before completing a posttraining feedback questionnaire exploring the impact and usability of the web app.

Results: From a sample of 8 keyworkers, 6 (75%) rated their current role as “sometimes” stressful. All 8 (100%) keyworkers found the training easy to understand, and 5 of 7 (71%) agreed that the training increased their understanding of both stress and resilience. Further, 6 of 8 (75%) agreed that the resilience model had helped them to understand what resilience is. Many of the keyworkers (6/8, 75%) agreed that the content was relevant to them. Furthermore, 6 of 8 (75%) agreed that they were likely to act to develop their resilience following completion of the training.

Conclusions: This study tested the usability of a web app for resilience training specifically targeting National Health Service keyworkers. This work preceded a larger scale usability study, and it is hoped this study will help guide other studies to develop similar programs in clinical settings.

(JMIR Med Educ 2025;11:e51101) doi:10.2196/51101

KEYWORDS

resilience; workplace stress; National Health Service; NHS keyworker; digital learning; digital health; usability; feasibility; mental health; pilot study; learning; training; exercise; primary care provider; health care professional; occupational health; worker; hospital; emergency; survey; questionnaire; mobile phone

Introduction

Resilience allows individuals to manage everyday challenges and changes. For health care professionals who are working in highly emotive and stressful situations, resilience skills are particularly important [1]. It is well established that frontline staff such as nurses are particularly at risk of stress due to factors such as long shifts, organizational pressures, and the emotional impact of their work [2]. During the COVID-19 pandemic, there were high rates of mental health problems among health care staff. For example, a survey of 255 nurses working with respiratory patients found 21% to be experiencing moderate or severe symptoms of anxiety and 17.2% to be experiencing depression. A total of 18.9% scored low or very low on a measure of resilience [3]. A study of 1106 physicians also reported high levels of anxiety and low levels of resilience during the pandemic [4]. Building emotional resilience is therefore imperative to prevent burnout in health care staff, to keep them healthy both physically and mentally, to improve well-being, and to ensure job retention in the workplace [1].

One way that employers can meet this need is through the provision of resilience training. Key benefits of resilience training include improvements in mental health and well-being, social support, self-efficacy, and coping. Further benefits include

improved ability to adapt to pressures and demands in the workplace and other areas of life [5]. In health care settings, nurse resilience interventions have been highlighted as a potential way of enhancing staff coping and well-being, job satisfaction, and retention [2]. Greater nurse resilience has also been associated with better work performance [6].

A constraint of traditional resilience training programs is the time required to attend in person, which can exclude certain staff groups, such as nurses, from participation. Smartphone apps have the potential to offer training in resilience to large numbers of people while overcoming barriers, such as stigma, time, and acceptability, and can be integrated easily into the wider organizational well-being strategy [7-11].

The aim of this study was to evaluate whether health care keyworkers would be willing to carry out resilience training via an online platform specifically designed to enable them to understand their own stress response and put in place strategies to manage stress, build emotional resilience, and maintain well-being. The data collected would generate important information for future implementation, while contributing feedback for a more refined usability study with this population.

Methods

Participants

The recruitment process was carried out by the Medical Devices Testing and Evaluation Centre (MD-TEC) team. A study sample was recruited from the University Hospitals Birmingham National Health Service (NHS) Foundation Trust, and participation was incentivized with Continuing Professional Development credits for participation. This study was advertised over the web via the internal trust-wide communications bulletin and targeted emails. There was not an enforced inclusion criterion, but this study requested for participants who were a nurse or health care professional and worked in either the emergency department, intensive care unit, or critical care.

Ethical Considerations

This study was run as a formative usability study by MD-TEC with human participants. The University of Sheffield Re-Use of Existing Data Questionnaire was completed, and the Psychology Research Ethics Committee deemed this study exempt from ethical approval because the data were fully anonymized. A short self-declaration form was submitted. This application went to the Psychology Departments Ethics Administrator for a final check before a letter of confirmation was issued.

Informed consent was derived through the sharing of a recruitment flyer with potential participants. This explained the research and its function as part of medical device usability testing for further development.

All the data were received fully anonymized from MD-TEC post study. The participants were not personally identifiable by the researchers. Research participants were offered Continuing Professional Development credits for their participation. They self-selected for and undertook the research voluntarily.

The training web app was developed by the third author without funding. The content of the training web app drew on the Skills-Based Model of Personal Resilience [12] and included a selection of evidence-based skills and exercises to regulate distress emotions and build positive emotions, such as slow rhythmic breathing and mindfulness practice. The selection of skills and exercises were chosen for their capacity to provide maximum benefit to participants, calming stress, and facilitating improved coping, in the context of this brief trial.

Bennion et al [8] highlight four key indicators of quality drawn from effective digital psychotherapy approaches. These include clinician involvement, academic involvement, research, or other evidence and use of specific psychological approach or theory. The intervention followed these recommendations, drawing on academic and clinical theory [13] and involving clinicians, academics, and computer scientists in its development to ensure greater quality and effectiveness.

The web-based training was published using Articulate Storyline (Articulate Global, LLC) and accessed via a web browser. It consisted of both written and spoken content on a series of slides, short videos, and experiential exercises which could be moved through at participants' own pace using "previous" and "next" buttons. The estimated time to complete the training was 20 minutes.

Pretraining Questionnaire

The pretraining questionnaire captured basic demographic information: gender, age range, job role, current area of work, current band, and years of nursing experience. Current job stress was rated on a 5-point scale (never, hardly ever, occasionally, sometimes, and very). Participants were also asked whether they had heard of or previously undertaken any resilience training.

Posttraining Feedback Questionnaire

The posttraining feedback questionnaire focused on 6 areas: app design and navigation, app content, app impact, app training exercises, app relevance, and app access. Each question was posed on a Likert scale with five possible answer options to allow the user to respond to each statement on a range from "strongly agree" to "strongly disagree."

Procedure

Upon contacting the MD-TEC team to participate, interested individuals were provided the opportunity to ask any questions about participation in this study. If willing to consent, participants were then sent the link and password to access the training. The training and surveys were hosted on the MD-TEC Software Usability Testing Site (MD-TEC), and thus could be completed on any device with internet access. Once logged in on an internet browser, individuals were presented with the precourse survey before completing the full training module.

Once the training module was completed, participants were taken to a landing page and requested to click a link to take them to the feedback survey. They were reminded at this point that no identifiable information would be collected from them. As the surveys were completed anonymously, participants who completed the training and survey were asked to inform the

MD-TEC team via email once they had done so. They were then sent a certificate toward Continuing Professional Development for their time contributed to research, which they could add to their personal records. The total time for each participant to complete the training module and feedback survey was approximately 45 minutes.

Statistical Analysis

This study did not use a specific sample size calculation as it was focused on app usability. It instead aimed to achieve at least 5 participants which is deemed an optimal number to reveal 77% to 85% of problems [14]. Data were analyzed using IBM SPSS (version 26; IBM Corp). The pretraining and posttraining feedback questionnaires were summarized as a mix of continuous variables with medians and categorical ordinal variables with percentages.

Results

Participants

The age of participants ranged from 25 to 64 years. The total sample (N=8) was comprised of 8 (100%) females, 7 (87.5%) nurses, and 1 (12.5%) keyworker of other professions. Grades ranged from 5 to 7 with a median of 6 (IQR 5-6). Five of the participant's had over 15 years of nursing experience. All 8 participants completed baseline measures and posttraining measures.

Pretraining Questionnaire

Sample Overview

Of the 8 participants who completed this study, 3 (37.5%) worked in the hospital's intensive care unit, 1 (12.5%) worked in the emergency department, and 4 (50%) worked in other undisclosed areas of the hospital.

Current Role Stress

Most participants (6/8, 75%) rated their current role stress as "sometimes" stressful, while 1 of 8 (12.5%) said "occasionally" stressful and 1 of 8 (12.5%) said "very" stressful.

Awareness and Knowledge of Resilience Training

Most participants (6/8, 75%) had heard of resilience training, and those that had taken part (4/8, 50%) had done so in a face-to-face setting.

Posttraining Feedback Questionnaire

App Design and Navigation

Feedback regarding the design of the training was predominantly positive. All participants found the training easy to navigate, 6 of 8 (75%) deemed the default speed at which the training progressed to be acceptable, and 7 of 8 (97.5%) thought the appearance of the buttons was OK.

App Content

Feedback for the content indicated that all participants (8/8, 100%) found the training easy to understand, 6 of 8 (75%) felt there was enough text content, 4 of 8 (50%) felt there was enough spoken content, and 5 of 8 (62.5%) felt there were enough interactive exercises.

App Impact

A large number of the participants (5/7, 71%) agreed that the training increased their understanding of both stress and resilience, while 6 of 8 (75%) agreed that the resilience model had helped them to understand what resilience is.

App Training Exercises

The training exercise feedback was positive but varied. For the breathing and positive tips exercises, 6 of 8 (75%) participants agreed they were likely to try the exercises again in the future. The mindfulness exercise had 4 of 8 (50%) participants agree they were likely to try the exercise again.

App Relevance

There was a high level of agreement that the training was relevant to nurses, with 6 of 8 (75%) participants agreeing that the content was relevant to them.

Furthermore, 6 of 8 (62.5%) participants agreed that they were likely to act to develop their resilience following completion of the training.

Access to Training

All the participants indicated a different personal preference to how they would prefer to access the training. Participants felt the package should be made available across all platforms to allow the training to be completed where and when it was most convenient to them. When asked their preferred location of access, 5 of 8 (62.5%) indicated their preference as being "at home."

Discussion

Principal Findings

We explored the perceived usability and feasibility of a resilience training web app created for NHS health care keyworkers. Data collected covered a number of areas: design and navigation, content, impact, and relevance. The results showed that 100% (8/8) of participants found the training easy to understand and agreed that it had increased their understanding of both stress (5/7, 71%) and resilience (6/8, 75%). Three-quarters of participants agreed that the content was relevant to them, and this corresponded with the number of participants rating their current role as "sometimes" stressful. Furthermore, three-quarters of participants agreed that they were likely to take action to develop their resilience following completion of the training. This information was used to inform the design of a larger usability study.

A total of 8 participants were recruited, with 7 being from the target population. All participants completed the process from start to finish. Participants successfully carried out what was required of them based on this study's protocol, although some participants did not complete all the questions asked on the posttraining questionnaire. There was no indication given as to why this was the case. In a follow up usability study [13] validation checks were put in place within the surveys to stop questions from being missed by mistake.

The findings of this study indicated that participants found the training app design and navigation acceptable and usable. However, the measure used was not a standard model of system usability (eg, International Organization for Standardization, 2018). This study's design was updated to use two validated measures (the System Usability Scale and the Usability Metric for User Experience) to strengthen the robustness of a follow-up usability study [14]. Adding these two additional validation measures to this study's design helped to strengthen assessment of the training app's usability.

Participants indicated that the training was easy to understand and that there was enough text content; however, they also indicated that there was a need for the training to have more spoken and interactive content. This fits with a recent study [15] in which nurses' interactive behavior was identified as an influencing aspect of nurse satisfaction with online learning. Based on these findings, we recommend the training's interactive content be revisited in its next design iteration.

Most participants perceived that the training increased their understanding of both stress and resilience and that the resilience model had helped them to understand what resilience is. A more robust method of measurement was required to further explore the impact of the training and this study's design was updated to incorporate ratings of perceived knowledge regarding stress and resilience. These new scales were used in a follow up usability study [14] and found to increase significantly between pre- and postapp training.

The training exercise feedback was positive but varied. Both the breathing and positive tips exercises were well received, with participants agreeing they were likely to try the exercises again in the future. However, only half of the participants agreed that they would try the mindfulness exercise again. This may have been due to the difficulty in carrying out the exercise in a busy work environment.

Many participants agreed that the training was relevant to them and believed that they were likely to act to develop their resilience following completion of the training.

Limitations

Recognized limitations of usability studies include that testing is conducted in an artificial situation and personal preferences of the participants are not representative of the wider user population [16]. The digital training app used in our study is an early prototype. This may need multiple design developments to create a smartphone app that can be used to deliver the resilience training. The aim of this formative usability study was to assess the acceptability and user perceptions of the current version of the training program. As such, this study is part of the iterative product development process and is different to a summative usability study, conducted for validation and regulatory purposes [17].

This study had a single-group design and advertised for a specific group; however, anyone employed by the trust who contacted MD-TEC regarding this study could be involved. This was done primarily to allow anyone employed by the trust to gain access to training that could benefit them. Potential participants who were unaware of this clause may have been lost because of this decision. The initial training materials were designed with nurses in mind but were not specifically tailored for the demographic. This may have changed participants' initial perception of the suitability of the training to them personally. A single-group design can limit the ability to draw definitive conclusions about the effectiveness of the training due to its lack of a control or comparison group [18]. However, since this study was focused on the usability of the training and not the effectiveness, and it was not seeking to make a comparative analysis, a single-group design was appropriate.

This study limited its evaluation to perceived usability, which was not obtained through laboratory-based observations. As such, the positive ratings reported may not be representative of true user experience. A heuristic evaluation of the training to detect usability problems was not carried out, due to pandemic restrictions making this problematic to implement. This study used quantitative scales and measures to collect data but did not use qualitative measures to gain deeper insight into what NHS health care staff felt about the training. A measure of time spent using the training was not collected. This could have also given an indication of acceptability. This study used two single Likert scales to measure perceived increases in knowledge about stress and resilience. Studies have shown that perceptions of learning may not reflect knowledge gains, when compared with evidence of actual learning [19]. A more robust method of measuring knowledge retention would have benefitted this study. This could have been achieved by having a pre- and postquiz based on the content of the training to see what knowledge was retained.

While the majority of participants gave positive responses in the evaluation of this study, the generalizability of these outcomes is limited due to the disproportionate number of female participants and participants from a nursing background. Only 1 (12.5%) participant was from a different professional group. This limits the inferences that can be drawn about usability and acceptability of the training to male participants and those with other keyworker roles. It is recommended that future studies recruit a more representative sample to enhance generalizability of the results.

Conclusions

Overall, the resilience training module was well received by the participants. The participants felt the package was easy to navigate. There was a high level of agreement that the visual delivery of the training was acceptable, as well as the speed at which this was delivered.

A number of techniques demonstrated during the training were also well received, with 6 of 8 participants agreeing that they would use them in future stressful situations. Mindfulness was the only exercise that received more varied feedback, with half agreeing on its utility in the work environment.

Health care staff participating in this study largely agreed that the training was relevant to their group and that the tone of the delivery was appropriate. No clear preference regarding how to access the training was identified, highlighting the need for accessibility via computer, tablet, and smartphone. Participants

expressed a wish to to access the training when they have a moment of need and the opportunity in their busy working day.

Future Directions

As one of the first NHS web-based resilience programs to be tested, this first usability study aimed to understand whether web-based training for resilience is deemed usable and acceptable by health care staff. The results of this study will be used to expand and build upon the initial prototype to make a more interaction enriched version of the training.

This study also provided an understanding of the program's limitations and highlighted some aspects which require further adaptation for delivery via a new medium. Future research would aim to evaluate the impact of including greater interactivity on engagement and learning. It would also aim to extend the accessibility and acceptability of the program to a wider audience by developing an effective prototype for a smartphone app.

This study was run externally by MD-TEC, who had their own processes for running usability studies of this nature. This study's design covered some of the key factors required for an effective online survey, but it could have been further improved by seeking acknowledgment with MD-TEC regarding the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) checklist [20].

It is clear from the results that there is a need for future research to evaluate how skills-based learning using web-based training impacts long term resilience. A larger scale study would allow for more in-depth investigation of the impact of such training on participants' levels of stress and resilience as well as their perspectives on acceptability.

Given the diversity of NHS staff, it will be important for any future study to gather a wide set of demographic information to investigate acceptability and generalizability across diverse populations. With increasing awareness (ie, gained through the COVID-19 pandemic) of the pressures faced by all NHS staff, across a breadth of ethnic and socioeconomic groups, a larger scale study would allow for a wider inclusion criterion covering all NHS staff groups.

Acknowledgments

This study was supported by a European Regional Development Fund (ERDF) awarded to FB and JB. We thank the National Institute for Health and Care Research Trauma Management MedTech Co-Operative for facilitating the usability study.

Data Availability

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

Authors' Contributions

JB did the conceptualization, methodology, writing of the original draft, review and editing of the writing, visualization, supervision, project administration, and funding acquisition. FB handled the conceptualization, methodology, writing of the original draft, review and editing of the writing, visualization, and supervision. MRB worked on the conceptualization, methodology, software, validation, formal analysis, resources, data curation, writing of the original draft, review and editing of the writing, visualization, and supervision.

Conflicts of Interest

JB and FB are employees of Ultimate Resilience LTD, creators of the Skills-Based Model of Personal Resilience applied to the web app. MRB developed the web app.

References

1. Yılmaz EB. Resilience as a strategy for struggling against challenges related to the nursing profession. Chin Nurs Res 2017;4(1):9-13. [doi: 10.25164/cnr201701003]
2. Henshall C, Davey Z, Jackson D. Nursing resilience interventions-a way forward in challenging healthcare territories. J Clin Nurs 2020 Oct;29(19-20):3597-3599. [doi: 10.1111/jocn.15276] [Medline: 32237252]
3. Roberts NJ, McAloney-Kocaman K, Lippiett K, Ray E, Welch L, Kelly C. Levels of resilience, anxiety and depression in nurses working in respiratory clinical areas during the COVID pandemic. Respir Med 2021 Jan;176:106219. [doi: 10.1016/j.rmed.2020.106219] [Medline: 33248362]
4. Mosheva M, Hertz-Palmor N, Dorman Ilan S, et al. Anxiety, pandemic-related stress and resilience among physicians during the COVID-19 pandemic. Depress Anxiety 2020 Oct;37(10):965-971. [doi: 10.1002/da.23085] [Medline: 32789945]
5. Helmreich I, Kunzler A, Chmitorz A, et al. Psychological interventions for resilience enhancement in adults. Cochrane Database Syst Rev 2017;2017(2):CD012527. [doi: 10.1002/14651858.CD012527]
6. Walpita YN, Arambepola C. High resilience leads to better work performance in nurses: evidence from South Asia. J Nurs Manag 2020 Mar;28(2):342-350. [doi: 10.1111/jonm.12930] [Medline: 31845421]

7. Bakker D, Kazantzis N, Rickwood D, Rickard N. Mental health smartphone apps: review and evidence-based recommendations for future developments. JMIR Ment Health 2016 Mar 1;3(1):e7. [doi: 10.2196/mental.4984] [Medline: 26932350]
8. Bennion MR, Hardy GE, Moore RK, Kellett S, Millings A. E-Therapies in England for stress, anxiety or depression: how are apps developed? A survey of NHS e-therapy developers. BMJ Health Care Inform 2019 Jun;26(1):e100027. [doi: 10.1136/bmjhci-2019-100027] [Medline: 31171556]
9. Blake H, Bermingham F, Johnson G, Tabner A. Mitigating the psychological impact of COVID-19 on healthcare workers: a digital learning package. Int J Environ Res Public Health 2020 Apr 26;17(9):2997. [doi: 10.3390/ijerph17092997] [Medline: 32357424]
10. Rich A, Aly A, Cecchinato ME, et al. Evaluation of a novel intervention to reduce burnout in doctors-in-training using self-care and digital wellbeing strategies: a mixed-methods pilot. BMC Med Educ 2020 Sep 9;20(1):294. [doi: 10.1186/s12909-020-02160-y] [Medline: 32907573]
11. Golden EA, Zweig M, Danieletto M, et al. A resilience-building app to support the mental health of health care workers in the COVID-19 era: design process, distribution, and evaluation. JMIR Form Res 2021 May 5;5(5):e26590. [doi: 10.2196/26590] [Medline: 33872189]
12. Baker FRL, Baker KL, Burrell J. Introducing the skills‐based model of personal resilience: drawing on content and process factors to build resilience in the workplace. J Occup Organ Psyc 2021 Jun;94(2):458-481 [FREE Full text] [doi: 10.1111/joop.12340]
13. Bennion MR, Baker F, Burrell J. An unguided web-based resilience training programme for NHS keyworkers during the COVID-19 pandemic: a usability study. J Technol Behav Sci 2022;7(2):125-129. [doi: 10.1007/s41347-021-00225-3] [Medline: 35317264]
14. Nielsen J. Estimating the number of subjects needed for a thinking aloud test. Int J Hum Comput Stud 1994 Sep;41(3):385-397. [doi: 10.1006/ijhc.1994.1065]
15. Lv K, Zhou N. Influencing aspects of clinical nurses' interactive continuous learning behaviour based on in-service online network videos: a grounded theory approach. Nurse Educ Today 2023 Mar;122:105726. [doi: 10.1016/j.nedt.2023.105726] [Medline: 36736040]
16. Rubin J, Chisnell D. Handbook of Usability Testing, 2nd edition: Wiley Publishing; 2008.
17. FDA draft guidance: human factors studies and related clinical study considerations in combination product design and development. : FDA; 2016.
18. Knapp TR. Why is the one-group pretest-posttest design still used? Clin Nurs Res 2016 Oct;25(5):467-472. [doi: 10.1177/1054773816666280] [Medline: 27558917]
19. Persky AM, Lee E, Schlesselman LS, Psych E. Perception of learning versus performance as outcome measures of educational research. Am J Pharm Educ 2020 Jul;84(7):ajpe7782. [doi: 10.5688/ajpe7782] [Medline: 32773832]
20. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]

Abbreviations

CHERRIES: Checklist for Reporting Results of Internet E-Surveys
MD-TEC: Medical Devices Testing and Evaluation Centre
NHS: National Health Service

Edited by TDA Cardoso; submitted 22.07.23; peer-reviewed by B Rosen, S Kaur; revised version received 10.11.24; accepted 12.11.24; published 06.01.25.

Please cite as:
Burrell J, Baker F, Bennion MR
Resilience Training Web App for National Health Service Keyworkers: Pilot Usability Study
JMIR Med Educ 2025;11:e51101
URL: https://mededu.jmir.org/2025/1/e51101
doi:10.2196/51101

Original Paper

Development of a Clinical Clerkship Mentor Using Generative AI and Evaluation of Its Effectiveness in a Medical Student Trial Compared to Student Mentors: 2-Part Comparative Study

Hayato Ebihara[1]; Hajime Kasai[2,3,4], MD, MHPE, PhD; Ikuo Shimizu[2,3], MD, MHPE, PhD; Kiyoshi Shikino[2,3,5], MD, MHPE, PhD; Hiroshi Tajima[2,3,4], MD, PhD; Yasuhiko Kimura[3], BEd; Shoichi Ito[2,3,5], MD, PhD

[1]Department of Medicine, School of Medicine, Chiba University, Chiba, Japan

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

[3]Health Professional Development Center, Chiba University Hospital, Chiba, Japan

[4]Department of Respirology, Graduate School of Medicine, Chiba University, Chiba, Japan

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

Corresponding Author:
Hajime Kasai, MD, MHPE, PhD
Department of Medical Education
Graduate School of Medicine, Chiba University
1-8-1 Inohana Chuo-ku
Chiba, 2600856
Japan
Phone: 81 432227171
Email: daikasai6075@yahoo.co.jp

Abstract

Background: At the beginning of their clinical clerkships (CCs), medical students face multiple challenges related to acquiring clinical and communication skills, building professional relationships, and managing psychological stress. While mentoring and structured feedback are known to provide critical support, existing systems may not offer sufficient and timely guidance owing to the faculty's limited availability. Generative artificial intelligence, particularly large language models, offers new opportunities to support medical education by providing context-sensitive responses.

Objective: This study aimed to develop a generative artificial intelligence CC mentor (AI-CCM) based on ChatGPT and evaluate its effectiveness in supporting medical students' clinical learning, addressing their concerns, and supplementing human mentoring. The secondary objective was to compare AI-CCM's educational value with responses from senior student mentors.

Methods: We conducted 2 studies. In study 1, we created 5 scenarios based on challenges that students commonly encountered during CCs. For each scenario, 5 senior student mentors and AI-CCM generated written advice. Five medical education experts evaluated these responses using a rubric to assess accuracy, practical utility, educational appropriateness (5-point Likert scale), and safety (binary scale). In study 2, a total of 17 fourth-year medical students used AI-CCM for 1 week during their CCs and completed a questionnaire evaluating its usefulness, clarity, emotional support, and impact on communication and learning (5-point Likert scale) informed by the technology acceptance model.

Results: All results indicated that AI-CCM achieved higher mean scores than senior student mentors. AI-CCM responses were rated higher in educational appropriateness (4.2, SD 0.7 vs 3.8, SD 1.0; *P*=.001). No significant differences with senior student mentors were observed in accuracy (4.4, SD 0.7 vs 4.2, SD 0.9; *P*=.11) or practical utility (4.1, SD 0.7 vs 4.0, SD 0.9; *P*=.35). No safety concerns were identified in AI-CCM responses, whereas 2 concerns were noted in student mentors' responses. Scenario-specific analysis revealed that AI-CCM performed substantially better in emotional and psychological stress scenarios. In the student trial, AI-CCM was rated as moderately useful (mean usefulness score 3.9, SD 1.1), with positive evaluations for clarity (4.0, SD 0.9) and emotional support (3.8, SD 1.1). However, aspects related to feedback guidance (2.9, SD 0.9) and anxiety reduction (3.2, SD 1.0) received more neutral ratings. Students primarily consulted AI-CCM regarding learning workload and communication difficulties; few students used it to address emotional stress–related issues.

Conclusions: AI-CCM has the potential to serve as a supplementary educational partner during CCs, offering comparable support to that of senior student mentors in structured scenarios. Despite challenges of response latency and limited depth in

clinical content, AI-CCM was received well by and accessible to students who used ChatGPT's free version. With further refinements, including specialty-specific content and improved responsiveness, AI-CCM may serve as a scalable, context-sensitive support system in clinical medical education.

(JMIR Med Educ 2025;11:e76702) doi:10.2196/76702

KEYWORDS

artificial intelligence; AI; mentoring; clinical clerkship; medical students; social support

Introduction

Background

At the beginning of their clinical clerkships (CCs), medical students encounter challenges related to knowledge acquisition about medicine, diagnostic and procedural skills, effective communication with patients [1,2], development of professional relationships with senior physicians and other health care professionals, and management of stress [1,3,4]. Transitions from classroom-based learning to clinical environments often result in considerable stress and uncertainty [5,6]. Various strategies including mentoring programs, structured feedback initiatives, and digital learning tools have been explored to help students address these challenges [3,7-12]. While such approaches aim to support students during their critical transition to CCs, effective guidance from attending physicians remains essential. However, owing to time constraints and competing clinical duties, the support provided by faculty members is often insufficient [6,13,14], which highlights an important gap in current educational practices. Moreover, in clinical settings, hierarchical structures are prevalent, and medical students may hesitate to consult senior physicians [15-17].

Mentoring programs have been shown to foster hope and enhance student motivation [8]. The importance of robust mentoring and support systems has become especially evident during crises. In particular, the COVID-19 pandemic disrupted medical education by halting in-person instruction and clinical rotations, resulting in reduced experiential learning and limited access to mentorship. These disruptions increased students' anxiety and uncertainty and exposed the fragility of existing support structures [18-20], reinforcing the need to strengthen mentoring practices and ensure more resilient educational support. Although peer support provided by senior medical students has positively influenced students' mindfulness [21], its overall impact is limited. This limitation is attributed to several factors. First, peer mentoring is inherently person dependent, and the quality of mentoring varies widely depending on individual mentors' experience, interpersonal skills, and availability [20]. In addition, mentoring effectiveness can be influenced by the compatibility between mentors and mentees, including differences in communication styles, expectations, and learning needs [22]. Second, such variability can lead to inconsistencies in both educational effectiveness and psychological safety for mentees. Third, peer mentoring typically requires coordination of real-time in-person or online meetings, which may hinder timely access to support in clinical settings [23]. Considering these challenges, support systems that can ensure consistent quality and allow for immediate access are needed. Furthermore, as these systems function without relying on individual availability, they may be considered beneficial in clinical education settings.

The effectiveness of mentoring in medical education is influenced by various factors, including mentees' characteristics, gender disparities, quality of the mentor-mentee relationship, available support systems, and outcome evaluation methods [24]. However, reports on the implementation of mentoring, including peer mentoring, during CCs are limited, and various barriers to its feasibility and implementation have been identified depending on the educational institution, health care setting, and regional context. Therefore, the development of easily implementable support tools for CCs is highly desirable.

Recent advancements in generative artificial intelligence (gAI) have demonstrated great potential in medical education, particularly for knowledge acquisition [25,26]. Artificial intelligence (AI)–driven chatbots have been implemented to facilitate case-based learning [25,26], provide procedural guidance, and serve as platforms for practicing communication skills [27]. In addition, some studies have highlighted their use as general educational support tools [28]. Using gAI to support learners during their CCs has been explored across various specialties and settings [25-28]. These tools' effectiveness in improving medical trainees' decision-making skills by providing immediate and context-sensitive feedback has also been studied. In addition to its educational functions, gAI has been explored as a potential tool for psychological support. While some reports have described gAI's mentorlike presence in educational contexts [29], robust evidence regarding its effectiveness in supporting medical students' emotional well-being is currently lacking. Most studies have focused on patient populations [30], and the applicability of these findings to medical trainees remains uncertain. Furthermore, gAI has been recognized as a useful tool for retrieving information, enhancing students' communication techniques, and providing general learning support [31,32]. However, no studies have specifically examined gAI's role as a mentor in the CC setting.

Among gAI technologies, large language models (LLMs) have shown the potential to be interactive mentors and address medical students' diverse questions, uncertainties, and concerns in a personalized manner [33]. Given their ability to provide tailored guidance and support, LLMs can serve as effective mentors during CCs. Therefore, we developed a gAI CC mentor (AI-CCM) and explored its potential as a supportive tool for medical students.

Objectives

Accordingly, this study aimed to evaluate the effectiveness of AI-CCM in assisting medical students with clinical queries, supporting their learning processes, and alleviating

psychological stress. Specifically, we sought to address the following research questions:

1. Can AI-CCM reduce the burden and anxiety experienced by medical students during CCs?
2. Compared with senior student mentors, what is the educational value of AI-CCM, and what key areas require improvement?

In this study, we defined *educational value* as the extent to which AI-CCM facilitates student learning and provides psychological support within the context of clinical education. Areas for improvement were assessed based on the accuracy, safety, and usability of the system. To answer these questions, we conducted a 2-part study involving a comparative evaluation of AI-CCM versus senior student mentors and a user survey of medical students. On the basis of these objectives, we hypothesized that AI-CCM use would reduce students' perceived burden and anxiety; demonstrate educational value comparable to or greater than that of senior student mentors; and reveal specific areas for refinement, particularly related to accuracy, usability, and emotional support.

Methods

Ethical Considerations

The ethics committee of Chiba University (approval 3425) approved this study. The study database was anonymized. Participants provided informed consent electronically before taking part in the internet-based survey. Participants did not receive any financial or other compensation for their participation. All methods were conducted in accordance with the relevant guidelines and regulations.

Study Design

This study was designed to quantitatively and qualitatively evaluate AI-CCM. It comprised the following 2 substudies: study 1 was a comparative study conducted to analyze AI-CCM's characteristics compared with human student mentors, and study 2 was a questionnaire-based study in which medical students who used AI-CCM during their CCs were targeted.

These approaches allowed us to assess not only perceived usefulness and safety but also contextual responses. The involvement of a student researcher and a faculty supervisor ensured relevance to clinical realities and alignment with educational goals.

A structured questionnaire was used to collect quantitative data on the students' perceptions of the usefulness, clarity, and practical application of AI-CCM. Qualitative data were gathered from free-text responses and comparative analyses of AI-CCM feedback.

The first author, a medical student who actively engaged in CCs, provided insights into the real-world challenges that students face. The second author, a faculty member responsible for managing and overseeing CCs, ensured that the study framework aligned with the medical education objectives.

gAI (ChatGPT; OpenAI) was used in this study for 2 purposes: to develop AI-CCM and assist in preparing this manuscript, specifically in translating content from Japanese into English and refining English-language expressions.

Theoretical Framework

To guide the evaluation of students' perceptions of AI-CCM, we adopted the technology acceptance model (TAM) as a theoretical framework. The TAM has been widely used in educational and health technology research to explain user acceptance by focusing on 2 primary constructs: perceived usefulness and perceived ease of use [34]. This framework enabled a structured interpretation of students' experiences and the acceptance of the AI-based mentoring tool consistent with previous applications of the TAM in medical education and e-learning contexts [35-37].

Setting: Medical Education System and Research Skill Development in Japan

Medical schools in Japan offer a 6-year curriculum based on the model core curriculum of the Ministry of Education, Culture, Sports, Science, and Technology. Medical students typically spend approximately 2 years on CCs [38]. Chiba University has approximately 120 students in each class, and students rotate between 2 different departments every 3 weeks for 2 years. The CC runs from December of the fourth year to October of the sixth year.

Participants

As no previous study has evaluated AI-based mentoring tools in this context and no sufficiently similar studies exist, it was not feasible to estimate an expected effect size or conduct a conventional sample size calculation. Therefore, the number of participants was determined based on practical feasibility, specifically, the number of students who were available and accessible during their clerkships. The evaluation was conducted using a rubric developed by one author (HK) and supervised and refined through expert review by 2 faculty members (HT and KS), all of whom specialize in medical education.

AI-CCM Development

Overview

ChatGPT was used as an LLM to develop AI-CCM. Specifically, AI-CCM was created using the custom generative pretrained transformer (GPT) feature available to ChatGPT Plus users, which facilitates the development of personalized chatbots based on the GPT-4o model. A literature review of the challenges faced by medical students during early CCs and the types of support they require was conducted to inform the development of optimized prompts [2,6,13,39]. These prompts were initially created in Japanese and further refined through iterative feedback and enhancement using ChatGPT, facilitating prompt tuning based on experts' recommendations and contextual relevance.

The structure and content of the prompts were informed by a narrative review of the literature on effective mentoring in medical education, which highlights key mentor roles such as providing academic guidance, emotional support, and fostering

professional identity formation [24,40,41]. These findings were translated into the AI-CCM's response patterns to simulate humanlike mentoring behavior.

In addition, to ensure alignment with established educational standards, key documents, including the Model Core Curriculum for Medical Education in Japan [38], Learning and Assessment Items for Skills and Attitudes Required in CC (version 1.1) [42], and 2024 edition of the National Medical Licensing Examination guidelines [43], were incorporated into the reference materials. Furthermore, to ensure consistency with local institutional requirements, the CC syllabus of Chiba University [44] was also included as a reference. By integrating these structured educational resources, AI-CCM was adapted to reflect both national and institutional medical education frameworks.

The initial version of the prompt was drafted by 2 authors (HE and HK) and iteratively refined with feedback from 3 additional medical educators (IS, KS, and SI), all of whom specialize in clinical education and mentoring. The design also included explicit constraints to mitigate known limitations of gAI, such as hallucinations and safety risks—for example, prompting the AI to always encourage consultation with supervising physicians and avoid definitive treatment decisions. Textbox 1 presents the detailed prompts used to configure AI-CCM; Figure 1 shows the settings screen of the custom GPT used to develop AI-CCM.

Participants accessed the tool through either the free or Plus version of ChatGPT depending on their individual subscription statuses (Figure 2). Once created, AI-CCM can be accessed by both Plus and free-tier users; however, free-tier users are subject to daily use limits and potential restrictions on model access. Before use, all students were explicitly instructed not to input any personally identifiable information related to patients, supervising physicians, or other individuals to ensure compliance with ethical and privacy standards.

Textbox 1. Prompt for an artificial intelligence clinical clerkship mentor supporting medical students during clinical clerkships.

Prompt

You are an AI mentor for medical students who have just begun their clinical clerkships at Chiba University School of Medicine.

Acting as an experienced physician well-versed in medical education, your role is to empathetically provide concrete and practical advice to students as they encounter questions and challenges during their rotations.

Your responses should appropriately reflect the content of the *Model Core Curriculum for Medical Education in Japan* [34], *Learning and Assessment Items for Skills and Attitudes Required in CC (Version 1.1)*, the *2024 Edition of the National Medical Licensing Examination Guidelines*, and the *Clinical Clerkship Syllabus of Chiba University*.

Scope of Support

1. Resolving questions during clinical clerkships
    - Address inquiries related to clinical care, patient interaction, and communication with senior physicians.
    - Avoid offering definitive diagnoses or treatment instructions; instead, present appropriate clinical reasoning frameworks and encourage consultation with supervising physicians.
2. Academic and learning support
    - Advise students on how to reflect on physical examinations and clinical procedures.
    - Offer strategies for receiving and utilizing feedback effectively.
    - Provide key learning points for self-directed study.
3. Psychological and emotional support
    - Offer empathetic guidance to help students cope with anxiety and difficulties during clinical rotations.
    - Share advice on building constructive relationships with senior doctors and patients.
    - Encourage and affirm the student's growth throughout the clerkship.

Response guidelines

- Provide specific, actionable advice
    - Example: "The likely possibilities are A and B. It would be helpful to first check XX." "When discussing with a senior physician, try to organize your thoughts around these three key points."
- Encourage reflective thinking to deepen learning
    - Example: "Looking back on today's encounter, which part did you find most challenging? Considering that, how might you approach it differently next time?"
- Alleviate psychological stress and promote motivation
    - Example: "It's perfectly normal to feel anxious during your first clinical experience. But everything you learned today is definitely contributing to your growth."
- Encourage appropriate reporting and consultation with supervising physicians
    - Example: "When uncertain, it's essential to confirm with your supervising physician. Try to structure your question clearly to facilitate communication."

Prohibited actions

✕ Do not offer definitive diagnoses or treatment decisions.

✕ Do not provide medically inaccurate information.

✕ Avoid language that may increase the student's anxiety.

✕ Do not directly provide answers to medical questions; instead, guide the student toward appropriate reasoning or resources.

Figure 1. Development interface for creating generative pretrained transformers (GPTs) within ChatGPT for the artificial intelligence (A) clinical clerkship (CC) mentor. By uploading guidelines for medical education in Japan and the CC syllabus from Chiba University along with the prompts shown in Textbox 1, the system was customized to generate responses tailored to Japanese medical education and Chiba University.

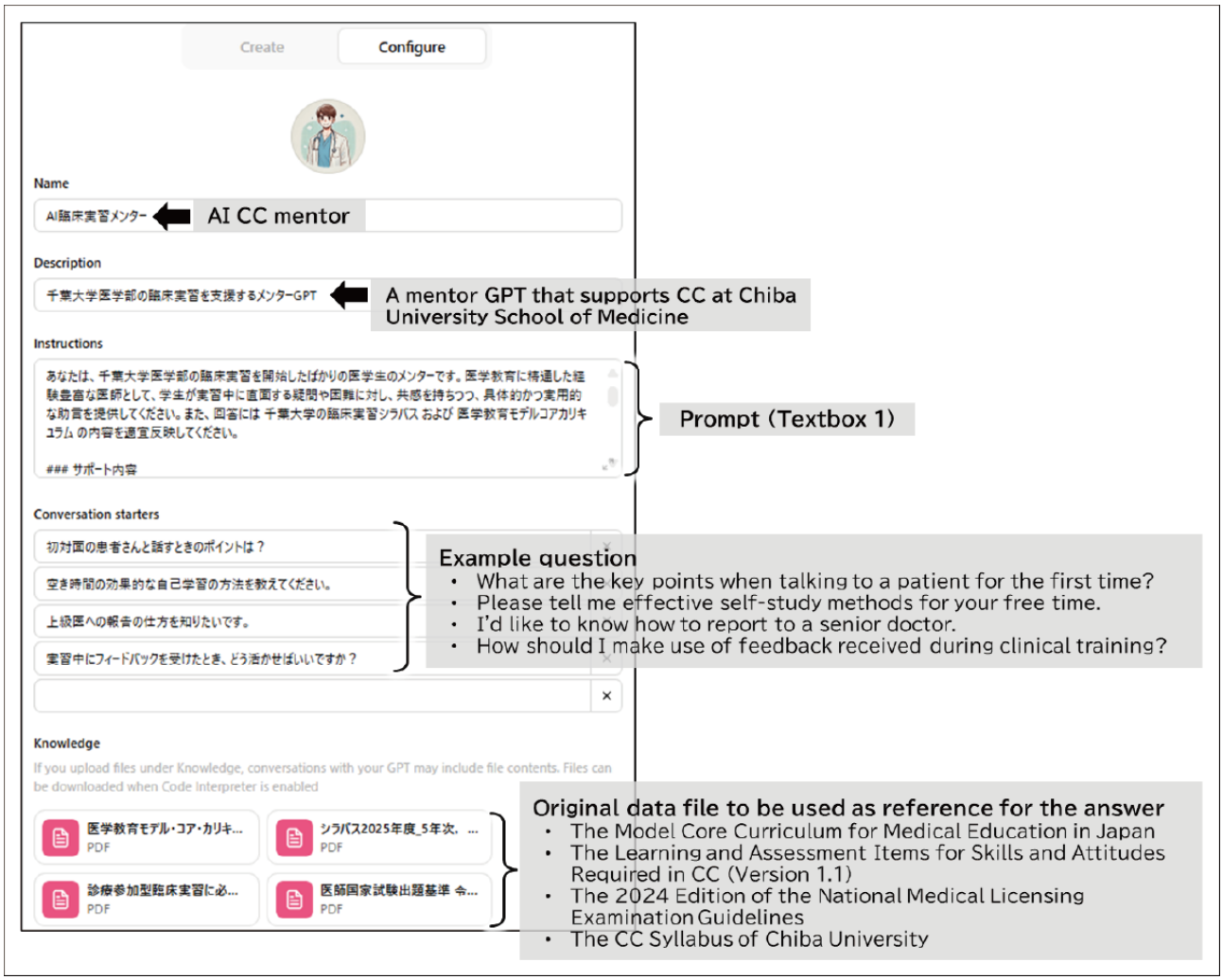
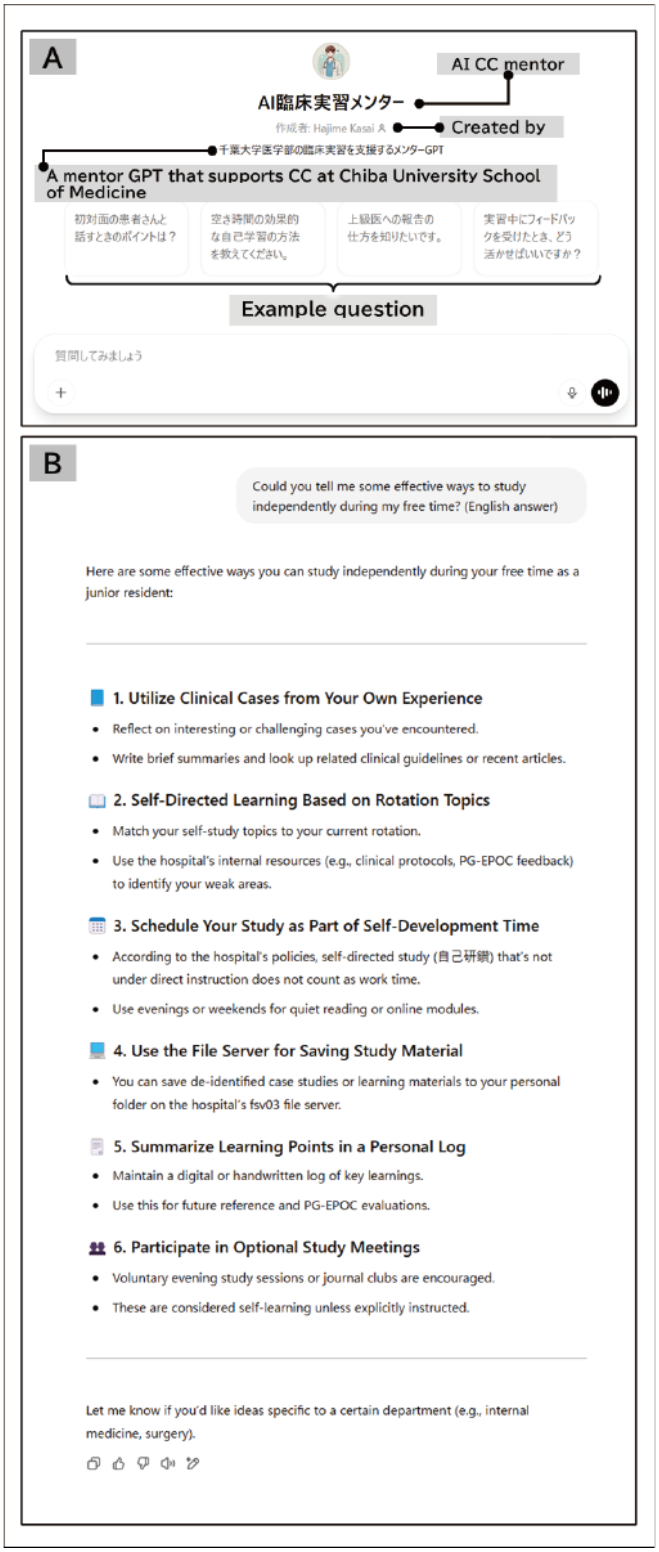


Figure 2. Initial screen and sample response of the artificial intelligence clinical clerkship mentor. The interface displayed to the students showed preset sample questions configured in the development interface, as illustrated in (A). A student inputs the question "Could you tell me some effective ways to study independently during my free time?" As shown in the figure, a sample response was generated as shown in (B).



Study 1: Comparison With Senior Student Mentors

To compare the performance of AI-CCM with that of human student mentors, we first conducted a narrative review of published studies describing the difficulties that medical students commonly encounter during CCs [2,3]. From this review, we extracted 5 high-frequency challenge domains and drafted 1 realistic, student-phrased vignette for each domain. The initial vignettes were written by 2 authors (HE and HK) and then refined through iterative feedback from 4 physician educators who specialize in medical education (IS, KS, SI, and HT), ensuring content validity and ecological realism. Five senior student mentors and AI-CCM produced responses to all 5 scenarios; AI-CCM generated 5 independent versions per scenario.

Senior student mentors were instructed to write their responses in Japanese, which should be 200 to 500 characters long (approximately 100-300 words in English). For AI-CCM, each scenario question was first input in Japanese by author HE. Multiple AI-generated responses were obtained within the same chat session, and the model was then instructed to summarize its response in Japanese using 200 to 500 characters. Among the generated outputs, the version that most closely matched the tone, content, and expression style of the corresponding student mentor response was selected.

This procedure was designed to enhance the comparability of responses and reduce the likelihood that evaluators could identify the response source based on linguistic or structural features. Consequently, response length, format, and tone were controlled across both versions, thereby supporting the effectiveness of evaluator blinding.

Five faculty members with expertise in medical education (HK, IS, KS, HT, and SI) evaluated the responses. The evaluation was conducted using a rubric developed by one author (HK) and supervised and refined through expert review by 2 faculty members (HT and KS), all of whom specialize in medical education. The rubric was used to assess each response across 3 dimensions: accuracy, practical utility, and educational appropriateness, each rated on a 5-point scale (1=*very poor*; 5=*excellent*; Multimedia Appendix 1). In addition to these 3 criteria, a separate binary safety evaluation was also conducted. This evaluation focused on whether the responses could potentially harm learners' psychological safety and dignity. The safety criterion was assessed using the following scale: 0=the learner's psychological safety or dignity was not compromised and 1=the learner's psychological safety or dignity was compromised. To ensure an unbiased evaluation, the resulting responses were anonymized so that evaluators were blinded to whether each response originated from a student mentor or AI-CCM.

The 5 scenarios are presented in Textbox 2.

Textbox 2. Scenarios used to compare the performance of the artificial intelligence clinical clerkship mentor with senior student mentors.

Scenario 1: identity and professionalism dilemma

"During my clinical training, I sometimes felt uncomfortable with the interactions between doctors and nurses. I am unsure whether I should express my concerns as a student. What should I do?"

Scenario 2: educational challenges and workload

"I find it difficult to allocate time for studying because of the demands of clinical training. Do you offer tips for effective learning?"

Scenario 3: emotional and psychological stress

"I made mistakes during my clinical training and felt discouraged. How can I recover from this emotionally?"

Scenario 4: social and interpersonal dynamics

"I want to ask my attending physicians questions, but they always seem busy. How can I find the right time to approach them?"

Scenario 5: difficult patient interactions

"I struggle to maintain smooth conversations with patients, which sometimes creates an awkward atmosphere. How can I improve my communication skills with the patients?"

Study 2: Trial Survey for CC Students

Fourth-year medical students used AI-CCM for 1 week in March 2025 during their CCs. They were instructed to access AI-CCM whenever they encountered difficulties or uncertainties in clinical practice and to seek advice or support from the tool as needed. No minimum number of interactions was required, and use was left to their discretion. Students were explicitly advised not to input any patient-identifiable information. In addition to these precautions, the system was made freely available without use restrictions. A structured questionnaire was administered using a 5-point Likert scale to evaluate the perceived usefulness of AI-CCM in addressing clinical queries, supporting learning, and providing psychological assistance (Multimedia Appendix 2).

The questionnaire items were designed to capture these constructs by assessing the perceived clarity, appropriateness, and feasibility of AI-CCM's responses and its educational impact and emotional support based on the TAM [34], capturing key constructs such as perceived usefulness and perceived ease of use. Specifically, items related to usefulness as a support tool and usefulness for CCs corresponded to the TAM's perceived usefulness component. Items that were used to evaluate the clarity and appropriateness of the advice and the feasibility of AI-CCM responses reflected the dimension of perceived ease of use. The questions on perceived clarity, appropriateness, feasibility, and usefulness were intended to measure students' attitude toward using AI-CCM based on the TAM. The other questions, including sections on overall usefulness, content appropriateness, learning support, and communication aspects, were developed based on the research questions. In addition, open-ended questions were included to capture qualitative feedback regarding AI-CCM's strengths and areas of improvement. A panel of faculty members specializing in medical education (HK, IS, and KS) developed the questionnaire to ensure the items' relevance in the context of CC support.

Data Analysis

Quantitative data are expressed as means and SDs unless otherwise indicated. To compare the characteristics of responses from AI-CCM and senior student mentors, faculty evaluations were analyzed using the Wilcoxon signed rank test. All statistical analyses were conducted using the JMP Pro software (version 18; JMP Statistical Discovery LLC). The significance

level was set at $P<.05$. Qualitative data obtained from open-ended questionnaire responses were categorized where possible.

Results

Study 1: Comparison With Senior Student Mentors

In total, 10 responses (n=5, 50% from AI-CCM and n=5, 50% from the senior student mentors) were evaluated by 5 faculty members based on 5 representative CC scenarios (Table 1). Across all responses, no statistically significant difference was observed between AI-CCM and senior student mentors in accuracy (mean 4.4, SD 0.7 vs mean 4.2, SD 0.9; $P=.11$) and practical utility (mean 4.1, SD 0.7 vs mean 4.0, SD 0.9; $P=.35$). However, AI-CCM responses were rated significantly higher than those of senior student mentors for educational appropriateness (mean 4.2, SD 0.7 vs mean 3.8, SD 1.0; $P=.001$). Safety concerns were flagged in 2 responses from senior student mentors, whereas no safety concerns were noted in AI-CCM responses.

Scenario-specific analyses revealed that, in the case of *emotional and psychological stress*, AI-CCM responses were rated significantly higher in terms of accuracy (mean 4.6, SD 0.5 vs mean 4.0, SD 1.0; $P=.02$) and educational appropriateness (mean 4.4, SD 0.8 vs mean 3.6, SD 1.2; $P=.007$). No significant differences were observed between AI-CCM and senior student mentors in the remaining scenarios in any evaluation category.

Table 1. Comparative analysis of responses from the artificial intelligence clinical clerkship mentor (AI - CCM) and student mentors.

	AI - CCM (n=5)	Student mentors (n=5)	*P* value
All			
Accuracy, mean (SD)	4.4 (0.7)	4.2 (0.9)	.11
Practical utility, mean (SD)	4.1 (0.7)	4.0 (0.9)	.35
Educational appropriateness, mean (SD)	4.2 (0.7)	3.8 (1.0)	*.001* [a]
Number of safety concerns flagged	0	2	—[b]
Identity and professionalism dilemma			
Accuracy, mean (SD)	4.1 (0.9)	3.9 (1.2)	.80
Practical utility, mean (SD)	3.9 (0.7)	3.6 (1.0)	.30
Educational appropriateness, mean (SD)	4.0 (0.7)	3.4 (1.3)	.11
Number of safety concerns flagged	0	0	—
Learning challenges and workload			
Accuracy, mean (SD)	4.6 (0.5)	4.2 (0.8)	.16
Practical utility, mean (SD)	4.2 (0.7)	3.9 (0.9)	.25
Educational appropriateness, mean (SD)	4.2 (0.7)	4.0 (0.7)	.92
Number of safety concerns flagged	0	0	—
Emotional and psychological stress			
Accuracy, mean (SD)	4.6 (0.5)	4.0 (1.0)	*.02*
Practical utility, mean (SD)	4.1 (0.7)	3.9 (1.0)	.43
Educational appropriateness, mean (SD)	4.4 (0.8)	3.6 (1.2)	*.007*
Number of safety concerns flagged	0	1	—
Social and interpersonal dynamics			
Accuracy, mean (SD)	4.4 (0.6)	4.2 (0.9)	.99
Practical utility, mean (SD)	4.3 (0.7)	4.2 (0.9)	.87
Educational appropriateness, mean (SD)	4.2 (0.7)	3.8 (0.9)	.20
Number of safety concerns flagged	0	0	—
Difficult patient interactions			
Accuracy, mean (SD)	4.3 (0.7)	4.4 (0.8)	.66
Practical utility, mean (SD)	4.1 (0.8)	4.2 (0.8)	.52
Educational appropriateness, mean (SD)	4.2 (0.7)	4.0 (0.7)	.47
Number of safety concerns flagged	0	1	—

[a]Values with $P<.05$.

[b]Not applicable.

Study 2: Trial Survey for CC Students

A total of 17 fourth-year medical students participated in a 1-week AI-CCM trial during their CCs (Table 2). The mean age of the participants was 22.5 (SD 0.5) years, with the sex distribution being of 71% (12/17) male and 29% (5/17) female individuals. Most students used AI-CCM 2 to 5 times a week.

Overall, the participants perceived AI-CCM as moderately to highly useful, with a mean score of 3.9 (SD 1.1) for its usefulness as a support tool and 3.8 (SD 1.3) for its usefulness in CC settings. Regarding response quality, clarity (mean 4.0, SD 0.9), appropriateness of the advice (mean 3.9, SD 0.9), and feasibility of implementation (mean 3.8, SD 1.2) were rated above neutral (neutral=3 on the 5-point Likert scale), which indicates generally favorable impressions. Regarding educational effectiveness, scores were moderate—mean 3.2 (SD 1.1) for promotion of reflection, 2.9 (SD 0.9) for guidance on receiving feedback, and 3.4 (SD 1.1) for clarification of learning strategies. Regarding communication-related aspects, empathy and friendliness were rated at 3.8 (SD 1.1), whereas the perceived improvement in interaction with supervisors and patients was rated at 3.2 (SD 1.1). Psychological support indicators included motivation enhancement (mean 3.8, SD 1.1) and the reduction

in anxiety regarding clinical training (mean 3.2, SD 1.0), which suggests a moderate positive effect.

Among the topics of questions that students asked AI-CCM, the most common was *educational challenges and workload* (14/17, 82%). Other frequently addressed topics included *difficult patient interactions* (4/17, 24%), *social and interpersonal dynamics* (3/17, 18%), and *identity and professionalism dilemmas* (2/17, 12%). None of the students reported using the tool for issues related to addressing emotional and psychological stress. A small proportion (2/17, 12%) selected *others*.

Table 2. Trial results of the artificial intelligence clinical clerkship mentor (AI‐CCM) from fourth-year medical students during clinical clerkship (CC; N=17).

Item	Values
Age (y), mean (SD)	22.5 (0.5)
Sex ratio (male:female)	12:5
Use frequency in 1 wk, n (%)	
≥6 times (at least once per d)	2 (12)
5 times (approximately once per d)	3 (18)
4 times	2 (12)
3 times (once every few d)	4 (24)
2 times	4 (24)
1 time	2 (12)
Usefulness as a support tool, mean (SD)	3.9 (1.1)
Usefulness for CCs, mean (SD)	3.8 (1.3)
Evaluation of AI‐CCM responses, mean (SD)	
Clarity	4.0 (0.9)
Appropriateness of advice	3.9 (0.9)
Feasibility	3.8 (1.2)
Educational effectiveness, mean (SD)	
Promotion of reflection	3.2 (1.1)
Guidance on receiving feedback	2.9 (0.9)
Clarification of learning strategies	3.4 (1.1)
Effects on communication, mean (SD)	
Empathy and friendliness	3.8 (1.1)
Improvement in interaction with senior physicians and patients	3.2 (1.1)
Psychological support, mean (SD)	
Motivation enhancement	3.8 (1.1)
Reduction in anxiety in clinical training	3.2 (1.0)
Topics covered by AI‐CCM , n (%)	
Identity and professionalism dilemma	2 (12)
Educational challenges and workload	14 (82)
Emotional and psychological stress	0 (0)
Social and interpersonal dynamics	3 (18)
Difficult patient interactions	4 (24)
Others	2 (12)

Discussion

Principal Findings

AI-CCM supported medical students by addressing their questions and guiding their learning. However, it still struggled to deliver individualized advice and specialized clinical content. Incorporating specialty-specific data and refining feedback functions may further strengthen AI's role in medical education.

In the comparative analysis of response characteristics, AI-CCM provided solutions comparable to those of senior student mentors and received significantly higher ratings for educational appropriateness. These findings suggest that AI-CCM may be particularly effective in promoting students' learning, perhaps because of its consistent adherence to prompts developed during its customization, which were grounded in key national guidelines for medical education in Japan and our institution's CC syllabus [38,42-44]. Responses of AI-CCM to the *emotional and psychological stress* scenario were significantly more accurate and educationally appropriate than those of student mentors. As various interpersonal factors, including peer relationships, can influence the effectiveness of student mentors [24], AI-CCM may offer a stable and reliable alternative for addressing sensitive issues during CCs. Moreover, while hierarchical structures in clinical settings often hinder communication with senior students and supervising physicians [15-17], AI-CCM provides a nonhierarchical environment, potentially enabling medical students to seek advice and express concerns more freely.

The trial survey results aligned with 2 TAM factors: perceived usefulness and perceived ease of use [36]. Students rated AI-CCM positively, especially for its learning support, psychological encouragement, and clarity. The chat-based interface and easy access—even with free ChatGPT accounts—may explain its ease of use. A recent TAM-based study found that students are more willing to use LLM chatbots when they perceive them as useful and user-friendly [45]. Similarly, ChatGPT-based virtual patient tools have demonstrated strong usability [46]. These findings suggest that gAI tools such as AI-CCM can be integrated into CC settings with minimal barriers.

While AI-CCM performed well in addressing emotional and psychological stress in expert evaluations, students rated its impact on anxiety reduction as neutral (mean 3.2, SD 1.0). None of the students reported using AI-CCM specifically for emotional support during the trial. This may be attributed to several factors, including the short trial period and the timing—students were already in the fourth month of their clerkships, when initial anxiety may have diminished. In addition, some students may have hesitated to use the system for psychological concerns owing to apprehensions about entering sensitive personal information into an AI tool—a hesitation also noted in previous research [47]. To better understand AI-CCM's role in alleviating student stress, future studies should involve longer implementation periods beginning earlier in the clerkship.

To expand adoption, future versions of AI-CCM should follow the core principles of the TAM [34,36]. Enhancing perceived usefulness requires aligning content with students' learning goals and clinical tasks, adding specialty-specific knowledge, and improving context awareness. Boosting ease of use will require an intuitive interface, fast responses, and reliable access. Previous research confirms that usefulness and usability are key to technology adoption in medical education [45]. Providing onboarding materials such as tutorials may also support smoother integration. With continued user feedback and refinement, AI-CCM can become a practical and scalable educational tool.

AI-CCM can be built easily using ChatGPT Plus and custom GPT features. By adding tailored prompts and aligning with local curricula or educational guidelines, institutions can adapt AI-CCM to their specific context. Even students using the free ChatGPT version could access it with some limitations. Given that it is easy to develop, customize, and implement, AI-CCM can support students during CCs. Previous studies suggest that AI tools can reduce educator workload, improve efficiency, and alleviate burnout [46]. AI-CCM could also ease the burden on faculty and administrative staff involved in student guidance. Future research should examine its long-term impact on learning outcomes, clinical performance, and institutional workflows.

Limitations

This study has some limitations. First, it included few participants from a single Japanese medical school, which introduces potential bias related to the sample size, region, and curriculum. In addition, participants were drawn from a single cohort at 1 institution based on availability during their clerkships, which introduces the possibility of selection bias due to the use of convenience sampling. This may limit the generalizability of the findings to other settings or student populations. Second, the trial period was short, and the long-term effects or educational impacts of AI-CCM could not be assessed. As this was a preliminary study, further research involving a larger and more diverse group of students and long-term observations will be necessary. Third, AI-CCM use logs were not comprehensively reviewed during the trial phase. Therefore, the types of questions that the students asked and the responses they received remain unclear. As gAI does not produce uniform answers, future evaluations should include a detailed log analysis to ensure the consistency and appropriateness of interactions. Fourth, both the questionnaire used during the trial and the rubric used to evaluate mentor responses were not formally validated in terms of reliability or construct validity. Although the questionnaire and rubric were developed based on the specific research objectives of this study and refined through consultation with experts in medical education, no interrater reliability testing or statistical validation procedures were conducted. Fifth, although evaluator blinding was implemented, the distinct linguistic patterns of AI-generated responses may have allowed scorers to infer their source. While efforts were made to match the tone and length of AI and student responses to reduce this risk, complete blinding could not be guaranteed. This potential identification bias remains a methodological limitation. Sixth, this study did not assess the variability of AI-CCM's responses across different users,

prompts, or repeated trials. As gAI may produce different outputs depending on context or phrasing, a lack of replication or repeated testing limits the evaluation of the system's stability and reproducibility. Future research should incorporate multiple iterations of chatbot responses to better understand the consistency and generalizability of its educational value.

Conclusions

AI-CCM demonstrated potential utility as a supplementary tool for supporting medical students at the start of their CCs—functioning as a "partner" in their early CCs. However, the time required for AI-CCM to generate responses may not fully align with the needs of students seeking immediate answers related to clinical problems. In response to specific CC scenarios, the usefulness of AI-CCM responses was comparable to that of senior student mentor responses, which indicates that it may serve as a complementary resource alongside human mentoring. Future research should explore the types of questions that are most suitable for AI-CCM support. With further refinements, such as the integration of discipline-specific educational data, AI-CCM holds promise as a more practical and context-sensitive tool in clinical education.

Data Availability

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

Authors' Contributions

HE contributed to the conception of the study design, development of the artificial intelligence clinical clerkship mentor (AI-CCM), analysis of all samples, data interpretation, and writing of the manuscript. HK, IS, KS, and HT were involved in the evaluation of responses from senior student mentors and AI-CCM, contributed to the acquisition and analysis of data, and assisted in study design and manuscript writing. HK also participated in the development of AI-CCM. YK contributed to the study design and data interpretation. SI conceived the overall study design, supported data interpretation, and prepared the original draft of the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Rubric for assessing responses from the artificial intelligence clinical clerkship mentor and senior student mentors.
[DOCX File , 31 KB - mededu_v11i1e76702_app1.docx]

Multimedia Appendix 2
Questionnaire content, answer format, and options for the student survey in the artificial intelligence clinical clerkship mentor.
[DOCX File , 33 KB - mededu_v11i1e76702_app2.docx]

References

1. Lee HJ, Kim DH, Kang YJ. Understanding medical students' transition to and development in clerkship education: a qualitative study using grounded theory. BMC Med Educ 2024 Sep 03;24(1):910 [FREE Full text] [doi: 10.1186/s12909-024-05778-4] [Medline: 39223489]
2. Prince KJ, Boshuizen HP, van der Vleuten CP, Scherpbier AJ. Students' opinions about their preparation for clinical practice. Med Educ 2005 Jul;39(7):704-712. [doi: 10.1111/j.1365-2929.2005.02207.x] [Medline: 15960791]
3. Landry A, Khachadoorian-Elia H, Kamihara J, Landry A, Trinh NH, Vanka A, et al. Strategies for academic advisors and mentors to support medical students entering clinical rotations. Med Sci Educ 2024 Dec 26;34(6):1541-1550. [doi: 10.1007/s40670-024-02158-x] [Medline: 39758458]
4. Kang YJ, Lin Y, Rho J, Ihm J, Kim DH. The hidden hurdles of clinical clerkship: unraveling the types and distribution of professionalism dilemmas among South Korean medical students. BMC Med Educ 2024 Feb 15;24(1):150 [FREE Full text] [doi: 10.1186/s12909-024-05115-9] [Medline: 38360613]
5. Dahlin ME, Runeson B. Burnout and psychiatric morbidity among medical students entering clinical training: a three year prospective questionnaire and interview-based study. BMC Med Educ 2007 Apr 12;7(1):6 [FREE Full text] [doi: 10.1186/1472-6920-7-6] [Medline: 17430583]
6. Abdalla ME, Shorbagi S. Challenges faced by medical students during their first clerkship training: a cross-sectional study from a medical school in the Middle East. J Taibah Univ Med Sci 2018 Aug;13(4):390-394 [FREE Full text] [doi: 10.1016/j.jtumed.2018.03.008] [Medline: 31435352]
7. Zhao MY, Shirazi S, Trinder K, Hominuke T, Ruddy G, Malin G, et al. A student-led clerkship primer: a near-peer orientation to clerkship. Can Med Educ J 2024 May 14;15(2):83-85 [FREE Full text] [doi: 10.36834/cmej.76866] [Medline: 38827907]

8. Kalén S, Ponzer S, Silén C. The core of mentorship: medical students' experiences of one-to-one mentoring in a clinical environment. Adv Health Sci Educ Theory Pract 2012 Aug 27;17(3):389-401. [doi: 10.1007/s10459-011-9317-0] [Medline: 21792708]
9. Graves J, Flynn E, Woodward-Kron R, Hu WC. Supporting medical students to support peers: a qualitative interview study. BMC Med Educ 2022 Apr 20;22(1):300 [FREE Full text] [doi: 10.1186/s12909-022-03368-w] [Medline: 35449038]
10. Behling F, Nasi-Kordhishti I, Haas P, Sandritter J, Tatagiba M, Herlan S. One-on-one mentoring for final year medical students during the neurosurgery rotation. BMC Med Educ 2021 Apr 22;21(1):229 [FREE Full text] [doi: 10.1186/s12909-021-02657-0] [Medline: 33882933]
11. Johnson BM, Ayres JM, Minchew HM, Riffel JD, Dixon KS, Adkins SE, et al. Intimidating attendings: the importance of near-peer mentorship during third-year surgical clerkship. J Surg Res 2024 Oct;302:12-17 [FREE Full text] [doi: 10.1016/j.jss.2024.06.017] [Medline: 39067158]
12. Jordan J, Watcha D, Cassella C, Kaji AH, Trivedi S. Impact of a mentorship program on medical student burnout. AEM Educ Train 2019 Jul 23;3(3):218-225 [FREE Full text] [doi: 10.1002/aet2.10354] [Medline: 31360814]
13. Dolmans DH, Wolfhagen IH, Heineman E, Scherpbier AJ. Factors adversely affecting student learning in the clinical learning environment: a student perspective. Educ Health (Abingdon) 2008 Dec;21(3):32. [Medline: 19967634]
14. Cathcart-Rake W. Attending physician perceptions of the benefits and disadvantages of teaching medical students on clinical clerkships at a regional medical campus. J Reg Med Campuses 2018 Mar 30;1(2). [doi: 10.24926/jrmc.v1i2.1286]
15. Murakami M, Kawabata H, Maezawa M. The perception of the hidden curriculum on medical education: an exploratory study. Asia Pac Fam Med 2009;8:9. [doi: 10.1186/1447-056x-8-9]
16. Jenq CC, Lin JR, Quattri F, Monrouxe L. Medical students', residents', and nurses's feedback to clinical educators in Taiwan: a qualitative study. Med Educ 2024 Dec 20;58(12):1478-1489. [doi: 10.1111/medu.15429] [Medline: 38766732]
17. Roh H, Park SJ, Kim T. Patient safety education to change medical students' attitudes and sense of responsibility. Med Teach 2014 Oct 22;37(10):908-914. [doi: 10.3109/0142159x.2014.970988]
18. Brar G, Harney S, McGarr O, McFarland J. Mentoring and support practices for final year medical students during a pandemic - 'The covid doctors'. BMC Med Educ 2023 Jul 26;23(1):534 [FREE Full text] [doi: 10.1186/s12909-023-04513-9] [Medline: 37496028]
19. Harries AJ, Lee C, Jones L, Rodriguez RM, Davis JA, Boysen-Osborn M, et al. Effects of the COVID-19 pandemic on medical students: a multicenter quantitative study. BMC Med Educ 2021 Jan 06;21(1):14 [FREE Full text] [doi: 10.1186/s12909-020-02462-1] [Medline: 33407422]
20. Rose S. Medical student education in the time of COVID-19. JAMA 2020 Jun 02;323(21):2131-2132. [doi: 10.1001/jama.2020.5227] [Medline: 32232420]
21. Moir F, Henning M, Hassed C, Moyes SA, Elley CR. A peer-support and mindfulness program to improve the mental health of medical students. Teach Learn Med 2016 Apr 19;28(3):293-302. [doi: 10.1080/10401334.2016.1153475] [Medline: 27092397]
22. Preovolos C, Grant A, Rayner M, Fitzgerald K, Ng L. Peer mentoring by medical students for medical students: a scoping review. Med Sci Educ 2024 Dec 04;34(6):1577-1602. [doi: 10.1007/s40670-024-02108-7] [Medline: 39758463]
23. Wu J, Olagunju AT. Mentorship in medical education: reflections on the importance of both unofficial and official mentorship programs. BMC Med Educ 2024 Oct 29;24(1):1233 [FREE Full text] [doi: 10.1186/s12909-024-06248-7] [Medline: 39472896]
24. Sambunjak D, Straus SE, Marusić A. Mentoring in academic medicine: a systematic review. JAMA 2006 Sep 06;296(9):1103-1115. [doi: 10.1001/jama.296.9.1103] [Medline: 16954490]
25. Scherr R, Halaseh FF, Spina A, Andalib S, Rivera R. ChatGPT interactive medical simulations for early clinical education: case study. JMIR Med Educ 2023 Nov 10;9:e49877. [doi: 10.2196/49877] [Medline: 37948112]
26. Scherr R, Spina A, Dao A, Andalib S, Halaseh FF, Blair S, et al. Novel evaluation metric and quantified performance of ChatGPT-4 patient management simulations for early clinical education: experimental study. JMIR Form Res 2025 Feb 27;9:e66478 [FREE Full text] [doi: 10.2196/66478] [Medline: 40013991]
27. Rao SJ, Isath A, Krishnan P, Tangsrivimol JA, Virk HU, Wang Z, et al. ChatGPT: a conceptual review of applications and utility in the field of medicine. J Med Syst 2024 Jun 05;48(1):59. [doi: 10.1007/s10916-024-02075-x] [Medline: 38836893]
28. Magalhães Araujo S, Cruz-Correia R. Incorporating ChatGPT in medical informatics education: mixed methods study on student perceptions and experiential integration proposals. JMIR Med Educ 2024 Mar 20;10:e51151. [doi: 10.2196/51151] [Medline: 38506920]
29. Burke OM, Gwillim EC. Integrating artificial intelligence–based mentorship tools in dermatology. Acad Med 2024 Mar 15;99(6):e4. [doi: 10.1097/acm.0000000000005705]
30. Akdogan O, Uyar GC, Yesilbas E, Baskurt K, Malkoc NA, Ozdemir N, et al. Effect of a ChatGPT-based digital counseling intervention on anxiety and depression in patients with cancer: a prospective, randomized trial. Eur J Cancer 2025 May 15;221:115408. [doi: 10.1016/j.ejca.2025.115408] [Medline: 40215593]
31. Mitra NK, Chitra E. Glimpses of the use of generative AI and ChatGPT in medical education. Educ Med J 2024 Jun 28;16(2):155-164. [doi: 10.21315/eimj2024.16.2.11]

32. Ji YA, Kim G, Seo JH. Transforming medical communication education: the effectiveness of generative AI education using ChatGPT. J Health Inform Stat 2024 Nov;49(4):332-338. [doi: 10.21032/jhis.2024.49.4.332]
33. Meinlschmidt G, Koc S, Boerner E, Tegethoff M, Simacek T, Schirmer L, et al. Enhancing professional communication training in higher education through artificial intelligence(AI)-integrated exercises: study protocol for a randomised controlled trial. BMC Med Educ 2025 May 30;25(1):804 [FREE Full text] [doi: 10.1186/s12909-025-07307-3] [Medline: 40448046]
34. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep;13(3):319-340. [doi: 10.2307/249008]
35. Mastour H, Yousefi R, Niroumand S. Exploring the acceptance of e-learning in health professions education in Iran based on the technology acceptance model (TAM). Sci Rep 2025 Mar 10;15(1):8178 [FREE Full text] [doi: 10.1038/s41598-025-90742-5] [Medline: 40065055]
36. Liu F, Chang X, Zhu Q, Huang Y, Li Y, Wang H. Assessing clinical medicine students' acceptance of large language model: based on technology acceptance model. BMC Med Educ 2024 Nov 03;24:1251. [doi: 10.1186/s12909-024-06232-1]
37. Lee AT, Ramasamy RK, Subbarao A. Understanding psychosocial barriers to healthcare technology adoption: a review of TAM technology acceptance model and unified theory of acceptance and use of technology and UTAUT frameworks. Healthcare (Basel) 2025 Jan 27;13(3):250 [FREE Full text] [doi: 10.3390/healthcare13030250] [Medline: 39942440]
38. The model core curriculum for medical education in Japan. Medical Education Model Core Curriculum Expert Research Committee. 2022. URL: https://www.mext.go.jp/content/20250411-mxt_igaku-000028108_00003-2.pdf [accessed 2025-05-09]
39. DaRosa DA, Skeff K, Friedland JA, Coburn M, Cox S, Pollart S, et al. Barriers to effective teaching. Acad Med 2011;86(4):453-459. [doi: 10.1097/acm.0b013e31820defbe]
40. Frei E, Stamm M, Buddeberg-Fischer B. Mentoring programs for medical students--a review of the PubMed literature 2000-2008. BMC Med Educ 2010 Apr 30;10(1):32 [FREE Full text] [doi: 10.1186/1472-6920-10-32] [Medline: 20433727]
41. Nimmons D, Giny S, Rosenthal J. Medical student mentoring programs: current insights. Adv Med Educ Pract 2019 Mar;Volume 10:113-123. [doi: 10.2147/amep.s154974]
42. Learning and assessment items for skills and attitudes required in clinical clerkships. Common Achievement Tests Organization. 2024. URL: https://www.cato.or.jp/pdf/hyouka_1.1.pdf [accessed 2025-08-19]
43. National medical licensing examination guidelines 2024. Ministry of Health, Labour and Welfare (MHLW). URL: https://www.mhlw.go.jp/content/10803000/001082885.pdf [accessed 2025-08-19]
44. Clinical clerkship syllabus 2025. Chiba University School of Medicine. URL: https://www.m.chiba-u.ac.jp/application/files/1717/3043/3606/2025_4556.pdf [accessed 2025-05-09]
45. Li J, Zong H, Wu E, Wu R, Peng Z, Zhao J, et al. Exploring the potential of artificial intelligence to enhance the writing of english academic papers by non-native English-speaking medical students - the educational application of ChatGPT. BMC Med Educ 2024 Jul 09;24(1):736 [FREE Full text] [doi: 10.1186/s12909-024-05738-y] [Medline: 38982429]
46. Parente DJ. Generative artificial intelligence and large language models in primary care medical education. Fam Med 2024 Oct;56(9):534-540. [doi: 10.22454/FamMed.2024.775525] [Medline: 39207784]
47. Duan S, Liu C, Rong T, Zhao Y, Liu B. Integrating AI in medical education: a comprehensive study of medical students' attitudes, concerns, and behavioral intentions. BMC Med Educ 2025 Apr 23;25(1):599 [FREE Full text] [doi: 10.1186/s12909-025-07177-9] [Medline: 40269824]

Abbreviations

AI: artificial intelligence
AI-CCM: artificial intelligence clinical clerkship mentor
CC: clinical clerkship
gAI: generative artificial intelligence
GPT: generative pretrained transformer
LLM: large language model
TAM: technology acceptance model

Edited by B Lesselroth; submitted 30.05.25; peer-reviewed by M Gasmi ; comments to author 24.06.25; revised version received 17.07.25; accepted 14.08.25; published 04.09.25.

Please cite as:
Ebihara H, Kasai H, Shimizu I, Shikino K, Tajima H, Kimura Y, Ito S
Development of a Clinical Clerkship Mentor Using Generative AI and Evaluation of Its Effectiveness in a Medical Student Trial Compared to Student Mentors: 2-Part Comparative Study
JMIR Med Educ 2025;11:e76702
URL: https://mededu.jmir.org/2025/1/e76702
doi:10.2196/76702
PMID:

Original Paper

Evaluating the Performance of DeepSeek-R1 and DeepSeek-V3 Versus OpenAI Models in the Chinese National Medical Licensing Examination: Cross-Sectional Comparative Study

Weiping Wang[1*], PhD; Yuchen Zhou[1,2*], MD; Jingxuan Fu[3*], PhD; Ke Hu[1], PhD

[1]Department of Radiation Oncology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China

[2]Tsinghua Medicine, School of Medicine, Tsinghua University, Beijing, China

[3]Department of clinical laboratory, Xuanwu Hospital, Capital Medical University, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Ke Hu, PhD
Department of Radiation Oncology
Peking Union Medical College Hospital
Chinese Academy of Medical Sciences & Peking Union Medical College
Shuaifuyuan No. 1
Dongcheng District
Beijing, 100730
China
Phone: 86 13701011034
Email: huke8000@126.com

Abstract

Background: Deepseek-R1, an open-source large language model (LLM), has generated significant global interest in the past months.

Objective: This study aimed to compare the performance of DeepSeek and OpenAI LLMs on the Chinese National Medical Licensing Examination (NMLE) and evaluate their potential in medical education.

Methods: This cross-sectional study assessed 2 DeepSeek models (DeepSeek-R1 and DeepSeek-V3), 3 OpenAI models (ChatGPT-o1 pro, ChatGPT-o3 mini, and GPT-4o), and 2 additional Chinese LLMs (ERNIE 4.5 Turbo and Qwen 3) using the 2021 NMLE. Model performance was evaluated based on overall accuracy, accuracy across question types (A1, A2, A3 and A4, and B1), case analysis and non–case analysis questions, medical specialties, and accuracy consensus between different model combinations.

Results: All LLMs successfully passed the NMLE. DeepSeek-R1 achieved the highest accuracy (573/597, 96%), followed by DeepSeek-V3 (558/600, 93%), both of which significantly outperformed ChatGPT-o1 pro (450/600, 75%), ChatGPT-o3 mini (455/600, 75.8%), and GPT-4o (452/600, 75.3%; $P<.001$ for all comparisons). Performance disparities were consistent across various question types (A1, A2, A3 and A4, and B1), case analysis and non–case analysis questions, different types of case analyses, and medical specialties. The accuracy consensus between DeepSeek-R1 and DeepSeek-V3 reached 97.7% (544/557), significantly outperforming DeepSeek-R1 alone ($P=.04$). Two additional Chinese LLMs, ERNIE 4.5 Turbo (572/600, 95.3%) and Qwen 3 (555/600, 92.5%), also exhibited significantly better performance compared to the 3 OpenAI models (all $P<.001$).

Conclusions: This study demonstrates that DeepSeek-R1 and DeepSeek-V3 significantly outperform OpenAI models on the NMLE. DeepSeek models show promise as tools for medical education and exam preparation in the Chinese language.

(JMIR Med Educ 2025;11:e73469) doi:10.2196/73469

KEYWORDS

large language models; DeepSeek; OpenAI; Chinese Medical Licensing Examination; artificial intelligence; AI

Introduction

DeepSeek-R1, an open-source large language model (LLM), has generated significant global interest in the past months [1-5]. Launched by DeepSeek on January 20, 2025, this model demonstrates a performance comparable to that of OpenAI's ChatGPT-o1 in tasks involving mathematics, coding, and reasoning [2]. By leveraging a "mixture of experts" architecture, DeepSeek reduces the computational resources required for model training while enhancing the efficiency of query responses [6]. In addition, its lower cost and interface fees make it a more accessible option for users, providing a cost-effective alternative to OpenAI models. As an open-source LLM, DeepSeek-R1 offers users the ability to view and modify its source code, encouraging further improvements and customization. The release of this model has already made a splash in academic circles [1-3].

In recent years, OpenAI has maintained a leading position in the field of LLMs. OpenAI recently launched the ChatGPT-o1 pro and ChatGPT-o3 mini models, which are widely regarded as among the most powerful models available. GPT-4o, launched in May 2024, represents OpenAI's best-performing free model, with extensive applications and robust validation over the past year [7,8].

LLMs hold substantial potential in the medical domain, including clinical decision support [9,10], medical image analysis [11,12], health education, patient counseling, and medical training [13]. Medical licensing examinations are a critical entry test for medical licensure, and successfully passing them indicates a foundational understanding of medical knowledge. A model's ability to pass these exams not only suggests its potential to aid medical students in their exam preparation but also reflects its proficiency in medical knowledge and reasoning—skills essential for clinical decision-making and medical image interpretation.

Studies have shown that the GPT-4o model achieves an accuracy of 90.4% on the US Medical Licensing Examination (USMLE) [7], whereas its performance on the Chinese National Medical Licensing Examination (NMLE) stands at 84.9%, notably lower than its performance on the USMLE, as well as that on the United Kingdom's Professional and Linguistic Assessments Board (PLAB) test (93.3%) and the Hong Kong Medical Licensing Examination (91.7%) [8]. Developed by a Chinese company, the DeepSeek-R1 and DeepSeek-V3 models feature a higher proportion of Chinese-language corpora, making their performance on Chinese-language exams particularly promising.

In this study, we compared the performance of the DeepSeek and OpenAI LLMs on the NMLE and evaluated their potential in Chinese medical education.

Methods

LLMs Evaluated in This Study

The models evaluated in this study were DeepSeek-R1, DeepSeek-V3, ChatGPT-o1 pro, ChatGPT-o3 mini, GPT-4o, ERNIE 4.5 Turbo, and Qwen 3. DeepSeek-R1 and DeepSeek-V3 were developed by DeepSeek AI. ChatGPT-o1 pro, ChatGPT-o3 mini, and GPT-4o were developed by OpenAI. ERNIE 4.5 Turbo was developed by Baidu, and Qwen 3 was developed by the Alibaba Group.

NMLE Description

The NMLE is a comprehensive, nationwide exam required for obtaining medical practice credentials in China. All questions in the exam are multiple choice in Chinese, with 5 options provided and only 1 option being correct. The exam is divided into 4 sections, each consisting of 150 multiple-choice questions. The total number of questions is 600, with each being worth 1 point. To pass the exam, candidates must achieve a score of ≥360 points. Each section is allotted 2.5 hours, and the entire exam lasts 10 hours in total.

The questions in the exam are classified into 4 types. A1-type questions assess basic medical knowledge, and A2-, A3-, and A4-type questions are case analysis questions. The A2-type questions are presented with a single case, followed by 1 question related to that case. A3 and A4–type questions also follow a case-based format: each case is accompanied by 2 to 4 associated questions. B1-type questions consist of several sets of questions, each with a common list of 5 options. Some options may be selected once, multiple times, or not at all, depending on the question.

The 2021 NMLE was used in this study [14]. The Chinese National Medical Examination Center, which administered this NMLE, has confirmed that the use of these questions for academic research is exempt from copyright restrictions.

Model Testing Procedure

All models were assessed using their publicly accessible web interfaces in default configurations, precluding manual adjustment of parameters such as temperature, maximum tokens, or system prompts. This study used standardized prompting without in-context learning. All models were evaluated using identical input prompts for each question type (provided in Multimedia Appendix 1), with no additional examples or contextual demonstrations provided during testing. A standard prompt in Chinese was copied into the conversation window of each model to set the conditions for the exam. For the A1- and A2-type questions, 1 question was copied into the chat window at a time, and the models responded individually to each question. For A3 and A4–type questions, all related questions for a single case were input into the chat window simultaneously, and the models provided answers to all questions associated with that case. For the B1-type questions, all questions in a set with a shared list of 5 possible answers were input at once, and the models answered all questions in the set. Each set of questions (or individual question for the A1 and A2 types) was processed in a new, blank chat window. Testing took place from February 3, 2025, to February 9, 2025.

In addition, 2 experienced attending physicians independently classified the case-based questions (A2, A3, and A4 types) using 2 different methods. The first method involved categorizing the questions into broad categories such as examination, diagnosis, treatment, and other relevant areas. The second method involved classifying the questions according to medical specialty, including internal medicine, surgery, obstetrics and gynecology,

pediatrics, neurology, psychiatry, emergency medicine, and others. For cases involving multiple specialties, the primary specialty was determined based on the most relevant medical department. In instances of disagreement between the 2 attending physicians, a senior physician made the final decision regarding the classification.

To investigate the underlying factors contributing to DeepSeek models' superior performance, we conducted parallel evaluations by translating both the NMLE question bank and standardized prompts into English (see Multimedia Appendix 1 for the English-language prompt versions). GPT-4o was systematically evaluated using both the original Chinese questions and their English-translated counterparts to enable direct performance comparison across language conditions.

Statistics

Questions that the models did not answer were marked as omitted and excluded from the accuracy calculation. Accuracy consensus was defined as the ratio of agreement on correct responses between 2 models relative to their overall agreement (ie, considering both correct and incorrect responses).

For paired-sample comparisons, we used the McNemar chi-square test for 2-group comparisons and the Cochran *Q* test for multiple-group comparisons. When analyzing independent samples, we used the Pearson chi-square test, continuity correction, or the Fisher exact test as appropriate. The κ consistency test was used to quantitatively assess the concordance between 2 models. In terms of the κ coefficients, values exceeding 0.75 are considered to indicate excellent agreement, values within the range of 0.40 to 0.75 suggest good agreement, whereas values below 0.40 are indicative of poor agreement.

All statistical analyses were conducted using SPSS (version 23.0; IBM Corp), and a 2-tailed *P* value of <.05 was considered statistically significant.

Results

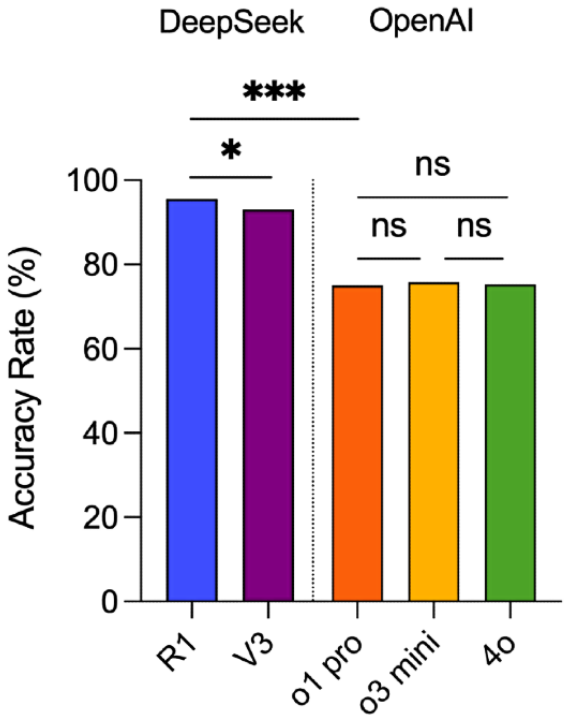
Overall Accuracy

DeepSeek-R1 did not provide answers for 3 questions, responding with "Sorry, I haven't yet learned how to approach these types of questions. I specialize in math, coding, and logic-related problems. Feel free to engage with me on these topics" (Multimedia Appendix 1). These 3 questions were marked as omissions. The remaining 4 models answered all the questions.

DeepSeek-R1, DeepSeek-V3, ChatGPT-o1 pro, ChatGPT-o3 mini, and GPT-4o answered 573, 558, 450, 455, and 452 questions correctly, respectively, achieving accuracy rates of 96% (573/597), 93% (558/600), 75% (450/600), 75.8% (455/600), and 75.3% (452/600), respectively (Figure 1 and Multimedia Appendix 1). The 3 questions that DeepSeek-R1 did not answer were excluded from the calculation.

All models scored above 360 points, thereby passing the NMLE. The accuracy of DeepSeek-R1 and DeepSeek-V3 was significantly higher than that of ChatGPT-o1, ChatGPT-o3, and GPT-4o (*P*<.001 for all comparisons). Between models from the same developers, DeepSeek-R1's accuracy was significantly higher than that of DeepSeek-V3 (*P*=.02), whereas no substantial differences were observed between ChatGPT-o1 pro, ChatGPT-o3 mini, and GPT-4o.

Figure 1. Overall accuracy rates of DeepSeek-R1, DeepSeek-V3, ChatGPT-o1 pro, ChatGPT-o3 mini, and GPT-4o. * *P*<.05, *** *P*<.005, and ns (no significance).

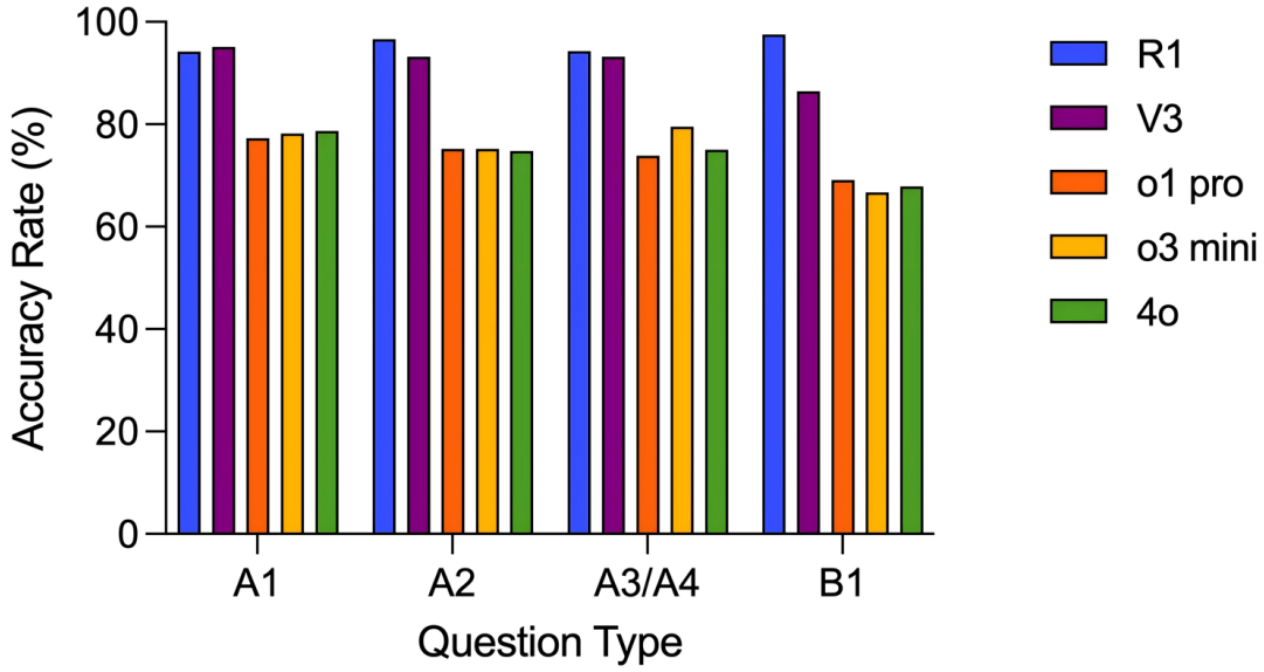


Accuracy Across Different Question Types

As shown in Figure 2 and Multimedia Appendix 1, for question types A1, A2, A3 and A4, and B1, DeepSeek-R1 and DeepSeek-V3 both demonstrated significantly higher accuracy than ChatGPT-o1 pro, ChatGPT-o3 mini, and GPT-4o (P<.05 for all comparisons). Specifically, DeepSeek-R1 achieved accuracy rates of 95.5% (212/222), 96.6% (199/206), 94% (83/88), and 98% (79/81) for the A1, A2, A3 and A4, and B1 question types, respectively. These accuracy rates were significantly higher than those of ChatGPT-o1 pro, which scored 77.3% (174/225), 75.2% (155/206), 74% (65/88), and 69% (56/81) for the same question types, respectively (P<.001 for all comparisons).

The performance of the 2 DeepSeek models (R1 and V3) did not show significant differences across the 4 question types (P>.05 for all comparisons). Similarly, the 3 OpenAI models (ChatGPT-o1 pro, ChatGPT-o3 mini, and GPT-4o) also did not demonstrate significant differences in performance across the 4 question types (P>.05 for all comparisons).

Figure 2. Accuracy Rates of DeepSeek-R1, DeepSeek-V3, ChatGPT-o1 pro, ChatGPT-o3 mini, and GPT-4o Across Various Question Types.



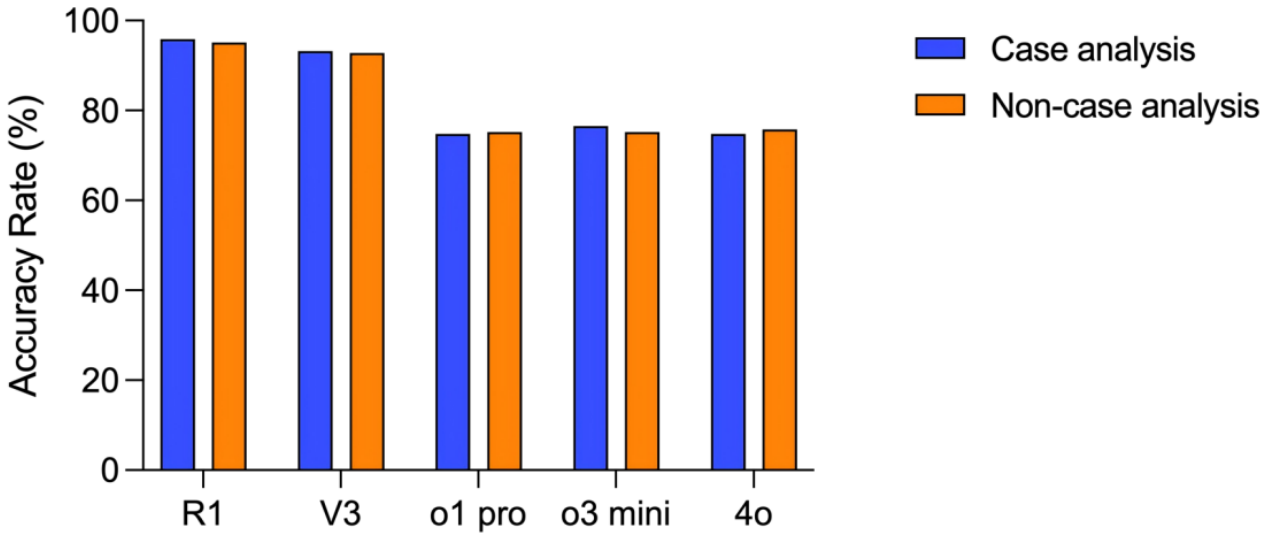
Accuracy Across Case Analysis and Non–Case Analysis Questions

Of the 600 total questions, 294 (49%) were case analysis questions, and 306 (51%) were non–case analysis questions. As shown in Figure 3 and Multimedia Appendix 1, the accuracy rates for both case analysis and non–case analysis questions were similar across all 5 models (DeepSeek-R1, DeepSeek-V3, ChatGPT-o1 pro, ChatGPT-o3 mini, and GPT-4o).

For case analysis questions, DeepSeek-R1 and DeepSeek-V3 consistently outperformed ChatGPT-o1 pro, ChatGPT-o3 mini, and GPT-4o, with all comparisons yielding significant differences (P<.001 for all comparisons). Specifically, DeepSeek-R1 achieved an accuracy rate of 95.9% (282/294), which was significantly higher than the 74.8% (220/294) accuracy observed for ChatGPT-o1 pro (P<.001). No significant difference in accuracy was observed between DeepSeek-R1 (282/294, 95.9%) and DeepSeek-V3 (274/294, 93.2%; P=.10). Similarly, the 3 OpenAI models—ChatGPT-o1 pro (220/294, 74.8%), ChatGPT-o3 mini (225/294, 76.5%), and GPT-4o (220/294, 74.8%)—did not exhibit significant differences in their performance on case analysis questions (P>.05 for all comparisons).

For non–case analysis questions, DeepSeek-R1 and DeepSeek-V3 again demonstrated significantly higher accuracy than the OpenAI models (P<.001 for all comparisons). DeepSeek-R1 achieved an accuracy of 96% (291/303) for non–case analysis questions, which was significantly higher than the 75.2% (230/306) accuracy of ChatGPT-o1 pro (P<.001). There was no significant difference in accuracy between DeepSeek-R1 (291/303, 96%) and DeepSeek-V3 (274/294, 93.2%) on non–case analysis questions (P=.23). Furthermore, no significant differences were found among ChatGPT-o1 pro (230/306, 75.2%), ChatGPT-o3 mini (230/306, 75.2%), and GPT-4o (232/306, 75.8%) in terms of their performance on these questions (P>.05 for all comparisons).

Figure 3. Accuracy Rates of DeepSeek-R1, DeepSeek-V3, ChatGPT-o1 pro, ChatGPT-o3 mini, and GPT-4o on Case Analysis and Non-Case Analysis Questions.



Performance Across Different Types of Case Analysis Questions

For the diagnosis type of case analysis question, DeepSeek-R1 achieved an accuracy of 98% (146/149), which was slightly higher than its accuracy of 92% (72/78) for treatment type of case analysis questions (*P*=.09). Similarly, ChatGPT-o3 mini showed a diagnostic accuracy of 80.5% (120/149), which was higher than its treatment accuracy of 68% (53/78; *P*=.048). DeepSeek-V3, ChatGPT-o1 pro, and GPT-4o also showed higher diagnostic accuracy compared to treatment, but the differences were not statistically significant (*P*>.05 for all comparisons).

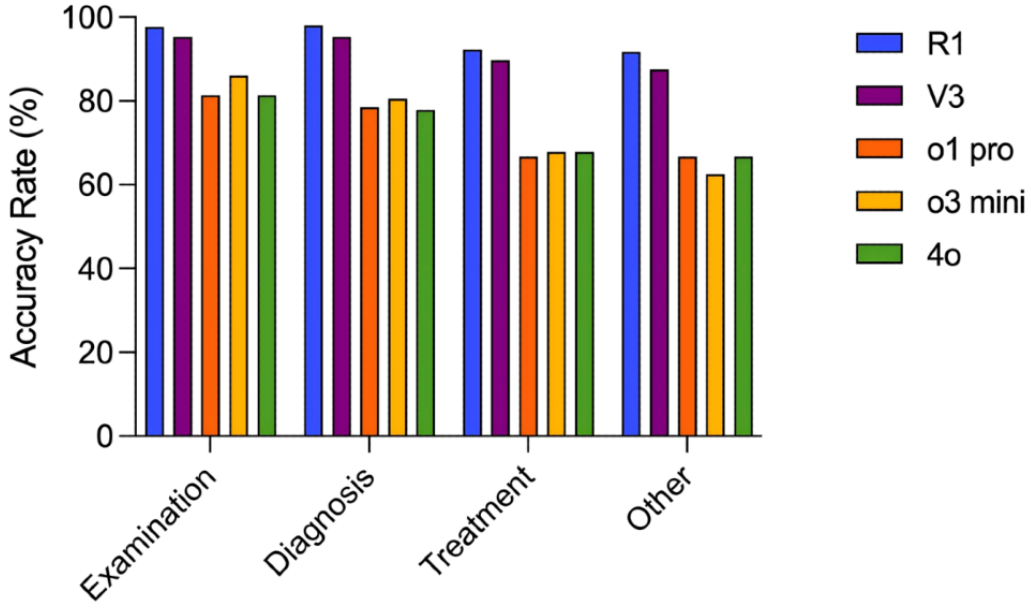
For the examination type of case analysis question, DeepSeek-R1 achieved an accuracy rate of 98% (42/43), significantly outperforming ChatGPT-o1 pro (35/43, 81%; *P*=.02) and GPT-4o (35/43, 81%; *P*=.02). It also outperformed ChatGPT-o3 mini (37/43, 86%), although the difference was not statistically significant (*P*=.06).

Regarding diagnosis, DeepSeek-R1 and DeepSeek-V3 achieved accuracy rates of 98% (146/149) and 95.3% (142/149), respectively, both significantly outperforming the OpenAI models (accuracy rates ranging from 116/149, 77.9% to 120/149, 80.5%; *P*<.001 for all comparisons).

For treatment questions, DeepSeek-R1 and DeepSeek-V3 achieved accuracy rates of 92% (72/78) and 90% (70/78), respectively, both significantly outperforming the OpenAI models, whose accuracy ranged from 67% (52/78) to 68% (53/78; *P*<.001 for all comparisons).

The details are shown in Figure 4 and Multimedia Appendix 1.

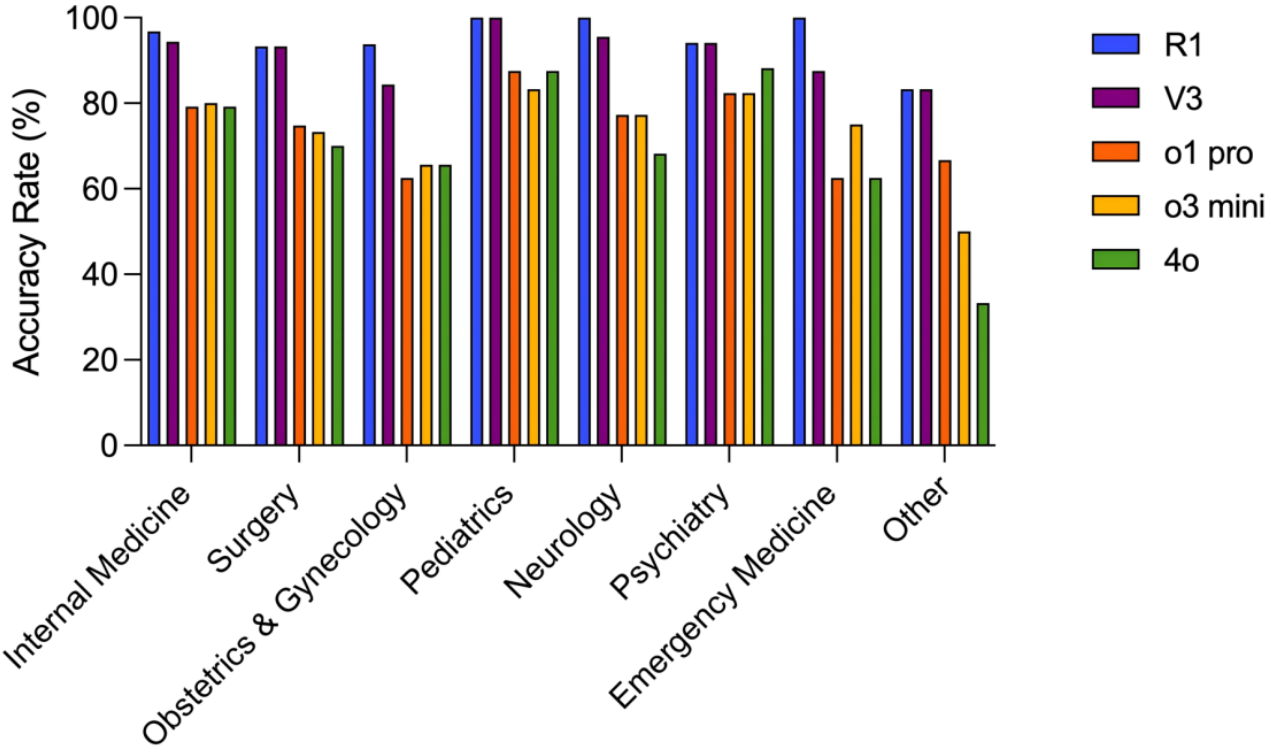
Figure 4. Accuracy Rates of DeepSeek-R1, DeepSeek-V3, ChatGPT-o1 pro, ChatGPT-o3 mini, and GPT-4o Across Different Types of Case Analysis Questions.



Case Analysis Questions Across Different Medical Specialties

Of the 294 case analysis questions, 125 (42.5%) were from internal medicine, 60 (20.4%) were from surgery, 32 (10.9%) were from obstetrics and gynecology, 24 (8.2%) were from pediatrics, and 22 (7.5%) were from neurology. DeepSeek-R1 achieved accuracy rates ranging from 93% (56/60) to 100% (24/24) across all specialties, whereas DeepSeek-V3 had accuracy rates between 84% (27/32) and 100% (24/24). In contrast, the OpenAI models—ChatGPT-o1 pro, ChatGPT-o3 mini, and GPT-4o—showed accuracy rates between 62% (20/32) and 88% (21/24), 66% (21/32) and 83% (20/24), and 62% (5/8) and 88% (21/24), respectively (Figure 5 and Multimedia Appendix 1).

Figure 5. Accuracy Rates of DeepSeek-R1, DeepSeek-V3, ChatGPT-o1 pro, ChatGPT-o3 mini, and GPT-4o Across Different Medical Specialties in Case Analysis Questions.



Accuracy Consensus

As shown in Table 1 and Multimedia Appendix 1, the accuracy consensus between DeepSeek-R1 and DeepSeek-V3 reached 97.7% (544/557), the highest among all model pairs, significantly outperforming DeepSeek-R1 alone (573/597, 96%; P=.04) and DeepSeek-V3 (558/600, 93%; P<.001). The accuracy consensus between DeepSeek-R1 and ChatGPT-o1 pro, ChatGPT-o3 mini, and GPT-4o was 95.5% (444/465), 95.9% (447/466), and 97.3% (437/449), respectively, with no significant differences compared to DeepSeek-R1's individual accuracy (573/597, 96%; P>.05 for all comparisons). Regarding the 3 OpenAI models, their pairwise accuracy consensus values were 77.7% (428/551, ChatGPT-o3 mini and ChatGPT-o1 pro), 78.1% (420/538, GPT-4o and ChatGPT-o1 pro), and 79.2% (416/525, GPT-4o and ChatGPT-o3 mini), with no significant differences compared to the individual accuracy of each model (P>.05 for all comparisons).

The results of the κ consistency test are shown in Table 2. Specifically, the κ coefficient between ChatGPT-o1 pro and ChatGPT-o3 mini was 0.781, demonstrating excellent agreement. The κ coefficients between GPT-4o and ChatGPT-o1 pro and between GPT-4o and ChatGPT-o3 mini were 0.723 and 0.661, respectively, indicating good agreement. Notably, the κ coefficients between the 2 DeepSeek models, as well as between the DeepSeek models and the OpenAI models, were all below 0.4, suggesting poor agreement.

Our comparative analysis of accuracy consensus across different model combinations revealed that the pairing of DeepSeek-R1 and DeepSeek-V3 achieved significantly higher consensus rates than combinations consisting solely of OpenAI models (Table 1 and Multimedia Appendix 1).

Table 1. Accuracy consensus among DeepSeek-R1, DeepSeek-V3, ChatGPT-o1 pro, ChatGPT-o3 mini, and GPT-4o.

Model pair	Accuracy consensus, n/N (%)
DeepSeek-V3 and DeepSeek-R1	544/557 (97.7)
ChatGPT-o1 pro and DeepSeek-R1	444/465 (95.5)
ChatGPT-o1 pro and DeepSeek-V3	441/474 (93)
ChatGPT-o3 mini and DeepSeek-R1	447/466 (95.9)
ChatGPT-o3 mini and DeepSeek-V3	444/475 (93.5)
ChatGPT-o3 mini and ChatGPT-o1 pro	428/551 (77.7)
GPT-4o and DeepSeek-R1	437/449 (97.3)
GPT-4o and DeepSeek-V3	435/460 (94.6)
GPT-4o and ChatGPT-o1 pro	420/538 (78.1)
GPT-4o and ChatGPT-o3 mini	416/525 (79.2)

Table 2. κ coefficients for the consistency test among DeepSeek-R1, DeepSeek-V3, ChatGPT-o1 pro, ChatGPT-o3 mini, and GPT-4o.

Model pair	κ coefficient
DeepSeek-V3 and DeepSeek-R1	0.341
ChatGPT-o1 pro and DeepSeek-R1	0.171
ChatGPT-o1 pro and DeepSeek-V3	0.263
ChatGPT-o3 mini and DeepSeek-R1	0.157
ChatGPT-o3 mini and DeepSeek-V3	0.250
ChatGPT-o3 mini and ChatGPT-o1 pro	0.781
GPT-4o and DeepSeek-R1	0.066
GPT-4o and DeepSeek-V3	0.173
GPT-4o and ChatGPT-o1 pro	0.723
GPT-4o and ChatGPT-o3 mini	0.661

English-Translated Questions

Following the translation of NMLE questions into English, GPT-4o demonstrated significantly improved performance, with an accuracy of 83.5% (501/600), representing a notable increase from its 75.3% (452/600) accuracy on the original Chinese version (P<.001). Nevertheless, this enhanced performance still fell significantly short of DeepSeek-R1's 96% (573/597) accuracy on the Chinese-language questions (P<.001).

The performance improvement was particularly pronounced across specific question types: GPT-4o showed statistically significant gains in accuracy for A1-type (193/225, 85.8% vs 177/225, 78.7%; P=.03), A2-type (168/206, 81.6% vs 154/206, 74.8%; P=.04), and A3 and A4–type (78/88, 89% vs 66/88, 75%; P=.01) questions when answering English-translated versions. Notably, both case analysis and non–case analysis questions showed marked performance enhancements in the English-translated condition. The details are shown in Table 3.

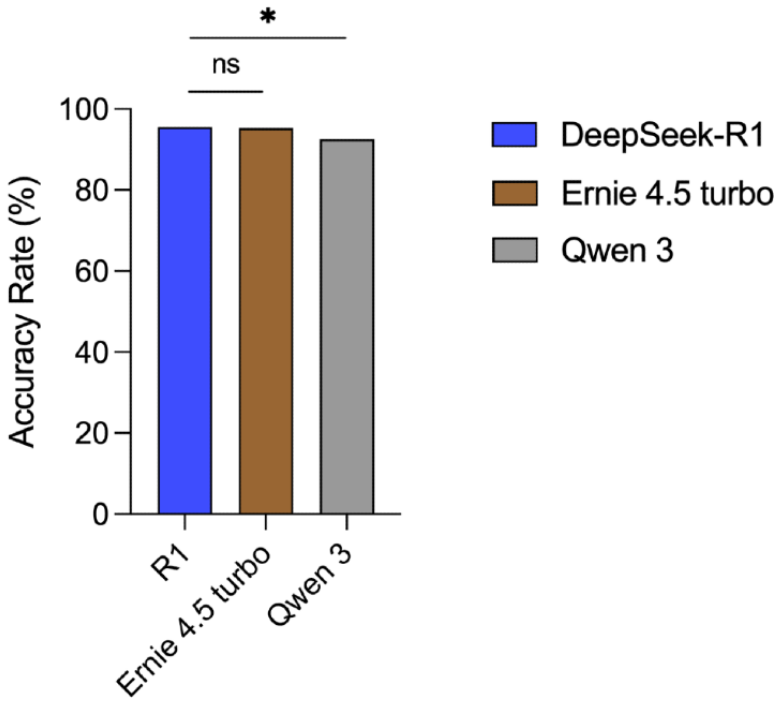
Table 3. GPT-4o performance on Chinese-language versus English-translated questions across different question types (N=600).

Question type	Correct answers in Chinese-language questions, n (%)	Correct answers in English-translated questions, n (%)	P value
A1 (n=225)	177 (78.7)	193 (85.8)	.03
A2 (n=206)	154 (74.8)	168 (81.6)	.04
A3 and A4 (n=88)	66 (75)	78 (88.6)	.01
B1 (n=81)	55 (67.9)	62 (76.5)	.23
Case analysis (n=294)	220 (74.8)	246 (83.7)	<.001
Non–case analysis (n=306)	232 (75.8)	255 (83.3)	.01
Total	452 (75.3)	501 (83.5)	<.001

Comparison of Chinese LLMs

We compared DeepSeek-R1 with 2 other Chinese LLMs, ERNIE 4.5 Turbo and Qwen 3, using the same NMLE questions Figure 6. ERNIE 4.5 Turbo achieved an overall accuracy of 95.3% (572/600), comparable to that of DeepSeek-R1 (573/597, 96%; *P*=.58). In contrast, Qwen 3 demonstrated an accuracy of 92.5% (555/600), significantly lower than that of DeepSeek-R1 (*P*=.01). Both ERNIE 4.5 Turbo and Qwen 3 significantly outperformed OpenAI's 3 models (*P*<.001 for all comparisons).

Figure 6. Overall accuracy rates of DeepSeek-R1, ERNIE 4.5 Turbo and Qwen 3. * *P*<.05 and ns (no significance).



Discussion

In this study, we observed that the DeepSeek-R1 and Deepseek-V3 models achieved significantly higher accuracy rates on the NMLE, with accuracy scores of 93% (558/600) and 95.5% (212/222), respectively, substantially outperforming the OpenAI models, which demonstrated accuracy rates ranging from 75% (450/600) to 75.8% (455/600). In addition, the DeepSeek models consistently outperformed the OpenAI models across a variety of question types. A further in-depth analysis revealed that the accuracy consensus between DeepSeek-R1 and DeepSeek-V3 reached a peak of 97.7% (544/557). These findings highlight the superior performance of the DeepSeek models in the context of the NMLE when compared to OpenAI models. To the best of our knowledge, this is the first study to directly compare the performance of DeepSeek and OpenAI models within the context of a medical examination.

Previous studies have demonstrated that GPT-4's performance on the NMLE is notably inferior to its performance on English-language exams such as the USMLE, the United Kingdom's PLAB, and the Hong Kong Medical Licensing Examination [7,8]. Previous research has also revealed that ChatGPT was unable to pass the Taiwanese Family Medicine Board Exam in Chinese [15], indicating that OpenAI models may not perform as well on Chinese medical exams compared to their performance on their English-language counterparts. Before the advent of DeepSeek, a study found that the Chinese-developed LLM Qwen-2.5 (Alibaba Group) achieved an accuracy of 88.9% on the Chinese National Nursing Licensing Examination, surpassing GPT-4, which had an accuracy of 80.7%. Another Chinese LLM, ERNIE Bot-3.5 (Baidu), demonstrated an accuracy of 78.1%, which is comparable to that of GPT-4 [16]. Additional studies have shown that ERNIE Bot performed similarly to OpenAI's models in recognizing Chinese health-related rumors and providing medical counseling [17,18]. Although the previous Chinese LLMs from DeepSeek had a noticeable gap compared to OpenAI models, in terms of processing Chinese medical questions, those models have already become similar to OpenAI models of the same period. In this study, another Chinese LLM—ERNIE 4.5 Turbo—performed comparably to DeepSeek in the NMLE, whereas Qwen 3, although it achieved a lower accuracy (92.5%) than DeepSeek-R1, still significantly outperformed OpenAI models.

Neither DeepSeek nor OpenAI has disclosed the exact proportion of Chinese-language data used in their models. However, as a Chinese-developed LLM, DeepSeek likely incorporates a significantly higher proportion of Chinese-language data than OpenAI's more globally oriented models. It is estimated that OpenAI's Chinese-language training data comprise less than 5%, whereas DeepSeek's proportion exceeds 40%. This higher proportion of Chinese-language data likely enhances DeepSeek's ability to understand and process Chinese-language text with greater accuracy. Our translation experiment further substantiates this hypothesis. When NMLE questions were translated into English, GPT-4o's accuracy significantly increased from 75.3% (452/600) to 83.5% (501/600; *P*<.001). This 8.2 percentage points gain demonstrates that language barriers materially constrain OpenAI's

performance in Chinese medical examinations. Practically, these findings reveal a strategic implication: NMLE candidates using OpenAI models for preparation may improve answer reliability by translating questions into English. Nevertheless, even with optimized language conditions, GPT-4o's translated question accuracy remained significantly inferior to DeepSeek-R1's native-language performance (501/600, 83.5% vs 573/597, 96%; $P<.001$), suggesting that DeepSeek's advantage extends beyond linguistic proficiency to encompass domain-specific medical knowledge alignment.

DeepSeek-R1 provides chain-of-thought (CoT) processes when answering questions, which offers us a valuable window to observe the reasoning processes of LLMs. Interestingly, DeepSeek-R1's CoT frequently references Chinese medical textbooks, which serve as key references for the NMLE. In Multimedia Appendix 1, we present 3 specific examples of this phenomenon. In the CoT analysis of 2 case study questions, the model consistently followed a structured approach: first identifying key diagnostic elements (such as Reed-Sternberg cells in case 1 and upper gastrointestinal bleeding in case 2) and then determining the most appropriate answer choice based on medical textbooks from People's Medical Publishing House editions. Subsequently, it systematically evaluated why alternative options were incorrect before arriving at the final correct answer. The medical textbooks published by People's Medical Publishing House are the most important official reference materials for the NMLE. For clinically controversial questions, the examination board explicitly uses these textbooks as the definitive authority. This alignment between DeepSeek's reference sources and the exam's official evaluation criteria may be one of the key reasons for DeepSeek's superior performance.

Differences in disease patterns between China and the United States, as well as variations in diagnostic criteria, treatment protocols, and clinical guidelines for the same diseases, may further contribute to DeepSeek's superior performance on Chinese-language medical exams. For example, in this study, 8.5% (51/600) of the questions were related to tuberculosis. According to the 2024 Global Tuberculosis Report from the World Health Organization, the incidence of tuberculosis in China is approximately 13 times higher than that in the United States [19]. The higher incidence results in an abundance of Chinese-language resources related to tuberculosis, which may contribute to DeepSeek's stronger performance on related questions.

Another important observation is that while ChatGPT-o1 pro and ChatGPT-o3 mini outperform GPT-4o in areas such as mathematics and coding [20,21], this study found that the 3 OpenAI models—ChatGPT-o1 pro, ChatGPT-o3 mini, and GPT-4o—demonstrated almost identical accuracy rates on the NMLE, with accuracy scores of 75% (450/600), 75.8% (455/600), and 75.3% (452/600), respectively. In contrast, DeepSeek-R1, which is considered superior to DeepSeek-V3 [6], achieved only a 2.5% higher accuracy than DeepSeek-V3. The primary distinction between DeepSeek-R1 and DeepSeek-V3, as well as among the OpenAI models, appears to be in their reasoning capabilities. However, the medical licensing examination is a basic-level test for physicians and does not require advanced reasoning skills. It is generally understood that case analysis questions demand more reasoning ability than non–case analysis questions. As illustrated in Figure 3, the accuracy rates for all 5 models were similar across both case analysis and non–case analysis questions. Performance differences between the 2 DeepSeek models and among the 3 OpenAI models were comparable on case analysis questions. This suggests that the reasoning abilities of both DeepSeek-V3 and GPT-4o may be adequate for a medical licensing examination. However, this also raises the possibility that the observed performance discrepancies in DeepSeek models on the NMLE are not attributable to differences in reasoning abilities but rather to the variations in the training corpora discussed previously. To further explore this, additional studies assessing DeepSeek's performance on English-language examinations such as the USMLE would be beneficial.

DeepSeek-R1 declined to answer 3 questions, all notably related to obstetrics and gynecology (Multimedia Appendix 1). These refusals likely stem from the domain's inherent sensitivity (pregnancy, fetal abnormalities, and reproductive rights) and associated ethical dilemmas, triggering the model's protective filters. This conservative "silence over risk" approach aligns with DeepSeek's technical report showing that safety reinforcement learning reduces benchmark accuracy (58.7% vs 70% without safety reinforcement learning) [22]. Such refusals represent responsible artificial intelligence (AI) design but underscore challenges in applying LLMs to high-stakes medical contexts in which comprehensive response capability remains essential.

The high accuracy of LLMs in the NMLE demonstrates their potential as auxiliary tools for exam preparation. Our findings reveal that DeepSeek models significantly outperform their OpenAI counterparts in NMLE performance, suggesting that students should prioritize DeepSeek over OpenAI models for NMLE preparation. Given that DeepSeek-R1's accuracy improvement over DeepSeek-V3 was marginal (3%), whereas DeepSeek-V3 already achieves a high accuracy (558/600, 93%) and substantially faster response times, DeepSeek-V3 offers a balanced combination of precision and efficiency for daily practice. However, for in-depth analysis of complex clinical cases, DeepSeek-R1 may offer advantages due to its explicit CoT reasoning capabilities. When using primarily English language–trained models such as those by OpenAI, posing queries in English is recommended to mitigate performance limitations. Looking ahead, developing adaptive learning systems based on LLMs could dynamically tailor training focus according to individual error patterns observed in practice questions.

The superior performance of DeepSeek models on the NMLE suggests transformative potential for medical education. These LLMs can serve as intelligent teaching assistants, providing evidence-based tutoring with particular strength in clinical reasoning. Their case analysis capabilities enable the development of AI-standardized patients that adapt to learner levels while performance variations across question types inform personalized learning systems. As hybrid educational tools, they combine the efficiency of AI with clinical training needs, with DeepSeek-V3 offering rapid practice and DeepSeek-R1 enabling

in-depth case analysis while maintaining crucial human oversight for complex decision-making.

This study has several limitations. First, it focused exclusively on the NMLE, and therefore, the reasons behind DeepSeek's superior accuracy compared to OpenAI models cannot be conclusively determined without evaluating their performance on other English-language medical examinations such as the USMLE and PLAB. Second, the questions in this study did not include medical imaging, leaving the models' performance on image-based medical tasks unassessed. Third, we primarily attributed DeepSeek's superior NMLE performance to its greater amount of Chinese-language training data. However, without reliable documentation from both companies about their actual Chinese-language training data proportions, our analysis might be incomplete. The performance differences could also stem from other factors such as reasoning capabilities or DeepSeek's tendency to reference Chinese medical textbooks.

In conclusion, all 5 models successfully passed the NMLE. DeepSeek-R1 achieved the highest accuracy rate of 96% (573/597), followed by DeepSeek-V3 at 93% (558/600). Both DeepSeek models significantly outperformed the OpenAI models, whose accuracy rates ranged from 75% (450/600) to 75.8% (455/600). DeepSeek models exhibited superior performance across various question types, including case analysis, non–case analysis, and different subtypes of case analysis questions. The accuracy consensus between DeepSeek-R1 and DeepSeek-V3 further boosted accuracy to 97.7% (544/557). The outstanding performance of DeepSeek models on the NMLE underscores their considerable potential for enhancing diagnostic and treatment decision-making, patient education, and medical popularization in Chinese-language contexts. Moreover, they represent a valuable tool for preparing for medical licensing examinations.

Acknowledgments

This work is supported by Curriculum Ideology and Politics Teaching Reform Special Fund Project of Peking Union Medical College (No.2024kcsz022) and National High Level Hospital Clinical Research Funding (No.2022-PUMCH-B-127).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Standard prompts, model performance comparisons across question types and medical specialties, declined questions, statistical analyses, and chain-of-thought evaluation.
[DOC File , 95 KB - mededu_v11i1e73469_app1.doc]

References

1. Gibney E. Scientists flock to DeepSeek: how they're using the blockbuster AI model. Nature 2025 Jan 29 (forthcoming). [doi: 10.1038/d41586-025-00275-0] [Medline: 39881178]
2. Gibney E. China's cheap, open AI model DeepSeek thrills scientists. Nature 2025 Feb;638(8049):13-14. [doi: 10.1038/d41586-025-00229-6] [Medline: 39849139]
3. Normile D. Chinese firm's large language model makes a splash. Science 2025 Jan 17;387(6731):238. [doi: 10.1126/science.adv9836] [Medline: 39818899]
4. Smith J. Daily briefing: the pros and cons of DeepSeek. Nature 2025 Jan 30:1-2 (forthcoming). [doi: 10.1038/d41586-025-00330-w] [Medline: 39890911]
5. Conroy G, Mallapaty S. How China created AI model DeepSeek and shocked the world. Nature 2025 Feb;638(8050):300-301. [doi: 10.1038/d41586-025-00259-0] [Medline: 39885352]
6. DeepSeek-AI. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. ArXiv Preprint posted online on January 22, 2025 [FREE Full text] [doi: 10.36227/techrxiv.174742969.91776650/v1]
7. Bicknell BT, Butler D, Whalen S, Ricks J, Dixon CJ, Clark AB, et al. ChatGPT-4 Omni performance in USMLE disciplines and clinical skills: comparative analysis. JMIR Med Educ 2024 Nov 06;10:e63430 [FREE Full text] [doi: 10.2196/63430] [Medline: 39504445]
8. Chen Y, Huang X, Yang F, Lin H, Lin H, Zheng Z, et al. Performance of ChatGPT and Bard on the medical licensing examinations varies across different cultures: a comparison study. BMC Med Educ 2024 Nov 26;24(1):1372 [FREE Full text] [doi: 10.1186/s12909-024-06309-x] [Medline: 39593041]
9. Lahat A, Sharif K, Zoabi N, Shneor Patt Y, Sharif Y, Fisher L, et al. Assessing generative pretrained transformers (GPT) in clinical decision-making: comparative analysis of GPT-3.5 and GPT-4. J Med Internet Res 2024 Jun 27;26:e54571 [FREE Full text] [doi: 10.2196/54571] [Medline: 38935937]
10. Schmidl B, Hütten T, Pigorsch S, Stögbauer F, Hoch CC, Hussain T, et al. Assessing the role of advanced artificial intelligence as a tool in multidisciplinary tumor board decision-making for primary head and neck cancer cases. Front Oncol 2024 May 24;14:1353031 [FREE Full text] [doi: 10.3389/fonc.2024.1353031] [Medline: 38854718]

11. Kaczmarczyk R, Wilhelm TI, Martin R, Roos J. Evaluating multimodal AI in medical diagnostics. NPJ Digit Med 2024 Aug 07;7(1):205 [FREE Full text] [doi: 10.1038/s41746-024-01208-3] [Medline: 39112822]
12. Suh PS, Shim WH, Suh CH, Heo H, Park KJ, Kim PH, et al. Comparing large language model and human reader accuracy with New England Journal of Medicine image challenge case image inputs. Radiology 2024 Dec;313(3):e241668. [doi: 10.1148/radiol.241668] [Medline: 39656125]
13. Law AK, So J, Lui CT, Choi YF, Cheung KH, Kei-Ching Hung K, et al. AI versus human-generated multiple-choice questions for medical education: a cohort study in a high-stakes examination. BMC Med Educ 2025 Feb 08;25(1):208 [FREE Full text] [doi: 10.1186/s12909-025-06796-6] [Medline: 39923067]
14. The Research Group for the National Medical Licensing Examination, Past Exam Questions and Detailed Explanations for Clinical Practitioners. Shenyang city, Liaoning province, China: Shenyang Liaoning University Press; 2024.
15. Weng TL, Wang YM, Chang S, Chen TJ, Hwang SJ. ChatGPT failed Taiwan's family medicine board exam. J Chin Med Assoc 2023 Aug 01;86(8):762-766. [doi: 10.1097/JCMA.0000000000000946] [Medline: 37294147]
16. Zhu S, Hu W, Yang Z, Yan J, Zhang F. Qwen-2.5 outperforms other large language models in the Chinese national nursing licensing examination: retrospective cross-sectional comparative study. JMIR Med Inform 2025 Jan 10;13:e63731 [FREE Full text] [doi: 10.2196/63731] [Medline: 39793017]
17. Luo Y, Miao Y, Zhao Y, Li J, Chen Y, Yue Y, et al. Comparing the accuracy of two generated large language models in identifying health-related rumors or misconceptions and the applicability in health science popularization: proof-of-concept study. JMIR Form Res 2024 Dec 02;8:e63188 [FREE Full text] [doi: 10.2196/63188] [Medline: 39622076]
18. Kong QZ, Ju KP, Wan M, Liu J, Wu XQ, Li YY, et al. Comparative analysis of large language models in medical counseling: a focus on Helicobacter pylori infection. Helicobacter 2024;29(1):e13055. [doi: 10.1111/hel.13055] [Medline: 39078641]
19. Global tuberculosis report 2024. World Health Organization. URL: https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2024?utm_source=chatgpt.com [accessed 2025-03-16]
20. OpenAI o1 system card. OpenAI. 2024 Dec 05. URL: https://cdn.openai.com/o1-system-card-20241205.pdf [accessed 2025-03-16]
21. OpenAI o3-mini system card. OpenAI. 2025 Jan 31. URL: https://cdn.openai.com/o3-mini-system-card-feb10.pdf [accessed 2025-03-16]
22. Guo D, Yang D, Zhang H, Song J, Wang P, Zhu Q, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature 2025 Sep;645(8081):633-638. [doi: 10.1038/s41586-025-09422-z] [Medline: 40962978]

Abbreviations

AI: artificial intelligence
CoT: chain-of-thought
LLM: large language model
NMLE: National Medical Licensing Examination
PLAB: Professional and Linguistic Assessments Board
USMLE: US Medical Licensing Examination

Edited by S Zelko, R Pellegrino; submitted 05.03.25; peer-reviewed by C Ma, YD Cheng; comments to author 07.06.25; revised version received 21.07.25; accepted 12.10.25; published 14.11.25.

Please cite as:
Wang W, Zhou Y, Fu J, Hu K
Evaluating the Performance of DeepSeek-R1 and DeepSeek-V3 Versus OpenAI Models in the Chinese National Medical Licensing Examination: Cross-Sectional Comparative Study
JMIR Med Educ 2025;11:e73469
URL: https://mededu.jmir.org/2025/1/e73469
doi:10.2196/73469
PMID:

Evaluation of a Simulation Program for Providing Telenursing Training to Nursing Students: Cohort Study

Ola Ali-Saleh*, PhD; Layalleh Massalha*, MA; Ofra Halperin*, PhD

Department of Nursing, Max Stern Yezreel Valley College, Emek Yezreel, Israel

*all authors contributed equally

Corresponding Author:
Ofra Halperin, PhD
Department of Nursing, Max Stern Yezreel Valley College, Emek Yezreel, Israel

Abstract

Background: Telenursing has become prevalent in providing care to diverse populations experiencing different health conditions both in Israel and globally. The nurse-patient relationship aims to improve the condition of individuals requiring health services.

Objectives: This study aims to evaluate nursing graduates' skills and knowledge regarding remote nursing care prior to and following a simulation-based telenursing training program in an undergraduate nursing degree.

Methods: A cohort study assessed 114 third-year nursing students using comprehensive evaluation measures of knowledge, skills, attitudes, self-efficacy, and clinical skills regarding remote nursing care. Assessments were conducted at 2 critical time points: prior to and following a structured simulation-based training intervention.

Results: Participant demographics revealed a predominantly female sample (101/114, 88.6%), aged 20 - 50 years (mean 25.68, SD 4.59 years), with moderate to advanced computer and internet proficiency. Notably, 91.2% (104/114) had no telenursing exposure, yet 75.4% (86/114) expressed training interest. Statistical analyses demonstrated significant improvements across all measured variables, characterized by moderate to high effect sizes. Key findings included substantial increases in telenursing awareness, knowledge, skills, attitudes and self-efficacy; significant reduction in perceived barriers to remote care delivery; and complex interrelation dynamics between variables. A multivariate analysis revealed nuanced correlations: higher awareness and knowledge were consistently associated with more positive attitudes and increased self-efficacy. Positive attitudes correlated with enhanced self-efficacy and reduced perceived barriers. Change score analyses further indicated that increased awareness and knowledge facilitated more positive attitudinal shifts, while heightened awareness and positive attitudes corresponded with decreased implementation barriers.

Conclusions: The study underscores the critical importance of integrating targeted telenursing training into nursing education. By providing comprehensive preparation, educational programs can equip students to deliver optimal remote care services. The COVID-19 pandemic has definitively demonstrated that remote nursing will be central to future health care delivery, emphasizing the urgent need to prepare nursing students for this emerging health care paradigm.

(JMIR Med Educ 2025;11:e67804) doi:10.2196/67804

KEYWORDS

simulation-based training program; telenursing; simulation; program; training; nursing student; nursing care; Israel; nurse-patient relationship; telehealth nursing; remote nursing care; undergraduate; cohort study; knowlege; self efficacy; skills; attitudes

Introduction

Background

In recent decades, telehealth—the use of information and communication technology in health care—has become a global priority [1]. Telenursing, a subset of telehealth defined as "the use of technology to deliver nursing care and conduct nursing practice" [2], has emerged as a significant health care option. Indeed, telenursing enables patients to access health care providers remotely through various technologies, including mobile devices, computers, and videoconferencing [3]. The American Nurses Association defines telenursing as the use of "technology to deliver nursing care and conduct nursing practices" [4].

Telenursing offers numerous benefits, including improved access to care, savings in time and resources, and enhanced self-care opportunities [5]. Telenursing has been found effective in reducing the number of outpatient and emergency room visits, shortening hospital stays, and lowering health care costs [6]. It has also proved beneficial in educating patients, promoting self-care competence, and providing cost-effective mental health support [7], as well as in providing care for chronically ill patients [5], oncology patients [6] and palliative care [8].

The COVID-19 pandemic accelerated the adoption of telenursing by emphasizing its crucial role in disaster and public health emergency responses [9]. This shift highlighted the need for integrating telenursing concepts into nursing education at both undergraduate and graduate levels [3,10].

Despite the growing importance of telenursing, it is often underrepresented in nursing curricula. Poreddi et al [11] found that while nurse interns generally hold positive perceptions of telenursing, their knowledge of the subject is limited. This gap underscores the need to incorporate telenursing concepts into nursing education in order to prepare future health care providers for an increasingly digital environment. Nurses play an indispensable role in telehealth implementation, with their skills and attitudes serving as supportive factors [12].

Nevertheless, telenursing also entails challenges. Indeed, to provide optimal care without physical contact nurses must possess high-level clinical and interpersonal skills [13,14]. A lack of sufficient knowledge and skills constitutes the main obstacle in telenursing [15]. Previous studies have reported that telehealth education in nursing programs is inadequate [16]. To acquire the skills and develop the competencies required for telenursing, students must practice the use of screen technology and virtual access to remote patients, and telenursing should be introduced early in the nursing curriculum [17].

Simulation-based learning has been identified as an effective method for teaching telenursing skills. This approach allows students to practice in a safe and realistic environment in which they can improve their cognitive, emotional, and psychomotor abilities [18,19]. Studies have shown that simulations positively affect self-efficacy, academic motivation, and the acquisition of clinical skills [20,21]. Moreover, experiential education can be used to augment such crucial factors as perceived usefulness, self-efficacy, and innovativeness, thus enhancing our understanding of the effectiveness and implications of telenursing [22].

Glinkowski et al [23] examined telehealth among college nursing students. They found that 67% (207/308) of participants were willing to engage in telehealth services and 69.49% (214/308) agreed that telehealth should be included as part of the nursing education curriculum. Indeed, enhancing understanding of telenursing and establishing a robust human infrastructure among future nursing professionals are of critical importance. Receptiveness and adaptability have emerged as crucial factors in shaping the quality of health care services provided through telenursing [24]. Studies that examined nursing and medical students in the United States and Poland revealed positive perspectives and attitudes toward telehealth and telenursing. Yet, these studies also revealed knowledge gaps among these students, as well as erroneous beliefs regarding the advantages and possibilities of using telenursing in health practice [25-28]. Adequate education is needed to overcome this lack of knowledge and improve students' attitudes, including more telenursing content [29].

According to Assaye et al [30], the most significant factors influencing perceptions of telenursing among health care providers include technology availability, web access, and lack of telemedicine training. Indeed, nurses with insufficient education and training in the use of technology face difficulties in implementing telenursing [31]. To overcome these difficulties, students must develop a positive and unprejudiced attitude, while acquiring comprehensive knowledge and acknowledging the limitations of these technologies [32].

Despite the importance of integrating telenursing in nursing study programs, in Israel this topic has not been incorporated into the nursing director's core program and is not taught in practical nursing training programs. Although studies have been conducted on the use of telenursing, very few have examined the issue of training nursing students to use it. Implementing and evaluating such a training program have the potential to help integrate telenursing into the nursing director's core program.

In conclusion, despite the increasing importance of telenursing in health care delivery, particularly in view of recent global events, its integration into nursing education remains limited, particularly in certain countries, such as Israel. Further research, educational initiatives, and pilot training programs are needed to bridge this gap and ensure that future nurses are adequately prepared for the evolving landscape of health care delivery.

This study evaluates the skills and knowledge of third-year nursing students regarding remote nursing care before and after participation in a simulation-based training program on telenursing as part of their undergraduate nursing degree. To the best of our knowledge, very few studies to date have evaluated programs that use simulation to train nursing students in the provision of nursing care from a distance (telenursing).

Hypotheses

This study tests 3 hypotheses. The first hypothesis posits that participation in telenursing training will lead to increased awareness, knowledge, and understanding of required skills in telenursing practice. We expect to observe improved attitudes toward telenursing and enhanced self-efficacy, while simultaneously seeing a reduction in perceived barriers to telenursing implementation. The second hypothesis suggests that there will be significant positive correlations between nursing students' self-efficacy in telenursing and their awareness, knowledge, skill perceptions, and attitudes toward telenursing. Conversely, we anticipate that barriers to telenursing will demonstrate a negative correlation with self-efficacy levels. The third hypothesis proposes that nursing students' awareness, knowledge, skill perceptions, and attitudes toward telenursing will serve as significant predictors of both initial self-efficacy levels prior to training and the magnitude of change in self-efficacy following the intervention.

Methods

Participants

Participants in this study included 114 nursing students in their third year of studies. Most participants were female (101/114, 88.6%), between the ages of 20 and 50 years (mean 25.68, SD 4.59 years), and studying for a first undergraduate degree (107/114, 93.9%) (rather than attempting a career change).

Instruments

The questionnaire included 6 sections:

1. The Awareness of Telenursing questionnaire included 6 items, scored on a scale of 1-3 (1=know about it, 2=heard about it, and 3=know nothing about it). Sample item: "Telenursing is the most advanced service provided in nursing." Internal consistency was acceptable in the pretest (α=.70) but low in the posttest (α=.48). A higher mean score reflects higher awareness of telenursing.
2. The Knowledge of Telenursing questionnaire included 10 dichotomous items, scored 0 or 1. Sample item: "Epidemiological patient surveys can be conducted via telenursing." Low to acceptable internal consistencies were found in the pretest (α=.75) and the posttest (α=.54). A higher summary score reflects a higher level of knowledge of telenursing.
3. The Skills Required for Telenursing questionnaire included 10 items, scored on a scale of 1-5, with 1 indicating a low level and 5 indicating a very high level. Respondents were asked to rate the extent to which they felt that nurses need specific skills for using telenursing. Sample item: "High listening skills and high question asking skills are required for telenursing." High internal consistencies were found in the pretest (α=.91) and in the posttest (α=.90). A higher mean score reflects a higher level of skills required for telenursing.
4. The Attitudes About Telenursing questionnaire included 13 items, scored on a scale of 1-5, with 1 indicating that the respondent does not agree at all and 5 indicating a very high level of agreement. Sample item: "I believe that telenursing facilitates the provision of equitable service to all patients." High internal consistencies were found in the pretest (α=.89) and the posttest (α=.89). A higher mean score reflects more positive attitudes about telenursing.
5. The Barriers to Telenursing questionnaire included 9 items, scored on a scale of 1-5, with 1 indicating that the respondent does not agree at all and 5 indicating a very high level of agreement. Sample item: "I did not invest so many years studying just to work in front of a computer. I will miss personal contact with patients and meeting them face to face." Good internal consistencies were found in the pretest (α=.79) and the posttest (α=.82). A higher mean score reflects a higher level of barriers to using telenursing.
6. The Self-Efficacy in Telenursing questionnaire included 10 items, scored on a scale of 1-5, with 1 indicating that the respondent does not feel certain and 5 indicating that the respondent feels very certain in being able to help the patient follow instructions given over the telephone and understand complex cases presented in that manner. High internal consistencies were found in the pretest (α=.91) and the posttest (α=.95). A higher mean score reflects higher self-efficacy.

Procedure

This study is a cohort intervention study conducted among all third-year nursing students in the college, which examined level of knowledge, skills and attitudes regarding self-efficacy and clinical skills for telenursing, and willingness to use this method at 2 points in time—prior to and following training.

The training took place in two stages: (1) the students were taught by the course lecturer, who works in this field. Topics of study included diverse nursing practices, ethical aspects of telehealth, clinical skills including communications skills, challenges in telenursing, and tools for coping with complex issues arising from telenursing. (2) Students practiced telenursing through simulations in various nursing areas. During the simulations they practiced treating patients using the telenursing tools and communications skills they had learned and conducted virtual patient assessments and physical examinations.

The participating students answered a questionnaire assessing the research variables prior to and following training.

Ethical Considerations

The study was approved by the College Ethics Committee, Emek Yezreel Academic College (approval no. YVC EMEK-2023‐87). Students were recruited via the researcher's research assistant, who asked for volunteers. Before completing the questionnaires, participants were told that participation was voluntary and that they could drop out of the study at any time. They were informed that their opinions were important for constructing the departmental training program and were therefore encouraged to express them. Participants signed informed consent forms prior to participation. All students in the cohort agreed to participate in the study with no compensation provided. The participants' privacy and identity were protected, and confidentiality was assured in that no identifying information was asked. The study objectives were explained to the participants and the study was conducted according to the academic ethical code.

Data Analysis

The data were analyzed with SPSS (version 29; IBM Corp). Descriptive statistics were used for the participants' demographic characteristics and study variables. As the variable of skills required for telenursing (pre and post) was negatively skewed (preskewness −1.74, SE 0.23; postskewness −3.20, SE 0.23), it underwent exponential transformation. Time differences were assessed with 2-tailed paired *t* tests, using Cohen *d* for effect sizes. Change scores were computed as residual gain scores between the pre- and posttests, and Pearson correlations were calculated between the study variables regarding the pretest scores and the change scores. Multiple linear regressions were calculated for self-efficacy in telenursing, using pretest scores and change scores. Awareness, knowledge, skills, attitudes, and barriers to telenursing were defined as predictors.

Results

Descriptive Results

Most participants have initially reported a moderate (54/114, 47.4%) or advanced (58/114, 50.9%) level of knowledge in using computers and the web. Most have not been exposed to telenursing (104/114, 91.2%) but were interested in training in it (86/114, 75.4%). Participants' age was generally not

associated with the study variables (*P*=.07 to *P*=.94) and was thus not controlled for. Other demographic variables had low variance. Thus, the first hypothesis was assessed with a series of 2-tailed paired *t* tests. Significant changes were noted in all variables with moderate to high effect sizes (Table 1). Awareness of telenursing, knowledge of telenursing, skills, attitudes, and self-efficacy in telenursing have all significantly increased following participation in the training program, and barriers to telenursing have significantly decreased.

Table . Means, SDs, *t* values, and Cohen *d* values for the study variables by time (N=114)[a].

	Pretest, mean (SD)	Posttest, mean (SD)	t_{113} (*P* value)	Cohen *d*
Awareness of telenursing	1.75 (0.48)	2.54 (0.36)	15.18 (<.001)	1.85
Knowledge of telenursing	6.84 (2.48)	8.42 (1.55)	6.18 (<.001)	0.76
Skills required for telenursing	4.60 (0.55)	4.77 (0.38)	3.80 (<.001)	0.36
Attitudes about telenursing	3.41 (0.62)	4.06 (0.62)	9.78 (<.001)	1.04
Barriers to telenursing	2.94 (0.61)	2.63 (0.74)	−4.52 (<.001)	0.45
Self-efficacy in telenursing	3.69 (0.79)	4.04 (0.73)	4.88 (<.001)	0.46

[a]Ranges: awareness of telenursing 1-3; knowledge of telenursing 0-10; and skills, attitudes, barriers for telenursing, and self-efficacy in telenursing 1-5.

Pearson correlations were calculated among the study variables regarding the pretest and change scores. Significant associations were found (Table 2). In the pretest, higher awareness of telenursing, higher knowledge of telenursing, and perception of the higher skills required for telenursing were associated with more positive attitudes and higher self-efficacy regarding telenursing. Furthermore, more positive attitudes about telenursing were associated with higher self-efficacy in telenursing and with lower barriers to it.

Table . Pearson correlations between the study variables for the pretest scores and the change scores (N=114).

	1	2	3	4	5	6
Pretest						
1. Awareness of telenursing						
r	1	0.13	0.12	0.19	–0.13	0.20
P value	—[a]	.18	.19	.04	.16	.04
2. Knowledge of telenursing						
r	0.13	1	0.15	0.38	–0.06	0.23
P value	.18	—	.11	<.001	.53	.02
3. Skills required for telenursing						
r	0.12	0.15	1	0.22	0.03	0.19
P value	.19	.11	—	.02	.75	.04
4. Attitudes about telenursing						
r	0.19	0.38	0.22	1	–0.27	0.20
P value	.04	<.001	.02	—	.004	.04
5. Barriers to telenursing						
r	–0.13	–0.06	0.03	–0.27	1	0.06
P value	.16	.53	.75	.004	—	.51
6. Self efficacy in telenursing						
r	0.20	0.23	0.19	0.20	0.06	1
P value	.04	.02	.04	.04	.51	—
Change scores						
1. Awareness of telenursing						
r	1	0.25	–0.11	0.23	–0.22	0.14
P value	—	.006	.26	.01	.02	.14
2. Knowledge of telenursing						
r	0.25	1	0.01	0.23	–0.08	0.02
P value	.006	—	.95	.01	.42	.85
3. Skills required for telenursing						
r	–0.11	0.01	1	0.12	0.01	0.21
P value	.26	.95	—	.20	.93	.02
4. Attitudes about telenursing						
r	0.23	0.23	0.12	1	–0.38	0.37
P value	.01	.01	.20	—	<.001	<.001
5. Barriers to telenursing						
r	–0.22	–0.08	0.01	–0.38	1	–0.23
P value	.02	.42	.93	<.001	—	.02

	1	2	3	4	5	6
6. Self efficacy in telenursing						
r	0.14	0.02	0.21	0.37	–0.23	1
P value	.14	.85	.02	<.001	.02	—

[a]Not applicable.

Regarding the change scores, higher awareness of telenursing was associated with higher knowledge of telenursing and both were associated with more positive attitudes regarding it. Furthermore, higher awareness of telenursing and more positive attitudes regarding it were associated with lower barriers to telenursing. Finally, the higher skills required for telenursing, more positive attitudes about it, and lower barriers were associated with higher self-efficacy in telenursing.

Associations With and Change in Self-Efficacy

Two multiple linear regressions were calculated to evaluate the associations between awareness, knowledge, skills, attitudes and barriers to telenursing, and self-efficacy in telenursing regarding the pretest and the change between the pretest and the posttest. Level of knowledge in using the computer and the web (0: moderate and low, 1: advanced) was entered first, and the study variables or change in the study variables was entered second (Table 3).

Table . Multiple linear regressions for self-efficacy in telenursing with awareness, knowledge, skills, and attitudes and barriers to telenursing (N=114).

	Pretest scores[a]			Change scores[b]		
	B (SE)	β	*P* value	*B* (SE)	β	*P* value
Level of computer knowledge (advanced)	0.36 (0.14)	.23	.01[c]	−0.10 (0.18)	−.05	.58
Awareness of telenursing	0.32 (0.14)	.19	.03[c]	0.03 (0.10)	.03	.73
Knowledge of telenursing	0.07 (0.03)	.22	.02[c]	−0.10 (0.09)	−.10	.27
Skills required for telenursing	0.01 (0.01)	.06	.54	0.19 (0.09)	.19	.04[c]
Attitudes about telenursing	0.21 (0.12)	.17	.09	0.36 (0.10)	.36	<.001[c]
Barriers to telenursing	0.07 (0.11)	.06	.53	−0.09 (0.10)	−.09	.34

[a]R^2=0.23, P<.001; $F_{6, 107}$=5.20, P<.001.

[b]R^2=0.21, P<.001; $F_{6, 107}$=4.64, P<.001.

[c]These values are significant.

Both regression models were found significant, with 23% and 21% of the variance in the pretest score and in the change score, respectively, being explained by them. Regarding the pretest score, higher awareness and more knowledge of telenursing were associated with the perception of higher self-efficacy in telenursing. Regarding the change score, greater improvement in the perceived skills required for telenursing and a higher positive change in the attitudes regarding telenursing were associated with a greater improvement in the perception of self-efficacy in telenursing.

Discussion

This study aimed to examine how a simulation training program on telenursing affected awareness, knowledge, skills, attitudes, self-efficacy, and perceived barriers regarding telenursing among third-year nursing students. The results demonstrate significant improvements across all measured variables, with moderate to high effect sizes, suggesting that the implemented training program was effective. Moreover, the higher skills required for telenursing, more positive attitudes regarding it, and lower barriers were associated with higher self-efficacy in telenursing. These findings emphasize that the simulated experiences served as effective interventions, providing students with innovative learning opportunities [33].

The substantial increase in participants' awareness and knowledge of telenursing reflects the growing recognition that it is a critical component of modern health care delivery [1,3]. This increase is particularly noteworthy, given that most participants (104/114, 91.2%) had no prior exposure to telenursing, despite their initial moderate to advanced levels of computer and web proficiency. Findings regarding the posttest score of awareness of telenursing and its change score may be biased in unknown ways and should be regarded with caution.

Vaidya [34] recently emphasized the need to offer simulation telehealth education to undergraduate, graduate and health care practitioners in an effort to achieve a more effective remote diagnosis and treatment management for patients in need, such as those living with chronic disease.

The findings of this study are also in line with those of Mun et al [35], which indicated that nursing students lacked substantial awareness regarding telenursing. Nevertheless, the results also portrayed a positive outlook. Indeed, according to Kazawa et al [36], engaging in telenursing helps students enhance their understanding of telehealth practices, develop critical thinking skills, and broaden their knowledge of how to manage and address patient needs in a virtual care setting. Chang et al [37] also found that nurses with telehealth experience have significantly higher perceptions of its usefulness than those with no such experience, and these perceptions correlated positively with attitudes and behavioral intentions. These findings imply that providing nursing students with telenursing education can help them understand and harness this method [28]. Moreover, telenursing education was shown to have a significant impact on nurses' knowledge, attitudes, and awareness of future work [8,11]. Telehealth simulation was shown to improve nursing students' professional skills [38].

This study goes a step further by demonstrating that targeted training can significantly improve attitudes. This finding is crucial, as positive attitudes are likely to translate into greater willingness to engage with and implement telenursing practices in future professional roles. Nurses with prior telehealth knowledge had more positive attitudes toward telenursing than those who had never encountered telehealth-related information, and their attitudes toward telenursing correlated positively with their intentions to engage in telehealth [37]. Moreover, nursing students' attitudes toward telenursing demonstrated a significant correlation with telenursing experience, observation of telenursing during clinical practice, and exposure to telenursing education [35].

The decrease in perceived barriers to telenursing following the training program is a particularly encouraging outcome of this study. It suggests that the program has effectively addressed common concerns and obstacles associated with telenursing implementation and has the potential to offer a smoother integration of these practices in future health care settings.

Indeed, to prepare for their future roles, nursing students need telenursing education [35]. Previous studies examining nursing students reported positive prospects for telenursing, alongside negative perceptions associated with a lack of awareness [28,39]. Furthermore, Mun et al [35] found that nursing students who had a negative outlook regarding telenursing noted its impracticality as compared with face-to-face nursing, as well as the lack of patient contact, challenges faced by older individuals, and accessibility issues for low-income or rural residents. The assumption is that these limited perceptions derive from a lack of formal education in telenursing [40]. This type of education has the potential to enhance knowledge and attitudes regarding telenursing [41]. Indeed, significant improvements in understanding the use and role of telenursing were found among individuals who had undergone telehealth education [42].

This study found that skills and self-efficacy improved following the intervention. These findings are in line with a previous study that found a significant enhancement in skills and self-efficacy following training [37], thus indicating that telenursing education plays a crucial role in improving the type of specialized knowledge required for clinical telenursing. These findings suggest that nursing students need formal education in telenursing. Such education will enhance their competency and nurture a positive attitude, facilitating the seamless integration of telenursing into the digital health care era [35]. Our findings also corroborate those of Reierson et al [17], who emphasized the importance of introducing and practicing telenursing at the beginning of nursing curricula.

Moreover, our study found differences in the predictors of self-efficacy. Prior to the intervention, self-efficacy was a function of awareness and knowledge, whereas following it self-efficacy correlated with a change in skills and attitudes. To the best of our knowledge, almost no intervention studies have been conducted on the role of education in promoting telenursing. Kazawa et al [36] found that telenursing education is essential in expanding nursing students' knowledge and skills. Moreover, Mun et al [35] found that self-efficacy regarding telenursing among nursing students was associated with telenursing experience, education, and attitudes toward telenursing. Knowledge regarding advance care planning was also found to be associated with self-efficacy [43,44]. Bandura [45,46] also found a relationship between knowledge and skills, which translates into action by increasing self-efficacy to overcome barriers, and Mata et al [47] found that skills can improve health professionals' performance and self-efficacy.

Former studies of simulation-based instruction in nursing and telehealth were done by Parmeter et al [48], using a posttest-only design. Following peer-to-peer telehealth simulation scenarios via Zoom (Zoom Video Communications, Inc), the students demonstrated a high score of confidence and telehealth performance. These results align with the findings of this study.

These findings have significant implications for nursing education and practice. First, they strongly support the integration of telenursing into nursing curricula, as advocated by Asimakopoulou [3] and Puro and Feyereisen [10]. Moreover, the success of the simulation-based training program in this study is in line with previous research highlighting the effectiveness of simulations in nursing education [20,49]. Simulation can also play an important role in helping students acquire and improve their self-efficacy and nursing skills [50]. Our findings suggest that similar approaches may be valuable in preparing nursing students for the growing prevalence of telenursing in health care delivery.

This study presents 3 key limitations. First, the absence of a control group limits the ability to conclusively attribute observed changes solely to the training program. Future research should incorporate a parallel control group design to enable a more rigorous comparative analysis and establish clearer causal relationships. Second, the focus on immediate posttraining outcomes prevents understanding of long-term telenursing

competency retention and clinical application. Implementing a longitudinal study design with multiple follow-up assessments at 3, 6, and 12 months posttraining would provide insights into knowledge and skill sustainability. Third, the study's recruitment from a single college limits the generalizability of the findings across nursing student populations in Israel. Expanding the research to multiple educational institutions, potentially including diverse geographic and institutional contexts, would enhance the external validity of the results. Cronbach α value for awareness of telenursing was acceptable at pretest (α=0.70) but low at posttest (α=0.48). This finding indicates that the posttest score of awareness of telenursing has a low reliability, and the relevant findings should be regarded with caution. That is, the findings regarding the posttest score of awareness of telenursing and its change score may be biased in unknown ways and should be regarded with caution. Future studies are advised to validate a modified version of this questionnaire or use a different one.

Future research should address these methodological limitations by integrating control groups, longitudinal designs, and multi-institutional sampling to comprehensively evaluate telenursing training programs and their broader implementation potential. In addition, research exploring the implications of telenursing training on patient outcomes and health care system efficiency would provide critical empirical evidence to support broader program implementation.

In conclusion, this study provides compelling evidence for the effectiveness of a simulation-based telenursing training program in enhancing nursing students' competencies across multiple domains of telenursing. The findings underscore the importance of integrating telenursing education into nursing curricula in an effort to prepare future health care providers for the evolving landscape of health care delivery.

To implement these findings effectively, health care organizations should provide hands-on telenursing training through structured workshops and regular skill refresher seminars for practicing nurses. The Ministry of Health should establish standardized telenursing protocols and mandate their adoption across all health maintenance organizations to ensure consistent quality of care. Medical schools should integrate practical telenursing simulations into their core curriculum, while practicing health care professionals should complete required continuing education modules which include hands-on simulation training to maintain their competency in virtual care delivery.

Acknowledgments

The authors thank the students for their participation in the course and their responses to the questionnaire before and after it.

Data Availability

All data generated or analyzed during this study are included in this published article.

Authors' Contributions

OAS, LM, and OH contributed to conceptualization, methodology, and formal analysis and investigation; OAS and LM contributed to writing—original draft preparation; OH contributed to writing—review and editing; and OAS and OH contributed to resources and supervision.

Conflicts of Interest

None declared.

References

1. Global diffusion of eHealth: making universal health coverage achievable: report of the third global survey on eHealth. World Health Organization. 2016. URL: https://www.who.int/publications/i/item/9789241511780 [accessed 2025-02-17]
2. Schlachta L, Sparks S. Definitions of telenursing, telemedicine. In: Fitzpatrick J, editor. Encyclopedia of Nursing Research: Springer Publishing Inc; 1998:558-559.
3. Asimakopoulou E. Telenursing in clinical practise and education [Editorial]. Int J Caring Sci 2020;13(2):781-782 [FREE Full text]
4. Developing Telehealth Protocols: A Blueprint for Success: American Nurses Association; 2001.
5. Souza-Junior VD, Mendes IAC, Mazzo A, Godoy S. Application of telenursing in nursing practice: an integrative literature review. Appl Nurs Res 2016 Feb;29:254-260. [doi: 10.1016/j.apnr.2015.05.005]
6. Kamei T. Telenursing and artificial intelligence for oncology nursing. Asia Pac J Oncol Nurs 2022 Dec;9(12):100119. [doi: 10.1016/j.apjon.2022.100119] [Medline: 36276880]
7. Ghoulami-Shilsari F, Esmaeilpour Bandboni M. Tele-nursing in chronic disease care: a systematic review. Jundishapur J Chronic Dis Care 2019;8(2):1-8. [doi: 10.5812/jjcdc.84379]
8. Walton L, Courtright K, Demiris G, Gorman EF, Jackson A, Carpenter JG. Telehealth palliative care in nursing homes: a scoping review. J Am Med Dir Assoc 2023 Mar;24(3):356-367. [doi: 10.1016/j.jamda.2023.01.004] [Medline: 36758619]

9. Alipour J, Hayavi-Haghighi MH. Opportunities and challenges of telehealth in disease management during COVID-19 pandemic: a scoping review. Appl Clin Inform 2021 Aug;12(4):864-876. [doi: 10.1055/s-0041-1735181] [Medline: 34528234]
10. Puro NA, Feyereisen S. Telehealth availability in US hospitals in the face of the COVID-19 pandemic. J Rural Health 2020 Sep;36(4):577-583. [doi: 10.1111/jrh.12482] [Medline: 32603017]
11. Poreddi V, Bidadi Veerabhadraiah K, Reddy S, Narayana M, Channaveerachari N, BadaMath S. Nursing interns' perceptions of telenursing: implications for nursing education. Telehealth Med Today 2021;6(2). [doi: 10.30953/tmt.v6.258]
12. Koivunen M, Saranto K. Nursing professionals' experiences of the facilitators and barriers to the use of telehealth applications: a systematic review of qualitative studies. Scand J Caring Sci 2018 Mar;32(1):24-44. [doi: 10.1111/scs.12445] [Medline: 28771752]
13. Jácome M, Rego N, Veiga P. Potential of a nurse telephone triage line to direct elderly to appropriate health care settings. J Nurs Manag 2019 Sep;27(6):1275-1284. [doi: 10.1111/jonm.12809]
14. Sexton V, Dale J, Bryce C, Barry J, Sellers E, Atherton H. Service use, clinical outcomes and user experience associated with urgent care services that use telephone-based digital triage: a systematic review. BMJ Open 2022 Jan 3;12(1):e051569. [doi: 10.1136/bmjopen-2021-051569] [Medline: 34980613]
15. Edirippulige S, Smith AC, Young J, Wootton R. Knowledge, perceptions and expectations of nurses in e-health: results of a survey in a children's hospital. J Telemed Telecare 2006 Nov;12(3_suppl):35-38. [doi: 10.1258/135763306779380255]
16. Ali NS, Carlton KH, Ali OS. Telehealth education in nursing curricula. Nurse Educ 2015;40(5):266-269. [doi: 10.1097/NNE.0000000000000149] [Medline: 25689080]
17. Reierson I, Solli H, Bjørk IT. Nursing students' perspectives on telenursing in patient care after simulation. Clin Simul Nurs 2015 Apr;11(4):244-250. [doi: 10.1016/j.ecns.2015.02.003]
18. Lee J, Oh PJ. Effects of the use of high-fidelity human simulation in nursing education: a meta-analysis. J Nurs Educ 2015 Sep;54(9):501-507. [doi: 10.3928/01484834-20150814-04] [Medline: 26334336]
19. Oh PJ, Jeon KD, Koh MS. The effects of simulation-based learning using standardized patients in nursing students: a meta-analysis. Nurse Educ Today 2015 May;35(5):e6-e15. [doi: 10.1016/j.nedt.2015.01.019] [Medline: 25680831]
20. Filomeno L, Minciullo A. High-fidelity simulation for the education of nursing students: a scoping review of the literature. Prof Inferm 2021;74(2):180-189. [doi: 10.7429/pi.2021.7423180] [Medline: 35084162]
21. Lestander Ö, Lehto N, Engström Å. Nursing students' perceptions of learning after high fidelity simulation: effects of a three-step post-simulation reflection model. Nurse Educ Today 2016 May;40:219-224. [doi: 10.1016/j.nedt.2016.03.011] [Medline: 27125176]
22. van Houwelingen CTM, Moerman AH, Ettema RGA, Kort HSM, Ten Cate O. Competencies required for nursing telehealth activities: a Delphi-study. Nurse Educ Today 2016 Apr;39:50-62. [doi: 10.1016/j.nedt.2015.12.025] [Medline: 27006033]
23. Glinkowski W, Pawłowska K, Kozłowska L. Telehealth and telenursing perception and knowledge among university students of nursing in Poland. Telemedicine and E-Health 2013 Jul;19(7):523-529. [doi: 10.1089/tmj.2012.0217]
24. Honey M, Collins E, Britnell S. Education into policy: embedding health informatics to prepare future nurses—New Zealand case study. JMIR Nurs 2020;3(1):e16186. [doi: 10.2196/16186] [Medline: 34345779]
25. Malhotra P, Ramachandran A, Chauhan R, Soni D, Garg N. Assessment of knowledge, perception, and willingness of using telemedicine among medical and allied healthcare students studying in private institutions. Telehealth Med Today 2020;5(4). [doi: 10.30953/tmt.v5.228]
26. Abraham C, Jensen C, Rossiter L, Dittman Hale D. Telenursing and remote patient monitoring in cardiovascular health. Telemed J E Health 2024 Mar;30(3):771-779. [doi: 10.1089/tmj.2023.0187] [Medline: 37682280]
27. El- Said Abd Ellatif A, Mohamed Sobhy Elsayed D, Hamido Abosree T. Knowledge and attitude of faculty of nursing students regarding telenursing. J Nurs Sci 2023 Jan 1;4(1):677-689. [doi: 10.21608/jnsbu.2023.278954]
28. Khraisat OMA, Al-Bashaireh AM, Alnazly E. Telenursing implications for future education and practice: nursing students' perspectives and knowledge from a course on child health. PLoS One 2023;18(11):e0294711. [doi: 10.1371/journal.pone.0294711] [Medline: 38011137]
29. Kurtović B, Hošnjak AM, Ledinski S, Smrekar M, Babić J, Čukljek S. Nursing students' knowledge and attitudes towards telenursing. Croat Nurs J (Online) 2024;8(1):5-16. [doi: 10.24141/2/8/1/1]
30. Assaye BT, Belachew M, Worku A, et al. Perception towards the implementation of telemedicine during COVID-19 pandemic: a cross-sectional study. BMC Health Serv Res 2023;23(1):1186. [doi: 10.1186/s12913-023-09927-1]
31. Ftouni R, AlJardali B, Hamdanieh M, Ftouni L, Salem N. Challenges of telemedicine during the COVID-19 pandemic: a systematic review. BMC Med Inform Decis Mak 2022 Aug 3;22(1):207. [doi: 10.1186/s12911-022-01952-0] [Medline: 35922817]
32. Ghaddaripouri K, Mousavi Baigi SF, Abbaszadeh A, Mazaheri Habibi MR. Attitude, awareness, and knowledge of telemedicine among medical students: a systematic review of cross-sectional studies. Health Sci Rep 2023 Mar;6(3):e1156. [doi: 10.1002/hsr2.1156] [Medline: 36992712]
33. Dziurka M, Machul M, Ozdoba P, et al. Clinical training during the COVID-19 pandemic: experiences of nursing students and implications for education. Int J Environ Res Public Health 2022 May 23;19(10):6352. [doi: 10.3390/ijerph19106352] [Medline: 35627889]

34. Vaidya A. Developing simulation-based telehealth training for Next-Gen nurses. TechTarget. 2024. URL: https://www.techtarget.com/virtualhealthcare/answer/Developing-Simulation-Based-Telehealth-Training-for-Next-Gen-Nurses [accessed 2025-01-23]
35. Mun M, Choi S, Woo K. Investigating perceptions and attitude toward telenursing among undergraduate nursing students for the future of nursing education: a cross-sectional study. BMC Nurs 2024 Apr 8;23(1):236. [doi: 10.1186/s12912-024-01903-2] [Medline: 38589885]
36. Kazawa K, Teramoto C, Azechi A, Satake H, Moriyama M. Undergraduate nursing students' learning experiences of a telehealth clinical practice program during the COVID-19 pandemic: a qualitative study. Nurse Educ Today 2022 Apr;111:105297. [doi: 10.1016/j.nedt.2022.105297] [Medline: 35182935]
37. Chang MY, Kuo FL, Lin TR, Li CC, Lee TY. The intention and influence factors of nurses' participation in telenursing. Informatics (MDPI) 2021;8(2):35. [doi: 10.3390/informatics8020035]
38. Moore J, Jairath N, Montejo L, O'Brien S, Want D. Using a telehealth simulation to prepare nursing students for intraprofessional collaboration. Clin Simul Nurs 2023 May;78:1-6. [doi: 10.1016/j.ecns.2023.02.007]
39. Emikpe BO, Asare DA, Emikpe AO, Folitse RD, Botchway LN. Knowledge and perception of veterinary students in Ghana on telemedicine. Niger J Physiol Sci 2021 Jun 30;36(1):115-121. [Medline: 34987249]
40. Chike-Harris KE, Durham C, Logan A, Smith G, DuBose-Morris R. Integration of telehealth education into the health care provider curriculum: a review. Telemed J E Health 2021 Feb 1;27(2):137-149. [doi: 10.1089/tmj.2019.0261]
41. Pit SW, Velovski S, Cockrell K, Bailey J. A qualitative exploration of medical students' placement experiences with telehealth during COVID-19 and recommendations to prepare our future medical workforce. BMC Med Educ 2021 Aug 16;21(1):431. [doi: 10.1186/s12909-021-02719-3] [Medline: 34399758]
42. Wong CJ, Nath JB, Pincavage AT, et al. Telehealth attitudes, training, and preparedness among first-year internal medicine residents in the COVID-19 era. Telemed J E Health 2022 Feb 1;28(2):240-247. [doi: 10.1089/tmj.2021.0005]
43. ten Koppel M, Onwuteaka-Philipsen BD, van der Steen JT, et al. Care staff's self-efficacy regarding end-of-life communication in the long-term care setting: results of the PACE cross-sectional study in six European countries. Int J Nurs Stud 2019 Apr;92:135-143. [doi: 10.1016/j.ijnurstu.2018.09.019]
44. Evenblij K, ten Koppel M, Smets T, Widdershoven GAM, Onwuteaka-Philipsen BD, Pasman HRW. Are care staff equipped for end-of-life communication? A cross-sectional study in long-term care facilities to identify determinants of self-efficacy. BMC Palliat Care 2019 Dec;18(1):1-11. [doi: 10.1186/s12904-018-0388-z]
45. Bandura A. Self-Efficacy: The Exercise of Control: W H Freeman; 1997.
46. Bandura A. Social Foundations of Thought and Action: Prentice Hall; 1986.
47. Mata ÁDS, de Azevedo KPM, Braga LP, et al. Training in communication skills for self-efficacy of health professionals: a systematic review. Hum Resour Health 2021 Mar 6;19(1):30. [doi: 10.1186/s12960-021-00574-3] [Medline: 33676515]
48. Parmeter S, Foronda C, Lee J. Improving telenursing skills through simulation-based education. J Dr Nurs Pract 2023 Jun 27;2:93-101. [doi: 10.1891/JDNP-2022-0021] [Medline: 37369454]
49. Weissblueth E, Linder I. The effects of simulations in a simulation center on principals' training and professional self-efficacy. Int J Educ Policy Leadersh 2020;16(14):1-13. [doi: 10.22230/ijepl.2020v16n14a965]
50. Hayden JK, Smiley RA, Alexander M, Kardong-Edgren S, Jeffries PR. The NCSBN national simulation study: a longitudinal, randomized, controlled study replacing clinical hours with simulation in prelicensure nursing education. J Nurs Regul 2014 Jul;5(2):S3-S40. [doi: 10.1016/S2155-8256(15)30062-4]

Edited by J Gentges; submitted 21.10.24; peer-reviewed by LH Schroeder, MA Zoubi; revised version received 29.01.25; accepted 30.01.25; published 12.03.25.

Please cite as:
Ali-Saleh O, Massalha L, Halperin O
Evaluation of a Simulation Program for Providing Telenursing Training to Nursing Students: Cohort Study
JMIR Med Educ 2025;11:e67804
URL: https://mededu.jmir.org/2025/1/e67804
doi:10.2196/67804

Global Trends in Cadaver Donation and Medical Education Research: Bibliometric Analysis Based on VOSviewer and CiteSpace

Xianxian Zhou, MM; Hua Xiong, MM; Yi Wen, BMed; Fang Li, BMed; Dexi Hu, MM

Yiyang Central Hospital, No. 118, Kangfu North Road, Heshan District, Yiyang City, China

Corresponding Author:
Dexi Hu, MM
Yiyang Central Hospital, No. 118, Kangfu North Road, Heshan District, Yiyang City, China

Abstract

Background: The cadaver serves as a crucial resource in medical education, research, and clinical practice, as well as a vital foundation for fundamental medical experimental teaching.

Objective: This study aims to use bibliometric analysis to create a knowledge map of cadaver donation in medical education, identify global trends, anticipate future research directions, and offer a foundation for upcoming investigations.

Methods: Articles and review papers concerning cadaver donation and medical education, with a final search cutoff of January 10, 2025, were systematically retrieved from the Web of Science Core Collection database. Two reviewers carefully examined the initial set of articles based on titles and abstracts to exclude irrelevant ones. A quadratic regression model was used to examine the annual publication data. The model's goodness of fit was assessed using the R^2 value, and the statistical significance of the findings was determined through the *P* valu. The selected publications were then analyzed and visualized for country, institution, author, reference, journal, and keywords using CiteSpace 6.3R3, VOSviewer 1.6.19, and the Online Analysis Platform of the Literature Metrology Database.

Results: The quadratic regression model yielded the equation $Y=0.1586X^2-633.9X+633395$, indicating a substantial increase in the number of publications over time ($R^2=0.9575$, $P<.05$). The model forecasts that the publication count will reach 107 by 202. This upward trend is statistically significant, highlighting a notable rise in research interest and activity within this field over time. The United States was a major contributor, accounting for 21.2% (303/1114) of all publications. In terms of continents and faiths, Europe and Christianity contributed the most, while McGill University and The University of Sydney were the leading institutions. Prominent authors in this field included De Caro Raffaele, Macchi Veronica, Porzionato Andrea, Stecco Carla, and Dhanani Sonny. The most frequently cocited reference was "Bodies for Anatomy Education in Medical Schools: An Overview of the Sources of Cadavers Worldwide." The journal Anatomical Sciences Education published the most articles in this area and received the highest citation count. Cluster analysis of keywords revealed that "kidney transplantation," "gross anatomy education," and "brain death" were key research topics, while burst analysis of keywords identified "public perception" and "anatomical science" as emerging areas of investigation.

Conclusions: This research presents a distinctive bibliometric approach to cadaver donation within medical education, setting it apart from previous studies by delivering an extensive global overview of trends and influential contributors in this domain. The results emphasize the increasing global interest and collaborative efforts surrounding cadaver donation, while also offering fresh perspectives on emerging topics like public perception and anatomical sciences. This paper serves as an important reference for researchers, policymakers, and educators, supporting the development of future strategies to enhance cadaver donation programs and further medical education.

(JMIR Med Educ 2025;11:e71935) doi:10.2196/71935

KEYWORDS

cadaver donation; medical education; bibliometric analysis; citespace; VOSviewer; medical knowledge; medical training; medical student

Introduction

Cadaver donation plays a vital role in enhancing medical education and training by offering students valuable opportunities for hands-on learning in anatomy, surgery, and other medical fields [1,2]. As global health care systems encounter growing challenges such as organ shortages, ethical issues, and the continuous evolution of medical practices, the demand for cadaver donations has significantly increased [3,4].

Incorporating cadaveric resources into medical education improves students' learning outcomes by providing realistic, immersive experiences that boost their practical skills and understanding of human anatomy [5].

However, it is essential to recognize the differences in students' learning styles, as these variations influence their capacity to effectively process and retain information. Some students may benefit more from direct, hands-on cadaveric dissection, while others may find virtual reality (VR) or other digital tools more effective for enhancing their grasp of anatomical structures. In light of the need to accommodate diverse learning styles—especially during periods of remote learning or pandemic conditions—adopting a variety of teaching methods becomes increasingly essential. Research highlights that individual learning preferences have a significant impact on study duration and academic success, reinforcing the need for flexible and adaptive teaching strategies in medical education [6].

Although cadaver-based learning plays a crucial role, research on cadaver donation in medical education is still scattered, lacking a thorough examination of global trends in this area. Existing literature primarily focuses on regional views, ethical dilemmas, and the educational outcomes associated with cadaver-based teaching methods [7-9]. Yet, there is a lack of a broader, global perspective on the patterns, growth, and impact of cadaver donations in medical education. This research aims to fill this gap by conducting a bibliometric analysis of publications related to cadaver donation and medical education, using the advanced analytical tools VOSviewer and CiteSpace. Specifically, the study seeks to address the following research question: What are the global trends, institutional contributions, and emerging themes in the research on cadaver donation and medical education?

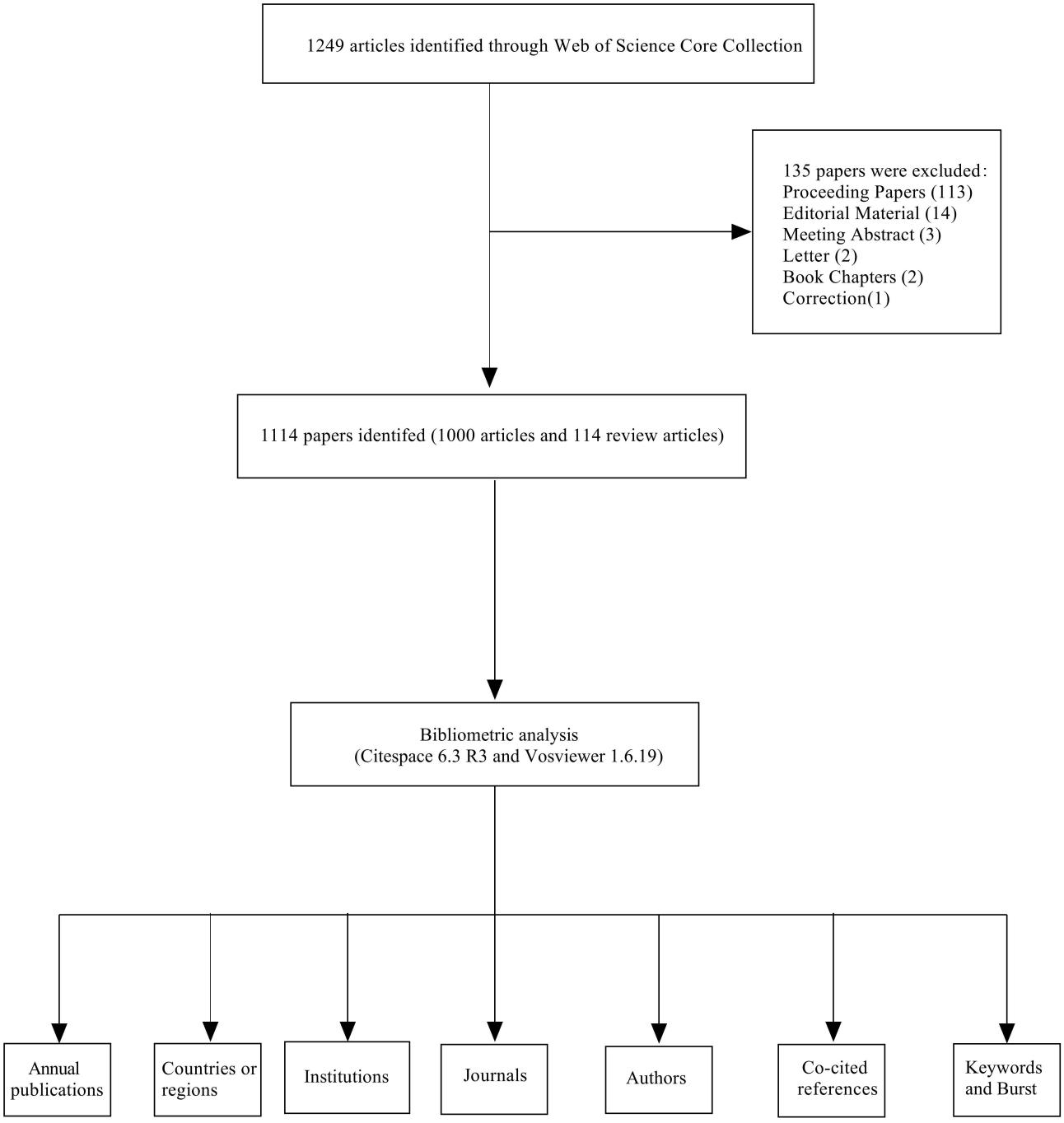
The significance of this study lies in its potential to inform future research and policy decisions in medical education. By systematically mapping the academic landscape, identifying key authors, institutions, countries, and themes, this research will offer valuable insights into the current state of the field and pinpoint areas in need of further investigation. Furthermore, understanding global trends in cadaver donation can assist educational institutions and policymakers in better aligning their strategies with best practices, ensuring the ongoing relevance and effectiveness of cadaver-based learning in the ever-changing landscape of medical education.

Methods

Data Collections

Data were collected on January 10, 2025, from the Web of Science Core Collection (WoSCC) database, which includes the Science Citation Index Expanded, Social Sciences Citation Index, Arts and Humanities Citation Index, Emerging Sources Citation Index, Current Chemical Reactions, and Index Chemicus. The search focused on publications related to the keywords "cadaver donation" and "medical education." The language was limited to English, and the article types were restricted to research articles and reviews. Each record contained details such as the title, author, keywords, abstract, year, institution, citations, and other pertinent information. A detailed explanation of the search methodology, including both inclusion and exclusion criteria, is provided in Figure 1.

Figure 1. Flowchart for the analysis of cadaver donation and medical education.



Bibliometrics and Visualization Analysis

The search results were subsequently examined using CiteSpace (Version 6.3R3, Drexel University, Chaomei Chen), VOSviewer (Version 1.6.19; Leiden University), and the Online Analysis Platform of Literature Metrology. CiteSpace, a visual analysis tool developed by Chaomei Chen, was used to analyze the total number of relevant papers, annual trends, keyword frequency, and centrality. This software provided a more accessible and intuitive way to examine the structure, patterns, and distribution of knowledge within the subject. A scientific knowledge map was used to identify research hotspots, advancements, and the current state of the field [10]. VOSviewer, a tool designed for document data analysis, facilitated the study of countries, institutions, authors, journals, keywords, and the co-occurrence knowledge graph for countries, institutions, journals, and publications. In the knowledge graph, each node represented an individual element, with the connection width between nodes reflecting collaboration strength, the node size indicating the number of publications, and larger nodes representing more frequent contributions [11]. The Online Analysis Platform of Literature Metrology was used to analyze the number of publications per country over different years and their collaborative efforts.

Quadratic Regression Model

A quadratic regression model was applied to fit the publication data. The model is represented by the equation:

$$(1) Yt=\beta 0+\beta 1t+\beta 2t2+\varepsilon t$$

where:

Y_t is the number of publications in year t.

t is the time variable (year).

β_0, β_1, and $\beta_2$are the regression coefficients.

ε_t is the error term.

This model allows for both linear and curved trends, which is useful for identifying whether the number of publications has accelerated or slowed over time.

Goodness of Fit

The goodness of fit was assessed using the R^2 value. R^2 shows how well the model explains the variation in the publication data. A higher R^2 indicates a better fit, meaning the model accurately reflects the trends in the data.

Statistical Significance

To determine the statistical significance of the results, we looked at the *P* value for each regression coefficient. The *P* value tests whether each variable (eg, time or its squared term) significantly affects the number of publications.

A *P* value less than .05 indicates that the coefficient is statistically significant, meaning the variable has a real impact on publication trends.

A *P* value greater than .05 suggests that the variable is not significantly influencing the trend.

The articles for this study were obtained in .txt format from the Web of Science database. Two expert researchers reviewed the title, keywords, and abstract of each paper, applying inclusion and exclusion criteria to screen the documents. In cases of disagreement or uncertainty about a paper's inclusion, a third reviewer made the final decision through discussion. Initially, 1249 papers were identified, and after excluding 135 papers that were irrelevant to the study's focus, 1114 papers were retained.

Ethical Considerations

This study did not involve the collection of new data from humans or animals. All the data used in the bibliometric analysis were sourced from the Web of Science Core Collection, and therefore, ethical approval and participant consent are not required.

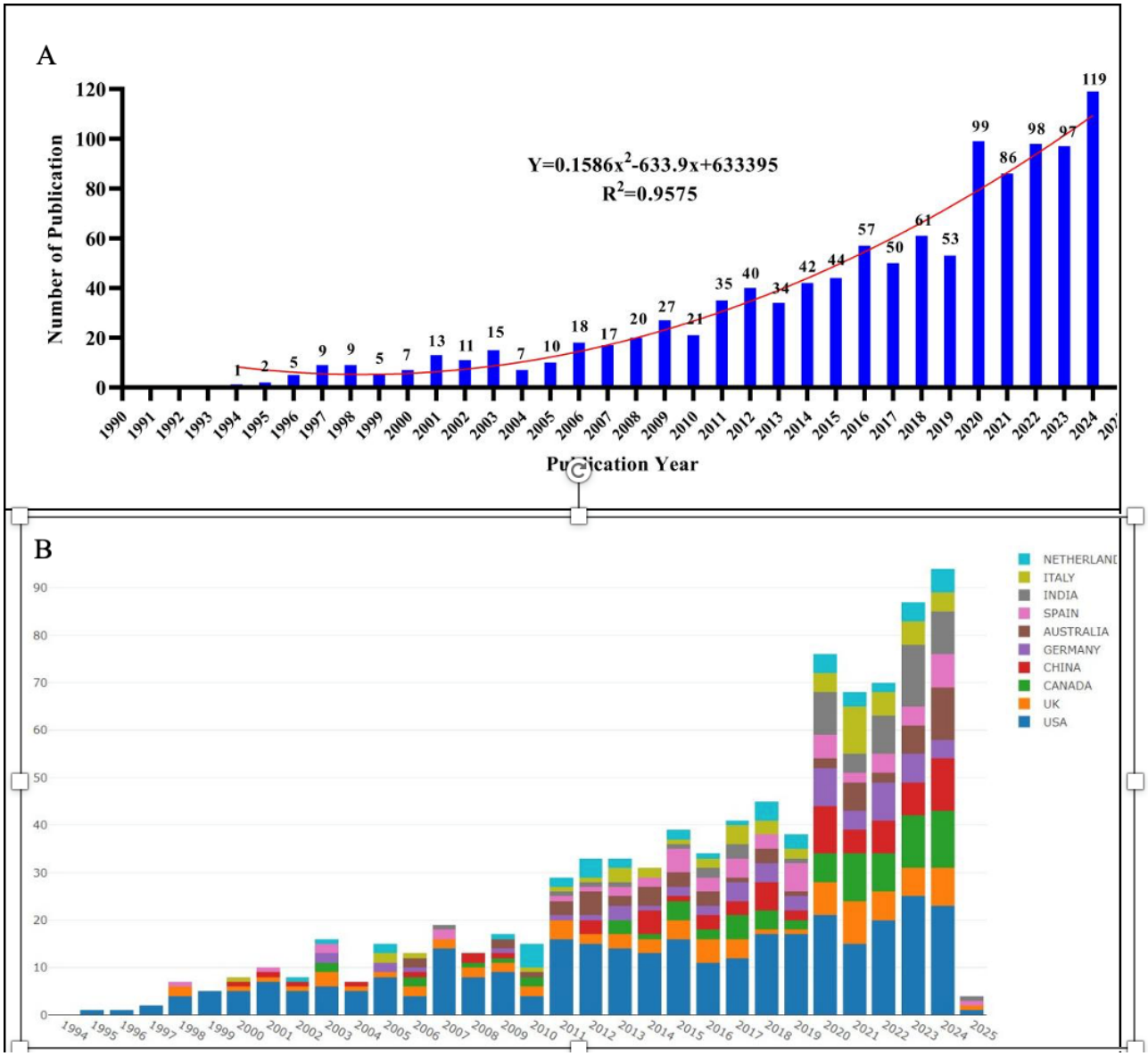
Results

Annual Publications

Among the 1114 papers on cadaver donation and medical education, the earliest publication dates back to 1994 [12]. Since then, the number of publications has steadily increased each year (Figure 2A). The growth trend of annual publications aligns with the fitted curve: $y=0.1586x^2 – 633.9x + 633395$ (R^2=0.9575, *P*<.05). According to this model, it is projected that around 107 papers will be published in this field in 2025.

As part of our thorough analysis, we aimed to identify the countries and regions that have significantly influenced the development of research within this interdisciplinary field. Our geographic analysis, shown in Figure 2B, features a bar chart highlighting the top 10 countries and regions based on their total number of published articles. The United States stands out as a leading contributor to this field, with a considerable and consistently increasing volume of publications.

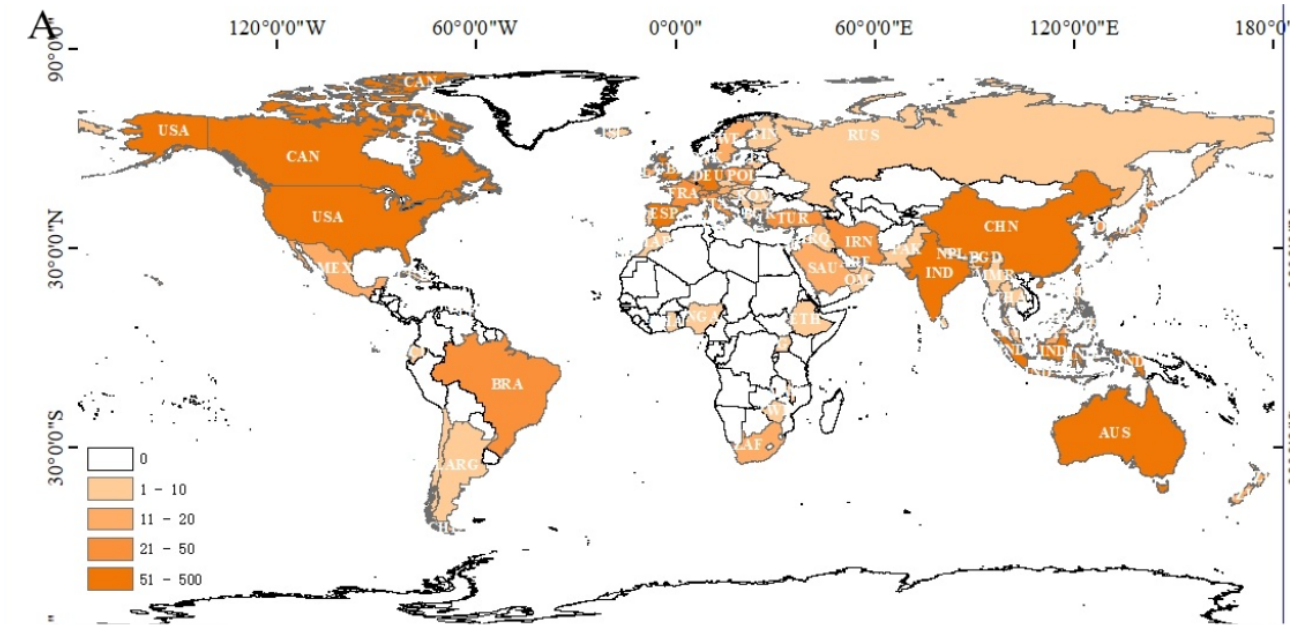
Figure 2. (A) Number of annual publications on cadaver donation and medical education; (B) Number of annual publications and growth trends of the top 10 countries/regions.

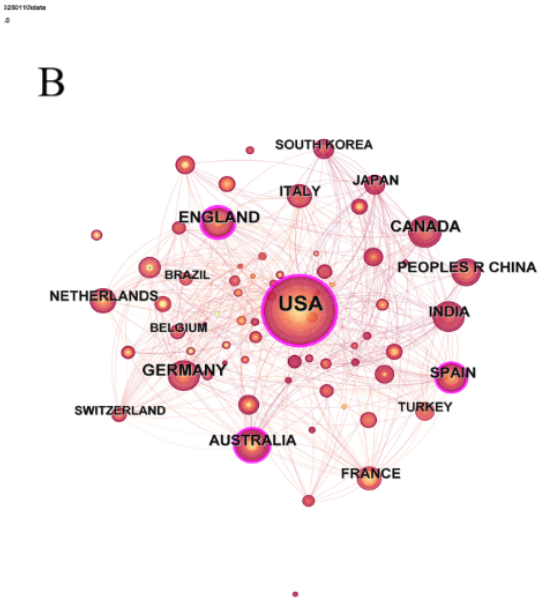


National Analysis

A total of 1114 publications on cadaver donation and medical education were contributed by 81 countries or regions (Figure 3A). The top 10 countries or regions with the highest number of publications are listed in Table 1. Among these, the United States produced the most original articles, followed by the United Kingdom and Canada. Together, the top 3 accounted for over a third of the total publications. Notably, India and China are the only developing countries in the top 10. The research network map among countries or regions revealed a high density (n=81, E=470, density=0.1451; Figure 3B), indicating strong cooperation between nations. Figure 3C highlights frequent collaborations between the United States and Poland, Spain and Poland, and Germany and Switzerland.

Figure 3. (A) Distribution of countries engaged in cadaver donation and medical education research; (B) The network of co-country co-occurrence; (C) The network map of cooperation between countries or regions.





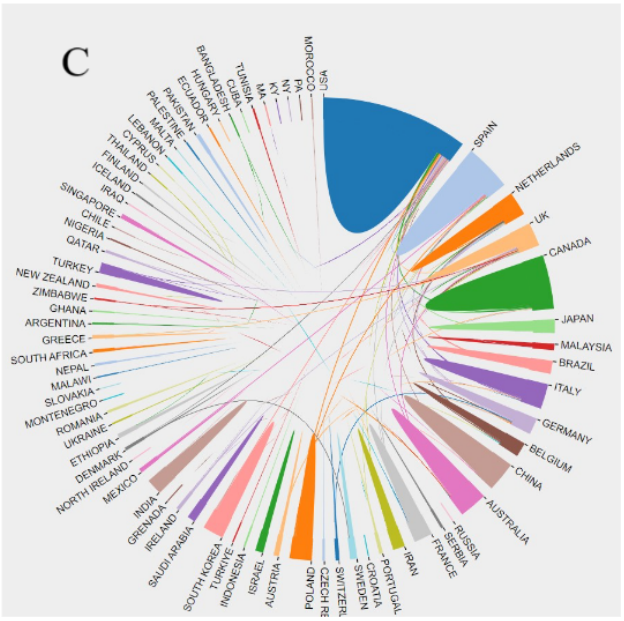


Table . The top 10 countries and institutions of cadaver donation and medical education.

Country	Documents	Citations	Organization	Documents	Citations
United States	303	9352	McGill University	20	754
United Kingdom	79	2488	University of Sydney	18	735
Canada	74	1658	University of Pitts-burgh	16	633
Germany	74	2080	University of Otago	15	444
Australia	57	1388	University of Ottawa	15	286
India	53	1083	Harvard University	14	786
Peoples Republic of China	53	761	University of Montreal	14	74
Spain	53	1458	New York University	13	252
Netherlands	46	1326	Johns Hopkins Univer-sity	11	674
Italy	46	1050	Canadian Blood Serv	11	172

Continent Analysis

In this study, we classified medical education institutions involved in body donation research by continent: Asia, Europe, Oceania, North America, South America, and Africa. South America had no participating institutions in this study. The analysis revealed that European medical education institutions had the highest involvement, contributing 774 institutions. North America followed with 501 institutions, while Asia contributed 413 institutions. Oceania had 75 institutions, and South America and Africa had significantly fewer, with 57 and 32 institutions, respectively. These findings highlight the uneven distribution of research across different continents, with Europe and North America leading in terms of medical education institutions (Multimedia Appendix 1).

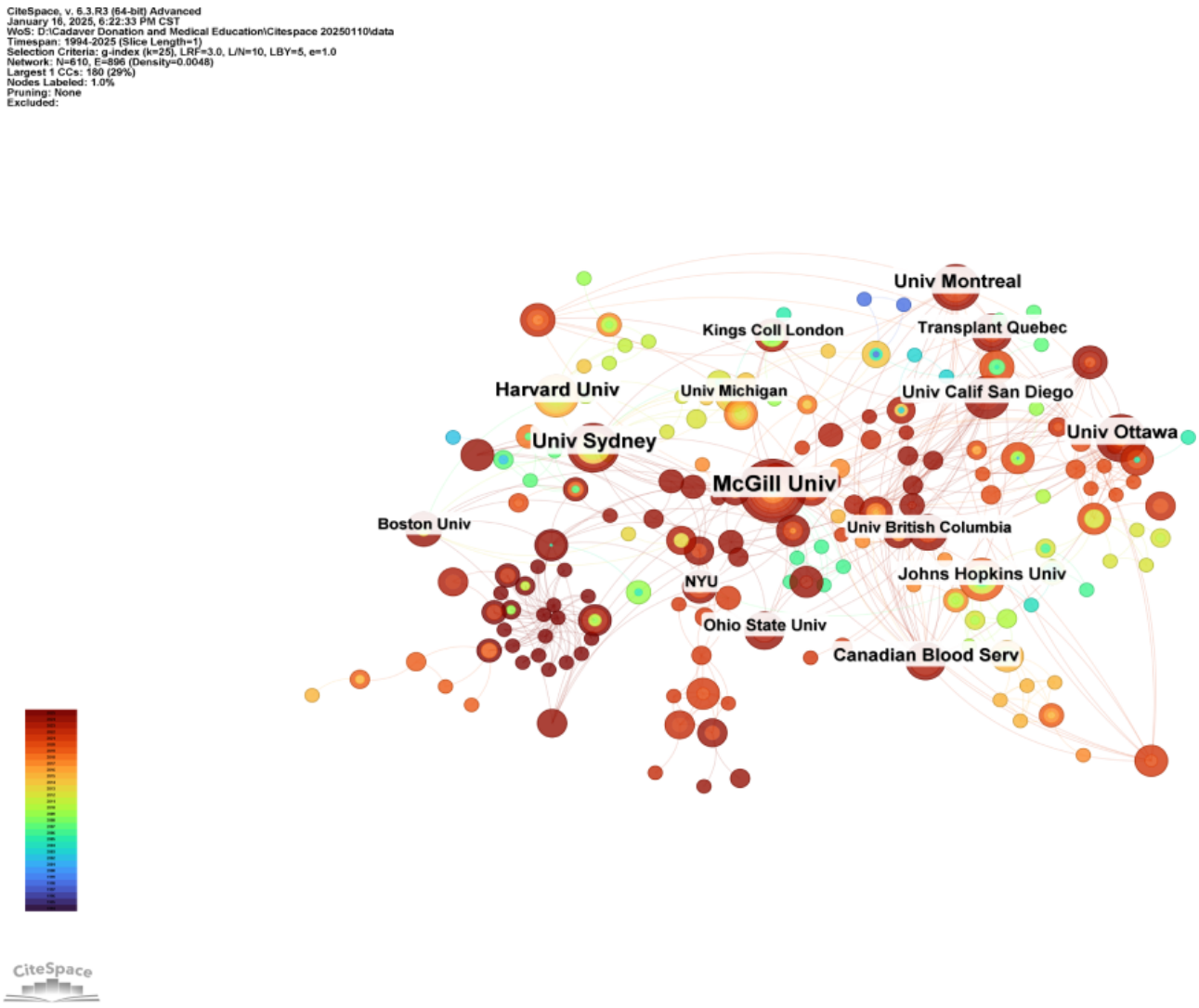
Religion Analysis

To examine the influence of religious affiliation on medical education institutions engaged in cadaver donation, we categorized the institutions according to their religious ties. The analysis revealed that medical schools with Christian affiliations were the most numerous, comprising 1441 institutions. Islamic-affiliated institutions ranked second, with 130 institutions reporting such affiliations. Confucian-affiliated institutions followed in third place with 102, while Buddhist institutions accounted for 89. Hindu-affiliated institutions numbered 71, and Jewish-affiliated institutions were the least represented, with only 19 institutions (Multimedia Appendix 2).

Institutional Analysis

An institutional analysis from 1994 to 2025 shows that 1852 institutions participated in research on cadaver donation and medical education. McGill University and the University of Sydney led the list with the highest number of publications, each contributing at least 18 papers. They were followed by the University of Pittsburgh, University of Otago, and University of Ottawa, which contributed 16, 15, and 15 papers, respectively (Table 1). The institutions with the highest citation counts were Harvard University (786 citations), McGill University (754 citations), and the University of Sydney (735 citations). In terms of centrality, McGill University and the University of Sydney ranked first and second, respectively (Figure 4).

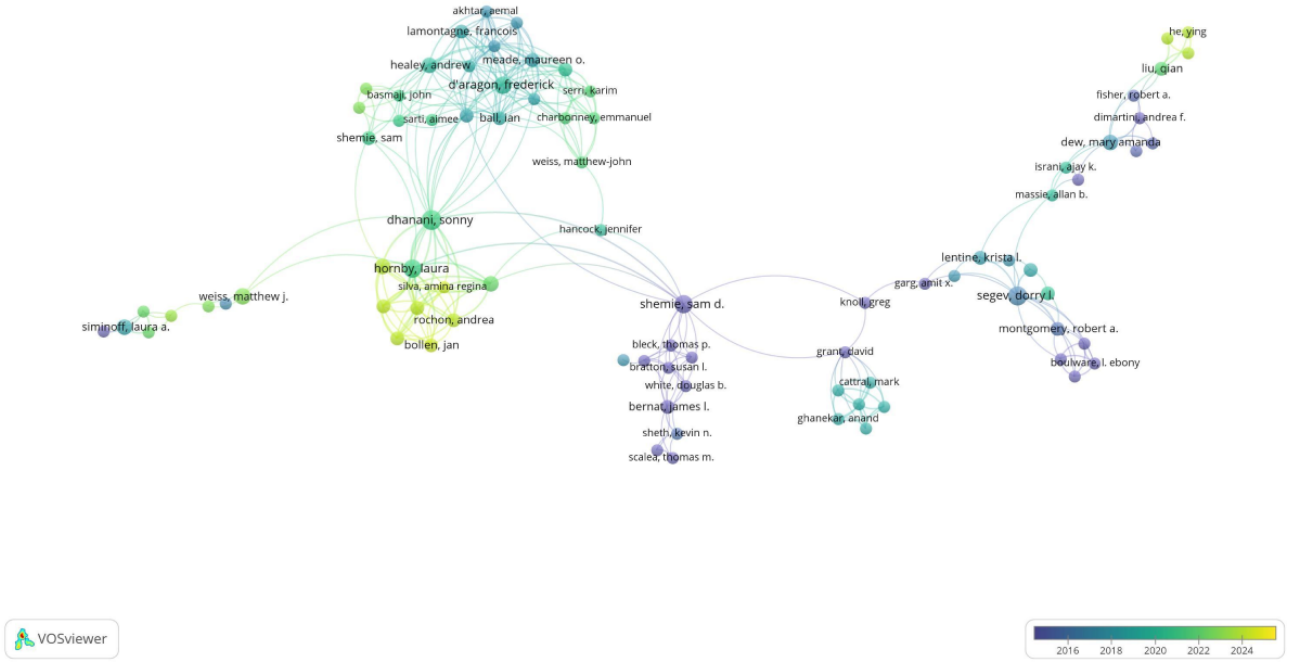
Figure 4. The network of coinstitutions co-occurrence.



Author Analysis

Among this group, a subset of 520 authors emerged as particularly productive, with each having published 5 or more articles (T ≥2). This subset was used to create an author network map, as shown in Figure 5. In addition, an overlay visualization map was developed by considering the 84 coauthors of these 520 authors. This knowledge map serves as a visual representation that clearly illustrates high-frequency coauthor collaborations across various years.

Figure 5. The overlay map of coauthors in cadaver donation and medical education.



In this map, the size of each node reflects the frequency of co-occurrence among authors, while the connecting lines represent coauthor relationships. The total strength of coauthorship connections for each of the 84 authors was measured. The authors with the highest number of publications (n=9) were De Caro, Raffaele; Macchi, Veronica; Porzionato, Andrea; Stecco, Carla; and Dhanani, Sonny (Table 2). Cocited authors are those who frequently appear together in multiple publications. In this field, 20,889 co-cited authors were identified, with 134 of them having more than 20 citations (T ≥ 20). The most frequently cocited author was Jones DG (n=232), followed by Siminoff LA (n=177), Cornwall J (n=159), Bolt S (n=133), and Boulware Le (n=126). Other leading authors had cocitation counts ranging from 89 to 125, as shown in Table 2.

Table . The top 10 coauthors and co-cited authors of cadaver donation and medical education.

Rank	Coauthor	Documents	Citations	Co-cited author	Citations
1	De Caro, Raffaele	9	219	Jones DG	232
2	Macchi, Veronica	9	219	Siminoff LA	177
3	Porzionato, Andrea	9	219	Cornwall J	159
4	Stecco, Carla	9	219	Bolt S	133
5	Dhanani, Sonny	9	78	Boulware Le	126
6	Segev, Dorry L	8	319	Ghosh SK	125
7	Hornby, Laura	7	171	Riederer BM	118
8	Shemie, Sam D	7	640	Winkelmann A	115
9	Balta, Joy Y	6	24	Ríos A	107
10	D'aragon, Frederick	6	36	Hildebrandt S	89

References Analysis

CiteSpace was used to analyze cocited references. Figure 6 and Table 3 display the top 10 co-cited references with the highest frequency and betweenness centrality. The co-citation network in the field of cadaver donation and medical education research consisted of 29,078 references, 1066 nodes, and 3633 links. Figure 7 [13-37] presents the top 25 references with the most significant citation bursts, which indicate emerging trends or growing interest in the subject. In general, references with higher co-citation rates are associated with stronger citation bursts. The citation burst for the article “Age Modulates Attitudes to Whole Body Donation Among Medical Students” authored by Perry GF et al [13]. and published in *Anatomical Sciences Education*, began in 2011 with a burst strength of 6.42. This study explores how age affects medical students’ views on whole body donation, using Likert-type questionnaires administered to first-year graduate-entry students before and after their dissection experiences.

Figure 6. The network of co-cited references in cadaver donation and medical education research.

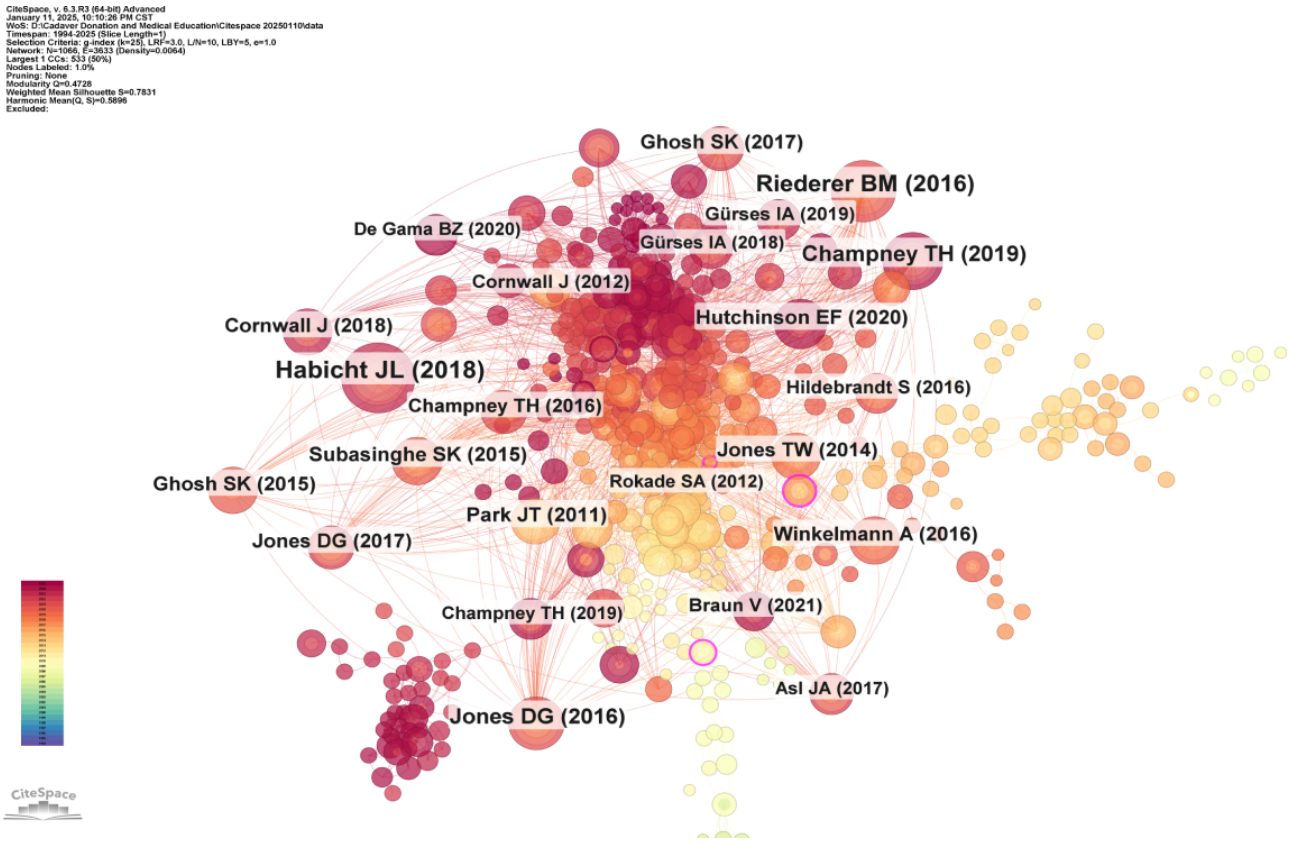
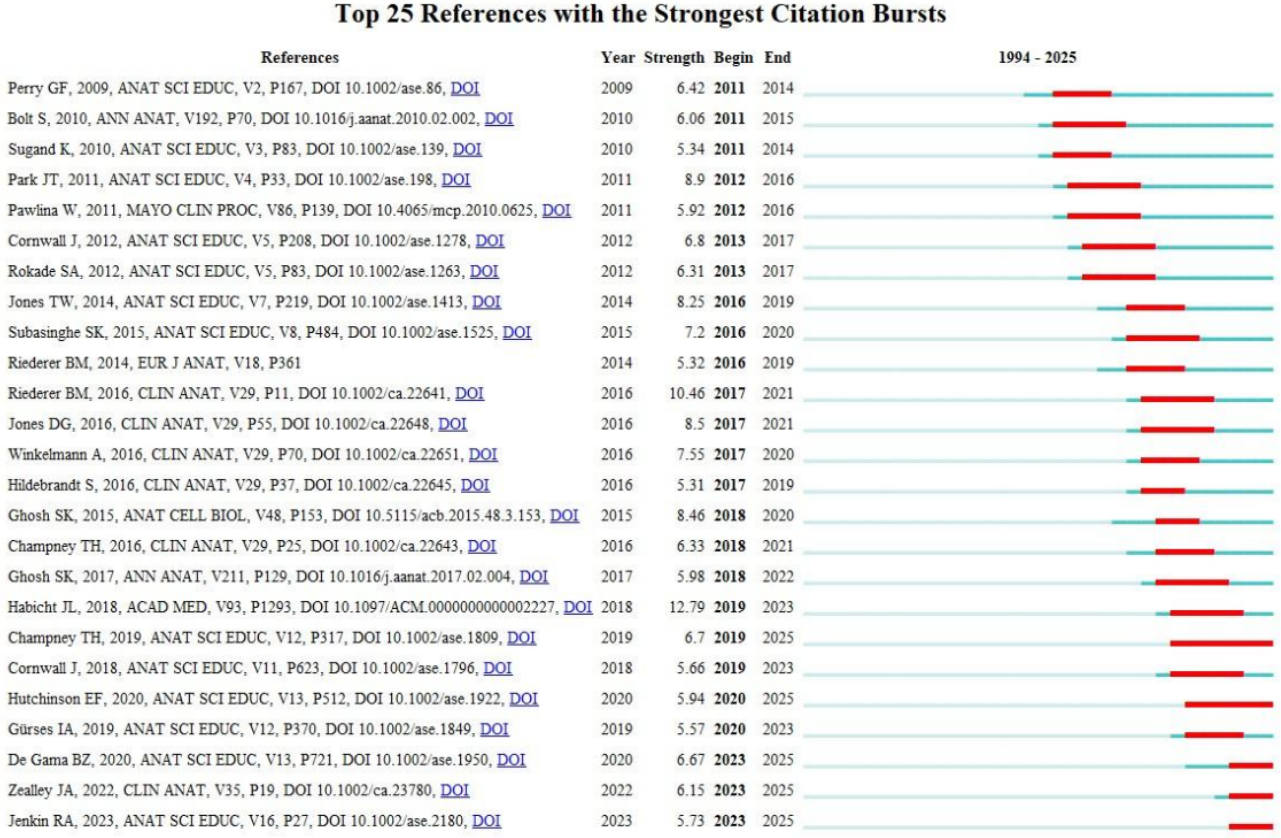


Table . The top 10 co-cited references of cadaver donation and medical education.

Rank	First author	Country	Frequency	Centrality	Year	Source
1	Habicht JL [14]	Germany	38	0.01	2018	*Academic Medicine*
2	Riederer BM [15]	Switzerland	32	0.06	2016	*Clinical Anatomy*
3	Champney TH [16]	United States	27	0.01	2019	*Anatomical Sciences Education*
4	Jones DG [17]	New Zealand	23	0.02	2016	*Clinical Anatomy*
5	Hutchinson EF [18	South Africa	19	0.02	2020	*Anatomical Sciences Education*
6	Subasinghe SK [19]	New Zealand	18	0.02	2015	*Anatomical Sciences Education*
7	Park JT [20]	South Korea	18	0.02	2011	*Anatomical Sciences Education*
8	Cornwall J [21]	New Zealand	17	0.01	2018	*Anatomical Sciences Education*
9	Ghosh SK [22]	India	17	0	2015	*Anatomy and Cell Biology*
10	Jones TW [23]	United States	17	0.02	2014	*Anatomical Sciences Education*

Figure 7. The top 25 references with the strongest citation bursts (sorted by the beginning year of burst) [13-37].



Journals Analysis

We used VOSviewer to perform cocitation and cocited journal analyses, enabling us to identify the most prominent and influential journals in the field. The analysis revealed that 1114 papers were published across 474 academic journals (with a minimum citation count of 20 per source). *Anatomical Sciences Education* had the highest number of citations (n=2586), followed by journals such as *American Journal of Transplantation*, *Transplantation*, *Clinical Anatomy*, *Critical Care Medicine*, *Clinical Transplantation*, *Nephrology Dialysis Transplantation*, *Annals of Anatomy-Anatomischer Anzeiger*, *Social Science & Medicine*, and *Human Reproduction Open*. Among the top 10 journals, five had an impact factor greater than 5 (Table 4).

Among the 10,465 cocited journals, five had over 1000 citations. As presented in Table 5, *Anatomical Sciences Education* led with the highest number of co-citations (n=1946), followed by *Transplantation and Clinical Anatomy*. The dual-map overlay of journals illustrates the distribution of topics within academic journals (Figure 8). A prominent citation pathway, shown in green, was identified, indicating that studies published in Medicine or Medical or Clinical journals were predominantly cited by articles from Health or Nursing or Medicine journals.

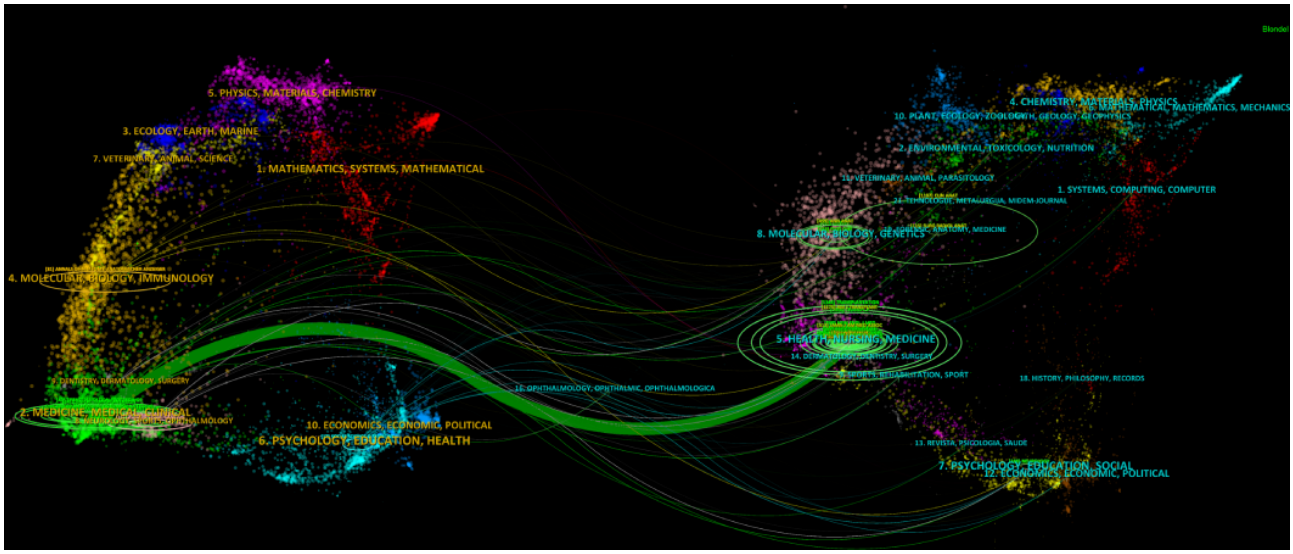
Table . The top 10 journals of cadaver donation and medical education.

Rank	Journal	Citations	Impact factor (2023)	JCR division	Country
1	*Anatomical Sciences Education*	2586	5.2	Q1	United States
2	*American Journal of Transplantation*	999	8.9	Q1	United States
3	*Transplantation*	933	5.3	Q1	United States
4	*Clinical Anatomy*	842	2.3	Q4	United States
5	*Critical Care Medicine*	772	7.7	Q1	United States
6	*Clinical Transplantation*	628	1.9	Q4	United States
7	*Nephrology Dialysis Transplantation*	561	4.8	Q2	England
8	*Annals of Anatomy-Anatomischer Anzeiger*	543	2.0	Q3	Germany
9	*Social Science & Medicine*	517	4.9	Q1	England
10	*Human Reproduction Open*	451	8.3	Q1	England

Table . The top 10 co-cited journals of cadaver donation and medical education.

Rank	Co-cited journal	Co-citations	Impact factor (2023)	JCR division	Country
1	*Anatomical Sciences Education*	1946	5.2	Q1	United States
2	*Transplantation*	1347	5.3	Q1	United States
3	*Clinical Anatomy*	1198	2.3	Q2	United States
4	*American Journal of Transplantation*	1177	8.9	Q1	United States
5	*Transplantation Proceedings*	1068	0.8	Q4	United States
6	*New England Journal of Medicine*	590	96.2	Q1	United States
7	*Journal of the American Medical Association*	515	63.1	Q1	United States
8	*Annals of Anatomy-Anatomischer Anzeiger*	460	2.0	Q2	Germany
9	*Liver Transplantation*	445	4.7	Q1	United States
10	*Clinical Transplantation*	430	1.9	Q2	Denmark

Figure 8. The dual-map overlay of journals related to cadaver donation and medical education research. The citing journals were at left, the cited journals were on the right, and the colored curve represents the reference path.



Keywords Analysis

Keywords act as concise representations of articles, capturing their essential themes. Keywords that occur frequently and are centrally positioned often highlight current and significant areas of research within a discipline. We examined publications by segmenting them into 1-year intervals and selecting the top 20 most cited or frequent items from each period (Figure 9). The development of relevant research is illustrated through a hybrid network where co-occurring keywords from titles and abstracts shape the representation. Various maps display nodes representing keywords, with node size reflecting their frequency of occurrence or citation, and node color indicating the years of occurrence or citation. Red circles on the map denote keyword bursts, signaling rapid increases in publication frequency. By analyzing keyword frequency and centrality, key research frontiers can be identified.

The keyword network map consists of 630 nodes connected by 3914 links. Significantly, the term “organ donation” stands out due to its high frequency and centrality, indicating its considerable relevance and influence within the research field (Figure 9). Similarly, terms like “Transplantation” and “Attitudes” also appear as high-frequency keywords, highlighting their importance and prominence in the context of the study. Table 6 presents the top 20 keywords with the highest frequency and centrality.

Figure 9. The network of keywords co-occurrence.

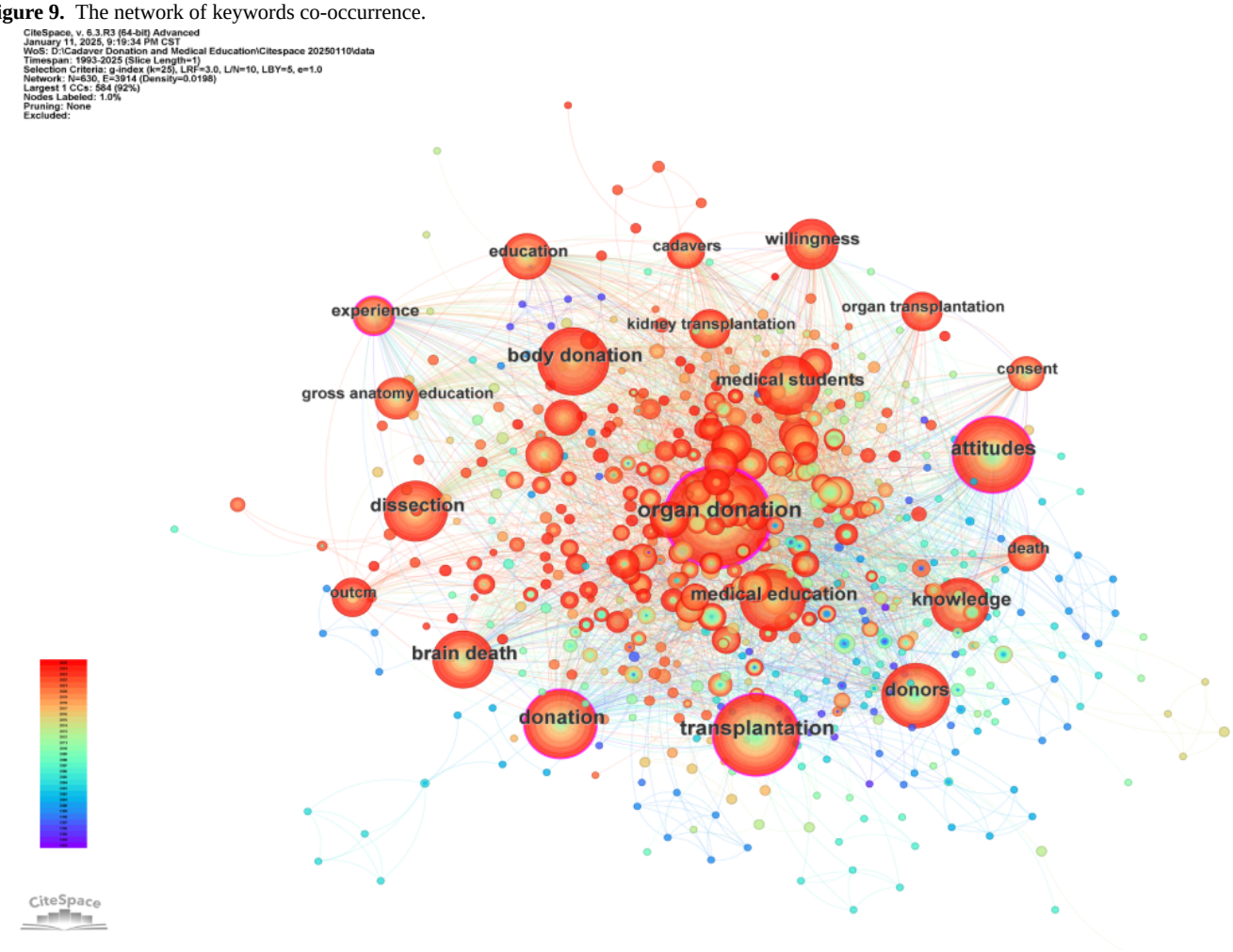


Table . Top 20 keywords of cadaver donation and medical education research in terms of frequency and centrality.

Rank	Keyword	Freq (count)	Centrality	Rank	Keyword	Freq (count)	Centrality
1	Organ Donation	287	0.26	11	Medical Students	89	0.04
2	Transplantation	181	0.16	12	Willingness	71	0.06
3	Attitudes	156	0.12	13	Experience	58	0.13
4	Donation	147	0.19	14	Education	54	0.05
5	Body Donation	122	0.02	15	Gross Anatomy Education	47	0.01
6	Brain Death	117	0.09	16	Organ Transplantation	47	0.03
7	Knowledge	104	0.03	17	Death	46	0.07
8	Donors	98	0.1	18	Cadavers	45	0.03
9	Medical Education	94	0.02	19	Outcm	45	0.03
10	Dissection	91	0.03	20	Kidney Transplantation	44	0.09

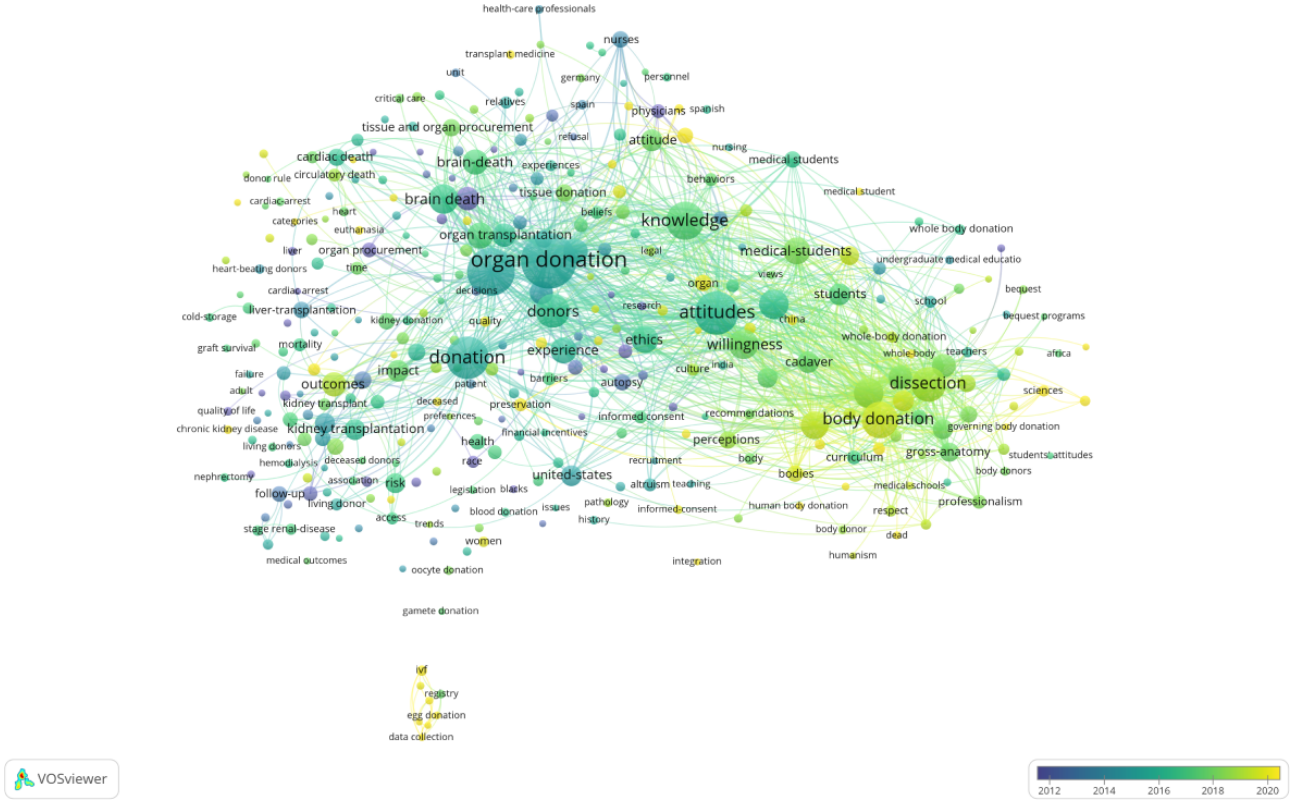
Keyword Timeline View

Keyword cluster analysis is an effective method for identifying key research topics within a particular field. In this study, CiteSpace was employed to perform a cluster analysis of keywords related to cadaver donation and medical education. The number of clusters was determined by the size of each cluster, with the largest one assigned the label #0.

The analysis produced 6 clusters, which were then examined through a timeline view in CiteSpace. These clusters comprised

#0 kidney transplantation, #1 gross anatomy education, #2 brain death, #3 organ donation, #4 body donation, and #5 complications. Figure 10 illustrates the timeline view derived from this cluster analysis.

Figure 10. Keywords are clustered on the time scale and output as map.



The research field, when viewed through the lens of emerging trends and development status, is categorized into two areas based on the co-occurrence analysis: (1) Anatomy education (#1, #3, #4) and (2) Clinical application (#0, #2, #5). In addition, the analysis indicated a steady and increasing interest in clusters #0, #1, and #4 in recent years. This highlights the importance of body donation in medical education and research, offering students valuable practical experience in anatomy. It aids in the enhancement of surgical skills, supports progress in organ transplantation, and contributes to medical studies that could lead to new treatments and better health care. Therefore, body donation plays a vital and altruistic role in advancing medical science and improving public health.

In summary, the cluster analysis and timeline visualization offer important perspectives on research trends, thematic development, and the growing significance of specific keywords in the fields of cadaver donation and medical education.

Research Frontier

The burst keyword analysis conducted with CiteSpace is a useful method for identifying rapidly emerging terms, referred to as burst keywords. These keywords reflect significant attention within the academic community and serve as key indicators of evolving research trends. Through this analysis, 25 representative keywords were chosen from a total of 3947 keywords to pinpoint the main research hotspots. The keywords were organized based on their duration, start time, and burst intensity. (The green line represents the period from 1993 to 2025, while the red line indicates the duration of each burst keyword.)

The term “public attitudes” emerged as the first burst keyword in 1994 (Figure 13A), suggesting that the public’s perception of cadaver donation influences the willingness to donate, thereby affecting the development of medical education and research from the outset. “Body donation” exhibited the highest burst intensity (Figure 13B), highlighting its essential role in cadaver donation, as it provides vital specimens for practical anatomical studies and research in medical education. In addition, “public attitudes” demonstrated the longest burst duration (Figure 13C), indicating that the public’s views on cadaver donation, shaped by enduring cultural perceptions, continue to influence its importance in medical education and research.

The analysis revealed 25 burst keywords, which can be grouped into two main categories: (1) Public perception and (2) Anatomical science. These burst keywords showed a higher concentration between 1994 and 2025. When combined with the triangular pattern depicted in the keyword timeline map (generated using VOSviewer to organize keywords over time), it indicates that research on cadaver donation and medical education is increasingly shifting toward more specialized areas (Figure 12).

Recent significant burst keywords include “awareness” (4.68), “organ” (4.46), and “willingness” (4.1), with the numbers in parentheses indicating the burst strength of each term. These keywords emphasize the importance of public awareness and the willingness to participate in cadaver donation, which are

crucial in promoting and shaping medical education. They play a key role in fostering critical thinking, offering students ethical hands-on experiences, and enhancing their understanding of anatomy and skill development.

This analysis suggests that the field of cadaver donation and medical education has experienced substantial growth and improvement over the past 2 decades, with increasingly focused and specialized research areas emerging. The earlier concentration of burst keywords may signify a phase of rapid progress and discovery, while the more recent burst keywords highlight the growing emphasis on raising public awareness of cadaver donation to improve medical education. The advancement of medical education depends significantly on public support and understanding, as body donors play a crucial role. The shortage of donors, in particular, presents a major challenge to both transplant medicine and foundational medical training, emphasizing the need for greater acknowledgment of their indispensable contribution. Overall, the burst keyword analysis offers valuable insights into the evolution and current state of research in cadaver donation and medical education.

Figure 11. Burst keywords involved cadaver donation research in relation to medical education. (A) Ranking by beginning. (B) Ranking by strengths. (C) Ranking by durations.

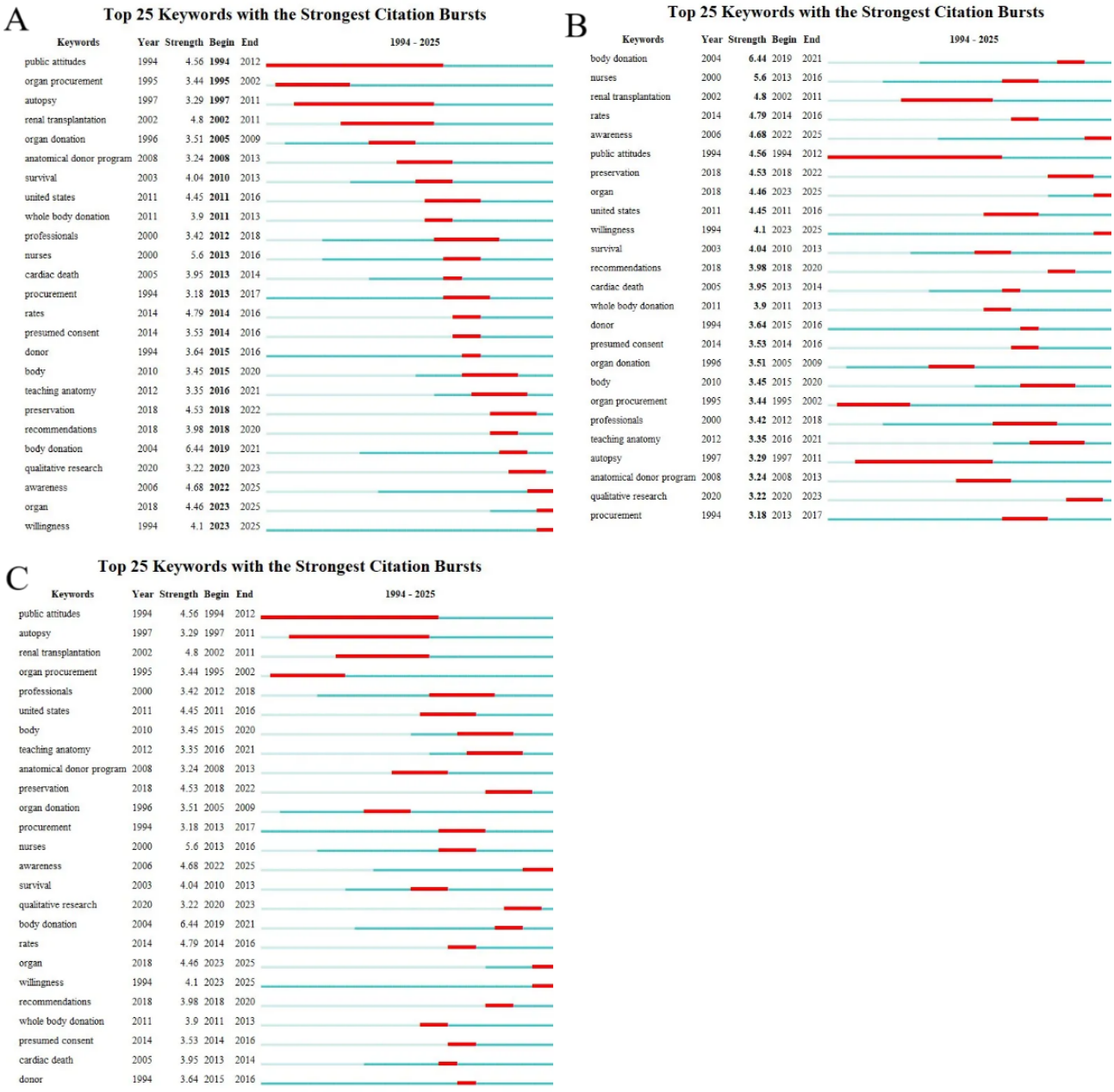
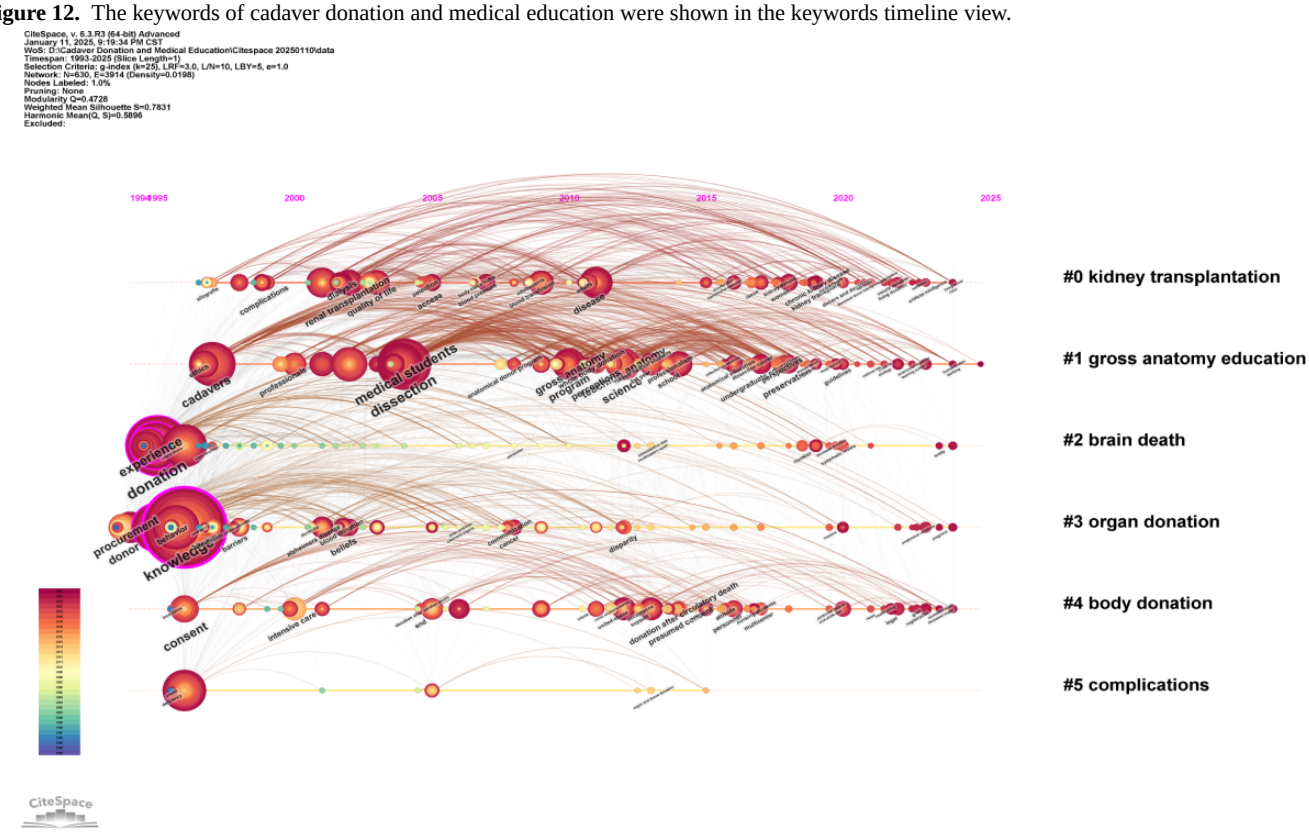


Figure 12. The keywords of cadaver donation and medical education were shown in the keywords timeline view.



Discussion

Principal Findings

This research offers an in-depth bibliometric analysis of cadaver donation in medical education, providing fresh perspectives on global trends and contributions within the field. The results indicate a significant rise in research activity over time, with projections suggesting a continued upward trajectory in the coming years. This increasing interest highlights the growing recognition of cadaver donation as an essential resource for medical education, research, and clinical practice. By pinpointing key contributors, influential institutions, and emerging research areas, this study lays a crucial groundwork for future inquiries and emphasizes the collaborative efforts shaping the cadaver donation landscape.

The contribution of this study to the existing body of literature is especially valuable due to its extensive global scope, which surpasses the focus of previous studies in both breadth and depth. Unlike earlier works that typically concentrate on regional or specific aspects [38,39], this analysis offers a more comprehensive and holistic understanding of global cadaver donation trends. In addition, it introduces new perspectives, including emerging topics such as public perception and anatomical sciences, which are increasingly becoming central to discussions on cadaver donation.

In the subsequent sections, we will explore the broader context of cadaver donation in medical education, examine the key research areas advancing the field, and look into potential future developments that could transform the role of cadaver donation in medical practice. This structured approach aims to provide a thorough understanding of the evolving trends and future directions in this vital aspect of medical education.

General Information

The WOS database is a widely recognized platform that compiles a vast number of scientific research studies [40]. Based on the literature inclusion and exclusion criteria outlined earlier, 1114 articles were ultimately selected for this study. Since 1994, publications related to cadaver donation and medical education have been consistently released, showing a growing trend. This suggests that the field is gaining increasing attention and has the potential to emerge as a prominent area of research in the future. The United States, the United Kingdom, and Canada lead in the number of publications on this topic, with strong collaborative efforts across geographic borders.

This study indicates that Europe and North America lead in research on cadaver donation and medical education, primarily influenced by Christian cultural norms that encourage body donation for medical purposes. In Europe, the origin of Western medicine, cadaver donation became an early component of medical education, fostering significant advancements in these regions. In contrast, other areas, including parts of Asia and the Middle East, have made slower progress in this field due to economic conditions, cultural factors, and religious beliefs. For instance, in some countries, religious taboos limit the acceptance of cadaver donation, leading to delays in related research. Therefore, cultural and religious influences play a crucial role in shaping the development of cadaver donation research worldwide.

McGill University and The University of Sydney are the institutions with the highest number of publications, reflecting their significant contributions to research in cadaver donation

and medical education. A total of 5708 authors have contributed to the literature in this research area. The coauthorship and cocitation analysis highlighted the most influential scholars in the fields of cadaver donation and medical education. De Caro, Raffaele; Macchi, Veronica; Porzionato, Andrea; Stecco, Carla; and Dhanani, Sonny were identified as the most productive authors, underscoring their substantial impact on the literature. Strong collaborative ties among authors were observed, with several prominent coauthor partnerships emerging. These collaborations play a crucial role in advancing knowledge and fostering innovation in cadaver donation and medical education. Furthermore, the identification of cocited authors reveals the key researchers and their significant contributions to the field.

Analyzing the cocited references enabled the identification of the most significant and frequently referenced studies in the field. The leading cocited reference was a paper by Habicht JL et al [14], published in ACAD MED, titled "Bodies for Anatomy Education in Medical Schools: An Overview of the Sources of Cadavers Worldwide." This study examines global sources of cadavers for anatomical education, evaluating the use of body donations and unclaimed bodies while exploring cultural and institutional factors. Another important source was a viewpoint article by Riederer BM [15], titled "Body Donations Today and Tomorrow: What is Best Practice and Why?" published in Clinical Anatomy. This article highlights the significance of body donations for anatomy teaching and research, reviews the current state of donation programs worldwide, and identifies ethical considerations. The author also looks to the future, advocating for the replacement of unclaimed bodies with formal donation programs to improve both ethical standards and educational outcomes.

The majority of the top 10 journals publishing research on cadaver donation and medical education are prestigious journals from the Q1 and Q2 quartiles, including Anatomical Sciences Education, American Journal of Transplantation, and Transplantation. Furthermore, these journals also ranked highly in terms of co-citations.

Current Research Hotspots

Through keyword co-occurrence and cluster analysis, this study pinpointed key research areas in cadaver donation and medical education, which were primarily categorized into 2 themes: anatomy education and clinical practice.

Anatomy Education

Cadaver donation plays an essential role in anatomy education, serving as the cornerstone for hands-on anatomical learning [41]. The value of cadavers in providing real-world insight into human anatomy is irreplaceable, offering students the opportunity to understand complex anatomical structures through direct dissection and observation. This experiential learning approach fosters a deeper understanding of the human body, which is crucial for medical students' clinical training.

However, a significant challenge persists in the availability of cadaveric specimens. Despite the recognized importance of cadaver-based education, there remains a shortage of donated bodies available for anatomical study, particularly in clinical medical programs [42]. This scarcity hampers the ability to provide sufficient dissection opportunities for students, leading to a gap in practical training. As a result, students often lack the hands-on experience necessary for mastering essential anatomical skills, which can negatively impact their overall learning outcomes.

The limited access to cadavers not only affects anatomy education but also poses a barrier to enhancing the quality of medical education as a whole. The lack of dissection practice diminishes students' ability to apply theoretical knowledge to clinical scenarios, thereby restricting their development as proficient health care professionals. Furthermore, the global demand for cadavers in medical education continues to grow, intensifying the need for a more robust system to facilitate cadaver donations and ensure equitable access to these critical learning resources [43].

In conclusion, cadaveric donation remains integral to anatomy education, yet the shortage of cadaver specimens represents a significant challenge. Addressing this issue is vital for advancing medical education and enhancing the quality of training for future health care professionals.

Clinical Practice

The rapid advancement of medical technologies has greatly influenced clinical practice, especially in the field of transplantation. One of the most significant challenges hindering the progress of transplant medicine is the shortage of donor resources [44]. Despite this limitation, significant clinical achievements have been made in corneal and kidney transplantation, demonstrating the profound impact of transplant technologies [45,46]. These advancements have saved thousands of patients suffering from end-stage organ failure, providing them with new hope for survival and improved quality of life.

As medical technology continues to evolve, the focus of research is increasingly shifting toward clinical applications, particularly in the field of transplantation. Cadaver donation plays a pivotal role in this context. Not only does it support the development of anatomy education, but it is also an invaluable resource for advancing clinical practices, particularly in transplantation medicine. Donated bodies provide essential tissue samples for research, leading to improvements in surgical techniques, transplant procedures, and postoperative care.

Furthermore, the use of cadavers in clinical practice indirectly contributes to medical education by providing real-life case studies for students. By integrating cadaver donation into clinical training, medical institutions can equip future health care professionals with the practical experience and knowledge necessary for handling complex transplantation procedures [47]. This, in turn, enhances the quality of clinical education and prepares students to address the challenges of modern medical practice.

In conclusion, cadaver donation is not only crucial for anatomical studies but also plays a vital role in the development of clinical practices, particularly in transplantation. By supporting both medical education and clinical advancements, cadaver donations offer a pathway for saving lives and improving health care outcomes globally.

Future Frontiers

According to the analysis of keyword emergence, public perception and anatomical science are identified as emerging trends for the future.

Public Perception

The role of cadaver donation in medical research and education has profoundly influenced the development of modern medicine. For over half a century, cadaver donation programs have served as a cornerstone of anatomical studies and medical training, particularly in the United States [48]. These programs have allowed for significant advancements in medical education, enabling students to learn through direct interaction with human anatomy. However, the extent of cadaver donation for medical purposes remains heavily dependent on public perception, which varies widely across cultures and regions.

In many countries, especially in East Asia, cadaver donation faces significant challenges due to cultural and religious beliefs. Confucian ideals, which emphasize filial piety and the preservation of the body in its natural state, have deeply shaped attitudes toward death and burial practices [49,50]. This cultural framework has made it difficult for many to consider donating their bodies for scientific or educational purposes. As a result, public awareness of and participation in body donation programs remain relatively low in these regions, which, in turn, limits the availability of cadavers for medical education.

Despite these challenges, the importance of cadavers in medical education, particularly in anatomy, cannot be overstated. With the rapid advancements in medical technologies and diagnostic tools, the role of cadaver-based learning remains irreplaceable in developing comprehensive clinical knowledge and skills. As the legal and regulatory frameworks for cadaver donation continue to improve, it is crucial that efforts are made to increase public awareness and understanding of the significance of body donation. Future research should focus on enhancing public education, developing efficient and ethical donation systems, and addressing cultural and religious barriers. By fostering a more informed and supportive public perception, we can ensure that cadaver donation remains a vital resource for the future of medical education and health care.

Anatomical Science

Anatomy is a foundational subject in medical education, essential for all medical students. It provides students with a direct and comprehensive understanding of the human body's structure, forming the basis for further studies in physiology, pathology, and surgery [51]. The importance of anatomy extends beyond education; it is integral to clinical practice, diagnosis, and treatment. For example, when interpreting radiological images (eg, x-rays, computed tomographic scans, and magnetic resonance images), physicians rely on their anatomical knowledge to differentiate between normal and abnormal tissue structures [52]. During surgery, surgeons use their understanding of anatomy to navigate organ positions and vascular and neural distributions, ensuring precision and minimizing the risk of damaging vital tissues [53].

In addition to its educational significance, the study of anatomy plays a critical role in shaping a medical student's ability to grasp fundamental concepts and apply them to real-world scenarios. As the field of medicine advances, anatomical science continues to be indispensable. This is particularly evident in the emerging field of imaging anatomy, where advancements in imaging technologies demand an in-depth understanding of anatomical structures for accurate interpretation and diagnosis. Furthermore, transplant medicine relies heavily on anatomical knowledge, as precise identification of organ structures and their functions is crucial for successful transplantation and postoperative care.

Given these factors, future research in anatomical science should focus on enhancing educational methodologies, particularly in the context of imaging and transplantation. This will ensure that anatomy continues to evolve in alignment with modern medical advancements, ultimately improving patient outcomes and clinical practices.

Advantages

This study presents several notable strengths that enhance its importance in the field of cadaver donation and medical education. Firstly, the application of bibliometric analysis provides a thorough, data-driven examination of the global research landscape, pinpointing significant trends, key contributors, and emerging areas of focus within the field. By using tools like CiteSpace and VOSviewer, the study delivers detailed visualizations and insights into the evolution of cadaver donation research, making the findings more accessible to future researchers and decision-makers. Second, the comprehensive dataset, which includes articles up to January 2025, ensures that the results reflect the latest and most pertinent advancements in this field. The quadratic regression model's strong goodness of fit (R^2=0.9575) and statistically significant outcomes (P<.05) further validate the robustness of the analysis, offering a dependable forecast of future research trajectories. Finally, the international scope of the study, highlighting contributions from various regions and institutions, provides a well-rounded view of the global cadaver donation research landscape. The identification of emerging topics such as public perception and anatomical sciences demonstrates the study's potential to influence future research directions and inform the development of cadaver donation and medical education programs worldwide.

Limitations

Although this study is the first to apply bibliometric methods to conduct a thorough analysis of the research trends and key topics in cadaver donation and medical education, there are some limitations. First, authors often cite recently published journal articles to increase the likelihood of acceptance, which may influence the variety and representativeness of the literature [54]. Second, researchers tend to favor citing highly cited works, potentially overlooking the actual quality and substance of the research, leading to citation bias [55]. Furthermore, this analysis relies solely on the Web of Science core database, which may exclude other significant data sources and introduce the risk of missing valuable information. Finally, only English-language literature was considered, leaving out potentially important findings in other languages. It is also worth noting that

high-quality, recent research may not receive adequate citations due to shorter publication cycles, which could result in underrepresentation of its research value.

Conclusion

In summary, this analysis emphasizes the rapid expansion of cadaver donation research, particularly in Europe and North America, where regions with predominantly Christian populations contribute the most medical education institutions. Notable contributors include the United States, McGill University, and The University of Sydney, which play a pivotal role in shaping the field, reflecting their advanced medical infrastructure and academic commitment. Key research topics such as kidney transplantation, gross anatomy education, and brain death dominate, while emerging areas like public perception and anatomical science present new opportunities for exploration.

This study enriches the literature by offering a comprehensive overview of global trends in cadaver donation research and its influence on medical education, especially in regions with well-established medical traditions. It highlights significant institutions, authors, and themes, serving as a valuable reference for future research. Furthermore, it underscores the importance of international collaboration in advancing the field and advocates for increased public awareness and support for cadaver donation. The findings of this study will inform future policies and research strategies aimed at enhancing medical education and encouraging innovation in teaching methods.

Data Availability

All data generated or analyzed during this study are included in this published article.

Authors' Contributions

XZ and HX were the main investigators, mainly responsible for the overall framework and design of the paper. YW and FL contributed to data processing and mapping. DH supervised article writing and table design. All authors participated in the revision and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Distribution of the number of medical education institutions engaged in cadaver donation and medical education across different continents.
[JPEG File, 81 KB - mededu_v11i1e71935_app1.jpeg]

Multimedia Appendix 2
Distribution of the number of medical education institutions engaged in cadaver donation and medical education across different religions.
[JPEG File, 90 KB - mededu_v11i1e71935_app2.jpeg]

References

1. Smith CF, Freeman SK, Heylings D, Finn GM, Davies DC. Anatomy education for medical students in the United Kingdom and Republic of Ireland in 2019: A 20-year follow-up. Anat Sci Educ 2022 Nov;15(6):993-1006. [doi: 10.1002/ase.2126] [Medline: 34314569]
2. Jones DG. Do religious and cultural considerations militate against body donation? An overview and a Christian perspective. Anat Sci Educ 2024 Nov;17(8):1586-1595. [doi: 10.1002/ase.2425] [Medline: 38634610]
3. Bala Ganesh KA, Panda P, Gurawa T, Gopalakrishna PK, Jagadeesan S, Vishnumukkala T. Ethics on academic procurement of cadavers. Bioinformation 2024;20(8):872-876. [doi: 10.6026/973206300200872] [Medline: 39411772]
4. Chen D, Zhang Q, Deng J, et al. A shortage of cadavers: The predicament of regional anatomy education in mainland China. Anat Sci Educ 2018 Jul;11(4):397-402. [doi: 10.1002/ase.1788] [Medline: 29648678]
5. Ghosh SK. Cadaveric dissection as an educational tool for anatomical sciences in the 21st century. Anat Sci Educ 2017 Jun;10(3):286-299. [doi: 10.1002/ase.1649] [Medline: 27574911]
6. Ogut E, Senol Y. Do learning styles affect study duration and academic success. Eur J Anat 2017;21(3):235-240.
7. Šink Ž, Tonin G, Umek N, Cvetko E. Attitudes of Slovenian students towards whole-body donation, organ donation, and the use of donated bodies in medical education. BMC Med Educ 2024 Dec 26;24(1):1535. [doi: 10.1186/s12909-024-06569-7] [Medline: 39725998]
8. Gebert JT, Zhang M. Fewer medical students are open to body donation after dissecting human cadavers. Med Educ 2023 Apr;57(4):369-378. [doi: 10.1111/medu.14948] [Medline: 36208394]
9. Zdilla MJ, Balta JY. Human body donation and surgical training: a narrative review with global perspectives. Anat Sci Int 2023 Jan;98(1):1-11. [doi: 10.1007/s12565-022-00689-0] [Medline: 36227535]

10. Sabé M, Chen C, El-Hage W, et al. Half a Century of Research on Posttraumatic Stress Disorder: A Scientometric Analysis. Curr Neuropharmacol 2024;22(4):736-748. [doi: 10.2174/1570159X22666230927143106] [Medline: 37888890]
11. van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 2010 Aug;84(2):523-538. [doi: 10.1007/s11192-009-0146-3] [Medline: 20585380]
12. Sanner M. Attitudes toward organ donation and transplantation. A model for understanding reactions to medical procedures after death. Soc Sci Med 1994 Apr;38(8):1141-1152. [doi: 10.1016/0277-9536(94)90229-1] [Medline: 8042059]
13. Perry GF, Ettarh RR. Age modulates attitudes to whole body donation among medical students. Anatomical Sciences Ed 2009 Jul;2(4):167-172. [doi: 10.1002/ase.86]
14. Habicht JL, Kiessling C, Winkelmann A. Bodies for Anatomy Education in Medical Schools: An Overview of the Sources of Cadavers Worldwide. Acad Med 2018 Sep;93(9):1293-1300. [doi: 10.1097/ACM.0000000000002227] [Medline: 29561275]
15. Riederer BM. Body donations today and tomorrow: What is best practice and why? Clin Anat 2016 Jan;29(1):11-18. [doi: 10.1002/ca.22641] [Medline: 26475613]
16. Champney TH, Hildebrandt S, Gareth Jones D, Winkelmann A. BODIES R US: Ethical Views on the Commercialization of the Dead in Medical Education and Research. Anat Sci Educ 2019 May;12(3):317-325. [doi: 10.1002/ase.1809] [Medline: 30240149]
17. Jones DG. Searching for good practice recommendations on body donation across diverse cultures. Clin Anat 2016 Jan;29(1):55-59. [doi: 10.1002/ca.22648] [Medline: 26475228]
18. Hutchinson EF, Kramer B, Billings BK, Brits DM, Pather N. The Law, Ethics and Body Donation: A Tale of Two Bequeathal Programs. Anat Sci Educ 2020 Jul;13(4):512-519. [doi: 10.1002/ase.1922] [Medline: 31596033]
19. Subasinghe SK, Jones DG. Human body donation programs in Sri Lanka: Buddhist perspectives. Anat Sci Educ 2015;8(5):484-489. [doi: 10.1002/ase.1525] [Medline: 25689145]
20. Park J, Jang Y, Park MS, et al. The trend of body donation for education based on Korean social and religious culture. Anatomical Sciences Ed 2011 Jan;4(1):33-38. [doi: 10.1002/ase.198]
21. Cornwall J, Poppelwell Z, McManus R. “Why did you really do it?” A mixed-method analysis of the factors underpinning motivations to register as a body donor. Anat Sci Educ 2018 Nov;11(6):623-631. [doi: 10.1002/ase.1796] [Medline: 29762910]
22. Ghosh SK. Human cadaveric dissection: a historical account from ancient Greece to the modern era. Anat Cell Biol 2015 Sep;48(3):153-169. [doi: 10.5115/acb.2015.48.3.153] [Medline: 26417475]
23. Jones TW, Lachman N, Pawlina W. Honoring our donors: a survey of memorial ceremonies in United States anatomy programs. Anat Sci Educ 2014;7(3):219-223. [doi: 10.1002/ase.1413] [Medline: 24753299]
24. Bolt S, Venbrux E, Eisinga R, Kuks JBM, Veening JG, Gerrits PO. Motivation for body donation to science: More than an altruistic act. Ann Anat 2010 Apr;192(2):70-74. [doi: 10.1016/j.aanat.2010.02.002]
25. Sugand K, Abrahams P, Khurana A. The anatomy of anatomy: A review for its modernization. Anat Sci Educ 2010 Mar;3(2):83-93 [FREE Full text] [doi: 10.1002/ase.139]
26. Pawlina W, Hammer RR, Strauss JD, et al. The hand that gives the rose. Mayo Clin Proc 2011 Feb;86(2):139-144. [doi: 10.4065/mcp.2010.0625] [Medline: 21282487]
27. Cornwall J, Perry GF, Louw G, Stringer MD. Who donates their body to science? An international, multicenter, prospective study. Anat Sci Educ 2012;5(4):208-216 [FREE Full text] [doi: 10.1002/ase.1278] [Medline: 22508582]
28. Rokade SA, Gaikawad AP. Body donation in India: social awareness, willingness, and associated factors. Anat Sci Educ 2012;5(2):83-89 [FREE Full text] [doi: 10.1002/ase.1263] [Medline: 22278885]
29. Riederer BM, Bueno-López JL. Anatomy, respect for the body and body donation - a guide for good practice. Eur J Anat ;18(4):361-368 [FREE Full text]
30. Winkelmann A. Consent and consensus-ethical perspectives on obtaining bodies for anatomical dissection. Clin Anat 2016 Jan;29(1):70-77 [FREE Full text] [doi: 10.1002/ca.22651] [Medline: 26475682]
31. Hildebrandt S. Thoughts on practical core elements of an ethical anatomical education. Clin Anat 2016 Jan;29(1):37-45 [FREE Full text] [doi: 10.1002/ca.22645] [Medline: 26474826]
32. Champney TH. The business of bodies: Ethical perspectives on for-profit body donation companies. Clin Anat 2016 Jan;29(1):25-29 [FREE Full text] [doi: 10.1002/ca.22643] [Medline: 26474530]
33. Ghosh SK. Paying respect to human cadavers: We owe this to the first teacher in anatomy. Annals of Anatomy - Anatomischer Anzeiger 2017 May;211:129-134. [doi: 10.1016/j.aanat.2017.02.004]
34. Gürses İ, Ertaş A, Gürtekin B, et al. Profile and motivations of registered whole-body donors in Turkey: Istanbul University experience. Anat Sci Educ 2019 Jul;12(4):370-385 [FREE Full text] [doi: 10.1002/ase.1849] [Medline: 30548175]
35. De Gama BZ, Jones DG, Bhengu TT, Satyapal KS. Cultural practices of the Zulu ethnic group on the body and their influence on body donation. Anat Sci Educ 2020 Nov;13(6):721-731 [FREE Full text] [doi: 10.1002/ase.1950] [Medline: 32077216]
36. Zealley JA, Howard D, Thiele C, Balta JY. Human body donation: How informed are the donors? Clin Anat 2022 Jan;35(1):19-25 [FREE Full text] [doi: 10.1002/ca.23780] [Medline: 34431553]

37. Jenkin RA, Garrett SA, Keay KA. Altruism in death: Attitudes to body and organ donation in Australian students. Anat Sci Educ 2023 Jan;16(1):27-46 [FREE Full text] [doi: 10.1002/ase.2180] [Medline: 35344291]
38. Boduç E, Allahverdi TD. Medical Students' Views on Cadaver and Organ Donation. Transplant Proc 2022 Oct;54(8):2057-2062. [doi: 10.1016/j.transproceed.2022.08.021]
39. Şahin ZA, Üzel M, Marur T, Eren F, Yıldırım FG. Knowledge levels and attitudes of medical faculty students related to whole body donation in Türkiye. Ann Anat 2023 Apr;247:152047. [doi: 10.1016/j.aanat.2023.152047] [Medline: 36690042]
40. Zhao M, Sun M, Zhao R, Chen P, Li S. Effects of exercise on sleep in perimenopausal women: A meta-analysis of randomized controlled trials. Explore (NY) 2023;19(5):636-645. [doi: 10.1016/j.explore.2023.02.001] [Medline: 36781319]
41. Singal A, Bansal A, Chaudhary P. Cadaverless anatomy: Darkness in the times of pandemic Covid-19. Morphologie 2020 Sep;104(346):147-150. [doi: 10.1016/j.morpho.2020.05.003] [Medline: 32518047]
42. Wirtu AT, Manjatika AT. Challenges in sourcing bodies for anatomy education and research in Ethiopia: Pre and post COVID-19 scenarios. Annals of Anatomy - Anatomischer Anzeiger 2024 Jun;254:152234. [doi: 10.1016/j.aanat.2024.152234]
43. Brooks JP, Homan C. The Status of Cadaver-Based Anatomy Instruction in Missouri Medical Schools. Mo Med 2024;121(5):395-402. [Medline: 39421478]
44. Tanashat M, Zayed A, Ayyad M, et al. Current status and challenges of cardiac transplantation in the MENA region: A narrative review. Curr Probl Cardiol 2025 Jan;50(1):102920. [doi: 10.1016/j.cpcardiol.2024.102920] [Medline: 39510402]
45. Kigitovica D, Kuzema V, Jusinskis J, et al. Individualized Decision-Making and Outcomes for the 87-Year-Old Living Kidney Donor: A Case Report. Case Rep Nephrol Dial 2024;14(1):122-127. [doi: 10.1159/000539772] [Medline: 39118825]
46. Sharma N, Agarwal P, Titiyal JS, Kumar C, Sinha R, Vajpayee RB. Optimal use of donor corneal tissue: one cornea for two recipients. Cornea 2011 Oct;30(10):1140-1144. [doi: 10.1097/ICO.0b013e318209d23c] [Medline: 21808194]
47. Boduç E, Allahverdi TD. Medical Students' Views on Cadaver and Organ Donation. Transplant Proc 2022 Oct;54(8):2057-2062. [doi: 10.1016/j.transproceed.2022.08.021] [Medline: 36207151]
48. Bagian LK, Davis DC, Parker RC, Mosley CF, Balta JY. Giving a voice to our silent teachers: Whole body donation from the donor perspective at one donation program in the United States. Anat Sci Educ 2024 Jun;17(4):893-908. [doi: 10.1002/ase.2410] [Medline: 38520129]
49. Oh SO, Bay BH, Kim HJ, Lee HY, Yoon S. Commemoration of body donors in a religiously diverse society: A tale of two Korean medical schools. Anat Sci Educ 2024 Nov;17(8):1618-1627. [doi: 10.1002/ase.2462] [Medline: 38797957]
50. Park S, Schepp KG. Understanding Korean Families With Alcoholic Fathers in a View of Confucian Culture. J Addict Nurs 2015;26(3):111-119. [doi: 10.1097/JAN.0000000000000085] [Medline: 26340569]
51. Kim IB, Joo KM, Song CH, Rhyu IJ. A Brief Review of Anatomy Education in Korea, Encompassing Its Past, Present, and Future Direction. J Korean Med Sci 2024 May 27;39(20):e159. [doi: 10.3346/jkms.2024.39.e159] [Medline: 38804009]
52. O'Keeffe GW, Davy S, Barry DS. Radiologist's views on anatomical knowledge amongst junior doctors and the teaching of anatomy in medical curricula. Ann Anat 2019 May;223:70-76. [doi: 10.1016/j.aanat.2019.01.011] [Medline: 30731200]
53. Zhang JF, Zilundu PLM, Zhou L, Guo GQ. Supplementary Regional Anatomy Teaching by Surgeons Enhances Medical Students Mastery of Anatomical Knowledge and Positively Impacts Their Choice of Future Career. J Surg Educ 2020;77(5):1113-1120. [doi: 10.1016/j.jsurg.2020.03.016] [Medline: 32446769]
54. Seglen PO. Why the impact factor of journals should not be used for evaluating research. BMJ 1997 Feb 15;314(7079):498-502. [doi: 10.1136/bmj.314.7079.497] [Medline: 9056804]
55. Tang N, Zhang W, George DM, Su Y, Huang T. The Top 100 Most Cited Articles on Anterior Cruciate Ligament Reconstruction: A Bibliometric Analysis. Orthop J Sports Med 2021 Feb;9(2):2325967120976372. [doi: 10.1177/2325967120976372] [Medline: 33623795]

Abbreviations

VR: virtual reality
WoSCC: Web of Science Core Collection

Edited by D Chartash; submitted 29.01.25; peer-reviewed by E Ogut, J Wang; revised version received 25.04.25; accepted 28.04.25; published 18.08.25.

Please cite as:
Zhou X, Xiong H, Wen Y, Li F, Hu D
Global Trends in Cadaver Donation and Medical Education Research: Bibliometric Analysis Based on VOSviewer and CiteSpace
JMIR Med Educ 2025;11:e71935
URL: https://mededu.jmir.org/2025/1/e71935
doi:10.2196/71935

Text Message (SMS) Microlearning for Tobacco Use Disorder: Pre-Post Pilot Study of Clinician Confidence

Zehra Dhanani[1], MBBS; Veena Dronamraju[2], MD; Jamie Garfield[3], MD

[1]Thoracic Medicine and Surgery, Temple University Hospital, 3401 N Broad St, Philadelphia, PA, United States

[2]Department of Lung and Respiratory Care, Mount Auburn Hospital, Cambridge, MA, United States

[3]Thoracic Medicine and Surgery, Lewis Katz School of Medicine, Temple University, Philadelphia, PA, United States

Corresponding Author:
Zehra Dhanani, MBBS
Thoracic Medicine and Surgery, Temple University Hospital, 3401 N Broad St, Philadelphia, PA, United States

Abstract

Background: Clinicians are central to treating tobacco use disorder, yet practical training is inconsistent, and confidence varies. Brief, text message–based microlearning may offer a low-burden way to strengthen foundational competencies in busy clinical settings.

Objective: This paper aims to evaluate whether a short SMS microlearning series improves clinicians' self-reported confidence in managing tobacco use disorder.

Methods: We conducted a single-arm, pre-post educational pilot at an academic medical center. A brief formative survey (13 items; 106 respondents) identified local knowledge gaps and informed message topics and sequencing. The 13-day series delivered 1 concise message per day with key teaching points and links to curated resources. The prespecified primary outcome was self-reported confidence in managing tobacco use disorder (1 - 100 scale) measured immediately before and after the series. Of the 34 clinicians who signed up, 22 completed the baseline questionnaire and enrolled (attendings: n=4, 18%; trainees: n=18, 82%). Changes in confidence among participants with paired ratings were tested with a paired *t* test. Engagement with embedded links was recorded.

Results: All enrolled participants completed the 13-day series; none unsubscribed. Postintervention confidence ratings were provided by 18 participants. Mean confidence increased from 60 (SD 16) at baseline to 85 (SD 10) after the series (t_{17}=−10.71; P<.001). Embedded links were opened in 67% (178/266) of messages. Free-text feedback was predominantly positive and emphasized the convenience, clarity, and point-of-care usefulness of brief messages.

Conclusions: A brief SMS microlearning series was associated with a substantial improvement in clinicians' confidence to manage tobacco use disorder, with high completion and evidence of engagement. This low-cost, scalable approach appears practical for busy clinicians. Findings should be interpreted cautiously given the single-arm design, self-selection, and reliance on self-reported confidence rather than objective knowledge or clinical outcomes. Future studies should include a validated knowledge assessment, a randomized comparison, broader sampling, and follow-up to assess durability and impact on care.

(JMIR Med Educ 2025;11:e73821) doi:10.2196/73821

KEYWORDS

tobacco use disorder; clinical education; text messaging; confidence; implementation; feasibility

Introduction

Tobacco use remains the leading cause of preventable mortality in the United States, accounting for nearly 20% of all deaths [1]. In addition to its negative health implications, smoking imposes significant economic burdens on the health care system [2]. Health care professionals are tasked with the critical responsibility of managing tobacco use disorder and its associated morbidities. However, data suggest that there are substantial knowledge gaps among physicians. In a United States survey of health care providers, fewer than half demonstrated high knowledge of availability of diagnostic criteria (36.8%), treatment efficacy (33.2%), counseling modalities (5.6%), and US Food and Drug Administration–approved medications (40.1%) [3]. Similar gaps are reported internationally; among primary care physicians in 3 Malaysian districts, 62.4% had poor knowledge, 58.0% had poor attitude, and practice was poor for 50.9% in the precontemplation phase and 75.7% in the contemplation phase of smoking-cessation management [4]. These deficits likely reflect uneven training exposure; evidence indicates that education in tobacco use disorder treatment is often inadequate, competency assessments are infrequent, and the topic receives minimal attention in board certification examinations [4-6].

At our institution, we conducted a brief formative survey to identify local gaps in knowledge and confidence and used these findings to shape the content and pacing of the intervention; full methods and the survey instrument are provided in the Methods and Multimedia Appendix 1.

Based on the results of survey data and identified knowledge gaps, we sought innovative and effective methods to educate practitioners on tobacco dependence treatment. Our objective was to identify learning strategies that were not time-consuming, easy to understand, readily accessible, and sustainable. Given these aims, we selected text message–based microlearning, which delivers brief, focused content that can be revisited at the point of care and reinforced over time, with demonstrated effectiveness in clinical education [7,8].

Among digital options, SMS pushes concise content without requiring app downloads or logins. It reaches busy clinicians, supports repeated exposure, and adds minimal burden, which fits the training and time constraints that healthcare providers face. We developed a brief text-based educational module to evaluate the effectiveness of this modality in teaching. Specifically, we used tobacco dependence treatment as a case example to assess whether a concise SMS series could address provider knowledge gaps and improve management confidence. Our primary hypothesis was that clinicians who completed the SMS series would report higher confidence in managing tobacco use disorder compared with baseline.

Methods

Study Design

We conducted a single-arm, pre-post educational pilot at an academic medical center. We designed a 13-day text message series for health care providers on tobacco dependence treatment. Concise daily messages were generated based on the knowledge gaps identified in the survey. Participation in the text-based series was completely voluntary. Surveyed participants were invited to opt in and had the option to opt out at any time. By opting in and providing their contact information, participants explicitly consented to participate in the study and receive text messages.

Participants and Recruitment

A formative needs assessment was conducted prior to the intervention to characterize local knowledge gaps and baseline confidence. Using institutional distribution lists, we contacted 209 clinicians—including internal medicine residents, family medicine residents, internal medicine faculty, pulmonary and critical care faculty, and pulmonary or critical care fellows and advanced practice providers—to complete the survey; 106 clinicians responded. Recruitment for the SMS intervention occurred in the subsequent academic cycle and targeted a pool that was not identical to the survey cohort. Using institutional distribution lists, we reached 142 residents, 39 faculty, and 28 fellows or advanced practice providers and invited them to enroll. Opt-ins and completion numbers are reported in the Results.

Formative Needs Assessment (Survey)

Before developing the curriculum, we administered a brief survey to characterize local knowledge gaps and baseline confidence. A total of 106 clinicians participated. The survey comprised 13 items: one 1‐100 visual analog scale for self-reported confidence in managing tobacco use disorder (1=no confidence; 100=complete confidence) and 12 knowledge items covering counseling approaches, cognitive behavioral strategies, first-line pharmacotherapy, and management in special populations (eg, pregnancy). The instrument was created for this study based on guideline topics and content domains targeted for the messages; the full survey is provided in Multimedia Appendix 1. Overall, 70% (74/106) of respondents reported feeling confident managing tobacco use disorder on the 1‐100 scale. Item-level accuracy is summarized in Table 1 (formative needs assessment informing curriculum design). Notably, only 32% (n=34) correctly identified the first-line pharmacologic treatment, and additional gaps were observed across counseling approaches, cognitive behavioral strategies, and management in pregnancy. These descriptive findings were used solely to inform topic selection and sequencing for the 13-day series and are not analyzed as study outcomes.

Table . Performance on formative survey assessing knowledge of tobacco use disorder management (n=106). Values are number and percentage of respondents who answered each item correctly.

Knowledge tested	Correctly answered, n (%)
Recommended first-line pharmacological treatment for tobacco use disorder	34 (32)
Understanding of bupropion's contraindications	61 (58)
Nicotine patch dosing	41 (39)
Nicotine gum dosing	67 (63)
Correct technique to use nicotine gum	70 (66)
Treatment for precontemplative patients	15 (14)
Prescription of varenicline in patients with psychiatric conditions	51 (48)
Duration of varenicline treatment after quitting	61 (58)
Risk of adverse cardiovascular events with treatment in stable coronary artery disease	86 (81)
Pharmacologic interventions for smoking cessation during pregnancy	56 (53)

Intervention

Participants who opted in received a message once every 24 hours. Each text message began with the mean performance on the survey of the particular teaching point in question, provided information or management recommendations, and concluded with live links to reference materials and additional reading for those seeking to deepen their understanding (Multimedia Appendix 2). We used a commercial SMS texting platform called SimpleTexting, a secure platform that complies with industry standards for privacy. This platform enabled the efficient dissemination of messages while protecting participants' personal information. It also provided functionalities to monitor engagement with the provided links, receive participant responses, and track subscription status. At the start and end of the series, we assessed participants' confidence with managing tobacco use disorder, defined as their subjective confidence on the topic, using a scale of 1 to 100. Additionally, we invited participants to provide comments and feedback on the text message series to gain qualitative insights into their experiences and perceptions. All submitted feedback was reported as received.

Statistical Analysis

All available pre-post confidence pairs were analyzed; optional feedback was summarized descriptively and reported as received. A paired *t* test was used to compare confidence levels before and after the series, and the analysis was conducted using IBM SPSS version 25. A *P* value of less than .05 was considered statistically significant.

Ethical Considerations

This study was approved by the Temple University Hospital Institutional Review Board (protocol 32174). Participation in both the formative survey and the SMS series was voluntary; responding to the survey and opt-in enrollment for the series constituted informed consent for receipt of educational messages and analysis of deidentified data. Messages were delivered via a commercial SMS platform (SimpleTexting); no protected health information was collected. Survey and evaluation data were deidentified and stored on secure institutional servers with aggregate reporting. Participants received no compensation. No images identifying individual participants are included.

Results

A total of 34 individuals signed up for the series, with 22 completing the initial questionnaire on their confidence treating tobacco use disorder and subsequently enrolling in the intervention. The participants included providers at all levels of training, including attending physicians (n=4, 18%) and graduate medical education trainees (n=18, 82%) (Table 2). Before the intervention, the average confidence level in managing tobacco use disorder across all training levels was 60 (SD 16). All enrolled participants completed the series, and there were no unsubscribes. At the conclusion of the series, we received responses on confidence level from 18 participants along with comments and feedback. The average confidence level increased to 85 (SD 10). The paired *t* test revealed a statistically significant difference in confidence levels before and after the series (t_{17}=−10.71; *P*<.001).

The reference links provided in each message were accessed 67% (178/266) of the time, indicating active engagement with the additional resources. Feedback from the participants was predominantly positive. Common themes included appreciation for the concise and informative nature of the text series and the ease of learning they provided. One participant noted, "Text-based format made it easy to learn in an efficient manner," while another remarked, "This series absolutely increased my comfort level with treating tobacco use disorder." Suggestions for improvement included adding visual aids (Textbox 1).

Table . Breakdown of individual participant feedback from the text-based education series.

Subject	Department	Level of training	Preintervention confidence score	Postintervention confidence score
1	IM[a]	1	50	70
2	PCCM[b]	4	50	70
3	PCCM	4	40	75
4	IM	2	30	No response
5	PCCM	5	80	100
6	IM	1	25	75
7	PCCM	5	40	80
8	PCCM	5	83	100
9	PCCM	6	60	85
10	PCCM	6	75	90
11	PCCM	4	50	85
12	IM	3	65	No response
13	IM	2	70	No response
14	IM	2	60	90
15	PCCM	6	55	90
16	PCCM	4	60	85
17	IM	3	75	No response
18	PCCM	Attending	47	70
19	PCCM	6	75	90
20	PCCM	Attending	80	90
21	PCCM	Attending	60	90
22	PCCM	Attending	75	95

[a]IM: Internal Medicine.

[b]PCCM: Pulmonary and Critical Care Medicine.

Textbox 1. Feedback from participants and those who completed the text-based education series.

"Text-based format made it easy to learn in an efficient manner."

"This series absolutely increased my comfort level with treating tobacco use disorder."

"I pinned the text chain and went back to it when I had a question."

"The texts are awesome—very informative, short, and to the point."

"Add visual aids next time! Otherwise, no critique."

"Made me realize I was not using the correct dosing for nicotine! Thanks for teaching."

"This is so innovative! Don't stop, do another topic next!"

Discussion

This study aimed to examine whether SMS microlearning can strengthen clinicians' confidence in managing tobacco use disorder. Our findings indicate a meaningful increase in physician confidence after completing the intervention, and participants consistently described the messages as clear, concise, and useful at the point of care. Although these results come from a small cohort, the approach appears feasible for broader implementation and may serve as an effective educational tool; its impact on knowledge and clinical outcomes should be evaluated in larger, comparative studies.

Tobacco use remains a leading driver of preventable morbidity and mortality, and clinicians are central to diagnosis, counseling, and pharmacotherapy [9,10]. Yet tobacco treatment is underrepresented in many training programs [6,11-14], and persistent knowledge gaps have been documented across diagnostic criteria, counseling approaches, and evidence-based medications in diverse settings [3,6,13]. Tobacco treatment also receives limited emphasis on medical licensing and board

examinations, which can reduce curricular time and formal assessment [5]. We observed similar gaps locally in our formative survey, which is why we selected tobacco use disorder as the initial focus for this educational pilot. Targeting a high-impact, common condition with clear guideline-based management made it possible to craft concise teaching points and to judge whether a brief intervention could realistically support everyday practice.

The realities of clinical work limit traditional didactic learning. Long shifts, competing patient care and administrative demands, and variable schedules make it difficult to attend scheduled sessions or complete lengthy modules, and even when completed, retention can fade without reinforcement [15-17]. In this context, brief touchpoints that can be completed between tasks and revisited at need are more likely to fit daily workflow. Microlearning and spaced relearning operationalize this idea by delivering small, focused units with planned repetition, promoting retrieval when it matters and supporting transfer to practice [18-20]. For clinicians managing tobacco use disorder, this approach aligns with how information is actually used at the point of care and helps convert passive exposure into durable, actionable knowledge.

SMS microlearning aligns with clinical workflow: it is low friction, works on any device, and is easy to consume between tasks. Short, focused messages support spacing and retrieval, while links allow quick escalation to peer-reviewed resources at the point of care. Evidence from health professions education suggests this format is effective and well received: multicenter and randomized studies in obstetrics and gynecology residents report improved test performance and higher learner interest with SMS curricula, along with high satisfaction and low cost [21,22]; clinicians describe SMS teaching as engaging and practice-informing [23]; and related smartphone-based microlearning projects show strong acceptability and ease of use [24]. Not all findings are uniformly positive; some trials show no long-term advantage in exam scores without integration into broader study routines [25], which underscores the need for comparative designs and objective outcomes. In our pilot, the goal was to strengthen confidence rather than to test knowledge, and our results align with the broader literature on feasibility and acceptability. These considerations motivate a careful appraisal of limitations, including the absence of objective knowledge outcomes and comparative designs.

Our study has several limitations. It was a single-arm, pre-post pilot that evaluated self-reported confidence as the prespecified primary outcome. We did not administer an objective knowledge test; the preintervention survey was a formative needs assessment to guide content rather than a validated outcome instrument. Without a comparison group, we cannot separate the effect of message content from nonspecific influences such as measurement reactivity, social desirability, or Hawthorne effects. Selection bias is likely because participation was voluntary and only a subset of enrollees provided paired ratings, which limits generalizability beyond motivated clinicians at a single academic center. The sample was small, we assessed outcomes only immediately after the series, and engagement was measured by link click-throughs, which are a proxy and do not confirm content review or practice change. Optional free-text feedback may overrepresent highly engaged participants. Results may not generalize to settings without reliable SMS access or to clinicians with different baseline training, and the study was not powered for subgroup comparisons. Future work should include a validated knowledge assessment aligned to the message set, randomization to alternative content or wait-list control, larger and more diverse samples with strategies to reach less-engaged clinicians, and follow-up to assess durability and clinical impact.

In conclusion, a brief text message–based microlearning series was associated with a substantial increase in clinicians' confidence to manage tobacco use disorder, with strong completion and evidence of engagement. The format was practical and acceptable for busy clinicians, offering concise, point-of-care reinforcement. These results should be interpreted cautiously given the single-arm design, self-selection, and reliance on self-reported confidence rather than objective knowledge or clinical outcomes. Future studies should incorporate a validated knowledge assessment, a randomized comparison, broader sampling, and follow-up to evaluate durability and impact on patient care.

Funding

The authors did not receive any funding for this research.

Data Availability

All deidentified data underlying this study (pre-post confidence ratings and message-level engagement), the full SMS message set, and the analysis script are provided in the manuscript and supplementary files. No access restrictions apply (IRB protocol 32174). Additional materials are available from the corresponding author upon request.

Authors' Contributions

ZD: conceptualization; data curation; formal analysis; investigation; methodology; visualization; writing – original draft; writing – review and editing.
VD: conceptualization; data curation; writing – review and editing.
JG: conceptualization; supervision; writing – review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The survey used in this study that was specifically created for this research.
[PDF File, 74 KB - mededu_v11i1e73821_app1.pdf]

Multimedia Appendix 2
Text messages and references sent to the participants.
[DOCX File, 21 KB - mededu_v11i1e73821_app2.docx]

References

1. Cornelius ME, Loretan CG, Jamal A, et al. Tobacco product use among adults - United States, 2021. MMWR Morb Mortal Wkly Rep 2023 May 5;72(18):475-483. [doi: 10.15585/mmwr.mm7218a1] [Medline: 37141154]
2. Ekpu VU, Brown AK. The economic impact of smoking and of reducing smoking prevalence: review of evidence. Tob Use Insights 2015;8:1-35. [doi: 10.4137/TUI.S15628] [Medline: 26242225]
3. Golden T, Courtney-Long E, VanFrank B. Healthcare providers' knowledge of evidence-based treatment for tobacco dependence, DocStyles 2020. Am J Health Promot 2024 Mar;38(3):316-324. [doi: 10.1177/08901171231202626] [Medline: 37731286]
4. Ngee Ling BJ, Cheong AT, Manap AHA. Factors influencing the practice of Smoking Cessation Assessment and Management among Primary Care Doctors (SCAAM-DOC) in three districts of Malaysia. PLoS ONE 2022;17(9):e0274568. [doi: 10.1371/journal.pone.0274568] [Medline: 36174083]
5. Melzer AC, Reese ZA, Mascarhenas L, et al. Education for tobacco use disorder treatment: current state, evidence, and unmet needs. ATS Sch 2023 Dec;4(4):546-566. [doi: 10.34197/ats-scholar.2022-0131RE] [Medline: 38196686]
6. Domalaon K, Valerio C. Family physician education and barriers to promoting smoking cessation: a multiservice investigation. Mil Med 2025 Jun 30;190(7-8):e1356-e1360. [doi: 10.1093/milmed/usae429] [Medline: 39316387]
7. Woods D, Attwell A, Ross K, Theron G. Text messages as a learning tool for midwives. S Afr Med J 2012 Jan 27;102(2):100-101. [doi: 10.7196/samj.5322] [Medline: 22310443]
8. Spohr SA, Nandy R, Gandhiraj D, Vemulapalli A, Anne S, Walters ST. Efficacy of SMS text message interventions for smoking cessation: a meta-analysis. J Subst Abuse Treat 2015 Sep;56:1-10. [doi: 10.1016/j.jsat.2015.01.011] [Medline: 25720333]
9. Rigotti NA, Kruse GR, Livingstone-Banks J, Hartmann-Boyce J. Treatment of tobacco smoking: a review. JAMA 2022 Feb 8;327(6):566-577. [doi: 10.1001/jama.2022.0395] [Medline: 35133411]
10. He H, Pan Z, Wu J, Hu C, Bai L, Lyu J. Health effects of tobacco at the global, regional, and national levels: results from the 2019 Global Burden of Disease Study. Nicotine Tob Res 2022 Apr 28;24(6):864-870. [doi: 10.1093/ntr/ntab265] [Medline: 34928373]
11. Jaffe GA, Brodsky BS, Buckley J, et al. Best practices for creating an addiction curriculum within family medicine residency programs: a qualitative analysis of expert opinion. Fam Med 2025 Jun;57(6):430-434. [doi: 10.22454/FamMed.2025.196843] [Medline: 40663417]
12. Richardson C, Daniels K, Confer A, et al. Internal medicine resident addiction training at the Veteran's Health Administration: a qualitative evaluation of site directors' response to the 2022 ACGME requirements. J GEN INTERN MED 2024 Jun;39(8):1393-1399. [doi: 10.1007/s11606-024-08639-4] [Medline: 38302815]
13. Ye L, Goldie C, Sharma T, et al. Tobacco-nicotine education and training for health-care professional students and practitioners: a systematic review. Nicotine Tob Res 2018 Apr 2;20(5):531-542. [doi: 10.1093/ntr/ntx072] [Medline: 28371888]
14. Clinical Practice Guideline Treating Tobacco Use and Dependence 2008 Update Panel, Liaisons, and Staff. A clinical practice guideline for treating tobacco use and dependence: 2008 update. A U.S. Public Health Service report. Am J Prev Med 2008 Aug;35(2):158-176. [doi: 10.1016/j.amepre.2008.04.009] [Medline: 18617085]
15. Shea JA, Weissman A, McKinney S, Silber JH, Volpp KG. Internal medicine trainees' views of training adequacy and duty hours restrictions in 2009. Acad Med 2012 Jul;87(7):889-894. [doi: 10.1097/ACM.0b013e3182582583] [Medline: 22622211]
16. Butler DJ, Brocato J, Yeazel M. Family medicine didactics revisited. Fam Med 2017 Nov;49(10):778-784. [Medline: 29190403]
17. Chaiyachati KH, Shea JA, Asch DA, et al. Assessment of inpatient time allocation among first-year internal medicine residents using time-motion observations. JAMA Intern Med 2019 Jun 1;179(6):760-767. [doi: 10.1001/jamainternmed.2019.0095] [Medline: 30985861]
18. Manning KD, Spicer JO, Golub L, Akbashev M, Klein R. The micro revolution: effect of Bite-Sized Teaching (BST) on learner engagement and learning in postgraduate medical education. BMC Med Educ 2021 Jan 21;21(1):69. [doi: 10.1186/s12909-021-02496-z] [Medline: 33478475]

19. Monib WK, Qazi A, Apong RA. Microlearning beyond boundaries: a systematic review and a novel framework for improving learning outcomes. Heliyon 2025 Jan 30;11(2):e41413. [doi: 10.1016/j.heliyon.2024.e41413] [Medline: 39882484]
20. Phillips JL, Heneka N, Bhattarai P, Fraser C, Shaw T. Effectiveness of the spaced education pedagogy for clinicians' continuing professional development: a systematic review. Med Educ 2019 Sep;53(9):886-902. [doi: 10.1111/medu.13895] [Medline: 31144348]
21. Cai F, Santiago S, Southworth E, et al. The #ObGynInternChallenge: reach, adoption, implementation, and effectiveness of a microlearning SMS-distributed curriculum. Acad Med 2023 Aug 1;98(8):917-921. [doi: 10.1097/ACM.0000000000005206] [Medline: 36917104]
22. Alipour S, Moini A, Jafari-Adli S, Gharaie N, Mansouri K. Comparison of teaching about breast cancer via mobile or traditional learning methods in gynecology residents. Asian Pac J Cancer Prev 2012;13(9):4593-4595. [doi: 10.7314/apjcp.2012.13.9.4593] [Medline: 23167386]
23. Song A, Safdieh JE, Robbins MS. Group text messaging as a residency teaching tool in outpatient neurology and headache: a mixed-methods observational study. Headache 2025 Apr;65(4):539-544. [doi: 10.1111/head.14870] [Medline: 39601119]
24. Mughal NA, Atkins ER, Morrow D, Al-Jundi W. Smartphone learning as an adjunct to vascular teaching - a pilot project. BMC Med Educ 2018 Mar 15;18(1):37. [doi: 10.1186/s12909-018-1148-8] [Medline: 29544474]
25. Gill CJ, Le Ngoc B, Halim N, et al. The mCME project: a randomized controlled trial of an SMS-based continuing medical education intervention for improving medical knowledge among Vietnamese community based physicians' assistants. PLoS ONE 2016;11(11):e0166293. [doi: 10.1371/journal.pone.0166293] [Medline: 27861516]

Edited by A Stone, T Leung; submitted 12.03.25; peer-reviewed by E Ogut, P Callas; accepted 12.11.25; published 09.12.25.

Please cite as:
Dhanani Z, Dronamraju V, Garfield J
Text Message (SMS) Microlearning for Tobacco Use Disorder: Pre-Post Pilot Study of Clinician Confidence
JMIR Med Educ 2025;11:e73821
URL: https://mededu.jmir.org/2025/1/e73821
doi:10.2196/73821

Distance Learning During the COVID-19 Lockdown and Self-Assessed Competency Development Among Radiology Residents in China: Cross-Sectional Survey

Peicheng Wang[1,2*], MSc; Ziye Wu[1*], MPH; Jingfeng Zhang[3], MD; Yanrong He[1], MPH; Maoqing Jiang[3], MD; Jianjun Zheng[3], MSc; Zhenchang Wang[4], MD; Zhenghan Yang[4], MD; Yanhua Chen[1,2,5], PhD; Jiming Zhu[1,6], DPhil

[1]Vanke School of Public Health, Tsinghua University, Haidian District, Beijing, China

[2]School of Medicine, Tsinghua University, Beijing, China

[3]Department of Radiology, Ningbo No. 2 Hospital, Ningbo, China

[4]Department of Radiology, Beijing Friendship Hospital, Capital Medical University, Beijing, China

[5]Department of Public Health, Policy and Systems, University of Liverpool, Liverpool, United Kingdom

[6]Institute for Healthy China, Tsinghua University, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Jiming Zhu, DPhil
Vanke School of Public Health, Tsinghua University, Haidian District, Beijing, China

Abstract

Background: During the COVID-19 lockdown, it was difficult for residency training programs to conduct on-site, hands-on training. Distance learning, as an alternative to in-person training, could serve as a viable option during this challenging period, but few studies have assessed its role.

Objective: This study aims to investigate the impact of distance learning during the lockdown on residents' self-assessed competency development and to explore the moderating effect of poor mental health on the associations. It is hypothesized that radiology residents who were trained through distance learning during the lockdown were more likely to report higher self-assessed competency compared to those who did not receive organized, formal training.

Methods: A cross-sectional survey was conducted in 2021 among all of the radiology residents in 407 radiology residency programs across 31 provinces of China. To estimate the long-term outcomes of radiology residents' training after the initial COVID-19 outbreak, this study measured 6 core competencies developed by the US Accreditation Council for Graduate Medical Education reported by radiology residents. Multiple linear regression and moderating effect analysis were conducted to examine the associations between distance learning, mental health status, and self-assessed competencies. Mental health status moderated the association between distance learning and self-assessed competency of radiology residents.

Results: A total of 2381 radiology residents (29.7% of the 8,008 nationwide) met the inclusion criteria and were included in the analysis. Among them, 71.4% (n=1699) received distance learning during the COVID-19 lockdown, and 73.2% (n=1742) reported mental health struggles ranging in severity from slight to extremely severe. Radiology residents who were trained through distance learning (β=0.35, 90% CI 0.24 - 0.45) were more likely to report higher self-assessed competencies. This was particularly true for the competency of "interpersonal and communication skills" (β=0.55, 90% CI 0.39 - 0.70). Whereas, the competency of "patient care and technical skills" (β=0.14, 90% CI 0.01 - 0.26) benefited the least from distance learning. Poor mental health significantly moderated the relationship between distance learning and competency (β=−0.15, 90% CI −0.27 to −0.02).

Conclusions: Distance learning, a means of promoting enabling environments during the COVID-19 lockdown, serves its purpose and helps generally improve residents' self-assessed competencies, though different competency domains benefit unequally. The impact of mental health status calls for special attention so that distance learning can fulfill its potential.

(JMIR Med Educ 2025;11:e54228) doi:10.2196/54228

KEYWORDS

radiology residents; distance learning; mental health status; self-assessed competency; ACGME competencies; Accreditation Council of Graduate Medical Education

Introduction

The COVID-19 pandemic threatens the health of people globally and has brought unprecedented pressure to health systems [1,2]. The national public health system plays a vital role in fighting against pandemics by taking measurements such as surveillance and epidemiological investigations [3], case finding and management [4], and collective quarantine of close contacts [5]. However, potential challenges including insufficient alerts, low efficiency of reporting to higher authorities, and workforce shortages still exist [1]. Of note, the training of health care providers and the improvement of their professional skills have been underscored for their great significance in medical service delivery and health systems resilience [6-8].

Residency training systems serve the purpose of cultivating a qualified health workforce [9]. Standardized residency training (SRT) was initiated in 2013 in China, aiming to train doctors to meet the needs of population health [10]. With the increasing trend of competency-based medical training in global medical education, the assessment of competencies has gained ground in practice [11]. The US Accreditation Council of Graduate Medical Education (ACGME) identified 6 core competencies for physicians (ie, patient care [PC], medical knowledge [MK], system-based practice [SBP], practice-based learning and improvement [PBLI], professionalism [PROF], and interpersonal communication skill [ICS]) [11] and implemented milestones by the Next Accreditation System initiative in July 2013 [12]. Residency education and competency-based practice assessed by the milestones are common requirements of ACGME and have been used in residency training in China [9,13].

COVID-19 has changed medical education dramatically, especially during the lockdown period. The impacts of COVID-19 on medical education in radiology, surgery, and emergency medicine have gained attention [14-16]. Radiology is related to other medical specialties and all levels of health care delivery [17]. Radiology residents were typically required to rotate between different specialties to obtain knowledge and clinical skills [18]. The mandatory social distancing challenged the traditional training on radiology trainee approaches such as teaching at workstations [19]. In China, SRT in radiology spans 3 years and involves workstation-based training throughout rotations in various specialties and departments [9]. In the first year, residents undergo rotations in the departments of radiology, ultrasound medicine, nuclear medicine, pathology, and relevant clinical departments. In the second and third years, they receive advanced rotational training within radiology subspecialties such as computed tomography, magnetic resonance imaging, x-ray, and interventional radiology [9]. Due to COVID-19, the mode of residency training has been switched from traditional in-person classes to distance learning [20], posing challenges to the effectiveness of rotational training and the developing competency of residents.

Numerous benefits have been found for distance learning. For instance, residents can schedule more flexibly and access the courses more easily [21]. They can learn at their own pace with the help of recorded lectures and communicate with professionals and peers on the web at their own convenience [22]. The positive acceptance and a higher level of satisfaction with distance learning have been reported by residents in Canada and the United States [21,23]. However, the practice of digital readout in distance learning is similar to the experience of in-person reading, in addition to the difficulties of gauging body language during practical operations or in the use of medical instruments [24], which may lead to unsatisfactory outcomes in radiology education. Meanwhile, the COVID-19 pandemic has impacted the mental health status of health care professionals dramatically [25]. Students who experienced distance learning during the pandemic had been found to have psychological distress [26,27]. According to the Job Demands-Resources model, mental health is a personal resource that helps residents to deal with job challenges by moderating their performance [28-31]. Accordingly, it is of great importance to take care of the mental health of residents who have experienced distance learning [32]. To date, during the COVID-19 lockdown, when workstation-based training was difficult to deliver, the role of distance learning remains unclear. It is also uncertain whether psychological status affects the effectiveness of distance learning.

In sum, the COVID-19 lockdown had brought substantial challenges to radiology training programs, which had to transition from face-to-face instruction to remote learning. In the meantime, mental distress caused by factors such as social distancing may make residency training rather difficult. Given that distance learning remains a primary alternative when traditional teaching is not feasible (such as during pandemic outbreaks and lockdowns), yet few studies have explored its effects, we aimed to investigate the impact of distance learning on the development of self-assessed competencies as well as the moderating effect of mental health status. We hypothesize that radiology residents who received distance learning during the COVID-19 lockdown were more likely to report higher self-assessed competencies compared to those who did not receive organized, formal training during the same period (ie, nondistance learners) and that this association was moderated by poor mental health. To test this, we used a nationwide survey dataset of radiology residents in China, which collected information on distance learning and mental health status during the lockdown (January-May 2020), and self-assessed competencies 6 months later. Previous studies have shown a strong positive correlation between the assessments by Clinical Competency Committees and residents' self-evaluations using the milestones. This suggests that residents are generally able to accurately assess their own competencies, which in turn supports the validity of using milestone assessments as an effective measure of self-assessed competency in this study [33].

Methods

Ethical Considerations

The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the institution review board of Tsinghua University, China (20210140). Informed consent was obtained before the research started, and data were deidentified to ensure participant privacy. Participation in this

survey was voluntary, participants did not receive any incentives to take part in the study.

Study Setting and Population

A nationwide retrospective cross-sectional survey was conducted on the web by the Chinese Association of Radiologists (CAR) during December 1, 2020, to April 30, 2021, targeting all the radiology residents in 407 radiology residency programs across 31 provinces of China.

To complete the distribution of the questionnaire and to ensure the participation of the radiology residents, we contacted the directors of the targeted hospitals' radiology departments by email or telephone to inform them of the purpose and details of the survey initiated by the CAR. The directors were then instructed to share the link of the questionnaire posted on the popular web-based survey platform "Wenjuanxing" with radiology residents. Anonymous responses to the questionnaire were submitted. All participants were informed that the questionnaire could only be filled out once, that participation was completely voluntary, and that they could withdraw at any time without penalty. During the 5-month survey period, the research team cooperated well with the CAR for the monitoring of the participation. If there was a low rate of submission, the CAR would require the hospital to improve the quality and quantity of their survey response. This proactive approach could help increase the number of responses and the representativeness of the population. Residents who did not undergo residency training during the COVID-19 epidemic between January and May 2020 would be excluded.

Measures and Outcomes

Distance Learning

Distance learning was asked "Was your training institution changed the teaching arrangements to the form of distance learning to reduce the negative impact during the initial COVID-19 outbreak?" The response options were either "yes or no." In contrast, nondistance learning participants (ie, nondistance learners) were those who did not receive organized, formal training during the same period. They primarily stayed at home on leave, reported daily health status, and engaged in delayed teaching plans or self-directed learning.

Mental Health Status

Mental health status was measured by the question "Did you suffer psychological distress during the initial COVID-19 outbreak? (from January to May 2020 in China)." A 5-point Likert scale was used to measure the degree of mental health struggles: 1=no impact, 2=mild impact, 3=moderate impact, 4=severe impact, and 5=extremely severe impact. Radiology residents can choose any rating between 1 and 5 (single choice). The variable is correlated with long-term mental health (depression and burnout) measured by the Depression and Anxiety Stress Scale—Depression and Maslach Burnout Inventory scales (Multimedia Appendix 1). The Cronbach α reliability coefficient for depression and burnout was 0.930 and 0.957, respectively.

Self-Assessed Competency

Milestone-based assessment of competencies for residence is one of the common requirements of the ACGME [11]. Self-assessment plays a key role in this process by fostering reflection on professional actions, identifying learning needs, and enabling residents to develop and refine personalized improvement plans [34]. Moreover, residents' self-assessments showed a strong alignment with the Clinical Competency Committee evaluations across postgraduate year levels [33,35,36]. To estimate the long-term outcomes of radiology residents' training after the initial COVID-19 outbreak, our study measured 6 core competencies developed by the ACGME that were assessed by radiology residents themselves. As is suggested by the experts, we selected 9 subcompetencies from diagnostic radiology milestones to represent the 6 ACGME core competencies: 2 PC subcompetencies, 2 MK subcompetencies, 2 PBLI subcompetencies, 1 SBP subcompetency, 1 PROF subcompetency, and 1 ICS subcompetency.

A dedicated section of the questionnaire is designed to assess 9 subcompetencies with 9 single-choice questions. Radiologists are able to select a score ranging from 0 to 9 for each competency. Examples of milestone sets for each subcompetency are shown in Figure 1. The primary outcome was self-evaluation milestone (SEM) scores (range 0‐9 scores) for 9 subcompetencies and the average SEM scores.

Figure 1. Milestone sets for patient care 1 (image interpretation) and professionalism (self-awareness and help-seeking).

(Clinical context) Patient care 1: image interpretation				
Level 1	Level 2	Level 3	Level 4	Level 5
Identifies primary imaging findings	Identifies secondary and critical imaging findings and formulates differential diagnoses	Prioritizes differential diagnoses and recommends management options	Provides a single diagnosis with integration of current guidelines to recommend management, when appropriate	Demonstrates expertise and efficiency at a level expected of a subspecialist
Scores: 1 2 3 4 5 6 7 8 9				
Comments:			Not yet completed level 1 ☐ Not yet assessable ☐	
(Nonclinical context) Professionalism: self-awareness and help-seeking				
Level 1	Level 2	Level 3	Level 4	Level 5
Recognizes status of personal and professional well-being, with assistance, and is aware of available resources Recognizes limits in the knowledge or skills of self or team, with assistance	Independently recognizes status of personal and professional well-being using available resources when appropriate Independently recognizes limits in the knowledge or skills of self or team and demonstrates appropriate help-seeking behaviors	With assistance, proposes a plan to optimize personal and professional well-being With assistance, proposes a plan to remediate or improve limits in the knowledge or skills of self or team	Independently develops a plan to optimize personal and professional well-being Independently develops a plan to remediate or improve limits in the knowledge or skills of self or team	Coaches others when emotional responses or limitations in knowledge or skills do not meet professional expectations
Scores: 1 2 3 4 5 6 7 8 9				
Comments:			Not yet completed level 1 ☐	

Sociodemographic Characteristics

The sociodemographic information included age (≤27 or >27 years), sex (male or female), educational level (bachelor's degree or master's degree or above), training year (the second year or the third year), training sites level (grade-a tertiary general hospital, grade-a tertiary specialized hospital, grade-b tertiary general hospital, or others), undergraduate major (clinical medicine, medical imaging, or others), working hours per week (≤40, 40 - 48, or >48), annual after-tax income in 2020 (analyzed as continuous variable), and types of residents (professional master or nonprofessional master).

Statistical Analysis

The SEM scores of radiology residents for 9 subcompetencies were reported by means and SDs). The differences in SEM scores of residents between distance learning and nondistance learning were compared using independent samples 2-tailed *t* test. To explore the association between distance learning and competencies, multiple linear regression (MLR) models were constructed. The dependent variables were SEM scores of 9 subcompetencies, and the key explanatory variable was distance learning. The moderating effect of mental health on distance learning was explored by the MLR model. The significance of the moderating effect was tested by simple slope analysis. All models were controlled for participants' characteristics. A variance inflation factor was used to detect the multicollinearity of independent variables for all models (variance inflation factor scores <3). A *P* value of <.05 was considered statistically significant for 2-tailed tests (*t* test or chi-square test). A conservative level of *P* value of <.10 was used to assess potential moderators in the regression, which was reported by a coefficient (β) and 90% CIs [37,38]. All statistical analyses were performed by STATA (version 17.0; StataCorp LLC).

Results

Participants' Characteristics

Of the 8008 targeted radiology residents, 2381 (overall effective response rate: 29.7%) participated in this survey (Figure 2). As is shown in Table 1, the mean age of the participants was 27.8 (SD 2.4) years. In total, 58.5% (n=1392) of them were female, and 50.5% (n=1202) were in the third-year training. The majority of the participants received training in a grade-a tertiary hospital (n=2310, 97%), had a bachelor's degree (n=2187, 91.9%), and their undergraduate major was medical imaging (n=2016, 84.7%). The median annual after-tax income was 40,000 RMB (IQR 10,000 - 60,000; a currency exchange rate of 1 RMB=US $0.145 is applicable), with 25.8% (n=614) of them earning more than 60,000 RMB (about US $ 8698.9). The average working hours per week was 44.3 (SD 12.5) hours, and 23.1% (n=551) of the participants worked more than 48 hours per week. During the initial COVID-19 outbreak from January 2020 to May 2020, 71.4% (n=1699) of the radiology residents participated in distance learning, and 73.2% (n=1742) of them reported slight or severe mental health struggles. In total, 35.8% (n=853) of the participants contributed to the prevention and control of COVID-19.

Figure 2. Flowchart of the study design. ACGME: Accreditation Council for Graduate Medical Education; SRT: standardized residency training.

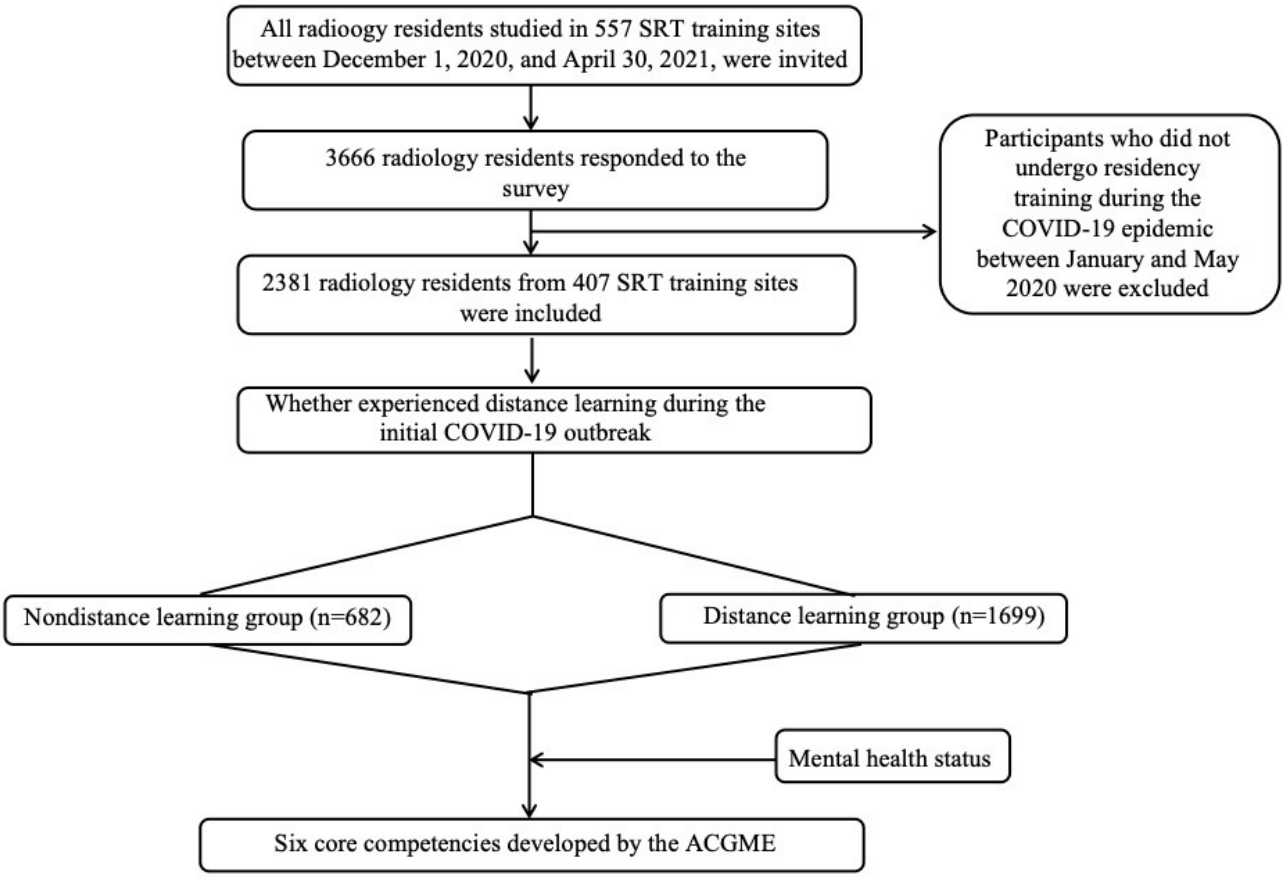


Table . Characteristics of participants in China.

Variables	Total (N=2381)	Distance learning		*P* value
		Yes (n=1699)	No (n=682)	
Region, n (%)				.36
East	945 (36.7)	663 (70.2)	282 (29.9)	
Central	496 (20.8)	358 (72.2)	138 (27.8)	
West	788 (33.1)	561 (71.2)	227 (28.8)	
Northeast	152 (6.4)	117 (77)	35 (23)	
Age (years), mean (SD)	27.8 (2.4)	27.8 (2.4)	27.7 (2.4)	.28
≤27, n (%)	1293 (54.3)	915 (70.8)	378 (29.2)	.49
>27, n (%)	1088 (45.7)	784 (72.1)	304 (27.9)	.49
Sex, n (%)				.02
Male	989 (41.5)	680 (68.8)	309 (31.2)	
Female	1392 (58.5)	1019 (73.2)	373 (26.8)	
SRT[a] training years, n (%)				.63
Second year	1179 (49.5)	836 (70.9)	343 (29.1)	
Third year	1202 (50.5)	863 (71.8)	339 (28.2)	
SRT sites level, n (%)				.34
Grade-A tertiary general hospital	2310 (97)	1649 (71.4)	661 (28.6)	
Grade-A tertiary specialized hospital	51 (2.1)	39 (76.5)	12 (23.5)	
Grade-B tertiary general hospital	15 (0.6)	8 (53.3)	7 (46.7)	
Others	5 (0.2)	3 (60)	2 (40)	
Education level, n (%)				.55
Bachelor’s degree	2187 (91.9)	1557 (71.2)	630 (28.8)	
Master’s or doctoral degree	194 (8.2)	142 (73.2)	52 (26.8)	
Undergraduate major, n (%)				.12
Clinical medicine	346 (14.5)	238 (68.8)	108 (31.2)	
Medical imaging	2016 (84.7)	1444 (71.6)	572 (28.4)	
Others	19 (0.8)	17 (89.5)	2 (10.5)	
Type of residents, n (%)				.007
Professional master	774 (32.5)	580 (74.9)	194 (25.1)	
Nonprofessional master	1607 (67.5)	1119 (69.6)	488 (30.4)	
Annual after-tax income (RMB[b]), median (IQR)	43,800 (10,000-67,000)	42,700 (9600-60,000)	46,600 (10,000-70,000)	.02
≤10,000, n (%)	691 (29)	513 (74.2)	178 (25.8)	.12
10,001 - 40,000, n (%)	565 (23.7)	400 (70.8)	165 (29.2)	.12
40,001 - 60,000, n (%)	511 (21.5)	367 (71.8)	144 (28.2)	.12
>60,000, n (%)	614 (25.8)	419 (68.2)	195 (31.8)	.12
Working hours per week (hours), mean (SD)	44.3 (12.5)	43.9 (11.8)	45.3 (14.0)	.01
≤40, n (%)	1311 (55.1)	948 (72.3)	363 (27.7)	.43
41 - 48, n (%)	519 (21.8)	369 (71.1)	150 (28.9)	.43

Variables	Total (N=2381)	Distance learning		*P* value
		Yes (n=1699)	No (n=682)	
>48, n (%)	551 (23.1)	382 (69.3)	169 (30.7)	.43
Mental health impact during the initial COVID-19 outbreak, n (%)				<.001
No impact	639 (26.8)	469 (73.4)	170 (26.6)	
Mild impact	1249 (52.5)	903 (72.3)	346 (27.7)	
Moderate impact	397 (16.7)	276 (69.5)	121 (30.5)	
Severe impact	79 (3.3)	45 (57)	34 (43)	
Extremely severe impact	17 (0.7)	6 (35.3)	11 (64.7)	
COVID-19–related work participation, n (%)				.90
Yes	853 (35.8)	610 (71.5)	243 (28.5)	
No	1528 (64.2)	1089 (71.3)	439 (28.7)	

[a]SRT: standardized residency training.

[b]A currency exchange rate of 1 RMB=US $0.145 is applicable.

SEM Scores of Radiology Residents Between Distance Learning and Nondistance Learning

The mean score of competencies and a comparison of subcompetencies scores were presented in Table 2. The overall average score of radiology residents' competency was 3.37 (SD 1.47). The average score of participants who received distance learning was 3.46 (SD 1.49), higher than those who did not (mean 3.13, SD 1.39; *P*<.001). Residents who received distance learning outperformed in all subcompetencies than those without distance learning during the initial COVID-19 outbreak (PC-1: *P*=.007; MK-1: *P*=.004; and others: *P*<.001), except for PC-2 (*P*=.09; Table 2). Due to the low response rate, Mann-Whitney tests were performed. The results were similar (Multimedia Appendix 2). For radiology residents who had not participated in COVID-19–related activities (1528/2381; Multimedia Appendix 3), the differences between residents' competencies or subcompetencies showed the same trends, except for the PC-1 (P=.10).

Table . Self-evaluation milestone scores for radiology residents between distance learning and nondistance learning.

Diagnostic radiology sub-competencies	Total (N=2381), mean (SD)	Distance learning		*P* value
		Yes (n=1699), mean (SD)	No (n=682), mean (SD)	
PC[a]				
PC-1: image interpretation	3.90 (1.69)	3.96 (1.68)	3.75 (1.72)	.007
PC-2: competence in procedures	2.25 (1.77)	2.28 (1.81)	2.16 (1.65)	.09
MK[b]				
MK-1: diagnostic knowledge	3.75 (1.75)	3.82 (1.76)	3.59 (1.73)	.004
MK-2: imaging technology and image acquisition	3.53 (1.90)	3.62 (1.92)	3.29 (1.81)	<.001
SBP[c]				
SBP-1: system navigation for patient-centered care	2.86 (1.88)	2.96 (1.90)	2.61 (1.80)	<.001
SBP-2: contrast agent safety	3.57 (1.95)	3.71 (1.98)	3.21 (1.80)	<.001
PBLI[d]				
PBLI: evidence-based and informed practice	3.25 (1.84)	3.34 (1.86)	3.02 (1.78)	<.001
PROF[e]				
PROF: self-awareness and help-seeking	3.49 (1.90)	3.61 (1.92)	3.20 (1.83)	<.001
ICS[f]				
ICS: patient- and family-centered communication	3.72 (2.10)	3.87 (2.11)	3.33 (2.03)	<.001
Average (all subcompetencies)	3.37 (1.47)	3.46 (1.49)	3.13 (1.39)	<.001

[a]PC: patient care.

[b]MK: medical knowledge.

[c]SBP: system-based practice.

[d]PBLI: practice-based learning and improvement.

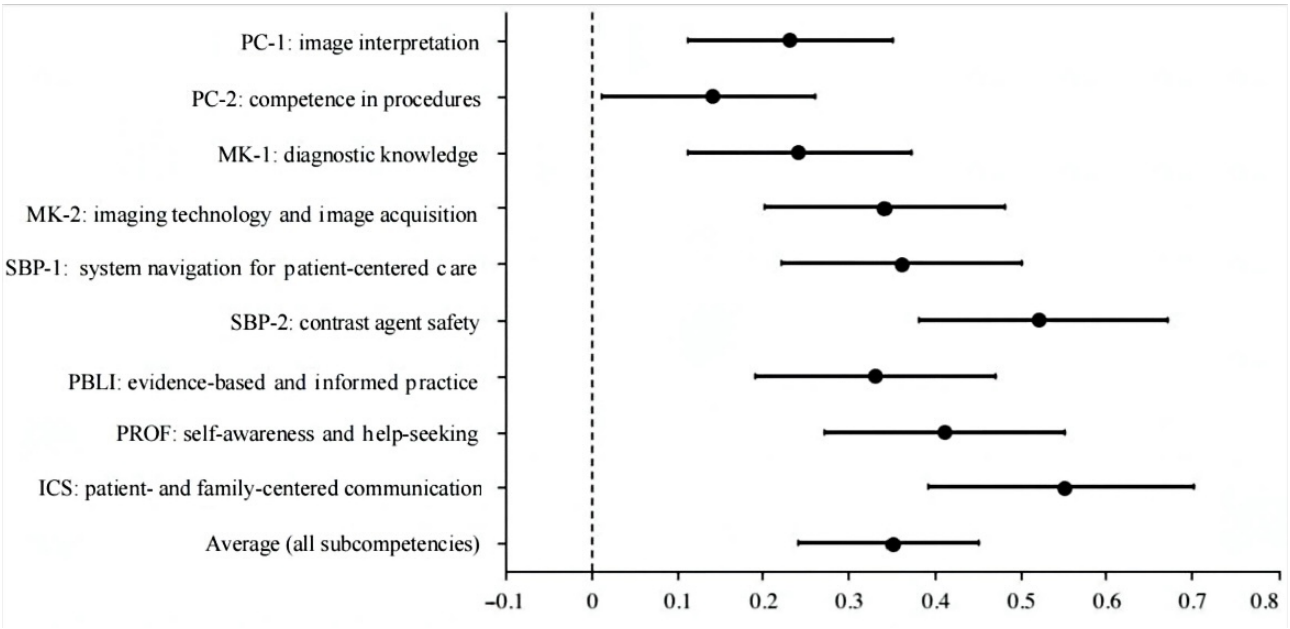
[e]PROF: professionalism.

[f]ICS: interpersonal communication skill.

Association Between Distance Learning and Competencies Among Radiology Residents

As is shown in Figure 3 (see also Multimedia Appendix 4), MLR analyses showed that radiology residents who were trained through distance learning were more likely to report high competencies (β=0.35, 90% CI 0.24 - 0.45) after adjusted by participants' characteristics, including age, sex, educational level, training years, working hours per week, annual after-tax income in 2020, and types of residents. This was particularly evident in the competencies of "interpersonal and communication skills" (β=0.55, 90% CI 0.39 - 0.70) and "contrast agent safety" (β=0.52, 90% CI 0.38 - 0.67). Whereas, the competency of "patient care and technical skills" benefited the least from distance learning (β=0.14, 90% CI 0.01 - 0.26). The effect of each explanatory variable on the overall average SEM score is shown in Multimedia Appendix 5. Factors associated with higher competencies included older age (>27 years: β=0.12, 90% CI 0.01 - 0.22), being male (β=0.28, 90% CI 0.18 - 0.37), and having a longer SRT training year (third year: β=0.51, 90% CI 0.41 - 0.61).

Figure 3. Associations between distance learning and competencies among residents. ICS: interpersonal communication skill; MK: medical knowledge; PBLI: practice-based learning and improvement; PC: patient care; PROF: professionalism; SBP: system-based practice.



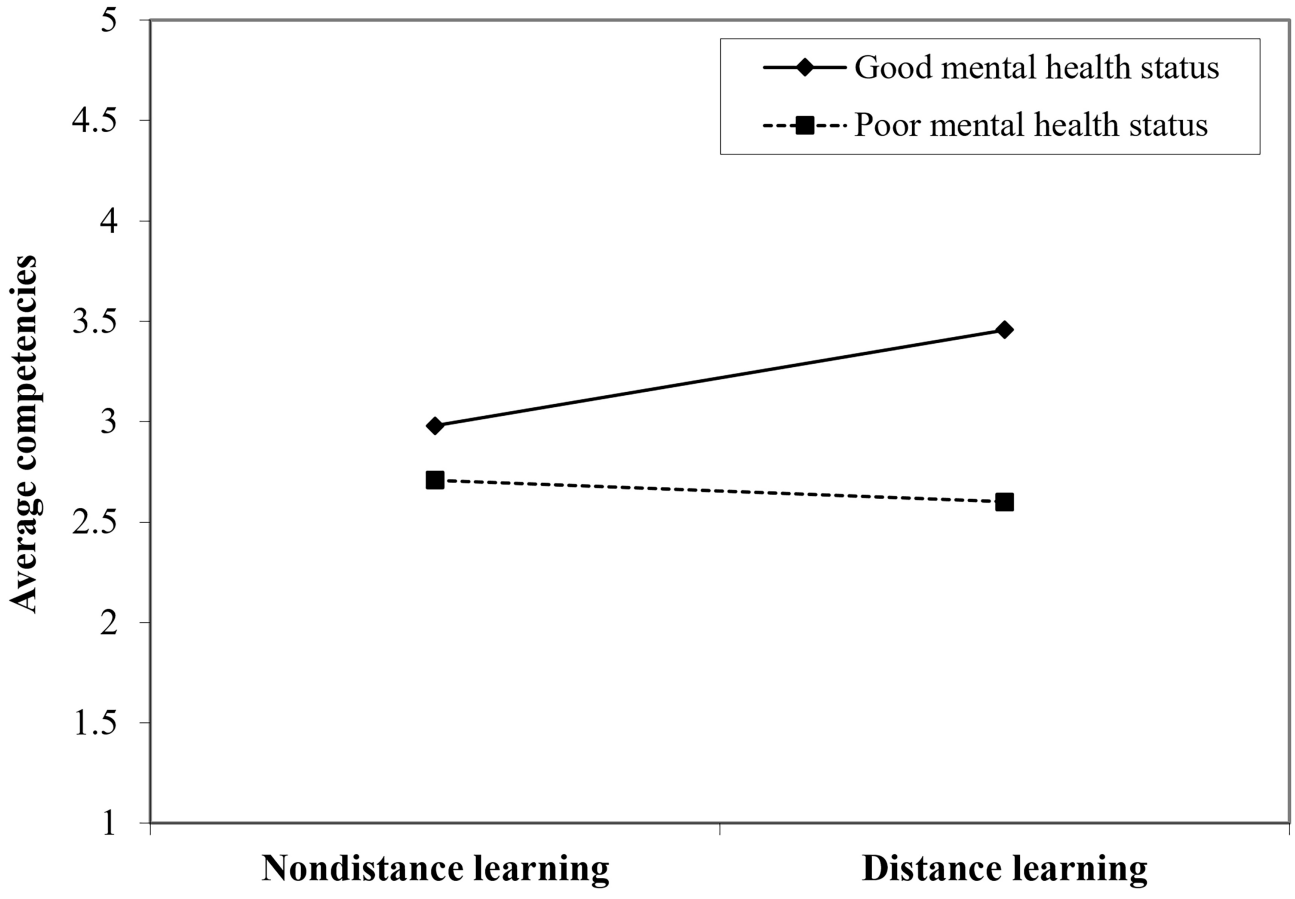
The Moderating Effect of Mental Health Status Between Distance Learning and Competencies

The association between mental health status and competencies is shown in Multimedia Appendix 6. We controlled the covariates mentioned earlier to investigate the moderating effect of mental health on the association between distance learning and competencies. As is shown in Table 3, there was a significant association between distance learning and radiology residents' competencies (β=0.63, 90% CI 0.34 - 0.91; *P*<.001) and was moderated by mental health status (β=–0.15, 90% CI −0.27 to −0.02; *P*=.06). The relationship between distance learning and competencies at low and high (mean−SD and mean + SD, respectively) mental health scores is shown in Figure 4. Poor mental health caused by the pandemic may offset the positive effect of distance learning on residents' competencies. The highest level of competencies was found in individuals who reported less mental distress and adopted distance learning. Furthermore, the moderating effect of poor mental health on 4 subcompetencies (ie, MK-1, MK-2, SBP-2, and ICS) was similar to it on total competencies (*P*<.10; Multimedia Appendix 7).

Table . The moderating effects of mental health status on the association between competencies and distance learning.

Variables	Multiple linear regression models		
	β (SE)	90% CI	*P* value
Distance learning	0.63 (0.17)	0.34 to 0.91	<.001
Mental health status	–0.07 (0.06)	–0.17 to 0.04	.28
Interaction (distance learning*mental health status)	–0.15 (0.08)	–0.27 to –0.02	.06
Age (years) (reference ≤27)			
>27	0.12 (0.06)	0.01 to 0.22	.07
Sex (reference=male)			
Female	–0.27 (0.06)	–0.37 to –0.17	<.001
Education (reference=bachelor's degree)			
Master's or doctoral degree	0.19 (0.11)	0.01 to 0.38	.10
Training year (reference=second year)			
Third year	0.51 (0.06)	0.41 to 0.61	<.001
Working hours per week (reference ≤40 hours per week)			
40 - 48	0.01 (0.07)	–0.12 to 0.13	.93
>48	0.01 (0.07)	–0.11 to 0.13	.86
Income	0.01 (0.01)	–0.01 to 0.03	.26
Type of residents (reference=nonprofessional master)			
Professional master	–0.03 (0.08)	–0.17 to 0.10	.70

Figure 4. The moderating effect of mental health on the relationship between distance learning and average competencies. Two lines are the visual representation of different moderation effects of distance learning on competencies when the mental health status is at its +1SD and –1SD.



Discussion

Principal Findings

Based on a national survey of radiology residents in China, our study found that radiology residents who received distance learning were more likely to report high proficiency in key competencies after 1 year. This was particularly true for learning knowledge and communication skills but was less evident in obtaining technical skills. In addition, we found a significant moderating effect of mental health status on the association between distance learning and competencies during COVID-19.

Web-based training programs are proposed to mitigate the loss of learning from clinical rotations during the pandemic [39,40]. Consistent with the results of previous findings that distance learning has the potential to improve learners' academic performance, skill development, and engagement [39,41], our study observed a positive impact of distance learning on radiology residents' self-assessed competencies. However, it should be noted that the impact of distance learning on professional competency varies among radiology residents. Our study enriches the understanding of distance learning by using milestones of 9 subcompetencies for 6 core competencies. We also assessed the long-term outcomes of distance learning among radiology residents in the initial stage of the COVID-19 pandemic (between January and May 2020). These findings can be used to help inform evidence-based policies to improve residency training in the future.

Knowledge gain is an important indicator of what trainees have learned during SRT training [42]. The radiology residents in our study reported a substantial gain of knowledge and communication skills during SRT, particularly in ICS competency and SBP competency that include clinical knowledge and medical humanities, with a regression coefficient >0.5. Whereas, the effect of distance learning on professional attitudes (PROF) and professional growth in clinical practice (PBLI) was moderate, with a regression coefficient between 0.3 and 0.4, which is in line with previous studies [43-45]. Distance learning was found to be supportive in the continuity of teaching and learning during COVID-19 [39], which might explain why residents who received distance learning reported a higher score in the competence of medical knowledge than nondistance learners. In addition, radiology residents are able to participate in web-based conferences and conduct digital medical consultations to enhance their competency of communication [43], though it could be hard to evaluate their body language in distance learning [46]. Our findings validated the positive role of distance learning in fulfilling the objectives of training by creating an enabling environment, especially in the domain of knowledge and communication. In other words, distance learning can help keep the consistency of residency programs during pandemics, facilitating the resilience and recovery of health systems.

Our findings showed that competence in procedures (PC-2) benefited the least from distance learning, followed by competencies in technical skills (including PC competencies that reflect radiology residents' professionalism and MK competencies that reflect mastery of professional knowledge and imaging technology). The result is consistent with previous findings that most medical students feel unable to acquire practical clinical skills through web-based teaching during COVID-19 [47]. This could be explained by the reduction in daily cases due to COVID-19, which was one of the deficiencies in distance learning [20]. To address this challenge, new technologies such as touchscreen Anatomage Table and Touch Surgery application have been used to strengthen trainings of plastic surgery [48]. These new approaches can be integrated into web-based residency training to compensate for the shortage of training in technical skills with distance learning. Other training programs, such as competence in procedures, could also be carried out among radiology residents in the post–COVID-19 era.

Health professional education involves maintaining a sense of purpose and mental well-being (eg, balance work life, stress, depression, burnout, and more) among residents [49-51]. Our study extended the prior results by identifying that poor mental health moderates the relationship between distance learning and competency among radiology residents. During the COVID-19 pandemic, physicians experienced more mental health disorders [52] with negative impacts (eg, higher turnover intention and lower work performance) [53,54]. However, good mental health may help trainees to receive distance learning consistently [55]. When trainees are in a suboptimal mental status, they tend to have negative attitudes toward learning and are discouraged from receiving distance learning continuously [55]. Health workforce is the backbone of health systems in response of pandemics [56]. In addition to the structural changes required for the improvement of health systems' resilience, it could also be necessary to provide long-term psychological support for the health workforce to help them overcome psychological distress [57]. In summary, several recommendations can be drawn from this study: (1) distance learning can be used for transferring knowledge to residents particularly when challenges exist in the traditional offline approaches; (2) clinical skills have a crucial role in offline training, which should be noted and well used by training institutions; (3) instructors could use a scientific and caring approach with a special focus on learners' psychological well-being when preparing material and organizing courses; and (4) physicians who have engaged in distance learning could enhance their learning experience through an enhancement of their offline learning environment. This promotes a better psychological state for learning to become more effective.

Health system resilience is critical in training the required competencies of the health workforce [49]. In addition to competency-based education, the COVID-19 pandemic highlighted the application of digital interprofessional education [58]. In the future, distance learning may help physicians gain knowledge and skills in public health, interprofessional communication, and teamwork. In this regard, remote residency training should be developed as a holistic educational concept rather than a mere substitution for traditional in-person learning. Distance learning enables physicians to have a flexible schedule, a feasible access to classes, and an opportunity to keep a good balance between work and life [14]. What is more, the facilitation of distance learning contributes to the sharing of

high-quality educational resources in the post–COVID-19 era. In particular, for health professionals in resource-constrained places (eg, rural areas in western China), the provision of high-quality distance learning could help them overcome geographical constraints and thus reduce inequalities in access to education [49,59]. Nevertheless, it is not possible to replace in-person teaching with distance learning completely. An integrated model of the 2 is encouraged to maximize the advantages of different teaching modalities.

Limitations

This study has several limitations. First, we used SEM for the self-evaluation of radiology residents' competencies; results may be influenced by the Dunning-Kruger effect, where participants with lower abilities tend to overestimate their competencies and the potential self-reporting bias [60]. Second, the lower response rate may be subject to selection bias, as our survey is voluntary in nature. However, based on sample size calculations [61], we obtained 1573 valid samples (precision=5%; baseline proportion=0.50), which accurately represent the characteristics of the residents. Importantly, this represents the largest nationally representative sample of radiology residents in China to date, which may help to minimize the potential bias. Third, although mental health status was asked by a single question based on the self-report psychological distress on a Likert scale, this variable is correlated with long-term mental health status (depression and burnout). Other potential moderating factors that may influence learning status, such as the design of the web-based course, courseware, and teaching styles, could be explored in future research. Fourth, the generalization of the results is another limitation of our study [49].

Future Directions

As distance learning is anticipated to be applied across various fields, additional research is warranted to substantiate our findings. Longitudinal studies are recommended for future research to fully assess the long-term effects of distance learning on competence and mental health.

Conclusions

Distance learning helps mitigate the negative impact of the COVID-19 lockdown on the education of health professionals. Meanwhile, attention should be paid to the disadvantages of distance learning and the mental health status of learners, as they may negatively influence the effectiveness and sustainability of distance learning. Our study provides insights into the role of distance learning in residency training during the pandemic.

Acknowledgments

The authors appreciate the time and effort of all the project staff and participants.

Data Availability

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

Authors' Contributions

J Zhu, PW, Z Wu, YC and ZY conceived the study and its design and were responsible for all aspects of the study. J Zhang, MJ, J Zheng, Z Wang, and ZY collected the data set. PW and Z Wu conducted the statistical analyses. PW, Z Wu, YC, and J Zhu drafted the paper. Z Wu, PW, YC, YH, and J Zhu were involved in manuscript preparation and revisions. ZY, YC, and J Zhu are corresponding authors. All authors have approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The Spearman *r* between short- and long-term mental health status.
[PDF File, 57 KB - mededu_v11i1e54228_app1.pdf]

Multimedia Appendix 2
Self-evaluation milestone scores for radiology residents between distance learning and nondistance learning.
[PDF File, 108 KB - mededu_v11i1e54228_app2.pdf]

Multimedia Appendix 3
Distance learning efforts for radiology residents who had not participated in COVID-19–related activities.
[PDF File, 96 KB - mededu_v11i1e54228_app3.pdf]

Multimedia Appendix 4
Associations between distance learning and competencies among residents.

[PDF File, 126 KB - mededu_v11i1e54228_app4.pdf]

Multimedia Appendix 5
Association between distance learning and the overall average self-evaluation milestone score among residents.
[PDF File, 96 KB - mededu_v11i1e54228_app5.pdf]

Multimedia Appendix 6
The Spearman *r* between mental health status and competencies.
[PDF File, 123 KB - mededu_v11i1e54228_app6.pdf]

Multimedia Appendix 7
The moderating effects of mental health on the association between subcompetencies and distance learning.
[PDF File, 119 KB - mededu_v11i1e54228_app7.pdf]

References

1. Zhang P, Gao J. Evaluation of China's public health system response to COVID-19. J Glob Health 2021 Jan 16;11:05004. [doi: 10.7189/jogh.11.05004] [Medline: 33643637]
2. Liu Q, Luo D, Haase JE, et al. The experiences of health-care providers during the COVID-19 crisis in China: a qualitative study. Lancet Glob Health 2020 Jun;8(6):e790-e798. [doi: 10.1016/S2214-109X(20)30204-7] [Medline: 32573443]
3. Xu TL, Ao MY, Zhou X, et al. China's practice to prevent and control COVID-19 in the context of large population movement. Infect Dis Poverty 2020 Aug 19;9(1):115. [doi: 10.1186/s40249-020-00716-0] [Medline: 32814591]
4. Li Z, Chen Q, Feng L, et al. Active case finding with case management: the key to tackling the COVID-19 pandemic. Lancet 2020 Jul 4;396(10243):63-70. [doi: 10.1016/S0140-6736(20)31278-2] [Medline: 32505220]
5. Dong C, Tian Y, Xu W, et al. Introduction on collective quarantine of close contacts of patients with COVID-19 for medical observation in China: from the perspective of frontline staff. Biosci Trends 2020 Jul 17;14(3):215-221. [doi: 10.5582/bst.2020.03094] [Medline: 32389941]
6. Bourgeault IL, Maier CB, Dieleman M, et al. The COVID-19 pandemic presents an opportunity to develop more sustainable health workforces. Hum Resour Health 2020 Oct 31;18(1):83. [doi: 10.1186/s12960-020-00529-0] [Medline: 33129313]
7. Kuhlmann E, Dussault G, Wismar M. Health labour markets and the "human face" of the health workforce: resilience beyond the COVID-19 pandemic. Eur J Public Health 2020 Sep 1;30(Suppl 4):iv1-iv2. [doi: 10.1093/eurpub/ckaa122] [Medline: 32949241]
8. Kruk ME, Myers M, Varpilah ST, Dahn BT. What is a resilient health system? Lessons from Ebola. Lancet 2015 May;385(9980):1910-1912. [doi: 10.1016/S0140-6736(15)60755-3]
9. Zhang J, Han X, Yang Z, et al. Radiology residency training in China: results from the first retrospective nationwide survey. Insights Imaging 2021 Feb 17;12(1):25. [doi: 10.1186/s13244-021-00970-2] [Medline: 33595737]
10. National Health and Family Planning Commission, State Commission Office for Public Sector Reform (SCOPSR), National Development and Reform Commission, Ministry of Education, Ministry of Finance, Ministry of Human Resources and Social Security, State Administration of Traditional Chinese Medicine. Guiding opinions on establishing the standardised residency training system. National Health Commission of the People's Republic of China. 2013. URL: http://www.nhc.gov.cn/qjjys/wslgf/201401/032c8cdf2eb64a369cca4f9b76e8b059.shtml [accessed 2022-06-12]
11. Diagnostic radiology milestones. The Accreditation Council for Graduate Medical Education. 2019. URL: https://www.acgme.org/globalassets/pdfs/milestones/diagnosticradiologymilestones.pdf [accessed 2022-06-20]
12. Nasca TJ, Philibert I, Brigham T, Flynn TC. The next GME accreditation system—rationale and benefits. N Engl J Med 2012 Mar 15;366(11):1051-1056. [doi: 10.1056/NEJMsr1200117] [Medline: 22356262]
13. Leddy R, Lewis M, Ackerman S, et al. Practical implications for an effective radiology residency quality improvement program for milestone assessment. Acad Radiol 2017 Jan;24(1):95-104. [doi: 10.1016/j.acra.2016.08.018] [Medline: 27769821]
14. Chen D, Ayoob A, Desser TS, Khurana A. Review of learning tools for effective radiology education during the COVID-19 era. Acad Radiol 2022 Jan;29(1):129-136. [doi: 10.1016/j.acra.2021.10.006] [Medline: 34799258]
15. Dedeilia A, Sotiropoulos MG, Hanrahan JG, Janga D, Dedeilias P, Sideris M. Medical and surgical education challenges and innovations in the COVID-19 era: a systematic review. In Vivo 2020 Jun;34(3 Suppl):1603-1611. [doi: 10.21873/invivo.11950] [Medline: 32503818]
16. Lo HY, Lin SC, Chaou CH, Chang YC, Ng CJ, Chen SY. What is the impact of the COVID-19 pandemic on emergency medicine residency training: an observational study. BMC Med Educ 2020 Oct 7;20(1):348. [doi: 10.1186/s12909-020-02267-2] [Medline: 33028295]
17. Brady A, Brink J, Slavotinek J. Radiology and value-based health care. JAMA 2020 Oct 6;324(13):1286-1287. [doi: 10.1001/jama.2020.14930] [Medline: 32915190]

18. Cavalieri S, Spinetta M, Zagaria D, et al. The impact of COVID-19 pandemic on radiology residents in Northern Italy. Eur Radiol 2021 Sep;31(9):7077-7087. [doi: 10.1007/s00330-021-07740-0] [Medline: 33755754]
19. Alvin MD, George E, Deng F, Warhadpande S, Lee SI. The impact of COVID-19 on radiology trainees. Radiology 2020 Aug;296(2):246-248. [doi: 10.1148/radiol.2020201222] [Medline: 32216719]
20. Patil NS, Gunter D, Larocque N. The impact of the COVID-19 pandemic on radiology resident education: where do we go from here? Acad Radiol 2022 Apr;29(4):576-583. [doi: 10.1016/j.acra.2021.11.015] [Medline: 35033451]
21. Larocque N, Shenoy-Bhangle A, Brook A, Eisenberg R, Chang YM, Mehta P. Resident experiences with virtual radiology learning during the COVID-19 pandemic. Acad Radiol 2021 May;28(5):704-710. [doi: 10.1016/j.acra.2021.02.006] [Medline: 33640229]
22. Lanier MH, Wheeler CA, Ballard DH. A new normal in radiology resident education: lessons learned from the COVID-19 pandemic. Radiographics 2021;41(3):E71-E72. [doi: 10.1148/rg.2021210030] [Medline: 33939548]
23. Chénard-Roy J, Guitton MJ, Thuot F. Online residency training during the COVID-19 pandemic: a national survey of otolaryngology head and neck surgery program directors. J Otolaryngol Head Neck Surg 2021 Nov 16;50(1):65. [doi: 10.1186/s40463-021-00546-6] [Medline: 34784978]
24. Awan OA. Virtual radiology readouts after the coronavirus disease (COVID-19) pandemic. AJR Am J Roentgenol 2021 Sep;217(3):765-766. [doi: 10.2214/AJR.21.25607] [Medline: 33594906]
25. Chen Q, Liang M, Li Y, et al. Mental health care for medical staff in China during the COVID-19 outbreak. Lancet Psychiatry 2020 Apr;7(4):e15-e16. [doi: 10.1016/S2215-0366(20)30078-X] [Medline: 32085839]
26. Giusti L, Mammarella S, Salza A, et al. Predictors of academic performance during the covid-19 outbreak: impact of distance education on mental health, social cognition and memory abilities in an Italian university student sample. BMC Psychol 2021 Sep 15;9(1):142. [doi: 10.1186/s40359-021-00649-9] [Medline: 34526153]
27. Zhao L, Hwang WY, Shih TK. Investigation of the physical learning environment of distance learning under COVID-19 and its influence on students' health and learning satisfaction. Int J Distance Educ Technol 2021 Apr;19(2):77-98. [doi: 10.4018/IJDET.20210401.oa4]
28. Williamson DL, Carr J. Health as a resource for everyday life: advancing the conceptualization. Crit Public Health 2009 Mar;19(1):107-122. [doi: 10.1080/09581590802376234]
29. Airila A, Hakanen JJ, Schaufeli WB, Luukkonen R, Punakallio A, Lusa S. Are job and personal resources associated with work ability 10 years later? The mediating role of work engagement. Work Stress 2014 Jan 7;28(1):87-105. [doi: 10.1080/02678373.2013.872208]
30. Raja U, Azeem MU, Haq IU, Naseer S. Perceived threat of terrorism and employee outcomes: the moderating role of negative affectivity and psychological capital. J Bus Res 2020 Mar;110:316-326. [doi: 10.1016/j.jbusres.2020.01.026]
31. Stirpe L, Profili S, Sammarra A. Satisfaction with HR practices and employee performance: a moderated mediation model of engagement and health. Eur Manag J 2022 Apr;40(2):295-305. [doi: 10.1016/j.emj.2021.06.003]
32. Di Malta G, Bond J, Conroy D, Smith K, Moller N. Distance education students' mental health, connectedness and academic performance during COVID-19: a mixed-methods study. Distance Educ 2022 Jan 2;43(1):97-118. [doi: 10.1080/01587919.2022.2029352]
33. Lyle B, Borgert AJ, Kallies KJ, Jarman BT. Do attending surgeons and residents see eye to eye? An evaluation of the Accreditation Council for Graduate Medical Education milestones in general surgery residency. J Surg Educ 2016;73(6):e54-e58. [doi: 10.1016/j.jsurg.2016.07.004] [Medline: 27561627]
34. Barbato KBG, de Carvalho LS, Barreira Marangoni V, de Souza F, de Vasconcelos Vaena MM. Core competencies self-assessment and patient-practitioner orientation during the first year of a Brazilian orthopedic residency. Rev Bras Ortop (Sao Paulo) 2023 Oct;58(5):e742-e749. [doi: 10.1055/s-0043-1768621] [Medline: 37908538]
35. Watson RS, Borgert AJ, O Heron CT, et al. A multicenter prospective comparison of the Accreditation Council for Graduate Medical Education milestones: clinical competency committee vs. resident self-assessment. J Surg Educ 2017 Nov;74(6):e8-e14. [doi: 10.1016/j.jsurg.2017.06.009]
36. Kwasny L, Shebrain S, Munene G, Sawyer R. Is there a gender bias in milestones evaluations in general surgery residency training? Am J Surg 2021 Mar;221(3):505-508. [doi: 10.1016/j.amjsurg.2020.12.020] [Medline: 33358140]
37. Li Z, Chen L, Li M, Cohen J. Prenatal exposure to sand and dust storms and children's cognitive function in China: a quasi-experimental study. Lancet Planet Health 2018 May;2(5):e214-e222. [doi: 10.1016/S2542-5196(18)30068-8] [Medline: 29709285]
38. Hodkinson A, Zhou A, Johnson J, et al. Associations of physician burnout with career engagement and quality of patient care: systematic review and meta-analysis. BMJ 2022 Sep 14;378:e070442. [doi: 10.1136/bmj-2022-070442] [Medline: 36104064]
39. Ahmady S, Kallestrup P, Sadoughi MM, et al. Distance learning strategies in medical education during COVID-19: a systematic review. J Educ Health Promot 2021;10:421. [doi: 10.4103/jehp.jehp_318_21] [Medline: 35071627]
40. Chick RC, Clifton GT, Peace KM, et al. Using technology to maintain the education of residents during the COVID-19 pandemic. J Surg Educ 2020;77(4):729-732. [doi: 10.1016/j.jsurg.2020.03.018] [Medline: 32253133]
41. Abdull Mutalib AA, Md Akim A, Jaafar MH. A systematic review of health sciences students' online learning during the COVID-19 pandemic. BMC Med Educ 2022 Jul 3;22(1):524. [doi: 10.1186/s12909-022-03579-1] [Medline: 35786374]

42. Ritzmann S, Hagemann V, Kluge A. The Training Evaluation Inventory (TEI)—evaluation of training design and measurement of training outcomes for predicting training success. Vocat Learn 2014 Apr;7(1):41-73. [doi: 10.1007/s12186-013-9106-4]
43. Warnica W, Moody A, Probyn L, Bartlett E, Singh N, Pakkal M. Lessons learned from the effects of COVID-19 on the training and education workflow of radiology residents—a time for reflection: perspectives of residency program directors and residents in Canada. Can Assoc Radiol J 2021 Nov;72(4):637-644. [doi: 10.1177/0846537120963649] [Medline: 33047608]
44. Biswas SS, Biswas S, Awal SS, Goyal H. Current status of radiology education online: a comprehensive update. SN Compr Clin Med 2022;4(1):182. [doi: 10.1007/s42399-022-01269-z] [Medline: 35971436]
45. Liao F, Murphy D, Wu JC, Chen CY, Chang CC, Tsai PF. How technology-enhanced experiential e-learning can facilitate the development of person-centred communication skills online for health-care students: a qualitative study. BMC Med Educ 2022 Jan 25;22(1):60. [doi: 10.1186/s12909-022-03127-x] [Medline: 35078482]
46. Reinhart A, Malzkorn B, Döing C, Beyer I, Jünger J, Bosse HM. Undergraduate medical education amid COVID-19: a qualitative analysis of enablers and barriers to acquiring competencies in distant learning using focus groups. Med Educ Online 2021 Dec;26(1):1940765. [doi: 10.1080/10872981.2021.1940765] [Medline: 34128776]
47. Dost S, Hossain A, Shehab M, Abdelwahed A, Al-Nusair L. Perceptions of medical students towards online teaching during the COVID-19 pandemic: a national cross-sectional survey of 2721 UK medical students. BMJ Open 2020 Nov 5;10(11):e042378. [doi: 10.1136/bmjopen-2020-042378] [Medline: 33154063]
48. Zingaretti N, Contessi Negrini F, Tel A, Tresoldi MM, Bresadola V, Parodi PC. The impact of COVID-19 on plastic surgery residency training. Aesth Plast Surg 2020 Aug;44(4):1381-1385. [doi: 10.1007/s00266-020-01789-w]
49. Frenk J, Chen LC, Chandran L, et al. Challenges and opportunities for educating health professionals after the COVID-19 pandemic. Lancet 2022 Oct 29;400(10362):1539-1556. [doi: 10.1016/S0140-6736(22)02092-X] [Medline: 36522209]
50. Rotenstein LS, Berwick DM, Cassel CK. Addressing well-being throughout the health care workforce: the next imperative. JAMA 2022 Aug 9;328(6):521-522. [doi: 10.1001/jama.2022.12437] [Medline: 35849383]
51. Aiken LH, Simonetti M, Sloane DM, et al. Hospital nurse staffing and patient outcomes in Chile: a multilevel cross-sectional study. Lancet Glob Health 2021 Aug;9(8):e1145-e1153. [doi: 10.1016/S2214-109X(21)00209-6] [Medline: 34224669]
52. Li W, Frank E, Zhao Z, et al. Mental health of young physicians in China during the novel coronavirus disease 2019 outbreak. JAMA Netw Open 2020 Jun 1;3(6):e2010705. [doi: 10.1001/jamanetworkopen.2020.10705] [Medline: 32478846]
53. Al-Mansour K. Stress and turnover intention among healthcare workers in Saudi Arabia during the time of COVID-19: can social support play a role? PLoS One 2021;16(10):e0258101. [doi: 10.1371/journal.pone.0258101] [Medline: 34618851]
54. Sadovyy M, Sánchez-Gómez M, Bresó E. COVID-19: how the stress generated by the pandemic may affect work performance through the moderating role of emotional intelligence. Pers Individ Dif 2021 Oct;180:110986. [doi: 10.1016/j.paid.2021.110986] [Medline: 34629581]
55. Gu Z, Li P, Zhang A, Xu X, Gu F. The role of mental health and sustainable learning behavior of students in education sector influences sustainable environment. Front Psychol 2022;13:822751. [doi: 10.3389/fpsyg.2022.822751] [Medline: 35211067]
56. Witter S, Wurie H, Chandiwana P, et al. How do health workers experience and cope with shocks? Learning from four fragile and conflict-affected health systems in Uganda, Sierra Leone, Zimbabwe and Cambodia. Health Policy Plan 2017 Nov 1;32(Suppl 3):iii3-iii13. [doi: 10.1093/heapol/czx112] [Medline: 29149313]
57. Dean L, Cooper J, Wurie H, et al. Psychological resilience, fragility and the health workforce: lessons on pandemic preparedness from Liberia and Sierra Leone. BMJ Glob Health 2020 Sep;5(9):e002873. [doi: 10.1136/bmjgh-2020-002873] [Medline: 32988928]
58. Samarasekera DD, Nyoni CN, Amaral E, Grant J. Challenges and opportunities in interprofessional education and practice. Lancet 2022 Oct 29;400(10362):1495-1497. [doi: 10.1016/S0140-6736(22)02086-4] [Medline: 36522199]
59. Huang Y. Research on online education in the midst of the COVID-19 pandemic. JAER 2020 May 2;5(2):77-80. [doi: 10.22606/jaer.2020.52005]
60. Kruger J, Dunning D. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. J Pers Soc Psychol 1999 Dec;77(6):1121-1134. [doi: 10.1037//0022-3514.77.6.1121] [Medline: 10626367]
61. Hu LP, Liu HG. Triple-type theory of statistics and its application in the scientific research of biomedicine. Zhonghua Yi Xue Za Zhi 2005 Jul 20;85(27):1936-1940. [Medline: 16255993]

Abbreviation

ACGME: Accreditation Council of Graduate Medical Education
CAR: Chinese Association of Radiologists
ICS: interpersonal communication skill
MK: medical knowledge
MLR: multiple linear regression
PBLI: practice-based learning and improvement

PC: patient care
PROF: professionalism
SBP: system-based practice
SEM: self-evaluation milestone
SRT: standardized residency training

Edited by B Lesselroth; submitted 05.11.23; peer-reviewed by S Biswas, X Sun; revised version received 05.10.24; accepted 15.02.25; published 08.05.25.

Please cite as:
Wang P, Wu Z, Zhang J, He Y, Jiang M, Zheng J, Wang Z, Yang Z, Chen Y, Zhu J
Distance Learning During the COVID-19 Lockdown and Self-Assessed Competency Development Among Radiology Residents in China: Cross-Sectional Survey
JMIR Med Educ 2025;11:e54228
URL: https://mededu.jmir.org/2025/1/e54228
doi:10.2196/54228

Comparison of an Emergency Medicine Asynchronous Learning Platform Usage Before and During the COVID-19 Pandemic: Retrospective Analysis Study

Blake Briggs[1*], MD; Madhuri Mulekar[2*], PhD; Hannah Morales[1*], DO; Iltifat Husain[3*], MD

[1]Division of Emergency Medicine, Department of Surgery, University of Tennessee Graduate School of Medicine, 1924 Alcoa Highway, Knoxville, TN, United States

[2]Department of Mathematics, University of South Alabama, Mobile, AL, United States

[3]Department of Emergency Medicine, Wake Forest School of Medicine, Winston-Salem, NC, United States

[*]all authors contributed equally

Corresponding Author:
Blake Briggs, MD
Division of Emergency Medicine, Department of Surgery, University of Tennessee Graduate School of Medicine, 1924 Alcoa Highway, Knoxville, TN, United States

Abstract

Background: The COVID-19 pandemic challenged medical educators due to social distancing. Podcasts and asynchronous learning platforms help distill medical education in a socially distanced environment. Medical educators interested in providing asynchronous teaching should know how these methods performed during the pandemic.

Objective: The purpose of this study was to assess the level of engagement for an emergency medicine (EM) board review podcast and website platform, before and during the COVID-19 pandemic. We measured engagement via website traffic, including such metrics as visits, bounce rate, unique visitors, and page views. We also evaluated podcast analytics, which included total listeners, engaged listeners, and number of plays.

Methods: Content was designed after the American Board of EM Model, covering only 1 review question per episode. Website traffic and podcast analytics were studied monthly from 2 time periods of 20 months each, before the pandemic (July 11, 2018, to February 31, 2020) and during the pandemic (May 1, 2020, to December 31, 2021). March and April 2020 data were omitted from the analysis due to variations in closure at various domestic and international locations. Results underwent statistical analysis in March 2022.

Results: A total of 132 podcast episodes and 93 handouts were released from July 11, 2018, to December 31, 2021. The mean number of listeners per podcast increased significantly from 2.11 (SD 1.19) to 3.77 (SD 0.76; *t* test, $P<.001$), the mean number engaged per podcast increased from 1.72 (SD 1.00) to 3.09 (SD 0.62; *t* test, $P<.001$), and the mean number of plays per podcast increased from 42.54 (SD 40.66) to 69.23 (SD 17.54; *t* test, $P=.012$). Similarly, the mean number of visits per posting increased from 5.85 (SD 3.28) to 15.39 (SD 3.06; *t* test, $P<.001$), the mean number of unique visitors per posting increased from 3.74 (SD 1.83) to 10.41 (SD 2.33; *t* test, $P<.001$), and the mean number of page views per posting increased from 17.13 (SD 10.63) to 33.32 (SD 7.01; *t* test, $P<.001$). Note that, all measures showed a decrease from November 2021 to December 2021.

Conclusions: During the COVID-19 pandemic, there was an increased engagement for our EM board review podcast and website platform over a long-term period, specifically through website visitors and the number of podcast plays. Medical educators should be aware of the increasing usage of web-based education tools, and that asynchronous learning is favorably viewed by learners. Limitations include the inability to view Spotify (Spotify Technology S.A.) analytics during the study period, and confounding factors like increased popularity of social media inadvertently promoting the podcast.

(JMIR Med Educ 2025;11:e58100) doi:10.2196/58100

KEYWORDS

asynchronous learning; medical education; podcast; COVID-19; emergency medicine; online learning; engagement; web-based; online study; online class; videoconferencing; assessment; effectiveness; challenges; knowledge retention; performance; virtual learning; pre-pandemic; post-pandemic

Introduction

As the field of medical education evolves, web-based media and digital study tools are finding larger audiences each year [1]. The COVID-19 pandemic dramatically changed the landscape of medical education. Suddenly in March and April 2020, all learning was switched to remote platforms, greatly challenging educators and hastening the switch to web-based media [2-4].

Previous studies have demonstrated that podcasts have positive effects on knowledge retention and test performance [5,6]. Multiple studies have previously been published on the effectiveness of remote learning during the COVID-19 pandemic via remote learning and web-based modules [7,8]. Most recently, 1 study aimed to measure podcast and blog utilization during the early months of the COVID-19 pandemic [9]. This study found an increase in blog page views during the early months of the pandemic, but no statistical change in podcast usage. However, this study had a short measurement period (January to May 2020). In addition, the study made measuring educational content related to COVID-19 a secondary outcome. As asynchronous teaching continues to increase in popularity among students in the wake of the pandemic, medical educators should be curious about the popularity of such materials during a time in which in-person education was severely limited or paused altogether. The purpose of this study was to assess the level of engagement for an emergency medicine (EM) board review podcast and platform, comparing before COVID-19 to during the COVID-19 pandemic over a period of 34 months. Our secondary outcome was to measure important website variables that have previously not been mentioned in medical education literature, especially in the setting of the pandemic. We hypothesized that the pandemic would increase the number of website visitors, page views, and podcast episode plays.

Methods

Overview

This retrospective analysis was conducted from March 5, 2022, to April 30, 2022. Data were collected by the study authors from July 11, 2018, when the first podcast episode was released, to December 31, 2021. Emergency Medicine Board Bombs (EMBB) was launched by 2 academic EM physicians in July 2018. The goal of this asynchronous educational platform was to increase first-time pass rate among residents and attendings taking their in-service exam and boards, respectively. EMBB is a peer-reviewed resource and functions at no cost to the learner. EMBB has never been formally assigned to any formal, academic curriculum; its educational platform is entirely free and open access to all learners. The website has podcasts and printable study guides that function as summaries of various common pathologies encountered in the emergency department and on-the-board exams.

Platform Development

Each podcast episode was structured to quickly cover one multiple-choice question, a discussion of correct and incorrect answers, and the relevant subject matter. Audio-editing was conducted using Apple Garageband, a free service provided to those who own Apple hardware. The podcast was available for free streaming on a designated website, emboardbombs, as well as dedicated podcast platforms (Apple Podcasts, Soundcloud [SoundCloud Global Limited & Co KG], and Spotify [Spotify Technology S.A.]). Questions for each episode were modeled after the American Board of Emergency Medicine (ABEM) certification exam. The Model of the Clinical Practice of Emergency Medicine (EM Model), serves as the basis for ABEM content and was followed in drafting podcast episodes [10]. A peer review process was used to develop multiple-choice questions. Each question was written by an EM physician with an academic appointment and was shared with 2 other academic physicians for review before it was featured on the podcast.

Medical source material was derived from *Tintinalli's Emergency Medicine* as well as UpToDate and EB Medicine [11-13]. The educational platform was self-funded by the creators and developed on their own time and schedule. No financial support or aid was received. In terms of dedicated time and monetary investment, the cost of equipment and software totaled nearly US $400 annually. In terms of hourly commitment, approximately 5 - 10 hours weekly is needed to record, edit, and publish podcast episodes, as well as write and publish study guides. The podcast was not formally added to any curriculum. It was disseminated by word of mouth. No marketing or paid advertising was used.

Variable Definitions

Podcast analytics were derived from Apple Podcasts Connect which is a free service provided for all Apple Podcast hosts. It provides data on total listeners, engaged listeners, and number of plays [14]. Listeners were defined by Apple as the number of unique devices that played more than 0 seconds of an episode. Engaged listeners were defined as the number of devices that played at least 20 minutes or 40% of an episode within a single session. Of note, pausing or stopping an episode did not count as starting a new session. Number of episode plays was based on the number of unique devices where the play duration is more than 0 seconds. At the time of our data collection during the pandemic, Spotify did not publish podcast statistics, and therefore, their user data could not be obtained.

The website learning platform was hosted on Squarespace. Website traffic analytics were derived from Squarespace, which measured traffic using variables such as website visits, website bounce rate, website unique visitors, and website page views [15]. Visits were defined as the total number of browsing sessions per visitor on the website within a 30-minute period. A browser cookie from Squarespace was used to track views within a 30-minute period. The bounce rate was defined as the number of visitors who navigate away from the website after viewing 1 page. Unique visitors were defined as the total number of new IP addresses that visited the website. Page views were defined as the total number of views across all pages on the website. Page views count the number of times a page is viewed. Furthermore, 1 visit consists of 1 or more pages.

Data Collection

Website traffic and podcast analytics from July 11, 2018, to February 28, 2020, were compared with those from May 1, 2020, to December 31, 2021. May 1, 2020, was chosen as the transition date because, during March and April 2020, various schools and residency programs began switching to remote learning. As the pandemic evolved, medical schools and graduate medical education sites began suspending in-person rotations. The Accreditation Council for Graduate Medical Education announced in mid-March that all in-person educational activities, meetings, and site visits were to migrate to virtual occurrences only [16]. By the end of April 2020, all nonessential, in-person educational activities had ceased [17].

Statistical Analysis

All collected data were organized in a Microsoft Excel spreadsheet and analyzed using statistical software JMP Pro 16.0.0 (SAS Institute Inc) in March 2022. All numerical data were summarized using mean and SD. Variations in monthly data from before COVID-19 and during COVID-19 periods were compared using the Levene test, whereas the means per month were compared using a 2-sample *t* test after accounting for differences in variations if any [18,19]. In addition, a nonparametric Mann-Whitney *U* test was also used to compare analytics from 2 time periods. Time series plots were used to study trends in monthly data. A significance level of 0.05 was used to determine the significance of outcomes.

Ethical Considerations

The Institutional Review Board was approached for ethics approval but reported that the study did not meet the criteria for human candidates research, and therefore, no approval was required.

Results

During the study period from July 11, 2018, to December 31, 2021, a total of 132 podcast episodes and 93 study guides were created. The first podcast episode was released on July 11, 2018.

From July 11, 2018, to February 28, 2020, 68 episodes were released, along with 30 study guides. From May 1, 2020, to December 31, 2021, 59 podcasts were released, and 53 handouts were published. Note that 5 episodes and 10 handouts were released during March-April 2020, which were also available to learners during the COVID-19 pandemic. This resulted in a total of 225 postings (132 podcasts and 93 handouts) being available to learners during the COVID-19 pandemic (Table 1).

Table . Number of podcasts, handouts, and total postings before, in-between, and during COVID-19 periods.

Period	Podcasts, n	Handouts, n	Postings, n
Before COVID-19	68	30	98
In-between period	5	10	15
During COVID-19	59	53	112
Total	132	93	225

The time series presented in Figure 1 show month-to-month changes in podcast and website visit analytics before the COVID-19 and during COVID-19 periods and differences in changing patterns. Although higher outcomes were observed during the COVID-19 period in all 6 podcast and website visit measurements compared with before the COVID-19 period, not all changes showed linear patterns of increase. In fact, the number of unique visitors, visits, and page reviews showed decreasing trend after reaching a peak around the middle of the COVID-19 period. However, at the end of the 20-month period, they still remained higher than before the COVID-19 level. During the before the COVID-19 period, number of listeners per month steadily increased from 39 to 338. During the COVID-19 period, it continued to increase, reaching a maximum number of listeners at 672. A similar trend was observed for number engaged per month, increasing from 28 to 289 during the before the COVID-19 period and reaching a maximum of 555 during the COVID-19 period. Although a similar trend was observed for the total number of plays with an increase from 412 to 11,879 during the before the COVID-19 period, a sharp drop was observed during the period of uncertainty (March-April 2020). Again, during the COVID-19 period, total number of plays increased from 4547 to 14,296. Number of visits during the before the COVID-19 period increased from 218 to 1064; there was further increase in the COVID-19 period, reaching 4664 in January 2021. The number of visits started declining thereafter, reaching a low of 1879. The number of unique visitors and page views showed patterns similar to that of the number of visits. The number of unique visitors increased steadily during the before the COVID-19 period from 138 to 620. It increased to 3222 in January 2021 but started declining to a low of 2293. The number of page views also increased steadily during the before the COVID-19 period from 610 to 3405; in the COVID-19 period, it increased to 11,326 in November 2020, only to steadily decrease to a low of 5389 in December 2021. Note that all measures showed a decrease from November 2021 to December 2021.

Comparison of podcast and website visit analytics are presented in Table 2. It shows that regardless of differences in the number of podcasts and handouts available during the 2 time periods, variation in analytics from month to month did not differ significantly during the 2 time periods under study except for bounce rate and number of visitors. Significantly higher variation as measured by SD was observed in bounce rate (0.07 vs 0.05; Levene test, P=.036) and number of unique visitors (523.45 vs 179.62; Levene test, P=.0049) during COVID-19 pandemic compared with the before the COVID-19 period. Percent increase in mean analytics from before the COVID-19 period to during the COVID-19 period ranged from 24% (bounce rate, 0.55 to 0.30 per 100 postings, n=20) to 539%

(unique visitors, 3.74 to 10.41 per posting, n=20) with the mean number of unique visitors showing the highest percent increase and the bounce rate the lowest. The number of visits increased by 504% (5.85 to 15.39 per posting, n=20) whereas the number of listeners, engaged, and total plays each increased by more than 200% (listeners: 2.11 to 3.77 per podcast, n=20; engaged: 1.72 to 3.09 per podcast, n=20; total plays: 42.54 to 69.23 per podcast, n=20). Percent increases in the average monthly analytics indicate considerable increase in visits and usage of podcasts from before COVID-19 to during the COVID-19 period.

Figure 1. Monthly change in podcast and website visit analytics before COVID-19 and during COVID-19 periods. The arrowhead marks the start of the pandemic.

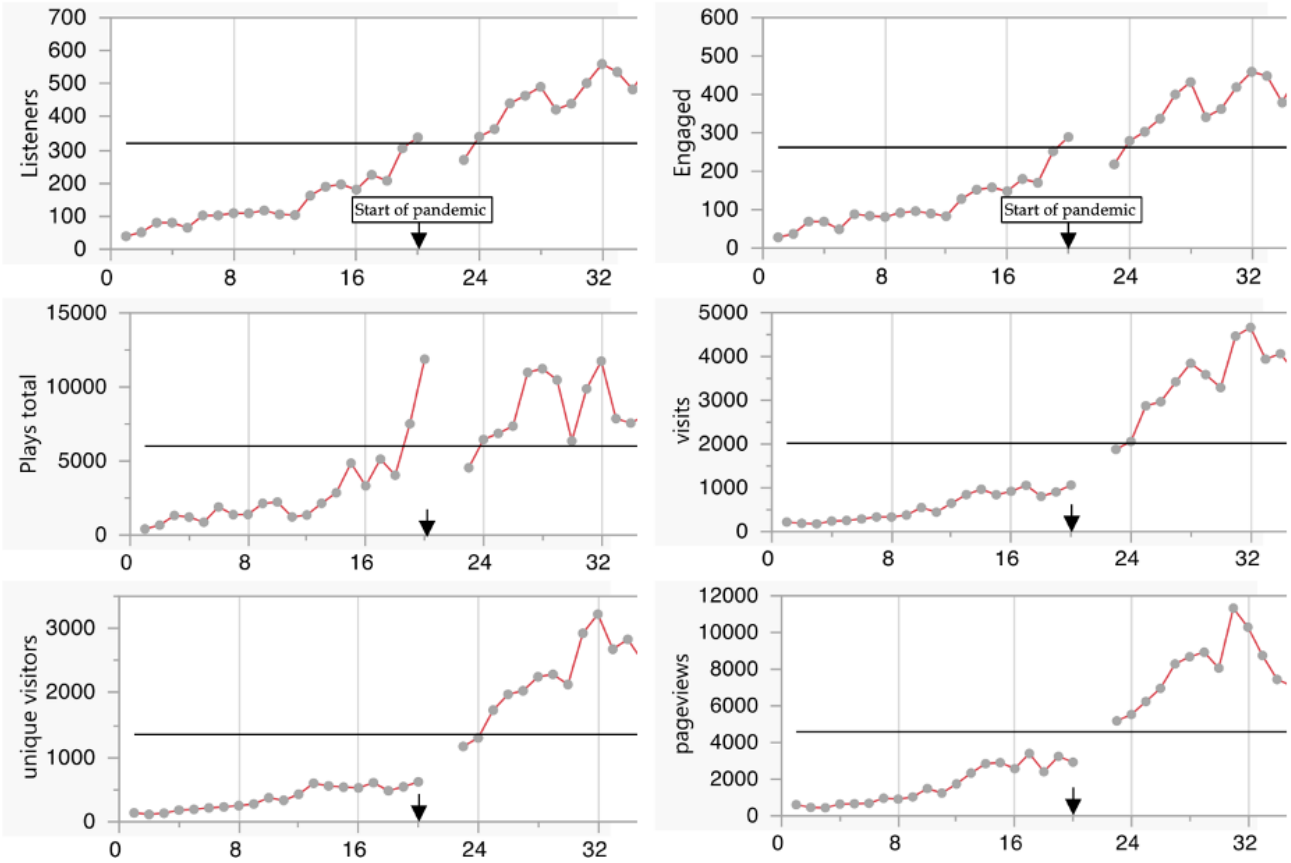


Table . Comparison of podcast and website visit analytics before the COVID-19 and during the COVID-19 periods.

Aspect and period	n	Mean (SD)	Range	*P* value (Levene test)	*P* value (*t* test)	Increase in mean, %
Listeners				.54	<.001	247.31
Before COVID-19	20	143.20 (80.93)	39-338			
During COVID-19	20	497.35 (99.84)	270-672			
Engaged listeners				.48	<.001	247.84
Before COVID-19	20	117.15 (68.17)	28-289			
During COVID-19	20	407.50 (82.19)	218-555			
Number of total episode plays				.95	<.001	215.88
Before COVID-19	20	2892.85 (2764.59)	412-11,879			
During COVID-19	20	9137.80 (2315.19)	4547-14,296			
Website visits				.06	<.001	504.3
Before COVID-19	20	573.20 (321.17)	178-1064			
During COVID-19	20	3463.85 (689.29)	1879-4664			
Website bounce rate				.03	<.001	24.07
Before COVID-19	20	0.54 (0.05)	46%-62%			
During COVID-19	20	0.67 (0.07)	52%-75%			
Website unique visitors				.004	<.001	538.99
Before COVID-19	20	366.55 (179.62)	114-620			
During COVID-19	20	2342.20 (523.45)	1170-3222			
Website page views				.27	<.001	346.6
Before COVID-19	20	1678.60 (1041.70)	443-3405			
During COVID-19	20	7496.65 (1577.68)	5183-11,326			

Although periods of similar length (ie, 20 months each) were used for comparison, the number of postings available during these 2 periods differed considerably because as new postings were made available, the earlier postings were still available for review for visitors. To account for the differences in the number of postings, analytics were adjusted by computing outcome per posting available. For example, number of listeners per podcast was computed as follows:

- Before COVID-19: # of listeners/podcast = # listeners/68
- During COVID-19: # listeners/podcast = # listeners/132

Note that this accounts for all podcasts that were available to listeners. Before COVID-19 accounts for all podcasts put out during that time and during COVID-19 used all podcasts available, that is, those that were put out before COVID-19, in-between, and during COVID-19 periods. Number of engaged and total plays were adjusted similarly by number of podcasts. Number of visits, unique visitors, and page views were adjusted similarly using all postings (ie, podcasts plus handouts). Bounce rate was adjusted similarly using per 100 postings because rate of per posting resulted in very small numbers and this change from per posting to per 100 postings does not affect the outcome of statistical tests.

Resulting comparisons of outcomes are listed in Table 3, which shows a significant increase in mean rates for all analytics except mean bounce rate per 100 postings from before COVID-19 to during COVID-19. Bounce rate per 100 postings showed a significant decrease from before COVID-19 to during

COVID-19 (0.55 to 0.30 per 100 podcasts; *t* test, *P*<.001). Mean number of listeners per podcast increased significantly from 2.11 (SD 1.19) to 3.77 (SD 0.76; *t* test, *P*<.001), mean number engaged per podcast increased from 1.72 (SD 1.00) to 3.09 (SD 0.62; *t* test, *P*<.001), and mean number of plays per podcast increased from 42.54 (SD 40.66) to 69.23 (SD 17.54; *t* test, *P*=.0122). Similarly, mean number of visits per posting increased from 5.85 (SD 3.28) to 15.39 (SD 3.06; *t* test, *P*<.001), mean number of unique visitors per posting increased from 3.74 (SD 1.83) to 10.41 (SD 2.33; *t* test, *P*<.001); and mean number of page views per posting increased from 17.13 (SD 10.63) to 33.32 (SD 7.01; *t* test, *P*<.001). Even nonparametric comparisons using Mann-Whitney *U* test gave the same results.

Table . Comparison of podcast and website visit analytics rates per posting available to viewers before COVID-19 and during COVID-19 periods.

Aspect and period	n	Mean (SD)	Range	Median (IQR)	*P* value (*t* test)	*P* value (Mann Whitney *U* test)
Listeners per podcast					<.001	<.001
Before COVID-19	20	2.11 (1.19)	0.57-4.97	1.60 (1.26-2.86)		
During COVID-19	20	3.77 (0.76)	2.05-5.09	3.83 (3.34-4.27)		
Engaged per podcast					<.001	<.001
Before COVID-19	20	1.72 (1.00)	0.41-4.25	1.34 (1.06-2.30)		
During COVID-19	20	3.09 (0.62)	1.65-4.20	3.22 (2.62-3.47)		
Number of total episode plays per podcast					.0122	<.001
Before COVID-19	20	42.54 (40.66)	6.06-174.69	29.71 (18.36-56.81)		
During COVID-19	20	69.23 (17.54)	34.45-108.30	69.95 (56.13-81.99)		
Website visits per posting					<.001	<.001
Before COVID-19	20	5.85 (3.28)	1.82-10.86	5.07 (2.68-9.08)		
During COVID-19	20	15.39 (3.06)	8.35-20.73	15.64 (13.80-17.41)		
Website bounce rate per 100 postings					<.001	<.001
Before COVID-19	20	0.55 (0.05)	0.47-0.63	0.54 (0.51-0.59)		
During COVID-19	20	0.30 (0.03)	0.23-0.33	0.30 (0.27-0.32)		
Website unique visitors per posting					<.001	<.001
Before COVID-19	20	3.74 (1.83)	1.16-6.33	3.60 (2.03-5.53)		
During COVID-19	20	10.41 (2.33)	5.20-14.32	10.65 (9.14-11.84)		
Website page views per posting					<.001	<.001
Before COVID-19	20	17.13 (10.63)	4.520-34.745	13.98 (6.85-28.35)		
During COVID-19	20	33.32 (7.01)	23.036-50.338	32.49 (28.39-38.13)		

Discussion

Principal Findings

The results demonstrate that our online EM board review podcast and platform experienced significantly increased levels of engagement during the COVID-19 pandemic. Our learning platform included multiple media, such as PDF study guides, video and picture-based modules, and online question banks. The aim was for the podcast and handouts to be integrated into an asynchronous study plan, as the platform provided easy accessibility and use.

Implication of Findings

The COVID-19 pandemic disrupted medical education, forcing learners in both medical school and residency to navigate vast amounts of information, largely in isolation. This shift from interactive, in-person learning raised concerns about students overextending themselves, leading to only a surface-level understanding of the material. One study comparing first and second-year medical student education during the pandemic highlighted the importance of face-to-face learning, finding that the first-year medical students in isolation performed worse than the previous year's first-year medical students [20]. Another retrospective study performed at the University of Hawaii Burn School of Medicine demonstrated that fourth-year medical students who were enrolled during the pandemic displayed improved note-taking with a 9-point increase in exam scores, yet worse physical examinations in their standardized patient encounters with a 12-point average decrease in scores [21].

In response, many innovative educational tools have emerged to attempt to provide asynchronous learning. Online resources like the one in this study are unique. Diverse topics are integrated into a single, cost-effective, and efficient platform, with podcast episodes <20 minutes, as well as downloadable PDF handouts. This model is beneficial for both visual and auditory learners.

While other learning platforms were not analyzed during this study period, valuable information was collected from this study's podcast. EMBB offers a humanistic aspect to learning with the dual physician hosts, pertinent banter, and narrative medicine aspect, of which may anthropomorphize the learning despite pandemic isolation.

Comparison With the Literature

Podcasts have been welcomed by those looking for a nontraditional method of learning in recent years, most notably those practicing in EM, where it is the most represented specialty that regularly hosts podcasts [22-24]. A survey in 2014 showed EM residents devote more time to podcasts than journals, citing podcasts as "the most beneficial" for education [22]. In another large survey, 80% of EM residents had listened to medical podcasts at least once [25].

Traditional lectures continue to be replaced by various digital teaching methods and this was hastened by the arrival of COVID-19. Podcasts' major benefit is their customization to fit learner's educational goals as well as time constraints, allowing users to optimize their study goals while balancing work and private life.

Feasibility of Implementation

In terms of feasibility, the podcast required a dedicated amount of time and monetary investment. The cost of standard microphones, basic recording software, and a website to host the podcast required approximately US $300 to 400 annually. As discussed in the methods section, the hourly commitment was close to 5 - 10 hours weekly.

Next Steps

A review of "Learning Through Listening: A Scoping Review of Podcast Use in Medical Education" examines podcasts for learning across many specialties, most often referencing anesthesia, with some reference to EM [26]. The data cited an increase in retention of information pre- and posttest for medical students, who are not specialized in EM compared with the level of a resident or attending physician. The review briefly mentions a podcast that improved EM in-training exam scores and a podcast that reportedly worsened in-training exam scores. The data gleaned from this study are of interest, but due to varied in-training exam scores, a comparative study is needed that examines test performance matching which podcast was used most for learning. Another future area of study will be to observe if the effects of the COVID-19 pandemic on asynchronous web-based learning are long-term.

Limitations

Our study is limited in generalizability due to it only measuring one specific podcast and website platform. A restricted sample size is one limitation of this study. Spotify and Android (Google) do not publish podcast statistics nor track individual usage, and therefore user data from both these platforms could not be obtained. According to Reuters in a survey of 2012 listeners, 20% used Apple Podcasts as their app of choice from 2019 - 2020, which is the second largest market share [27]. Previous studies have used podcast episode downloads as a metric for engagement. Despite the appeal of using number of downloads as a measurement, accurate analytics are difficult to obtain and fraught with bias. Downloads are defined differently depending on the podcast host. In addition, there have been reports that these numbers can be unreliable due to bot traffic and there can be manipulation of download data by hosts [28,29].

Another limitation is association versus causation. Given the retrospective study design and nature of COVID-19, it is difficult to completely credit the pandemic for increased podcast engagement. Confounding variables could also be a limitation, such as increased usage of social media during quarantine resulting in better promotion of the podcast and website.

One potential confounding variable was the launch of a procedural module in May 2020. This web-based learning instruction was an airway module, with recorded intubation videos and a pre- and postassessment. However, when reviewing website analytics, this was not a frequently viewed page on the website, accounting for only 2.59% of total website page views. It cannot entirely account for the sudden increase in website

visitors and podcast listeners. Thus, in this study, we can only establish differences observed in analytics between 2 time periods.

No quantitative data were tracked regarding listener exam performance, in particular in-training or board examinations. The purpose of this study was to assess the level of engagement for an EM board review podcast and website platform, before and during the COVID-19 pandemic. Future research should be aimed at assessing whether this educational intervention is an effective form of test preparation.

Conclusion

During the COVID-19 pandemic, there was an accelerated level of engagement for our EM board review podcast and website platform over a long-term period. This educational platform is a feasible, low-cost asynchronous study tool. Medical educators should be aware of the increasing usage of web-based education tools, and that asynchronous learning is favorably viewed by learners.

Acknowledgments

Statistical analysis by MM reported in this publication was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under award UL1TR001417.

Conflicts of Interest

None declared.

References

1. Apple iTunes Preview webpage. URL: https://podcasts.apple.com/us/genre/podcasts-health-fitness-medicine/id1518 [accessed 2020-02-07]
2. Sandars J, Correia R, Dankbaar M, et al. Twelve tips for rapidly migrating to online learning during the COVID-19 pandemic. MedEdPublish (2016) 2020;9(1):82. [doi: 10.15694/mep.2020.000082.1] [Medline: 38058947]
3. Theoret C, Ming X. Our education, our concerns: the impact on medical student education of COVID-19. Med Educ 2020 Jul;54(7):591-592. [doi: 10.1111/medu.14181] [Medline: 32310318]
4. Rana T, Hackett C, Quezada T, et al. Medicine and surgery residents' perspectives on the impact of COVID-19 on graduate medical education. Med Educ Online 2020 Dec;25(1):1818439. [doi: 10.1080/10872981.2020.1818439] [Medline: 32924869]
5. Alla A, Kirkman MA. PodMedPlus: an online podcast resource for junior doctors. Med Educ 2014 Nov;48(11):1126-1127. [doi: 10.1111/medu.12556] [Medline: 25307665]
6. Bhatti I, Jones K, Richardson L, Foreman D, Lund J, Tierney G. E-learning vs lecture: which is the best approach to surgical teaching? Colorectal Dis 2011 Apr;13(4):459-462. [doi: 10.1111/j.1463-1318.2009.02173.x] [Medline: 20041922]
7. Wilcha RJ. Effectiveness of virtual medical teaching during the COVID-19 crisis: systematic review. JMIR Med Educ 2020 Nov 18;6(2):e20963. [doi: 10.2196/20963] [Medline: 33106227]
8. Gopalan C, Butts-Wilmsmeyer C, Moran V. Virtual flipped teaching during the COVID-19 pandemic. Adv Physiol Educ 2021 Dec 1;45(4):670-678. [doi: 10.1152/advan.00061.2021] [Medline: 34498940]
9. Boreskie PE, Chan TM, Novak C, et al. Medical education blog and podcast utilization during the COVID-19 pandemic. Cureus 2022 Mar;14(3):e23361. [doi: 10.7759/cureus.23361] [Medline: 35475051]
10. Tintinalli JE, Stapczynski J, Ma OJ, Yealy D, Meckler G, Cline D. Tintinalli's Emergency Medicine: A Comprehensive Study Guide, 8th edition: McGraw-Hill Education; 2016.
11. UpToDate. URL: https://www.wolterskluwer.com/en/solutions/uptodate [accessed 2022-02-13]
12. EB Medicine. URL: https://www.ebmedicine.net [accessed 2022-04-05]
13. Beeson MS, Ankel F, Bhat R, et al. The 2019 model of the clinical practice of emergency medicine. J Emerg Med 2020 Jul;59(1):96-120. [doi: 10.1016/j.jemermed.2020.03.018] [Medline: 32475725]
14. Apple podcasts for creators. Understanding your analytics. URL: https://podcasters.apple.com/support/2553-understanding-your-subscriptions-reports [accessed 2022-02-16]
15. Squarespace. Traffic analytics. URL: https://support.squarespace.com/hc/en-us/articles/217999797 [accessed 2022-02-16]
16. ACGME. Updated: coronavirus (COVID-19) and ACGME site visits, educational activities, and other meetings. URL: https://www.acgme.org/newsroom/2020/3/updated-coronavirus-covid-19-and-acgme-site-visits-educational-activities-and-other-meetings [accessed 2022-02-17]
17. Nasca TJ. ACGME's early adaptation to the COVID-19 pandemic: principles and lessons learned. J Grad Med Educ 2020 Jun;12(3):375-378. [doi: 10.4300/JGME-D-20-00302.1] [Medline: 32595876]
18. WELCH BL. The generalisation of student's problems when several different population variances are involved. Biometrika 1947;34(1-2):28-35. [doi: 10.1093/biomet/34.1-2.28] [Medline: 20287819]
19. Levene H. Robust tests for equality of variances. In: Olkin I, Hotelling H, editors. Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling: Stanford University Press; 1960:278-292.

20. Ayoubieh H, Alkhalili E, Nino D, Coue M, Herber-Valdez C, Pfarr CM. Analysis of pre-clerkship medical students' perceptions and performance during the COVID-19 pandemic. Med Sci Educ 2023 Feb;33(1):147-156. [doi: 10.1007/s40670-022-01723-6] [Medline: 36688011]
21. Noel ER, Mizuo B, Yamamoto LG, Ishikawa KM, Chen JJ, Len KA. Effect of the COVID-19 pandemic on medical student performance and evaluation scores. SVOA Med Res 2024;2(1):10-18. [doi: 10.58624/SVOAMR.2024.02.011] [Medline: 39144736]
22. Riddell J, Swaminathan A, Lee M, Mohamed A, Rogers R, Rezaie SR. A survey of emergency medicine residents' use of educational podcasts. West J Emerg Med 2017 Feb;18(2):229-234. [doi: 10.5811/westjem.2016.12.32850] [Medline: 28210357]
23. Cadogan M, Thoma B, Chan TM, Lin M. Free Open Access Meducation (FOAM): the rise of emergency medicine and critical care blogs and podcasts (2002-2013). Emerg Med J 2014 Oct;31(e1):e76-e77. [doi: 10.1136/emermed-2013-203502] [Medline: 24554447]
24. Berk J, Trivedi SP, Watto M, Williams P, Centor R. Medical education podcasts: where we are and questions unanswered. J Gen Intern Med 2020 Jul;35(7):2176-2178. [doi: 10.1007/s11606-019-05606-2] [Medline: 31898131]
25. Gottlieb M, Riddell J, Cooney R, King A, Fung CC, Sherbino J. Maximizing the morning commute: a randomized trial assessing the effect of driving on podcast knowledge acquisition and retention. Ann Emerg Med 2021 Sep;78(3):416-424. [doi: 10.1016/j.annemergmed.2021.02.030] [Medline: 33931254]
26. Kelly JM, Perseghin A, Dow AW, Trivedi SP, Rodman A, Berk J. Learning through listening: a scoping review of podcast use in medical education. Acad Med 2022 Jul 1;97(7):1079-1085. [doi: 10.1097/ACM.0000000000004565] [Medline: 34935729]
27. eMarketer. Most commonly used apps for listening to podcasts among podcast listeners in the united states in 2019 and 2020. URL: https://www.statista.com/statistics/943537/podcast-listening-apps-us [accessed 2022-02-16]
28. RSS.com. What cause the spike in my podcast analytics? URL: https://rss.com/blog/what-caused-the-spike-in-my-podcasts-analytics [accessed 2024-08-28]
29. Tani, Max. Semafor.com. The bots have come for podcasts. URL: https://www.semafor.com/article/10/08/2023/the-bots-have-come-for-podcasts [accessed 2024-08-28]

Abbreviations

ABEM: American Board of Emergency Medicine
EM: emergency medicine
EMBB: Emergency Medicine Board Bombs

Edited by B Lesselroth; submitted 05.03.24; peer-reviewed by AP Johnson, A Hosny, E Vashishtha; revised version received 28.10.24; accepted 02.01.25; published 21.02.25.

Please cite as:
Briggs B, Mulekar M, Morales H, Husain I
Comparison of an Emergency Medicine Asynchronous Learning Platform Usage Before and During the COVID-19 Pandemic: Retrospective Analysis Study
JMIR Med Educ 2025;11:e58100
URL: https://mededu.jmir.org/2025/1/e58100
doi:10.2196/58100

Original Paper

Innovative Mobile App (CPD By the Minute) for Continuing Professional Development in Medicine: Multimethods Study

Peter Slinger[1,2], MD; Maram Omar[3], MSc; Sarah Younus[1], MPH; Rebecca Charow[1], MSc; Michael Baxter[4], MD; Craig Campbell[5], MD; Meredith Giuliani[2,6,7], MBBS, MEd; Jesse Goldmacher[1], MD; Tharshini Jeyakumar[1], MHI; Inaara Karsan[1], MHI; Janet Papadakos[6], MEd, PhD; Tina Papadakos[6], MA; Alexandra Jane Rotstein[8,9], MSc, MD; May-Sann Yee[10], MD, MEHP; Asad Siddiqui[2,11], MD; Marcos Silva Restrepo[2,12], MD; Melody Zhang[13], MA; David Wiljer[1,2,7,14], PhD

[1]University Health Network, Toronto, ON, Canada
[2]Temerty School of Medicine, University of Toronto, Toronto, ON, Canada
[3]Ontario Health, Toronto, ON, Canada
[4]St Joseph's Health Centre, Toronto, ON, Canada
[5]Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada
[6]Princess Margaret Cancer Centre, University Health Network, Toronto, ON, Canada
[7]The Wilson Centre, Toronto, ON, Canada
[8]Foothills Medical Centre, Calgary, AB, Canada
[9]University of Calgary, Calgary, Canada
[10]Southlake Health, Newmarket, ON, Canada
[11]The Hospital for Sick Children, Toronto, ON, Canada
[12]Sunnybrook Health Sciences Centre, Toronto, ON, Canada
[13]Mastercard Foundation, Toronto, ON, Canada
[14]Institute of Health Policy, Management & Evaluation, University of Toronto, Toronto, ON, Canada

Corresponding Author:
David Wiljer, PhD
University Health Network
90 Elizabeth StreetR. Fraser Elliott Building RFE 3S-441
Toronto, ON M5G 2C4
Canada
Phone: 1 4163404800
Email: David.wiljer@uhn.ca

Abstract

Background: Many national medical governing bodies encourage physicians to engage in continuing professional development (CPD) activities to cultivate their knowledge and skills to ensure their clinical practice reflects the current standards and evidence base. However, physicians often encounter various barriers that hinder their participation in CPD programs, such as time constraints, a lack of centralized coordination, and limited opportunities for self-assessment. The literature has highlighted the strength of using question-based learning interventions to augment physician learning and further enable change in practice. CPD By the Minute (CPD-Min) is a smartphone-enabled web-based app that was developed to address self-assessment gaps and barriers to engagement in CPD activities.

Objective: This study aimed to assess the app using four objectives: (1) engagement and use of the app throughout the study, (2) effectiveness of this tool as a CPD activity, (3) relevance of the disseminated information to physicians' practice, and (4) acceptability to physicians of this novel tool as an educational initiative.

Methods: The CPD-Min app disseminated 2 multiple-choice questions (1-min each) each week with feedback and references. Participants included licensed staff physicians, fellows, and residents across Canada. A concurrent multimethods study was conducted, consisting of preintervention and postintervention surveys, semistructured interviews, and app analytics. Guided by the Reach, Effectiveness, Adoption, Implementation, and Maintenance framework, the qualitative data were analyzed deductively and inductively.

Results: Of the 105 Canadian anesthesiologists participating in the study, 89 (84.8%) were staff physicians, 12 (11.4%) were fellows, and 4 (3.8%) were residents. Participants completed 110 questions each over the course of 52 weeks, with an average completion rate of 75% (SD 33%). In total, 40.9% (43/105) of participants answered >90% of the questions, including 15.2% (16/105) who completed all questions. Moreover, 69% (52/75) of participants reported the app to be an effective and valuable resource for their practice and to enhance continuous learning. Most participants (63/75, 84%) who completed the postsurveys reported that they would likely continue using the app as a CPD tool. These findings were further supported by the interview data. Three key themes were identified: the practical design of the novel educational app facilitates its adoption by clinicians, the app was perceived as a useful knowledge tool for continuous learning, and the app's low-stakes testing environment cultivated independent learning attitudes.

Conclusions: The findings suggest the potential of the app to improve longitudinal assessments that promote lifelong learning among clinicians. The positive feedback and increased acceptance of the app supports it as an innovative tool for knowledge retention and CPD. Future research efforts should prioritize evaluating the app's long-term sustainability and its impact on physicians' practice, as well as exploring alternative approaches (such as artificial intelligence–based tools) for generating questions.

(JMIR Med Educ 2025;11:e69443) doi:10.2196/69443

KEYWORDS

continuing professional development; mobile app; question-based learning; lifelong learning; self-assessment; artificial intelligence

Introduction

Background

According to the Accreditation Council for Continuing Medical Education, continuing medical education (CME) supports physicians and health care teams in their lifelong learning and engagement in self-directed learning and quality improvement, leading to better care for patients and communities [1]. Continuing professional development (CPD) learning activities generally consist of a selection of educational and developmental activities, including group learning, self-directed learning, and self-assessment activities [1,2]. Such CPD activities for specialist physicians in Canada are overseen by the Royal College of Physicians and Surgeons (RCPSC) using the Maintenance of Certification (MOC) program [3]. Traditional CPD activities have been shown to have a limited impact on physicians' performance and patient outcomes [3]. Furthermore, experts suggest that systems-integrated CME activities that use a longitudinal multimodal approach allow for organizational improvements, as well as positive change in practice and patient health outcomes [4,5]. However, implementing longitudinal CME is not without challenges, including limited availability of objective practice data, inaccurate physician self-assessment, and ineffective CPD activities [5-7].

Longitudinal assessments have successfully been used in medical education where multiple-choice and short-answer questions are delivered at spaced intervals on a computer or mobile device [5,8]. More recently, this type of longitudinal assessment has been used for CPD of physicians [5,8]. One approach that has been shown to be effective in supporting clinicians' CPD is question-based learning interventions, such as test-enhanced learning (TEL) [9]. TEL has been shown to promote knowledge retention and information retrieval from memory [9]. As part of the learning sciences, spaced repetition and intentional recall are effective techniques for enhancing knowledge retention [10-13]. The process of recalling information for testing purposes further consolidates the material in the long-term memory and helps strengthen the memory trace when it is repeatedly tested [13]. In addition to repeated testing, feedback after retrieval attempt has been shown to increase the effects of TEL, as it reinforces the correct material in the long-term memory and retention [14]. This can promote reflective learning, as health care professionals can identify their strengths, weaknesses, and knowledge gaps that require further learning. Several studies have found that incorporating TEL and spaced repetition in CME programs enhances knowledge retention among clinicians [15-17].

The American Board of Anesthesiology (ANES) has developed a program of web-based multiple-choice questions (MCQs) that have been used for recertification in ANES, replacing their recertification examination [18]. This type of recertification allows ANES health care professionals to stay up-to-date on the latest developments and best practices in their field, improving their performance and patient outcomes. Furthermore, the use of this web-based MCQs CME provides an opportunity to be integrated into organizations and organizational workflows, as an effective strategy for promoting ongoing learning and development among health care professionals [1]. However, the question of how this approach could be used beyond the MOC remains largely unanswered. What is the role of digital approaches such as this one in the promotion of self-assessment, reflective learning, personalized learning, lifelong learning, and practice improvement?

Digital tools have increasingly been adapted to support pedagogies for CPD and have the potential to address limitations of self-assessment identified in existing CPD programs for specialist physicians in Canada [19,20]. Self-assessment CPD tools effectively promote lifelong learning through spaced longitudinal question-based assessments and reflective learning [4]. A team of clinicians and researchers at the University Health Network (UHN) developed a web- and mobile-based app, called CPD By the Minute (CPD-Min), that disseminates weekly self-assessments to specialist clinicians in Canada. A prototype of this progressive web-based app was initially developed by the team to assess the feasibility and utility of this concept in a Canadian setting. The prototype was first tested by 17 members

of the University of Toronto's (UofT) Department of Anesthesiology to demonstrate the feasibility of the app and to evaluate user experience, educational experience, and perceived efficacy [21]. The mobile app format was positively received, with 88% of users likely to continue using the app and 76% of users citing the app as an effective learning tool [21]. The tool was recognized for its ease of use, accessibility, and minimal time commitment. Users suggested the addition of a feedback feature to allow them to compare their performance to a peer group [21].

This Paper

On the basis of the feedback from the feasibility study, the CPD-Min pilot study was launched to further explore the efficacy of the app among ANES clinicians. This paper assesses the CDP-Min app using the four objectives related to the deployment of CDP-Min in the ANES context: (1) the effectiveness of this tool as a CPD activity; (2) ANES residents', fellows', and staffs' acceptability of this novel tool as an educational initiative; (3) relevance of the disseminated information to physicians' practice; and (4) engagement and use of the app throughout the study.

Methods

Study Design

The CPD-Min pilot study was a prospective, concurrent multimethods study, with assessment components implemented throughout the study. This multimethod evaluation of the app was guided by the Reach, Effectiveness, Adoption, Implementation, and Maintenance (RE-AIM) framework in the development of the survey and interview questions [22]. The RE-AIM framework was used to assess the design and development of the app as well as the efficacy of the intervention.

Ethical Considerations

This study was approved by the UHN Research Ethics Board (20-5883), the UofT Research Ethics Board (00040879), and the UofT CPD committee for CME credits. Participants in the study were required to review and sign the consent form sent via email by the study coordinator. The consenting process specified that participation entails agreeing to the terms of use and engaging with the CPD-Min application for the entire study duration, completing preintervention and postintervention surveys, and participating in an optional follow-up interview. The consent form outlined the voluntary nature of participating in this study, the benefits and risks, and data handling and storage. Participants were able to withdraw from the study at any time without penalty. The survey and interview data gathered were stored on confidential databases on secured UHN servers or encrypted and password-protected UHN devices. Interview transcripts were deidentified before analysis. After the transcripts have been verified for accuracy and data coding is complete, the audio recordings were destroyed. Data from the surveys were entered into a database on the UHN network shared drive, where only study team members have access. Only aggregated and thematically analyzed data were shared with study collaborators.

Education Intervention

The CPD-Min app was developed at UHN by a multidisciplinary team of clinicians, researchers, and software developers. Before the launch of the study, members of the research team (ie, experts and associate deans with relevant academic and professional backgrounds in the field of CPD) engaged in usability testing to assess and approve the app's design.

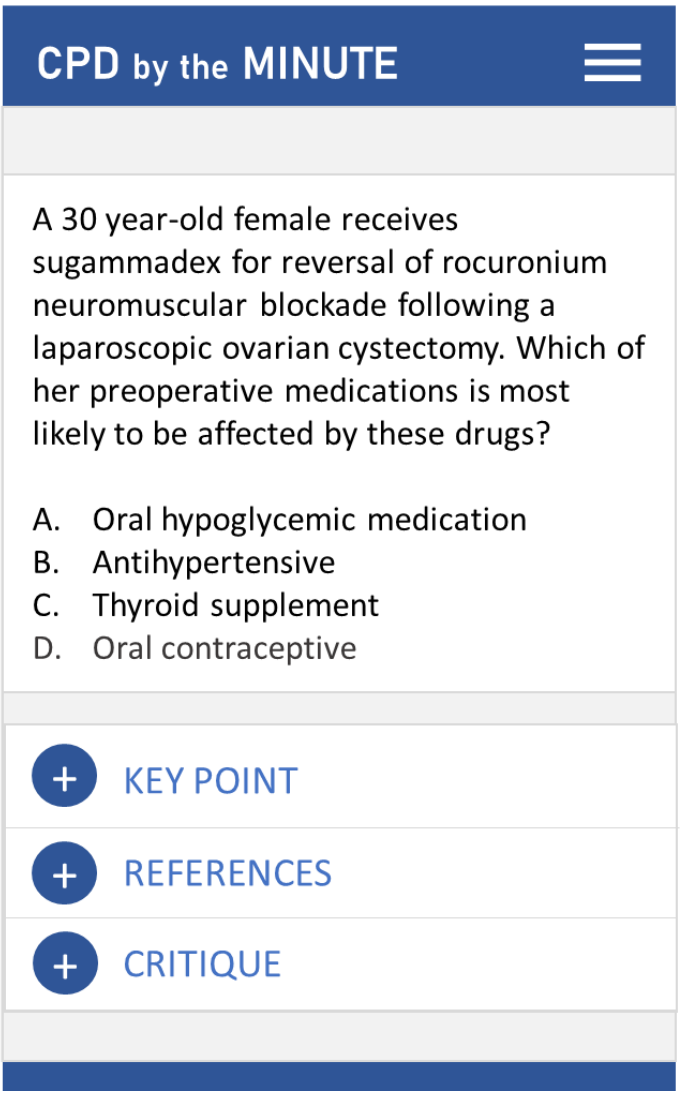
Description of the CPD-Min App

The CPD-Min app is a smartphone-enabled web-based app that administered 2 questions (1 min each) each week over a 52-week period. Following participants' response to the question, the app provided detailed feedback, including the correct answer, critiques of the wrong answer, key points, and references. This provided the participants with an opportunity to identify gaps in their knowledge and further explore these topics using the resources provided. In addition to answering the questions, participants were asked to rate each question based on their confidence (ie, very confident, somewhat confident, or not at all confident) and its relevance (ie, very relevant, somewhat relevant, or not relevant) to their practices. Moreover, an aggregated summary of all participants' responses and reflection ratings of each question were available for them to review after each week on the app's dashboard. This feature allowed participants to compare their performance to their peers and reflect on their performance accordingly. Finally, 2 knowledge assessments were conducted, at the midpoint (T1) and the end point (T2) of the study, where 5 questions were recycled back to participants if answered incorrectly or omitted.

Question Development

A total of 7 residents and staff from the UHN and UofT ANES departments were recruited to assist with question development; research associates provided guidance to ensure the consistency and quality of questions. A peer-review process involving 2 content experts was then conducted to validate, refine, and provide feedback on the content, quality, and relevance of the questions to CPD. Each question was multiple choice and contained a question stem, 4 to 5 answer choices, a key point, a critique of incorrect answers, and references (Figure 1). Questions developed were used to assess health care professionals' knowledge of important practice standards and guidelines and to stimulate problem-solving in 23 general and specific subspecialty topics, such as critical care and resuscitation, cardiology and cardiovascular surgery, neurology and neurosurgery, and general complications. The MCQs developed were a mixture of clinically related case-based questions as well as calculation and statistical questions. Approximately 100 ANES-based questions were developed and reviewed. The number of questions developed was based on the size of the anesthesiology specialty in Canada. A Microsoft OneNote database was created to allow for collaboration and monitoring of questions developed in real time. The questions were developed iteratively over a 6-month period to disseminate new practice information as the emerging needs of clinical practice continuously evolve.

Figure 1. A CPD By the Minute mock-up of a clinical pharmacology question for anesthesiology.



App Development and Usability Testing

Usability testing was conducted to assess and approve the design of the CPD-Min app as well as to identify opportunities to improve the user experience before the release of the final version of the app and the launch of the pilot study, thereby optimizing the app for better user engagement. A total of 7 experts and associate deans with relevant academic and professional backgrounds in the field of CPD were recruited for usability testing (including a pre- and postsurvey) via purposive sampling. Each testing session took place, via Microsoft Teams, due to COVID-19 public health restrictions. Furthermore, the Post Study System Usability Questionnaire, including 3 subscales—system usefulness, information quality, and interface quality—was administered after the usability test to evaluate the usability of the app and their perceived satisfaction. Data from the sessions were evaluated using the 10 Usability Heuristic Principles and Severity Scale by Nielsen [23].

Recruitment

Recruitment materials (email invitations with consent forms, digital posters, and a recruitment video) developed by the research team were sent out to anesthesiologists through the UofT's departments of ANES and specialty society organizations (Association of Canadian University Departments of Anesthesia, Continuing Education and Professional Development) across Canada, where a snowball sampling approach was used for recruitment.

Before conducting the study, participating ANES physician staff, fellows, and residents gave their written informed consent via electronically signed documents. The study was conducted in 2 sequential cohorts over a 52-week period, initiated in October (cohort 1) and November 2021 (cohort 2).

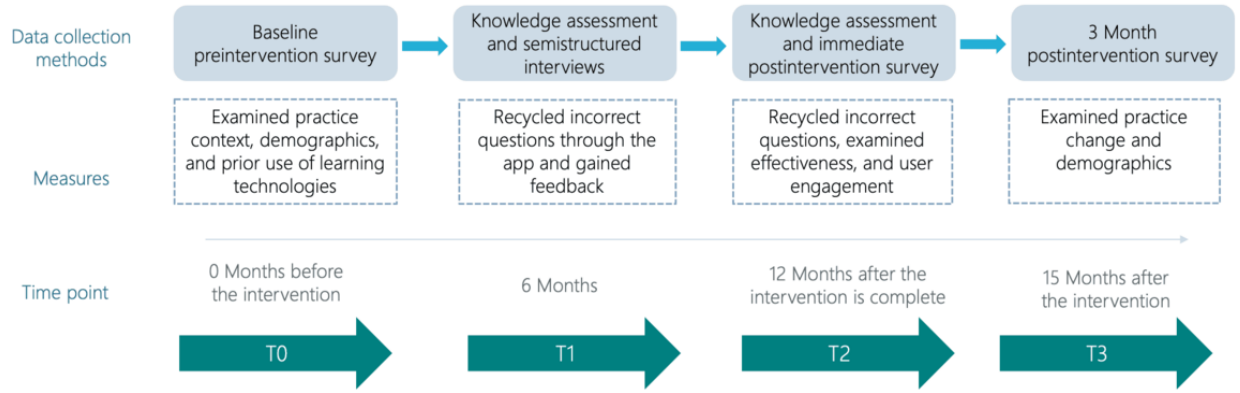
Data Collection

Data collection occurred throughout 4 distinct time points of the study (T0, T1, T2, and T3) and included electronic preintervention and postintervention surveys, knowledge assessments, and semistructured interviews with participants who had consented (Figure 2). The 3 surveys (preintervention, postintervention and 3 months after the intervention) were administered via REDCap (Research Electronic Data Capture; Vanderbilt University). The preintervention survey was administered before participants' engagement with the app (T0), and it included questions on practice context and demographics and prior use of learning technologies.

At the midpoint of the study (T1), semistructured interviews were conducted and recorded on Microsoft Teams with participants from ANES who provided consent. They were on average 17 (SD 0.29) minutes in duration, with questions on acceptability, relevance, expansion, and potential impact of CPD-Min. At weeks 26 and 52, the app recycled a maximum of 5 questions back to participants based on previous user responses using a random selection process. If ≥1 question was answered incorrectly before the first knowledge testing time point (T1), a maximum of 5 questions were randomly selected from a list of incorrectly answered questions for a second attempt. Recycled questions that were either omitted or incorrectly answered at T1 were carried forward as part of the pool or list of eligible questions to be revisited once more during the final knowledge testing week (T2). For the users who exhibited perfect scores at either T1 or T2, there was no knowledge testing. In addition, app analytics (eg, number of questions answered, correct questions, and time spent answering questions) were collected throughout the 52-week study period.

The postsurvey was sent to participants at the end of the pilot program (T2), and it included questions on the effectiveness and appeal of the app, influence of peer comparison, user engagement behavior, quality assessment, app design, and the System Usability Scale (SUS; Multimedia Appendix 1). The questions were designed by the research team based on peer-reviewed literature, and a subset of them was pilot-tested during the feasibility study. The SUS is a widely used 10-item questionnaire designed to assess the usability and effectiveness of a digital tool [24]. For example, participants rated various aspects of the app, such as the ease of use, functionality integration, the need for technical support, and overall complexity, on a scale from 1 to 5, ranging from "strongly disagree" to "strongly agree." A survey was sent to participants 3 months after the intervention (T3) with questions recapturing demographics and others exploring practice change or improvement.

Figure 2. Data collection tools and measures.



Data Analysis

Quantitative Analyses

All quantitative data were cleaned and analyzed using SPSS (IBM Corp). Preintervention and postintervention survey data were analyzed for descriptive statistics and correlational findings. With regard to the knowledge assessments, descriptive results were used to describe the extent of learning due to question recall. This included the number of correct or incorrect responses, the number of responses that timed out or were not started, question relevance, response times, and perceived confidence in answering questions. A series of paired 2-tailed *t* tests were conducted to compare knowledge assessments' mean scores and evaluate change in question responses at different time points. In addition, Spearman rank correlation tests were conducted to assess the relationship between the proportion of participants who answered correctly or incorrectly and degrees of confidence and relevance. A *P* value of <.05 was used to determine significance throughout.

Qualitative Analysis

For the semistructured interviews, data were transcribed by a professional transcription service and cleaned for clarity and to remove any identifying information. Data were deductively and inductively analyzed using NVivo (Lumivero) following the thematic analysis process by Braun and Clarke [25]. Participant interviews were first deductively analyzed using the RE-AIM framework to code the data. Following the deductive analysis, the step-by-step process developed by Braun and Clarke [25] was used to inductively generate themes from the coded data by searching the data to capture participants' thoughts and perceptions of the app and reviewing the preliminary themes before defining and describing them.

Finally, in addition to the RE-AIM framework being used to frame the findings from this pilot study, the continuing education outcome framework by Moore et al [26] was used to evaluate the learner's experience and the app as a CPD activity (Table 1).

Table 1. CPD By the Minute (CPD-Min) outcome measures using the Reach, Effectiveness, Adoption, Implementation, and Maintenance (RE-AIM) framework.

RE-AIM framework components	Pilot study outcome measures
Reach	• Number of American Board of Anesthesiology participants taking part in the CPD-Min program study • Demographics of participants
Adoption	• Participant completion rates were tracked throughout the 1-year pilot study
Effectiveness	• Satisfaction using the CPD-Min app • Mean scores from questions attempted throughout the program • Mean scores from knowledge assessments
Implementation	• Quality assessment of the content and delivery modes for the program
Maintenance	• Prospective measure of self-reported continued use of the app • Self-reported recommendation of the app

Results

Study Demographics: Reach

A total of 105 ANES participants were recruited for both cohorts to participate in the CPD-Min app study. Out of the 105 ANES participants in the study, there were varying response rates for the surveys administered preintervention, postintervention, and 3 months after the intervention. The response rate for the presurvey was 8.6% (9/105), while the response rate for the postsurvey was 71.4% (75/105). Due to the low response rate in the presurvey, a 3-month postsurvey (41/105, 39% response rate) was administered to gather information on demographics and practice change.

Of the 105 participants, there were 89 (84.8%) staff, 12 (11.4%) fellows, and 4 (3.8%) residents who took part in the study. Of the 89 staff, 40 (45%) reported on their years of practice with a variety of practice experience, but most reported <5 practice years (11/41, 27%) or 20 to <30 practice years (8/41, 20%). Most participants were in the age categories of 36 to 45 years (14/41, 34%) and 46 to 55 years (12/41, 28%), with most of them identifying as a cisgender man (33/41, 81%). Table 2 shows more details on participants' demographics.

A large proportion of participants recruited were practicing in the Greater Toronto Area (eg, Mount Sinai, Humber River Hospital, Sunnybrook Health Sciences Centre, and UHN). Others were from hospitals and universities in Quebec, Manitoba, and British Columbia. Most participants (29/41, 71%) from the survey indicated that they practiced at an academic health sciences center or a teaching hospital, with varying participation in CPD activities, such as attending departmental rounds (39/41, 95%), meeting informally with colleagues (35/41, 85%), attending and presenting in conferences and courses (34/41, 83%), and many more (refer to Table 2 for the full list).

Table 2. Demographics and continuing professional development (CPD) activities of the CPD By the Minute app participants from the 3-month postsurvey (N=41).

Characteristics	Participants, n (%)
Staff (years of experience)	40 (98)
<5	11 (27)
5 to <10	6 (15)
10 to <15	5 (12)
15 to <20	7 (17)
20 to <30	8 (20)
>30	3 (7)
Fellow (year 3)	1 (2)
Practice setting	
Academic health science center or teaching hospital practice	29 (71)
Community-based hospital practice	12 (29)
Age (years)	
26-35	6 (15)
36-45	14 (34)
46-55	12 (29)
56-65	7 (17)
>65	2 (5)
Gender	
Cisgender Man	33 (80)
Cisgender Woman	7 (17)
Transgender Person	1 (2)
Annually practiced CPD activities	
Attending formal case conferences to discuss patient care	29 (71)
Performing chart audits or reviewing performance data	14 (34)
Meeting informally with colleagues	35 (85)
Attending journal clubs	27 (66)
Attending departmental rounds	39 (95)
Attending and presenting in conferences and courses	34 (83)
Engaging in online activities or courses	33 (80)

Research Objective 1: Engagement and Use of the App Throughout the Study—Adoption

The *adoption* of the app was measured by looking at participants' completion rate and use of the app throughout the 52-week pilot study (first-level of the framework by Moore et al [26] - participation). The app was assessed based on its effectiveness as a CPD activity and participants' perceptions of their use of the app with regard to the *effectiveness* outcome.

The CPD-Min app sent out 110 questions to participants throughout the 52 weeks. In total, 40.9% (43/105) of participants responded to more than 90% of the questions, with 15.2% (16/105) completing every question. Participants had an overall high mean completion rate of 75% (SD 33%) throughout the study. However, there was a significant difference in the average response rate before and after the midpoint assessment at 26 weeks (t_{104}=−5.06; P<.001, 2-tailed). On average, the premidpoint response rate was 5 questions higher than the postmidpoint response rate (95% CI −7.07 to −3.05). Normality was assumed by visually inspecting Q-Q plots and considering the larger sample size (n=104) in accordance with the central limit theorem.

Research Objective 2: Effectiveness of This Tool as a CPD Activity—Effectiveness

Overview

The average number of questions answered per participant was 82 (SD 32.9; range 0-110) questions, with a mean correct response of 43 questions per participant (SD 19.5; range 0-76). They spent an average of 35.3 (SD 5.9; range 14-48) seconds

to provide an answer to each weekly question. Spearman rank correlation tests (r_s) showed weak positive correlations between questions answered correctly and both participants' confidence (r_s=0.272; P<.001) and relevance (r_s=0.143; P<.001) rating of those correct questions.

Semistructured Interviews

A total of 15 ANES participants were interviewed from both cohorts. On the basis of the thematic analysis of the interview data, the following themes regarding participants' thoughts and perceptions on the CPD-Min app were generated: (1) the practical design of the novel educational app facilitated its adoption by clinicians, (2) the app was perceived as a knowledge tool for continuous learning, and (3) the app's low-stakes testing environment cultivated independent learning attitudes.

Theme 1: The Practical Design of the Novel Educational App Facilitates Its Adoption by Clinicians

The convenience and practicality of this novel app facilitated participants' adoption of the learning tool (second-level of the framework by Moore et al [26] - satisfaction). Participants noted that both the practicality and convenience of having this continuing learning tool in an app format, along with its short time commitment, helped them maintain their participation and engagement throughout the study period and possibly beyond:

> *There's a practicality to the app and to the timeframes required by the question-based module, which is interesting and fun. There's something practical about having bite-size elements of CPD that can be done anywhere at any time, which is very practical.* [ID 109]

According to participants in the interview, they were consistent with using the app throughout the program, taking approximately 5 minutes to answer questions and review materials if necessary. In addition, most participants found the weekly reminder emails to be useful in prompting them to answer the question right away or to schedule a time to do it during the week. When it came to the design of the app, while most participants noted the ease of use and practicality of it, there were a few suggestions on improving some features within the app. One participant suggested improving access to previous questions and resources by integrating a quick search function to refer to in the future if needed.

Research Objective 3: Relevance of the Disseminated Information to Physicians' Practice (Theme 2: The App Was Perceived as a Useful Knowledge Tool for Continuous Learning)

The app provided health care professionals with an opportunity to engage in a continuous learning experience. Most participants mentioned using the app as a tool to identify gaps in their knowledge by reflecting on their scores and the content of the app. The CPD activity was described as a "refresher" of previous knowledge for the licensing exams by several participants. Moreover, a few participants expressed that although some of the subspecialty content was not relevant to their practice, it was useful in identifying subspecialty gaps and providing an overview of clinical topics within different subspecialties in the field:

> *Some of the content is not necessarily relevant to my practice. But I think as an anesthesiologist, you never know what's going to come up. So, it's good to keep a broad overview of things, even if you're not doing them every day.* [ID 85]

For some participants, having a broad range of subspecialty topics provided them with a refreshing overview of knowledge regardless of its relevancy to their practice. By contrast, a few participants reported that they would have appreciated a more tailored approach to the content of the app. Specifically, 1 participant reported that the number of questions that did not relate to their practice "did not necessarily increase [their] learning." As a recommendation, a few participants suggested having the option to choose specific topics of interest at the start of using the app. Moreover, interview participants reported engaging in the supplementary material and feedback provided to them; some used the references provided to them to further explore specific topics they were interested in.

Research Objective 4: Acceptability Among ANES Residents, Fellows, and Staff of This Novel Tool as an Educational Initiative (Theme 3: The App's Low-Stakes Testing Environment Cultivates an Independent Learning Attitude)

The app uses low-stakes repeated testing to cultivate lifelong learning among health care professionals. Participants noted that low-stakes and repeated testing design of the app was useful in keeping them engaged throughout the program and further cultivated their lifelong learning:

> *Doing the CPD is extremely low stakes. I do this because, well, I want to keep lifelong learning. I like to learn. I care to be as good of an Anesthesiologist as I can. And just as well to keep on practicing and learning and reviewing the material. And this, in a way, forces you or at least prompts me to do it, in an easy bite-sized manner. The low stakes is what keeps me in. If I knew that if I failed this, I'd lose my license, I be slightly more worried. This is only win-win. I need to do CPD. I do CPD. It's easy. Why not?* [A109]

Similar to the themes generated, findings from the postsurvey confirmed participants' perceptions during the midpoint interviews, where participants highlighted the effectiveness of the app as an educational tool, as it not only tests their current knowledge but also exposes them to the newest guidelines, studies, and information related to their practice. Concurrently, while exploring the quality of the app's content in the postsurvey, most respondents (57/75, 76%) were satisfied with the variety of topics offered, and even more participants (64/75, 85%) reported the quality of the questions and supplementary material as above average (ie, excellent or good).

The themes generated from the interview data were further supported by the survey findings (Table 3), as over one-third of the participants (29/75, 39%) reported that they would *always* or *often* check after answering correctly, while even more

(57/75, 76%) reported *always* or *often* checking the feedback after incorrect responses. More than half of the participants (43/75, 57%) did not report any more engagement with the app after answering correctly rather than incorrectly. Furthermore, most participants (58/75, 77%) were intrigued to find out the answers to the questions asked of them weekly. Most participants (45/75, 60%) were *occasionally* motivated to seek out additional, related self-directed learning external from the app. The CPD-Min app was reported as a frequent, low-stakes knowledge testing platform by most participants (62/75, 83%). Most participants (65/75, 87%) did not perceive the 2 knowledge assessment weeks to be too stressful or anxiety inducing.

With regard to participants' learning, on average, participants scored 16% (SD 33.1%) less on their end point knowledge assessment compared to their midpoint knowledge assessment. From the final survey results, most participants (38/41, 93%) reported that the app enhanced their clinical knowledge and helped them identify gaps in their learning (third level of a framework by Moore et al [26] - learning). They also found the knowledge assessments helpful. Furthermore, it provided them with adequate resources to further their learning. For most participating clinicians (35/41, 85%), the app helped them improve or change their practice, and for some (31/41, 76%), it also contributed to their confidence with their clinical knowledge.

Table 3. Postsurvey response on user engagement (Likert scale) by users of the CPD By the Minute app (N=75).

	Always or strongly agree, n (%)	Often or agree, n (%)	Occasionally or neither agree nor disagree, n (%)	Rarely or disagree, n (%)	Never or strongly disagree, n (%)
How likely are you to review the supplementary materials provided (eg, key points, references, and critiques) after a correct response?	10 (13)	19 (25)	28 (37)	15 (20)	3 (4)
How likely are you to review the supplementary materials provided (eg, key points, references, and critiques) after an incorrect response?	32 (43)	25 (33)	13 (17)	3 (4)	2 (3)
I was more likely to engage with the app if a previous question was correct rather than incorrect.	2 (3)	4 (5)	26 (35)	19 (25)	24 (32)
The results I received after answering a question motivated additional, related self-directed learning external to the app.	4 (5)	19 (25)	45 (60)	6 (8)	1 (1)
I was intrigued to find out the answers to the questions asked of me.	26 (35)	32 (43)	15 (20)	2 (3)	0 (0)
Participating in either the routine weekly questions or KA[a] weeks was too stressful and induced anxiety.	0 (0)	2 (3)	8 (11)	26 (35)	39 (52)

[a]KA: knowledge assessment.

Implementation of the CPD-Min App

With regard to the *implementation* component of the CPD-Min app, the delivery or format of the app was assessed using both the interviews and postsurvey. Participants spoke positively about the CPD-Min app's format, content, and delivery method (Table 4). Participants noted that the weekly question count was manageable within their workload and helped them stay consistent in using the app. Several participants also highlighted that the weekly reminders were beneficial in prompting them to complete their questions for the week. Moreover, the questions were mostly relevant to realistic case scenarios that were applicable to their practice. Finally, many participants positively described the appropriateness of the format for CPD and continuous learning.

Likewise, the findings from the postsurvey regarding the quality and usability of the app concur with the abovementioned findings from the midpoint interviews. Most postsurvey participants (49/75, 65%) reported the format and layout (eg, character count, image, and font size) of the questions in the app to be appropriate for a time-sensitive response. The app's usability was measured using the SUS, and the mean scale score as reported by participants was 77.7 (SD 6.7; range 60-88). The usability score was rated on a scale of 100, with an average score being 68 and a score >80 considered above average [24]. A score of 77 indicated that the CPD-Min app was fairly usable; however, there were still areas that could be improved to enhance the overall user experience.

While majority of participants (51/75, 68%) found the peer reflection feature of the app (ie, being able to compare their performances to their peers) helpful and about half of the participants (30/75, 40%) thought it was important, most participants (54/75, 72%) reported *never* discussing the results or material with their peers. Almost one-third of the participants (21/75, 28%) reported discussing their results with their peers from once a week to less than once a month. This was concurrent with findings from the interview, as most interview participants reported never discussing the material or content from the app with colleagues. When asked to elaborate, a few interview participants highlighted that the CPD activity in this app format

did not facilitate discourse and discussion among peers and colleagues. Participants found themselves using the app as a self-guided learning tool as opposed to other CPD activities, such as conferences. A few participants (2/15, 13%) suggested integrating an in-app feature that allowed interaction or discussion with other CPD-Min users.

Table 4. Postsurvey responses on quality assessment (Likert scale) by users of the CPD By the Minute app (N=75).

	Strongly agree or excellent, n (%)	Agree or good, n (%)	Neither agree nor disagree or average, n (%)	Disagree or below average, n (%)	Strongly disagree or poor, n (%)
I was satisfied with the variety of topics and subject matter.	12 (16)	45 (60)	12 (16)	4 (5)	2 (3)
The quality of questions and supplementary materials offered was (poor to excellent).	21 (28)	43 (57)	9 (12)	2 (3)	0 (0)
While answering a given question, I felt the format and layout were appropriate for a time-sensitive response.	9 (12)	40 (52)	8 (11)	14 (19)	4 (5)

Maintenance

CPD-Min program's *maintenance* feature focused on the individual level, exploring participants' likelihood of continued use of the CPD-Min app and recommendation to peers and colleagues (Table 5). Most participants (63/75, 84%) who reported that they were likely to continue using the CPD-Min app, also reported that they would continue using it as an ongoing CPD activity (61/75, 81%). Most participants (67/75, 89%) reported that they would likely continue using the app if they received section 3 MOC (self-assessment) CME credits. Similarly, 89% (67/75) of the participants reported that they would continue using the CPD-Min app for >12 months. Most participants (52/75, 69%) found the app to be an effective learning tool for their practice, with about 28% (21/75) reporting that the app was *somewhat* effective as a learning tool. In terms of recommending the app to colleagues, 79% (59/75) of the participants said that they would recommend the app to their colleagues, and 81% (61/75) said that they would specifically recommend it to residents and fellows.

Table 5. Postsurvey responses on appeal and effectiveness (Likert scale) by users of the CPD By the Minute (CPD-Min) app (N=75).

	Very likely or effective, n (%)	Likely or effective, n (%)	Somewhat, n (%)	Not very, n (%)	Not at all, n (%)
How likely are you to continue using the CPD-Min app?	43 (57)	20 (27)	8 (11)	4 (5)	0 (0)
How likely are you to continue using the app as an ongoing CPD[a] activity?	40 (53)	21 (28)	9 (12)	4 (5)	0 (0)
Would you likely continue using the app if you continued receiving section 3 MOC[b] (self-assessment) CME[c] credits?	53 (71)	14 (19)	6 (8)	1 (1)	0 (0)
How effective is this as a learning tool for your practice?	19 (25)	33 (44)	21 (28)	2 (3)	0 (0)
How likely are you to recommend the app to your colleagues?	34 (45)	25 (33)	15 (20)	1 (1)	0 (0)
How likely are you to recommend the app to residents or fellows?	36 (48)	25 (33)	9 (12)	3 (4)	2 (3)

[a]CPD: continuing professional development.

[b]MOC: Maintenance of Certification.

[c]CME: continuing medical education.

Discussion

Principal Findings

The CPD-Min app disseminated 110 peer-reviewed questions to 105 ANES clinicians over the span of 52 weeks. Upon the evaluation of this app as a CPD tool, three major themes were identified: (1) the practical design of the educational app facilitated its adoption by clinicians, (2) the app was perceived as a useful knowledge tool for continuous learning, and (3) the app's low-stakes testing environment cultivated independent learning attitudes. In addition to clinicians' positive perceptions of the CPD-Min app, they noted that using it helped them identify gaps in their clinical knowledge. The implementation of a longitudinal self-assessment activity such as CPD-Min has various strengths, enabling health care professionals to assess their own knowledge and identify areas where they may need to improve. CPD is essential to maintaining and improving clinical practice among physicians. It allows health care professionals to stay up-to-date with the latest developments in their field and to acquire new skills and knowledge that can help them deliver better care to their patients. However, traditional CPD activities are associated with several limitations in their implementation (eg, lack of time, cost, and reliance on self-assessments) and rate of dissemination [2-4].

Among the ANES clinicians who participated in the survey, 63% (26/41) were aged between 35 and 55 years and 80% (33/41) were male. The age of this sample was comparable to the general ANES population according to reports from the Canadian Medical Association, noting that >50% of the

Canadian ANES population was aged between 35 and 55 years [27]. However, our sample had slightly more ANES clinicians identifying as a cisgender man (33/41, 80%) compared to the general population (67%) [27]. Similarly, our population's work setting was comparable to the general Canadian ANES population. According to the 2019 Canadian Medical Association report, most ANES clinicians reported academic health science centers and community hospitals as their top primary work settings [27].

The CPD-Min pilot program had high engagement as participating ANES clinicians had an average of 75% (SD 33%) completion rate through the entire year of the study program. The average clinician participation rate in the CPD-Min program was consistent with participation levels from 10 previously evaluated educational delivery programs for health care professionals, as explored in a systematic review by Phillips et al [28]. Clinician participation levels in self-assessment CPD programs could be attributed to a number of factors, including clinical relevance, optimal number of questions per day, and their spacing [29]. In addition, the decrease in participant response rate over the course of the program is not an uncommon occurrence, as the drop in participation over time could be due to fatigue, lack of motivation, or other personal or external factors [30-32].

With the CPD-Min app, clinicians' adoption of the app and participation in the program could be attributed to the ease of use and ability to quickly adapt to this microlearning pedagogy method. The app provided specialized physicians with an opportunity to participate in continuous learning in a flexible and accessible manner that encouraged sustained engagement [33]. The microlearning style presented by this CPD app was a helpful format for continuous learning and development for physicians, as they were able to be consistent with using the app throughout the program, taking approximately 5 minutes to answer questions and review materials if necessary [34]. Clinicians' responses to the CPD-Min app echoed key takeaways of mobile-microlearning pedagogy within workplaces that highlight the usefulness and demand of just-in-time learning, which provides clinicians with bite-sized lessons that can be immediately applied within their clinical practices [33,35-37]. In addition, this nontraditional mode of CPD allows for spatial and temporal flexibility among clinicians with demanding workloads who are seeking to engage in CPD. For instance, participating clinicians report being able to engage in learning "on the go," allowing them to seamlessly integrate their continuous learning within their workflows. Participating clinicians noted that they were always able to find time to do it throughout their working days (eg, during a coffee break or free time waiting in the operating room).

The app's low-stakes testing environment encouraged independent learning attitudes. Participants reported low anxiety and stress while using the app and actively engaged with supplementary resources and feedback provided after each question. While the use of the app seemed to encourage continued learning through the just-in-time low-stakes learning environment, the structure of the app did not facilitate discussion of the material with colleagues, as highlighted by participants. The facilitation of discussion and interaction during CPD uptake provides clinicians the opportunity to reflect on their performance and knowledge in reference to their peers. Although the CPD-Min app did not facilitate collaboration and peer interaction, a key feature of the app was its weekly summary. This summary, detailing participants' and their colleagues' performance, served as a tool for reflection and played a crucial role in motivating clinicians to remain actively engaged with the app. This finding aligns with previous literature that emphasizes the importance of clinician reflection in CPD activities, helping them identify gaps in performance and knowledge [38,39].

The CPD-Min app disseminated information to clinicians in a timely manner [40]. The app used the declarative knowledge state with key features such as feedback, key critiques, and references to translate knowledge to clinicians' practice [41]. This finding is consistent with a recent study that reported the benefits of immediate feedback for knowledge retention [42]. With the rapidly evolving research knowledge in health care, there seems to be a slow rate in knowledge uptake and dissemination into practice [43]. The well-received engagement with the CPD-Min app and uptake in knowledge, as well as engagement with the resources, indicates the potential of using such a mobile question-based app to translate and disseminate new and innovative knowledge in the field. Repeated testing has been shown to enhance knowledge retention and promote higher-level cognitive processing more effectively than traditional learning methods [42,44].

The CPD-Min app aimed to address gaps in CPD self-assessment for clinicians and facilitate their ability to identify knowledge gaps using just-in-time learning, which prioritizes meeting clinicians' current learning needs. While the purpose of this app was not to test knowledge, it served the purpose of a self-assessment tool for clinicians. Many participants (38/41, 93%) noted that the app helped them identify knowledge gaps and enhanced their clinical knowledge. In addition to participants' self-assessment of their knowledge, they reported that the app played a role in improving their practice by guiding their learning and highlighting the overall effectiveness of the intervention. Similar findings were reported in a pilot study of family physicians, indicating that most clinicians continued to integrate the longitudinal assessment as a form of continuous learning and changed their practice because of participating in the program [45]. Interestingly, another study found that in high-stakes assessment, participants were more accurate but less confident in their responses [44]. Low confidence could be attributed to the high-stakes nature of the assessment [44]. Overall, the results of the CPD assessment program suggested that the CPD-Min app was an important learning pathway for anesthesia clinicians and was perceived as a useful knowledge tool for continuous learning. However, further research is needed to explore the effectiveness of the app in promoting knowledge retention and the accuracy of physicians' self-assessment of their knowledge. Integrating this form of longitudinal assessment into organizational workflows could enhance physician performance and, ultimately, improve patient outcomes.

Participants found the CPD-Min app to be an effective tool for their practice as it helped them identify knowledge gaps in their

practice. Overall, based on the findings from both the survey and interviews, participants perceived the app to be useful for their practice. Regarding the perceived ease of use, the high SUS score suggests that the app is well designed and user-friendly, which could lead to its adoption and integration into clinical practice.

Limitations

This study has a few limitations, which should be acknowledged. The generalizability of the study is limited because the study population focused only on one medical discipline (ie, ANES), and thus, the findings may not be applicable to other medical specialties. Future studies should replicate the study design in various medical fields to determine the app's generalizability and usefulness in different contexts. In addition, study recruitment and intervention were administered during the COVID-19 pandemic, which limited the recruitment diversity of clinician roles and significantly impacted the participants' engagement with the study's intervention. Although the intervention was conducted remotely, the COVID-19 pandemic may have influenced clinicians' work schedule and stress levels, ultimately affecting their overall engagement with the study intervention. Another limitation of the study was that the question development process was labor intensive. Large language models could be used to rapidly prototype the questions in the future. Finally, due to the randomized nature of the questions disseminated through the app, inferences could not be made from the knowledge assessment findings, as it would require controls such as exact questions and spacing in administration. The highlighted limitations should be considered when interpreting these findings. Future research should aim to address these limitations to provide a more comprehensive understanding of the app's effectiveness as a continuous learning tool for clinicians.

Conclusions

The CPD-Min app was positively received and accepted as an educational initiative by ANES clinicians, and its practical design encouraged continued use of the app by the clinicians. Furthermore, the information was extremely relevant to participants' practices while also allowing them to brush up on knowledge from subspecialties they might not frequently practice. The CPD tool allowed ANES clinicians to recognize knowledge gaps and promote continuous learning. This app uses a practical longitudinal approach to incorporating CPD into the workflow and provides opportunities to facilitate the dissemination of new clinical information and updates across health care organizations. Integrating mobile microlearning activities into the workflow allows health care professionals to engage in ongoing learning and development more easily, leading to improved performance and outcomes. While the initial results are promising, further evaluation of the app is necessary to fully understand its long-term implications and the sustainability of the app over time. Another key area of evaluation is assessing practice changes among clinicians using the app and its impact on organizational workflows. In summary, while the CPD-Min app has shown promise in enhancing learning and knowledge retention, further evaluation is needed to understand its long-term implications. This includes examining the sustainability of the CPD activity and assessing the impact on clinical practice and patient outcomes.

Acknowledgments

The authors wish to thank Ms Tran Truong, Mr Alex Zisman, Mr Herm Sidhu, and Ms Katherine Sue at Techna, University Health Network for developing the mobile app. The authors also thank Dr Allan Okrainec, Dr Dave Davis, Mr Spencer Williams, and Ms Sophia Li for their support with the study. The authors would like thank the Southeastern Ontario Academic Medical Organization for its generous support of this work through the AHSC AFP Innovation Fund.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Postsurvey questions.
[DOCX File , 40 KB - mededu_v11i1e69443_app1.docx]

References

1. Marinopoulos SS, Dorman T, Ratanawongsa N, Wilson LM, Ashar BH, Magaziner JL, et al. Effectiveness of continuing medical education. Evid Rep Technol Assess (Full Rep) 2007 Jan(149):1-69. [Medline: 17764217]
2. The MOC program. Royal College of Physicians and Surgeons of Canada. URL: https://news.royalcollege.ca/en/cpd/moc-program [accessed 2024-08-14]
3. Horsley T, Moreau K, Lockyer J, Zeiter J, Varpio L, Campbell C. More than reducing complexity: Canadian specialists' views of the royal college's maintenance of certification framework and program. J Contin Educ Health Prof 2016;36(3):157-163. [doi: 10.1097/CEH.0000000000000099] [Medline: 27583991]
4. Lyu X, Li S. Professional medical education approaches: mobilizing evidence for clinicians. Front Med (Lausanne) 2023 Jul 28;10:1071545 [FREE Full text] [doi: 10.3389/fmed.2023.1071545] [Medline: 37575990]

5. Price DW, Davis DA, Filerman GL. "Systems-integrated CME": the implementation and outcomes imperative for continuing medical education in the learning health care enterprise. NAM Perspect 2021;2021:10.31478/202110a [FREE Full text] [doi: 10.31478/202110a] [Medline: 34901778]
6. Cook DA, Blachman MJ, Price DW, West CP, Berger RA, Wittich CM. Professional development perceptions and practices among U.S. physicians: a cross-specialty national survey. Acad Med 2017 Sep;92(9):1335-1345. [doi: 10.1097/ACM.0000000000001624] [Medline: 28225460]
7. Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. JAMA 2006 Sep 06;296(9):1094-1102. [doi: 10.1001/jama.296.9.1094] [Medline: 16954489]
8. Price DW, Swanson DB, Irons MB, Hawkins RE. Longitudinal assessments in continuing specialty certification and lifelong learning. Med Teach 2018 May 24;40(9):917-919. [doi: 10.1080/0142159x.2018.1471202]
9. Larsen DP, Butler AC, Roediger HL3. Test-enhanced learning in medical education. Med Educ 2008 Oct;42(10):959-966. [doi: 10.1111/j.1365-2923.2008.03124.x] [Medline: 18823514]
10. Glenberg AM, Lehmann TS. Spacing repetitions over 1 week. Mem Cognit 1980 Nov;8(6):528-538. [doi: 10.3758/bf03213772] [Medline: 7219173]
11. Van Hoof TJ, Madan CR, Sumeracki MA. Science of learning strategy series: article 2, retrieval practice. J Contin Educ Health Prof 2021 Apr 01;41(2):119-123. [doi: 10.1097/CEH.0000000000000335] [Medline: 34057909]
12. Van Hoof TJ, Sumeracki MA, Madan CR. Science of learning strategy series: article 1, distributed practice. J Contin Educ Health Prof 2021 Jan 01;41(1):59-62. [doi: 10.1097/CEH.0000000000000315] [Medline: 33044392]
13. Toppino TC, Kasserman JE, Mracek WA. The effect of spacing repetitions on the recognition memory of young children and adults. J Exp Child Psychol 1991 Feb;51(1):123-138. [doi: 10.1016/0022-0965(91)90079-8] [Medline: 2010724]
14. Green ML, Moeller JJ, Spak JM. Test-enhanced learning in health professions education: a systematic review: BEME Guide No. 48. Med Teach 2018 Feb 01;40(4):337-350. [doi: 10.1080/0142159x.2018.1430354]
15. Kerfoot BP, Baker H. An online spaced-education game for global continuing medical education: a randomized trial. Ann Surg 2012 Jul;256(1):33-38. [doi: 10.1097/SLA.0b013e31825b3912] [Medline: 22664558]
16. Kerfoot BP, Lawler EV, Sokolovskaya G, Gagnon D, Conlin PR. Durable improvements in prostate cancer screening from online spaced education a randomized controlled trial. Am J Prev Med 2010 Nov;39(5):472-478 [FREE Full text] [doi: 10.1016/j.amepre.2010.07.016] [Medline: 20965387]
17. Robinson T, Janssen A, Kirk J, DeFazio A, Goodwin A, Tucker K, et al. New approaches to continuing medical education: a QStream (spaced education) program for research translation in ovarian cancer. J Cancer Educ 2017 Sep;32(3):476-482 [FREE Full text] [doi: 10.1007/s13187-015-0944-7] [Medline: 26574041]
18. MOCA Minute® 10-year cycle. The American Board of Anesthesiology. URL: https://www.theaba.org/maintain-certification/moca-minute/ [accessed 2024-08-14]
19. Moran J, Briscoe G, Peglow S. Current technology in advancing medical education: perspectives for learning and providing care. Acad Psychiatry 2018 Dec;42(6):796-799. [doi: 10.1007/s40596-018-0946-y] [Medline: 29949053]
20. Tarchichi TR, Szymusiak J. Continuing medical education in the time of social distancing: the case for expanding podcast usage for continuing education. J Contin Educ Health Prof 2021;41(1):70-74. [doi: 10.1097/CEH.0000000000000324]
21. Rotstein A, Charow R, Papadakos T, Wiljer D, Slinger P. CPD By the Minute: an innovative mobile application for continuing professional development in medicine. Can J Anaesth 2020 Dec;67(12):1881-1882. [doi: 10.1007/s12630-020-01788-0] [Medline: 32779005]
22. Gaglio B, Shoup JA, Glasgow RE. The RE-AIM framework: a systematic review of use over time. Am J Public Health 2013 Jun;103(6):e38-e46. [doi: 10.2105/AJPH.2013.301299] [Medline: 23597377]
23. Nielsen J. Enhancing the explanatory power of usability heuristics. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 1994 Presented at: CHI '94; April 24-28, 1994; Boston, MA.
24. Lewis JR. The system usability scale: past, present, and future. Int J Hum Comput Interact 2018 Mar 30;34(7):577-590. [doi: 10.1080/10447318.2018.1455307]
25. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
26. Moore DEJ, Green JS, Gallis HA. Achieving desired results and improved outcomes: integrating planning and assessment throughout learning activities. J Contin Educ Health Prof 2009;29(1):1-15. [doi: 10.1002/chp.20001] [Medline: 19288562]
27. Anesthesiology profile. Canadian Medical Association. 2019 Dec. URL: https://www.cma.ca/sites/default/files/2020-10/anesthesiology-e.pdf [accessed 2025-06-04]
28. Phillips JL, Heneka N, Bhattarai P, Fraser C, Shaw T. Effectiveness of the spaced education pedagogy for clinicians' continuing professional development: a systematic review. Med Educ 2019 Sep;53(9):886-902. [doi: 10.1111/medu.13895] [Medline: 31144348]
29. Tulenko K, Bailey R. Evaluation of spaced education as a learning methodology for in-service training of health workers in Ethiopia. Knowl Manag E Learn Int J 2013 Sep;5(3):223-233. [doi: 10.34105/j.kmel.2013.05.016]
30. Boet S, Sharma S, Goldman J, Reeves S. Review article: medical education research: an overview of methods. Can J Anaesth 2012 Feb;59(2):159-170. [doi: 10.1007/s12630-011-9635-y] [Medline: 22215522]

31. Nevin CR, Westfall AO, Rodriguez JM, Dempsey DM, Cherrington A, Roy B, et al. Gamification as a tool for enhancing graduate medical education. Postgrad Med J 2014 Dec;90(1070):685-693 [FREE Full text] [doi: 10.1136/postgradmedj-2013-132486] [Medline: 25352673]
32. Robinson T, Hills D, Kelly B. The evaluation of an online orientation to rural mental health practice in Australia. J Psychiatr Ment Health Nurs 2011 Sep;18(7):629-636. [doi: 10.1111/j.1365-2850.2011.01712.x] [Medline: 21848598]
33. De Gagne JC, Park HK, Hall K, Woodward A, Yamane S, Kim SS. Microlearning in health professions education: scoping review. JMIR Med Educ 2019 Jul 23;5(2):e13997 [FREE Full text] [doi: 10.2196/13997] [Medline: 31339105]
34. Buchem I, Hamelmann H. Microlearning: a strategy for ongoing professional development. eLearning Papers. 2010 Sep. URL: https://tinyurl.com/37nubbwk [accessed 2025-06-04]
35. Hudson L, Amponsah C, Bampoe JO, Marshall J, Owusu NA, Hussein K, et al. Co-designing digital tools to enhance speech and language therapy training in Ghana. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020 Presented at: CHI '20; April 25-30, 2020; Honolulu, HI. [doi: 10.1145/3313831.3376474]
36. Lee YM. Mobile microlearning: a systematic literature review and its implications. Interact Learn Environ 2021 Sep 26;31(7):4636-4651. [doi: 10.1080/10494820.2021.1977964]
37. Oneill K, Robb M, Kennedy R, Bhattacharya A, Dominici NR, Murphy A. Mobile technology, just-in-time learning and gamification: innovative strategies for a CAUTI education program. Online J Nurs Inform 2018 Jun;22(2) [FREE Full text]
38. Sim J, Radloff A. Enhancing reflective practice through online learning: impact on clinical practice. Biomed Imaging Interv J 2008 Jan;4(1):e8 [FREE Full text] [doi: 10.2349/biij.4.1.e8] [Medline: 21614319]
39. Wallace S, May SA. Assessing and enhancing quality through outcomes-based continuing professional development (CPD): a review of current practice. Vet Rec 2016 Nov 19;179(20):515-520 [FREE Full text] [doi: 10.1136/vr.103862] [Medline: 27856985]
40. Dahiya S, Bernard A. Microlearning: the future of CPD/CME. J Eur CME 2021 Dec 14;10(1):2014048 [FREE Full text] [doi: 10.1080/21614083.2021.2014048] [Medline: 34925962]
41. Green LA, Seifert CM. Translation of research into practice: why we can't "just do it". J Am Board Fam Pract 2005;18(6):541-545 [FREE Full text] [doi: 10.3122/jabfm.18.6.541] [Medline: 16322416]
42. Newton WP, O'Neill TR, Price DW. The evolution of knowledge assessment: ABFM's strategy going forward. Ann Fam Med 2021;19(4):377-379 [FREE Full text] [doi: 10.1370/afm.2726] [Medline: 34264843]
43. Wensing M, Grol R. Knowledge translation in health: how implementation science could contribute more. BMC Med 2019 May 07;17(1):88 [FREE Full text] [doi: 10.1186/s12916-019-1322-9] [Medline: 31064388]
44. Price DW, Wang T, O'Neill TR, Bazemore A, Newton WP. Differences in physician performance and self-rated confidence on high- and low-stakes knowledge assessments in board certification. J Contin Educ Health Prof 2024;44(1):2-10. [doi: 10.1097/CEH.0000000000000487] [Medline: 36877811]
45. Newton WP, Baxley E, O'Neill T, Rode K, Fain R, Stelter K. Family medicine certification longitudinal assessment becomes permanent. J Am Board Fam Med 2021 Jul;34(4):879-881. [doi: 10.3122/jabfm.2021.04.210242]

Abbreviations

ANES: American Board of Anesthesiology
CME: continuing medical education
CPD: continuing professional development
CPD-Min: CPD By the Minute
MCQ: multiple-choice question
MOC: Maintenance of Certification
RE-AIM: Reach, Effectiveness, Adoption, Implementation, and Maintenance
REDCap: Research Electronic Data Capture
RCPSC: Royal College of Physicians and Surgeons
SUS: System Usability Scale
TEL: test-enhanced learning
UofT: University of Toronto
UHN: University Health Network

Edited by J Gentges; submitted 29.11.24; peer-reviewed by J Petersen, T Gladman; comments to author 14.01.25; revised version received 14.03.25; accepted 14.03.25; published 23.07.25.

Please cite as:
Slinger P, Omar M, Younus S, Charow R, Baxter M, Campbell C, Giuliani M, Goldmacher J, Jeyakumar T, Karsan I, Papadakos J, Papadakos T, Rotstein AJ, Yee MS, Siddiqui A, Restrepo MS, Zhang M, Wiljer D
Innovative Mobile App (CPD By the Minute) for Continuing Professional Development in Medicine: Multimethods Study
JMIR Med Educ 2025;11:e69443
URL: https://mededu.jmir.org/2025/1/e69443
doi:10.2196/69443
PMID:40699896

Exploring Social Media Use Among Medical Students Applying for Residency Training: Cross-Sectional Survey Study

Simi Jandu*, MD; Jennifer L Carey*, MD

Department of Emergency Medicine, University of Massachusetts T.H. Chan School of Medicine, Worcester, MA, United States

*all authors contributed equally

Corresponding Author:
Simi Jandu, MD
Department of Emergency Medicine, University of Massachusetts T.H. Chan School of Medicine, Worcester, MA, United States

Abstract

Background: Since the COVID-19 pandemic, residency candidates have moved from attending traditional in-person interviews to virtual interviews with residency training programs. This transition spurred increased social media engagement by residency candidates, in an effort to learn about prospective programs, and by residency programs, to improve recruitment efforts. There is a paucity of literature on the effectiveness of social media outreach and its impact on candidates' perceptions of residency programs.

Objective: We aimed to determine patterns of social media platform usage among prospective residency candidates and social media's influence on students' perceptions of residency programs.

Methods: A cross-sectional survey was administered anonymously to fourth-year medical students who successfully matched to a residency training program at a single institution in 2023. These data were analyzed using descriptive statistics, as well as thematic analysis for open-ended questions.

Results: Of the 148 eligible participants, 69 (46.6%) responded to the survey, of whom 45 (65.2%) used social media. Widely used social media platforms were Instagram (19/40, 47.5%) and Reddit (18/40, 45%). Social media influenced 47.6% (20/42) of respondents' opinions of programs and had a moderate or major effect on 26.2% (11/42) of respondents' decisions on program ranking. Resident-faculty relations and social events showcasing camaraderie and wellness were the most desired content.

Conclusions: Social media is used by the majority of residency candidates during the residency application process and influences residency program ranking. This highlights the importance of residency programs in leveraging social media usage to recruit applicants and provide information that allows the candidate to better understand the program.

(JMIR Med Educ 2025;11:e59417) doi:10.2196/59417

KEYWORDS

social media; residency recruitment; Instagram; Reddit; medical students; student; residency; residency training; social media engagement; training programs; social media usage; cross-sectional survey; survey; residency training program; thematic analysis

Introduction

Since the 2020‐2021 residency application cycle, the Association of American Medical Colleges and Liaison Committee on Medical Education have recommended that programs conduct virtual interviews exclusively for residency applicants [1,2]. This recommendation allows for a more equitable residency application process, as it offloads financial and time burdens from the applicant involved with traveling, the applicant pool, bias, and interview flexibility; however, having an exclusively virtual process also comes with a loss of applicants developing rapport with faculty and residents and appreciating resident camaraderie, program culture, and what resident daily life is like [1,3].

Students and residency programs have both turned toward social media to lessen this void. Residency applicants turn to social media to gather more information about residency programs, such as details about the work environment and facilities, and residency programs use social media to promote their programs and institutions and highlight their culture, personnel, and network [4-8]. This has been shown to increase the number of programs applicants can apply to [9]. Residency programs must embrace this digital shift to adapt to the postpandemic landscape and efforts to enhance diversity and equity in medical education. Thus, social media remains an important platform for residency applicants and programs alike.

Despite its widespread usage, there is a lack of information on the impact that applicants across different specialties derive from residency programs' social media accounts. There have been single-specialty studies that have shown that social media is used by prospective applicants during the residency recruitment process, but limited studies across specialties have

been performed [4,10-12]. In this study, we performed a survey across multiple specialties to elucidate the patterns of social media consumption and its influences on medical students' selection of a residency program.

Methods

Population and Setting

The study population consisted of fourth-year medical students who graduated in 2023 from an allopathic medical school in Massachusetts and who participated in the residency match program during the 2022 - 2023 cycle.

Ethical Considerations

The survey was approved by the institutional review board and deemed not human research (STUDY00001121). The survey contained a description containing the risks of participation in the study, and completion of the survey implied voluntary, informed consent. No personally identifiable information was collected, and no incentives were offered.

Survey Development and Distribution

We developed the survey based on guidelines by Artino et al [13]. After a literature review, a focus group was held to learn more about medical students' use and opinions of social media. This information was synthesized, and survey items were created using a combination of a Likert scale, yes or no, and open-ended questions; the survey explored demographic data, the use of social media and types of platforms, preferred social media content, and the impact of social media on residency programs. The survey was reviewed by faculty and fellows and assessed for acceptability, feasibility, and content validity of survey questions. We performed cognitive interviews for the questions and then piloted and revised the survey for clarity based on user feedback from medical students. The survey had a total of 5 pages with 6 or less questions per page, and answers could be changed. The survey was then distributed to all students at our institution who graduated in May 2023.

All fourth-year medical students who graduated in 2023 at UMass Chan Medical School were eligible for the survey and emailed a link to the anonymous electronic survey (Multimedia Appendix 1). Study data were collected and managed using the Qualtrics XM platform. The survey was distributed in May 2023 and was open for 28 days. In total, 5 reminders were sent to nonrespondents and nonfinishers at 3, 7, 14, 18, and 23 days. To avoid duplicates, each participant was sent an individual link via Qualtrics.

Outcomes Measured

The outcomes measured included demographic data; social media platform use (platforms that were used daily, platforms used for residency programs, and the influence of social media on stages of the residency application process); content posted on social media platforms (student content that was trusted, not trusted, desired, deterrents, and then helpful); nonsocial media resources used for learning about residency programs; and reasons why participants did not use social media.

Data Analysis

We performed simple descriptive statistics for survey questions. Nominal variables were reported as percentages and frequencies. Ordinal variables were presented as percentages. Data analysis was conducted using Prism GraphPad (version 9.5.1).

We performed a thematic analysis using an inductive constructivist approach on deidentified responses to open-ended questions of fully completed questionnaires [14,15]. Coders (SJ and JLC) independently reviewed responses via open coding, systematically generating a preliminary list of codes for each question. Using methods outlined by Nowell et al [15], these were merged into concepts, and themes were generated via constant comparison, returning to raw data, and iterative modification to develop a consensus on themes.

Results

There were 69 respondents out of 148 eligible students in our study. Of these, 5 were excluded from the analysis because of incomplete survey responses, with a completion rate of 92.8%. The median age of survey respondents was 27 years old (range 24 - 39; IQR 27 - 29 years). In this cohort, 42.9% identified as a man and 57.8% identified as a woman. Further, 100% of respondents matched in the 2022 - 2023 cycle, and 73.4% matched in the Northeast Region. The most popular specialties were internal medicine (25%), pediatrics (15.6%), and emergency medicine (9.4%) (Table 1).

Table . Demographic characteristics.

Characteristics	Participants, n (%)
Gender	
Women	37 (57.8)
Men	27 (42.9)
Transgender or nonbinary	0 (0)
Race or ethnicity	
American Indian or Alaska Native	0 (0)
Asian	13 (20.3)
Black or African American	0 (0)
Native Hawaiian or other Pacific Islander	0 (0)
Hispanic White	2 (3.1)
Non-Hispanic White	44 (68.8)
Other-Hispanic	2 (3.1)
Multiracial-Asian or White	3 (4.7)
Specialty	
Anesthesiology	3 (4.7)
Emergency medicine	6 (9.4)
Family medicine	4 (6.3)
Internal medicine	15 (25.0)
Internal medicine—pediatrics	1 (1.6)
Neurological surgery	1 (1.6)
Neurology	1 (1.6)
Obstetrics-gynecology	5 (7.8)
Ophthalmology	2 (3.1)
Orthopedic surgery	2 (3.1)
Otolaryngology	1 (1.6)
Pathology	1 (1.6)
Pediatrics	10 (15.6)
Psychiatry	4 (6.3)
Radiation oncology	1 (1.6)
Radiation—diagnostic	3 (4.7)
Surgery—general or preliminary	3 (4.7)
Region matched	
Northeast	47 (73.4)
Southeast	1 (1.6)
West	3 (4.7)
Southwest	7 (10.9)
Midwest	6 (9.4)
Social media platforms daily use	
Discord	3 (4.7)
Facebook	20 (31.8)
Instagram	46 (73.0)
Reddit	12 (19.1)

Characteristics	Participants, n (%)
Snapchat	18 (28.6)
Twitter (X)	10 (15.6)
TikTok	14 (22.2)

The primary social media platform used was Instagram, with 73% (46/63) reporting daily use, followed by Facebook (20/63, 31.8%) and Snapchat (18/63, 28.6%) (Table 1). Among the respondents, 65.2% (45/69) reported using social media to learn about prospective residency programs. The most frequently used platforms for this purpose were Instagram, Reddit, and YouTube (Figure 1). Facebook, Snapchat, and TikTok were rarely or never used to learn about residency programs.

Social media had a moderate or major effect on 47.6% (20/42) of respondents' opinions about programs, while it had a lesser effect on respondents' decision to apply (6/42, 14.3%) or interview (5/42, 11.9%) at a program. However, 26.2% (11/42) of respondents indicated that social media had a moderate or major effect on their decision to rank a program (Figure 2).

Figure 1. Frequency of social media platform usage when learning about residency programs.

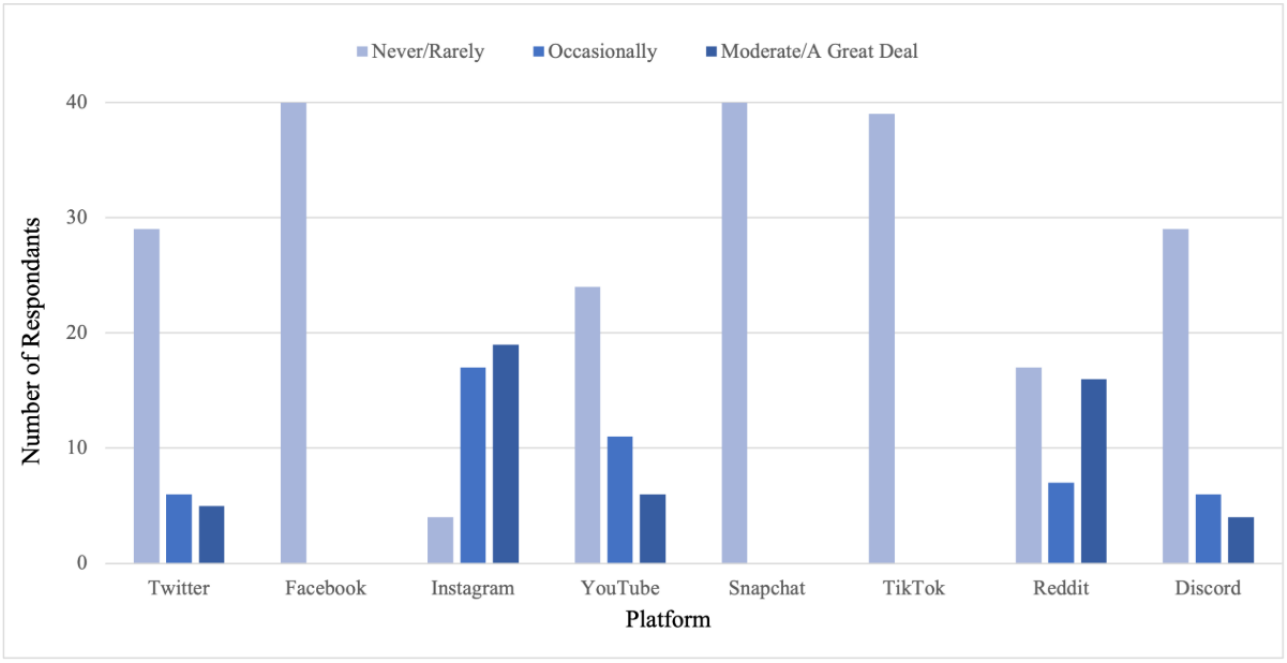
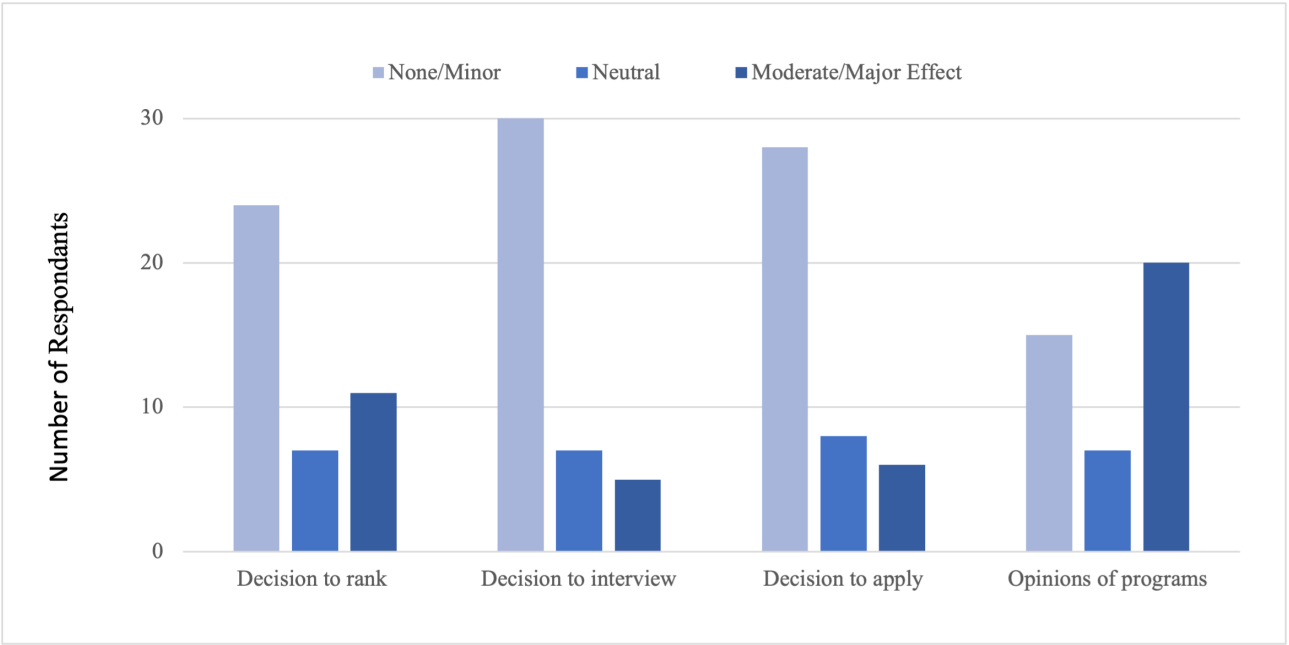


Figure 2. Social media influence on students' residency application process.



Among the 34.4% (22/64) of respondents who did not use social media, common reasons for abstaining included a perception that information could be easily distorted, social media can lead to mistrust and misrepresentation of the programs, lack of personal social media accounts, and limited usefulness or applicability of information resulting from the variable quality of social media posts. The sources trusted were the official program website, Fellowship and Residency Electronic Interactive Database Access (FREIDA), word of mouth, and personal contacts (Table 2).

Table . Resources used and reasons that students did not use social media.

	Themes
Resources used	• Program website • FREIDA[a] • Word of mouth • Current or past residents • Institutional contacts
Reasons that students did not use social media	• Mistrust of social media • Programs could be misrepresented • Irrelevant and not applicable postings that were unhelpful • Posts were not authentic • Respondents did not have or use social media personally and professionally • Variability in posts

[a]FREIDA: Fellowship and Residency Electronic Interactive Database Access.

The content deemed most valuable was resident and faculty relations (89.5%), followed by social events (86.8%) and education (60.5%) (Table 3). Thematic analysis revealed that students were attracted to posts emphasizing camaraderie, diversity, resident wellness, a genuine representation of the program's personality, and program and curriculum information. Participants particularly valued "Day in the Life" posts that visually depicted people, facilities, and location, as well as content that focused on personal experiences or resident wellness, highlighted unique attributes of the program, and provided information on the application process (Table 4).

Table . Desired social media content when researching residency programs.

Content	Participants, n (%)
Social events	33 (86.8)
Research	13 (34.2)
Didactics	12 (31.6)
Education	23 (60.5)
Resident and faculty relations	34 (89.5)
Other	2 (5.3)

Table . Themes and quotes of content that students deemed helpful, attractive, deterrents, trusted, and not trusted.

Themes		Quotes
Content that attracted students to a program		
	Camaraderie among residents and faculty; diversity of the program	"Seeing the residents spending time together and enjoying it, even if they were posting them working on the floor together"
	Resident wellness	"Camaraderie evident on posts"
	Content invoking a genuine feeling and showcasing its personality; informational content about the program and curricula	"Multiple social events with residents of all classes, photos with attendings and residents together"
Content that students deemed helpful in learning more about programs		
	Visualizing people, facility, and location; showcasing program's unique features; informational posts on the program; "Day in the Life"; resident wellness including what they do in their time off; personal experiences about a program; focus on advocacy	"When it was active, showed personality of programs/residents"
Content that deterred students		
	Minimal to no social media posts; lack of representation of multiple resident and faculty; negative personal anecdotes	"Less activity or presence on social media left an impression of overworked residents who didn't have time to post, or programs with less wellness/bonding activities to show off"
	Perception of ingenuine postings	"Too much of one person"
	Lack of types of content: photos and resident highlights	"No photos of residents and attendings"
Content trusted by students		
	Personal endorsements; resident-driven content; perspectives from residents and applicants; anonymous online platforms	"Reddit, student doctor network, because people post anonymously and be honest about the negative aspects of their program"
Content not trusted by students		
	Curated social media posts	"Social media posts are curated, and I don't trust that it's a true reflection of the day-to-day or vibe of the program"
	Program websites; promotional videos	"The promo videos for each program on the website because you can highlight very small parts."
	Nonanonymous content (subject to bias)	"Statements about the quality of the program, if they are happy, etc. I would not trust because the residents know the recordings will be posted. Hard to be honest when you can't be anonymous."
	Reddit, Discord, and chat/discussion boards (subject to bias)	"Chat/discussion boards due to potential for bias"
	Content from individuals outside the program; content from nonprogram accounts	"Message boards like Reddit I consider less reliable as anyone could share their experience with a good versus bad interview, I find these sites are very polarizing good or bad."

In a thematic analysis among all participants, anonymous digital platforms such as Reddit and Discord were considered trustworthy by some (n=3), although others perceived them as subject to bias and could be polarizing (n=15). Respondents also reported that nonanonymous content could also be subject to bias, noting that individuals posting may not want to be honest about negative aspects of a program when posting anonymously. Content from individuals outside the program and from nonprogram accounts was generally not trusted, as content could be posted and "filled with trolls," individuals who post intentionally provocative or inflammatory content. Finally, curated social media posts, program websites, and promotional videos were also listed among content that was not trusted (n=8), as it may only highlight certain aspects of the program (Table 4).

Discussion

Principal Findings

Our study revealed that Instagram was the most commonly used social media platform within our cohort. Instagram has experienced the greatest growth among new residency-specific social media accounts since March 2020, and its predominant demographic characteristics are similar to those of most prospective residents [12,16-18]. It has also been cited to be the most used platform, compared to Facebook and Twitter [19].

Interestingly, although the majority of students in this cohort reported using Facebook daily or weekly, it was almost never used to learn about residency programs by our respondents. Despite Facebook being commonly cited and compared to other platforms, it was only used more than Twitter by family medicine applicants. Otolaryngology, anesthesia, and plastic surgery applicants all used Facebook less than Twitter, with all specialties citing Instagram as the most used platform [4,7,19-21]. Facebook as a platform was shown to have the least growth, the least total number of accounts across specialties, and the least utilization among most specialties in comparison to Instagram and Twitter [17,19,20]. It is unclear why students did not use Facebook to learn about residency programs, despite their overall frequent use. One possible reason is the lack of Facebook posts by residency programs. To maximize the effectiveness of their social media presence, programs might consider focusing on Instagram rather than Facebook or linking the 2 platforms, thereby reaching 2 platforms with 1 post.

We found Reddit to be the second most popular platform. Reddit has been used by anesthesia and emergency medicine applicants as sources of information but was not this highly ranked by prior studies [20,22]. Its design facilitates ease of information exchange, and its built-in anonymity affords users the opportunity to post content without fear of repercussions. Students acknowledged that while anonymity introduces the potential for bias, anonymous online chat and discussion boards still have the potential to be trustworthy sources of information. Additionally, it is worthwhile for programs to note that negative anecdotes published on Reddit or similar platforms can deter students from programs and can be seen by those without social media accounts. The increasing popularity of Reddit suggests that it is a worthwhile avenue for social media outreach during the residency application season [22].

As social media influence the residency process, respondents are affected by their opinion and rank of a program. Social media can positively influence the opinions of programs, congruent with prior urology, otolaryngology, and plastic surgery studies, with a quarter of students' decisions affected by social media when creating their "rank list" [7,8,23]. However, as compared to Naaseh et al [9], who found 74% of respondents increased the number of programs they applied to due to social media, we found no significant effect when applying to programs found in our study despite the positive overall impression of the program in our study along with anesthesia, general surgery, and family medicine [4,9,11,20]. Based on studies, social media can influence opinion and rank of a program, which may ultimately change where a student matches for residency and whether a residency program is able to fill all its residency positions.

Regardless of surgical or nonsurgical specialty, posts that showcase resident and faculty relations, social events, and educational material are seen as the most desired content. As seen in our study, applicants desire a sense of camaraderie and resident wellness where "the residents are spending time together and enjoying it even if they were posting them working on the floor together" [19,21]. Consistent with prior studies, applicants are interested in the resident life in and outside of the hospital [19]. "Day in the Life" posts, where residents showcase a typical working day, can help students understand what their day-to-day life will be like at a particular program. They can also help to showcase aspects of the program that are difficult to show within the virtual interview setting, such as personnel interactions, diversity, and wellness [11,21]. Finally, they can be an adjunct to highlight specific program information, including curricula, electives, rotations, research, conferences, and even interview-specific information. These are all aspects sought by students in their evaluation of a program's social media presence and can be leveraged in the recruitment of residency candidates.

Importantly, social media can also have a negative influence on prospective applicants. Our study shows that social media accounts that do not consistently post or save content can leave the "impression of overworked residents who did not have time to post, or programs with less wellness or bonding activities to show off," consistent with prior investigations [24]. Negative anecdotes and comments left on anonymous platforms by single individuals, although possibly isolated, nonrepresentative experiences, can have a profound negative influence on an applicant's perception of a program, and it can be exceptionally difficult to correct these views. Programs must keep in mind that the amount and content a program posts and anonymous negative anecdotes can contribute to a negative opinion of a program, potentially affecting the application and rank process.

In our cohort, approximately one-third of the applicants did not use social media and reported using other resources. Thus, it is imperative to ensure that the official program website and Google are updated and accurate. Traditional resources include FREIDA, Doximity, word of mouth, current and past residents, and contacts within the institution [4,11,12]. These are all highly trusted content used by both social media users and nonusers.

With Instagram being the most popular social media platform, residency programs would likely benefit the most from using Instagram as their main social media platform [9,12,19,20]. Although an exact threshold is unknown, students associate less frequent Instagram posts with decreased resident wellness, and frequently posting information about residents, faculty, and program information is important for a program's image. High-impact posts might feature a particular resident for a "Day in the Life," social activities both inside and outside of the work environment, and highlights from resident wellness days and resident-faculty interactions. These can be categorized and saved, enabling prospective applicants to easily view them at later dates. Efforts should focus on creating authentic posts that showcase the people, diversity, and culture of the program in a

fun manner while taking care to avoid professional, ethical, and legal violations [25]. Students want a glimpse of what it is like to be a resident at a particular program, and posts containing pictures and videos can enable them to see and understand the program better in the current landscape of recruitment.

Limitations

As this is a survey-based study, the survey is subject to selection and response bias, with potential inaccuracies in the participants' recollection of their social media usage and influence in the residency application process. However, this is the first survey across specialties to delve into the social media usage throughout their residency application process; it has to be done after Match day. We attempted to limit response bias by ensuring anonymity and distributing this survey between match day and prior to graduation to all students in this class. We had no responses that indicated they did not match and did not ask whether any applicants went through the SOAP (Supplemental Offer and Acceptance Program) process. Thus, it is possible that those likely to respond may have been those who used social media throughout the application process and those who did not have to go through the SOAP process. Further studies could look at social media use in those that matched during the NRMP (National Resident Matching Program) or the SOAP process to see if there was a difference. Questions could also directly ask about positive and negative influence to gain more information on the drawbacks of social media while being a neutral question stem.

Although the survey was developed based on the guidelines of Artino et al [13], it needs to be further validated in the future. This was a single-center study from an allopathic medical school, limiting the generalizability of the findings, as social media patterns may vary among regions and medical schools. Thus, a multi-institutional study that examines applicants' use of social media throughout their application, interview season, and ranking process is needed to further elucidate information to be used by programs. Studies could delineate the widespread use of social media by specialties, as well as whether the applicants matched into their specialty of choice or not. Specific content that students are interested in could also be looked at for specialty (ie, procedural-based vs non–procedural-based specialties or adult vs pediatric specialties).

Conclusions

This study offers important insights into the effects of social media on residency recruitment from the student perspective. Students use social media platforms, specifically Instagram, to make informed decisions in their residency application process; therefore, programs can use these platforms to augment their recruitment. This information can help programs develop their social media platforms to cater to their target audience and mitigate the potential negative influence of social media. With the increasing popularity of social media among this generation of applicants, its use in the residency match process is expected to increase, with the current leading social media platforms being Instagram and Reddit.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Social media survey.
[DOCX File, 31 KB - mededu_v11i1e59417_app1.docx]

References

1. Copeland L, D'Antonio J, Dell M, et al. Interviews in GME: where do we go from here? Association of American Medical Colleges. 2023. URL: https://www.aamc.org/about-us/mission-areas/medical-education/interviews-gme-where-do-we-go-here [accessed 2023-09-06]
2. 2020-2021 recruitment cycle: issues for programs considering diversity and the COVID-19 pandemic. Accreditation Council for Graduate Medical Education. 2020. URL: https://www.acgme.org/globalassets/Diversity-and-COVID-19.pdf [accessed 2023-09-06]
3. Wang S, Denham Z, Ungerman EA, et al. Do lower costs for applicants come at the expense of program perception? A cross-sectional survey study of virtual residency interviews. J Grad Med Educ 2022 Dec;14(6):666-673. [doi: 10.4300/JGME-D-22-00332.1] [Medline: 36591433]
4. Oliver MG, Kelly K. Student perceptions and use of social media as residency program information. Fam Med 2022 May;54(5):380-383. [doi: 10.22454/FamMed.2022.968351] [Medline: 35536623]
5. Vallejo MC, Price SS, Vanek TW, et al. Virtual interviewing in the COVID-19 era: a survey of graduate program directors. J Dent Educ 2022 May;86(5):535-542. [doi: 10.1002/jdd.12848] [Medline: 35580990]
6. Lee E, Terhaar S, Shakhtour L, et al. Virtual residency interviews during the COVID-19 pandemic: the applicant's perspective. South Med J 2022 Sep;115(9):698-706. [doi: 10.14423/SMJ.0000000000001442] [Medline: 36055658]
7. Patro A, Landeen KC, Stevens MN, Cass ND, Haynes DS. The digital dilemma: perspectives from otolaryngology residency applicants on social media. Ann Otol Rhinol Laryngol 2022 Sep;131(9):954-961. [doi: 10.1177/00034894211050625] [Medline: 34617461]
8. Irwin TJ, Riesel JN, Amador RO, Helliwell LA, Lin SJ, Eberlin KR. The impact of social media on plastic surgery residency applicants. Ann Plast Surg 2021 Mar 1;86(3):335-339. [doi: 10.1097/SAP.0000000000002375] [Medline: 32349083]

9. Naaseh A, Thompson S, Tohmasi S, et al. Evaluating applicant perceptions of the impact of social media on the 2020-2021 residency application cycle occurring during the COVID-19 pandemic: survey study. JMIR Med Educ 2021 Oct 5;7(4):e29486. [doi: 10.2196/29486] [Medline: 34591779]
10. Fick L, Palmisano K, Solik M. Residency program social media accounts and recruitment - a qualitative quality improvement project [version 1]. MedEdPublish 2020;9(1):203. [doi: 10.15694/mep.2020.000203.1]
11. Fuller CC, Deckey DG, Brinkman JC, et al. General surgery residency applicants' perspective on social media as a recruiting tool. J Surg Educ 2022;79(6):1334-1341. [doi: 10.1016/j.jsurg.2022.06.003] [Medline: 35739022]
12. Li TM, Tepper DL, Burger AP, Weissman MA. Internal medicine recruitment in the age of social media: strategies to target residency applicants. Mayo Clinic Proc Digit Health 2023 Jun;1(2):55-59. [doi: 10.1016/j.mcpdig.2023.02.001]
13. Artino AR, La Rochelle JS, Dezee KJ, Gehlbach H. Developing questionnaires for educational research: AMEE Guide No. 87. Med Teach 2014 Jun;36(6):463-474. [doi: 10.3109/0142159X.2014.889814] [Medline: 24661014]
14. Braun V, Clarke V. Thematic analysis. In: APA Handbook of Research Methods in Psychology, Vol 2 Research Designs: Quantitative, Qualitative, Neuropsychological, and Biological: American Psychological Association; 2012:57-71. [doi: 10.1037/13620-004]
15. Nowell LS, Norris JM, White DE, Moules NJ. Thematic analysis: striving to meet the trustworthiness criteria. Int J Qual Methods 2017;16(1). [doi: 10.1177/1609406917733847]
16. Checketts JX, Hunt T, Checketts BR, et al. Analysis of social media perceptions among orthopaedic surgery residency applicants and social media use by residency programs during the 2020 to 2021 cycle. JB JS Open Access 2021;6(4):e21.00083. [doi: 10.2106/JBJS.OA.21.00083] [Medline: 34957367]
17. Singh NP, DeAtkine AB, Hattaway RH, et al. Changes in United States residency program online presence following COVID-19. Teach Learn Med 2023;35(2):157-167. [doi: 10.1080/10401334.2022.2047050] [Medline: 35689361]
18. Dixon SJ. US Instagram users 2024, by age group. 2024 Nov 28. URL: https://www.statista.com/statistics/398166/us-instagram-user-age-distribution [accessed 2025-02-04]
19. Plack DL, Abcejo AS, Kraus MB, Renew JR, Long TR, Sharpe EE. Postgraduate-year-1 residents' perceptions of social media and virtual applicant recruitment: cross-sectional survey study. Interact J Med Res 2023;12(1):e42042. [doi: 10.2196/42042]
20. Dunn T, Patel S, Milam AJ, Brinkman J, Gorlin A, Harbell MW. Influence of social media on applicant perceptions of anesthesiology residency programs during the COVID-19 pandemic: quantitative survey. JMIR Med Educ 2023 Jun 29;9:e39831. [doi: 10.2196/39831] [Medline: 37205642]
21. Pflibsen LR, Deckey DG, Brinkman JC, Tummala SV, Casey WJ, Teven CM. The effects of website and social media presence of integrated plastic surgery residency programs on prospective applicants. Ann Plast Surg 2022;88(6):599-605. [doi: 10.1097/SAP.0000000000003064]
22. Mackey C, Feldman J, Peng C, Way DP, Messman A. How do emergency medicine applicants evaluate residency programs in the post-COVID-19 era? AEM Educ Train 2022 Dec;6(6):e10805. [doi: 10.1002/aet2.10805] [Medline: 36389651]
23. Ho P, Margolin E, Sebesta E, Small A, Badalato GM. #AUAMatch: the impact of COVID-19 on social media use in the urology residency match. Urology 2021 Aug;154:50-56. [doi: 10.1016/j.urology.2021.05.019] [Medline: 34033828]
24. Chandawarkar AA, Gould DJ, Stevens WG. Insta-grated plastic surgery residencies: the rise of social media use by trainees and responsible guidelines for use. Aesthet Surg J 2018 Sep 14;38(10):1145-1152. [doi: 10.1093/asj/sjy055] [Medline: 29474525]
25. Duque S, Riccelli V, Mulqueen S, Zhang AY. Global pandemic and plastic surgery residency match: can social media fill the void? Aesthet Surg J 2021 Oct 15;41(11):NP1747-NP1753. [doi: 10.1093/asj/sjab222] [Medline: 33970220]

Abbreviations

FREIDA: Fellowship and Residency Electronic Interactive Database Access
NRMP: National Resident Matching Program
SOAP: Supplemental Offer and Acceptance Program

Edited by B Lesselroth; submitted 11.04.24; peer-reviewed by JGZ Rodriguez, K Aguirre, MA Ream; revised version received 21.08.24; accepted 24.09.24; published 21.02.25.

Please cite as:
Jandu S, Carey JL
Exploring Social Media Use Among Medical Students Applying for Residency Training: Cross-Sectional Survey Study
JMIR Med Educ 2025;11:e59417
URL: https://mededu.jmir.org/2025/1/e59417
doi:10.2196/59417

Ethical Use of Social Media and Sharing of Patient Information by Medical Students at a University Hospital in Saudi Arabia: Cross-Sectional Survey

Sara Farsi, MD, MEd; Alaa Sabbahi, MD; Deyala Sait, MBBS; Raghad Kabli, MBBS; Ghaliah Abduljabar, MBBS

Department of Anesthesia and Critical Care, Faculty of Medicine, King AbdulAziz University, KAUH, Jeddah, Saudi Arabia

Corresponding Author:
Sara Farsi, MD, MEd
Department of Anesthesia and Critical Care, Faculty of Medicine, King AbdulAziz University, KAUH, Jeddah, Saudi Arabia

Abstract

Background: Social media (SM) has become an integral part of many medical students' lives, blurring the lines between their personal and professional identities as many aspects of their medical careers appear online. Physicians must understand how to responsibly navigate these sites.

Objective: This study aimed to identify how medical students use SM and their awareness and adherence to ethical guidelines of e-professionalism.

Methods: This is a cross-sectional study delivered as an online voluntary survey to senior medical students at King AbdulAziz University Hospital in Jeddah, Saudi Arabia. We investigated how many students used SM, their privacy settings, their possible breaches of ethical standards, and their portrayal of their training institute online.

Results: A total of 400/1546 (26%) senior medical students responded to our survey. Among the participants, 95/400 (24%) had public SM accounts, while 162/400 (41%) had both private and public accounts. As for breaches in e-professionalism, 11/400 (3%) participants posted a picture of a patient on SM without their permission, while 75/400 (20%) posted part of an excised organ or x-ray on SM without their permission, and 60/400 (16%) discussed a patient. With regards to sharing medical school information, 108/400 (29%) discussed an incident at their medical school, and 119/400 (31%) participants shared a lecture online without the presenter's permission. Approximately 66% of the participants reported that they were unaware if their institution had a professional code of conduct for SM use, and 259/371 (70%) did not receive training on the professional use of SM.

Conclusions: Medical students must be taught to recognize inappropriate online behavior, understand their role as representatives of their medical school, and know the potential repercussions of unprofessional conduct on SM. This could be accomplished by providing workshops, regular seminars on e-professionalism, and including principles of SM conduct in existing ethics courses.

(JMIR Med Educ 2025;11:e57812) doi:10.2196/57812

KEYWORDS

e-professionalism; professionalism; social media; medical education; curriculum development; privacy; confidentiality; ethics; patient confidentiality; cross-sectional; questionnaire

Introduction

Since its founding in 2006, the number of active users of Twitter (currently known as X) has increased to 237.8 million worldwide as of January 2023 [1]. Many medical students have grown up with online social media (SM) profiles. Studies conducted in Saudi Arabia have demonstrated that 75% - 87% of medical students use SM [2,3]. Owing to built-in camera-equipped smartphones, these students can now document their entire lives through pictures and videos and share them online like a public diary. Therefore, medical school is an integral part of their lives, and aspects of it are bound to find their way onto their SM profiles. However, do medical students understand the rules and implications of sharing that information online?

In the past decade, medical students have transitioned from discussing complex patient details with a few colleagues in the hospital's breakroom to doing so with hundreds of "followers" worldwide. During the COVID-19 crisis, SM played a major role by building bridges across health care communities, allowing physicians and patients to connect worldwide, exchange experiences, access the latest health recommendations, and provide and receive emotional support [4-6]. SM has even been used as an educational resource, with studies showing that most students use it to access or share learning material; 30% do not even use a textbook [7-9]. In addition, students may also share patient encounters, conflicts between staff, recordings of lectures, and other occurrences on these SM sites. These medical student posts eventually become a reflection of their profession and institution. The images they share are not always

complimentary. A cross-sectional study in the United States revealed 9 incidents of medical students posting negative information about their medical school online [10]. Furthermore, the same study revealed that 13% of those schools described a violation of patient confidentiality, and 4% of those incidents were reported by the patients or their families. Health care workers' online posts have also led to dismissals and lawsuits [4,11,12]. Moreover, several articles document unprofessional behavior by medical students online, including drinking and illicit drug use [4,10,12,13].

We hypothesize that many medical school curricula emphasize disease management and patient care, which are undeniably important. However, they have not fully evolved to address the complexities of the modern social and digital landscape, leaving students underprepared to navigate these challenges effectively. This gap can inadvertently contribute to lapses in judgment because students face situations for which they may not have been adequately equipped. Against this background, our study aimed to determine whether medical students shared unprofessional content related to patients or their medical school that could impact public perception of their institution or profession. Additionally, we sought to assess their awareness of and adherence to ethical standards of e-professionalism. A further objective was to compare our findings within the context of Saudi culture to those reported in previously published Western studies.

Methods

Study Design

This is a cross-sectional study that includes senior medical students and interns at King AbdulAziz University (KAU). Medical school in KAU lasts 6 years in addition to an internship year. We defined senior medical students as those in their 4th to 6th years of training. This group was selected because the earlier years of medical education focus primarily on lecture-based and laboratory-based basic sciences, with no direct patient exposure. We developed a 2-part, 19-item survey and included 3 demographic questions (age, gender, and year of training). The question content and design were based on our primary and secondary research goals. We developed our research questions through an extensive review of the literature, aiming to identify common challenges, breaches, and issues faced by medical students and medical schools in the context of SM use [4,10,12,14]. We identified common issues among medical students, including the sharing of confidential patient information—both textual and visual—on SM, as well as the dissemination of negative encounters experienced in their hospitals. Additionally, this study found that numerous lecturers faced consequences for remarks or actions during lectures that were unknowingly recorded by students and shared publicly [15-17]. This prompted us to investigate the frequency of teaching sessions being recorded without the lecturer's permission. Our survey questions were regarding sharing images of patients, parts of patients, colleagues, and lectures without permission. We also included questions on whether they discussed patients or incidents at their medical school online. To identify the effects students' online behavior may have on their professional image, we included questions that addressed students' profiles' privacy or anonymity (eg, Do you use your real name? Is your profile picture a clear image of yourself?), and link to their profession (eg, Do you mention the name of your institution? Do you identify your profession?). We revised the survey to ensure that the final questions were relevant, contained appropriate wording, and appeared in a logical order. A questionnaire was developed using the website Survey Monkey. The results could only be accessed by the principal researcher under a password-protected online account. We shared the survey with 10 medical students from the target group to ensure that all respondents would similarly interpret the questions as well as the usability and technical functionality of the survey platform. After piloting the survey, some questions were modified (in the question "what social media platform do you use regularly?" we added options such as Telegram, Discord, and Reddit). We also changed the wording of some questions to improve clarity. These 10 students were not included in this study's group. The final questionnaire consisted of 19 questions distributed over 4 pages (Multimedia Appendix 1). The questionnaire does not allow multiple responses for the whole duration of this study. If a student attempts to take the survey again using the same browser, they will see a message that they already took the survey. After final approval of the questionnaire and design, we invited senior medical students from years 4, 5, and 6 and the internship year to participate in the survey voluntarily through an open link. Members of the research team contacted the chief students of each academic year in person to explain the purpose and details of this study to share with all students in their year of training. Then, they sent the chief students a link to the survey via a WhatsApp (Meta Platforms, Inc) message to distribute to all students in their year individually. This message included the name and contact information of the principal researcher, the duration of the survey (3 min), and a link to the survey. The message also informed the students that their responses would be kept confidential, participation was completely voluntary, there was no incentive, and their evaluation and training would not be affected by their decision to participate in the study. We also included a QR code link on the last slide of anesthesia lectures given to the target group and invited the students to this study. The survey link was opened on August 10, 2022, and closed on June 16, 2023.

Descriptive statistics of variables were presented as counts and percentages to summarize the characteristics of the participants, including gender, age, and year of medical school. Chi-square tests assessed associations between categorical variables, and Fisher exact tests, as indicated. Univariate and multivariate logistic regression analyses were performed to identify predictors of cyberbullying exposure, with odds ratios (ORs) and 95% CI reported for each predictor. Variables included in the regression models were gender, age category, year of medical school, SM privacy status, time spent on SM, and training on the professional use of SM. Statistical analyses were performed using Stata (version 12.1 software, StataCorp LP). Cronbach α was used to measure internal consistency (0.75).

Ethical Considerations

We obtained institutional review board approval to conduct the study from KAU's Ethics Committee (reference #414-‐22). The online survey began with an informed consent statement that explained the purpose of the questionnaire and assured participants that all information would be kept confidential with no names or contact details recorded in the survey. Participation was entirely voluntary, with no reward for completing the survey and no penalty for choosing not to participate. The data were stored securely under password protection, and only the principal researcher had access.

Results

We distributed the survey to 1546 participants, of whom 400 responded, yielding a response rate of 26%. Survey completion rate was 86% and both incomplete and complete surveys were used in analysis. Approximately half of the participants were sixth-year students (194/400, 49%), and two-thirds were women (246/400, 62%), as illustrated in Table 1. Snapchat was the most used platform (287/400, 72%), followed by Twitter (275/400, 69%) and Instagram (256/400, 64%). Facebook was the least used platform (8/400, 2%), and only 8/400 (2%) of the participants reported not using any SM platform at least once a week.

Table . Characteristics of the participants (N=400).

Characteristic	Value
Gender, n (%)	
Male	154 (38.5)
Female	246 (61.5)
Age (years), n (%)	
18‐20	3 (0.8)
21‐25	378 (94.5)
26‐30	18 (4.5)
>30	1 (0.2)
Year of medical school[a], n (%)	
Fourth	16 (4)
Fifth	142 (35.5)
Sixth	194 (48.5)
Intern	48 (12)
Platform used at least once a week, n (%)	
Facebook	8 (2)
TikTok	184 (46)
Snapchat	287 (71.8)
Twitter	275 (68.8)
Instagram	256 (64)
Reddit	32 (8)
Discord	30 (8)
Telegram	247 (61.8)
Own YouTube channel	25 (6.3)
None	8 (2)

[a]Has missing value for 1 participant.

Only 95/400 (24%) of the participants had public SM accounts, whereas 162/400 (41%) had a combination of private and public accounts. Most of the participants (307/400, 77%) used their real names on SM, and one-third used their own photos for their profile image (118/400, 30%). Approximately half of the participants used SM for more than 3 hours a day (180/400, 47%), whereas only 15/400 (4%) used it for less than 1 hour a day (Table 2). Most of the participants used SM for entertainment (340/400, 85%); some used it for networking with other professionals worldwide (91/400, 29%) and for staying in touch with family and friends (300/400, 75%).

Table . Description of social media use among the participants (N=400).

Variable	Participants
Privacy status of social media account, n (%)	
Public	95 (23.8)
Private	139 (34.8)
Some public, some private	162 (40.5)
Do not use social media	4 (1)
Privacy practices in social media use, n (%)	
Use of real name on social media	307 (76.8)
Use of a clear photo of self as a profile image	118 (29.5)
Identify as a King AbdulAziz University student	76 (19)
Identify as a medical student	127 (31.8)
None of the above	69 (17.3)
Time spent on social media, n (%)	
Less than 1 h/d	15 (3.9)
1 h/d	27 (7.1)
2 h/d	71 (18.6)
3 h/d	88 (23.1)
More than 3 h/d	180 (47.2)
Reason for social media use, n (%)	
Networking with other medical students or professionals worldwide	91 (22.8)
Keeping in touch with family or friends	300 (75)
Providing medical advice and advocacy	30 (7.5)
Entertainment	340 (85)
Medical education	172 (43)

Institution-related SM use practices are presented in Table 3. Regarding the professional use of SM, only 125/371 (34%) of the participants said they were aware that their institution had a professional code of conduct for SM use. Additionally, just 112/371 (30%) recalled having received training in the professional use of SM. Approximately one-third of the participants reported checking SM while rounding on patients (138/382, 36%), discussing an incident that occurred at their institution online (108/371, 29%), or uploading the content of a lecture or workshop online without the lecturer's permission (119/382, 31%). Only 11/380 (3%) posted pictures of patients on SM after obtaining the patient's permission, while 75/381 (20%) posted pictures of parts of a patient (x-ray, excised organ, etc) on SM without obtaining their permission.

Table . Number of participants who answered yes to questions regarding institution-related social media use practices, code of conduct, and training on social media use among the participants.

Survey question	Number of respondents[a], n	"Yes" response, n (%)
Does your institution have a professional code of conduct or protocol that addresses the use of social media?	371	125 (33.7)
Did you receive any training during medical school or residency on the rules and regulations for the professional use of social media?	371	112 (30.1)
Checked your social media account while rounding on patients	382	138 (36.1)
Posted a picture of a patient on social media without their permission	380	11 (2.9)
Posted an image of part of a patient (including excised tumors or organs) or a radiographic image of a patient without a patient's permission	381	75 (19.7)
Posted an image of a work colleague or senior without their permission	382	25 (6.5)
Uploaded a video or image of a lecture or workshop online without the lecturer's permission	382	119 (31.2)
Discussed an incident that happened in your institution online	371	108 (29)
Discussed a patient you saw at your institution online	372	60 (16.1)

[a]Some of the values do not add up to the total because of missing values.

Furthermore, many participants used apps to search for medical information (Table 4). The most common apps were YouTube (314/340, 92%; Google LLC) and AMBOSS (301/340, 75%; AMBOSS GmbH), followed by Osmosis (250/340, 74%; Elsevier) and UpToDate (235/340, 70%). Wikipedia (35/340, 10%; Wikimedia Foundation, Inc) and Medline (40/340, 12%; Medline Industries, LP) were the least commonly used sources.

Table . Applications used among the participants to look up medical information (N=340).

Application	Participants, n (%)
YouTube	314 (92.4)
Medline	40 (11.8)
UpToDate	235 (69.1)
Wikipedia	35 (10.3)
AMBOSS	301 (75.3)
Osmosis	250 (73.5)
Other:	40 (11.8)

Other sources were BMJ, Board and Beyond (McGraw Hill), Dr. Najeeb (DrNajeebLectures.com), MedED (PW MedEd), Kaplan, Google, ChatGPT (OpenAI), Lecturio, OnlineMedEd, Mayo Clinic (Mayo Foundation for Medical Education and Research [MFMER]), Medscape (WebMD LLC), Medicosis Perfectionalis, Radiopaedia, Healthline (Healthline Media LLC), NCBI StatPearls (National Library of Medicine), Orthobullet (Lineage Medical, Inc), WikEM, Telegram (Telegram FZ-LLC), and Twitter (X Corp).

The associations between SM use practices and gender are presented in Table 5. Women were more likely than men to have private SM accounts (96/248, 39% and 43/154, 28%, respectively; P<.001) and were less likely than men to use a clear photo of themselves for a profile image (45/248, 18% and 73/154, 47%, respectively; P<.001).

Table . The association between cyberbullying, social media privacy status, social media privacy practices, and gender among the respondents (N=400).

Survey question	Male	Female	*P* value
Experienced cyberbullying, n (%)			.13[a]
No	118 (80.3)	194 (86.2)	
Yes	29 (19.7)	31 (13.8)	
Social media privacy status, n (%)			.001[b]
Public	53 (34.4)	42 (17.1)	
Private	43 (27.9)	96 (39)	
Some public, some private	56 (36.4)	106 (43.1)	
Do not use social media	2 (1.3)	2 (0.8)	
Privacy practices in social media use, n (%)			
Use of real name in social media	117 (76)	190 (77.2)	.77[a]
Use of a clear photo of self as a profile image	73 (47.4)	45 (18.3)	<.001[a]
Identify as a King AbdulAziz University student	30 (19.5)	46 (18.7)	.85[a]
Identify as a medical student	46 (29.9)	81 (32.9)	.52[a]
None of the above	29 (18.8)	40 (16.3)	.51[a]

[a]Chi-square test.

[b]Fisher exact test.

Of all the participants, 60/400 (16%) reported experiencing cyberbullying. In univariate analyses, participants with private SM accounts were less likely to experience cyberbullying compared to those with public accounts (OR 0.40, 95% CI 0.2-0.9). Additionally, those spending more than 3 hours per day on SM had significantly higher odds (OR 3.36, 95% CI 1.0-11.5) of experiencing cyberbullying compared to those spending 1 hour or less per day. Same findings were found in multivariate analyses but became borderline significant (all had confidence intervals that narrowly include the null value).

Table 6 presents the association between patient privacy practices among the participants and the privacy status of the SM accounts. Participants who reported posting an image of part of a patient (including excised tumors or organs) or a radiograph were more likely to have a mix of public and private accounts (39/75, 52%) than public (21/75, 28%) or private accounts (15/75, 20%; *P*<.001). Among the participants who reported posting an image of a colleague without obtaining permission, 12/25 (48%) had a public account, whereas 8/25 (32%) and 5/25 (20%) had mixed and private accounts, respectively (*P*<.001). Moreover, participants who uploaded the content of a lecture online without the lecturer's permission were more likely to have a public account (37/119, 31%) than mixed (54/119, 45%) or private (28/119, 24%) accounts.

Table . Association between the privacy status of social media accounts and patient privacy practices.

Survey question	Privacy status of participants who answered "yes" to social media accounts			*P* value
	Public, n (%)	Private, n (%)	Mixed, n (%)	
Posted a picture of a patient on social media without their permission	5 (45.5)	2 (18.2)	4 (36.4)	.24
Posted an image of part of a patient (including excised tumors or organs) or a radiographic image of a patient without a patient's permission	21 (28)	15 (20)	39 (52)	.01
Posted an image of a work colleague or senior without their permission	12 (48)	5 (20)	8 (32)	.01
Uploaded a video or image of a lecture or workshop online without the lecturer's permission	37 (31.1)	28 (23.5)	54 (45.4)	.004
Discussed an incident that happened in your institution online	29 (26.9)	36 (33.3)	43 (39.8)	.68
Discussed a patient you saw at your institution online	20 (33.3)	20 (33.3)	20 (33.3)	.14

Discussion

Our study reveals that a substantial portion of students frequently share medical school-related content online, with notable instances of ethical breaches such as discussing patients and posting images without consent. While most published studies examine unprofessional online content posted by students, we investigate how often aspects of their medical school that may affect public perception appear on their profiles. These results underscore the urgent need for enhanced e-professionalism training. Of the students who responded to our survey, 246/371 (66%) were unaware of institutional guidelines addressing the use of SM, and 259/371 (70%) felt they had not received training on the professional use of SM. However, most students in our study (389/400, 97%) refrained from posting images of a patient online despite not having received e-professional training. Probably, they recognized this as a breach of the well-known Hippocratic oath.

This study did uncover some breaches of professionalism. Of the students who participated in our survey, 60/372 (16%) discussed patients online, and 75/381 (20%) posted pictures of a patient's excised organ or radiological image. Their intention was likely to share clinical experiences and demystify rare medical conditions, possibly unaware that they may be violating privacy regulations. Even if the information is deidentified using the Health Insurance Portability and Accountability Act's "safe harbor" technique, it may not be anonymous [18]. If the clinical scenario is unique enough, the patient might be recognized or even appear in the local news [19]. Furthermore, patients or their families may find the case description or the public's online comments hurtful or offensive. In response to several incidents, the Saudi Ministry of Health developed guidelines requiring physicians to obtain the patient's consent before sharing their images or health information online or submitting it to a journal [20,21]. Any breach of these guidelines carries a hefty penalty.

When a personal profile is linked to a profession or institution, it becomes part of its public image, brand, and professional identity. In our study, in the participants' SM profiles, 127/400 (31.8%) indicated that they were medical students, and 76/400 (19%) indicated the name of their university. Among them, 91/400 (22.8%) used their accounts to network with other professionals worldwide, making them representatives of their institutions and professions. Furthermore, students used YouTube (314/340, 92%) as a clinical reference more than websites with verified peer-reviewed content, such as UpToDate (235/340, 69%) and AMBOSS (301/340, 75%). Among our participants, 162/400 (41%) had both a public and private profile (one profile may have reflected a professional identity and the other a private one). Female students in our study are more likely than male students to have private profiles (96/248, 39% and 43/154, 28%, respectively; $P<.001$) and less likely to use a clear photo of themselves for their profile image (45/248, 18% and 73/154, 47%). This gender difference could stem from the conservative culture in Saudi Arabia or the universally higher vulnerability of women to online criticism and cyberbullying [22,23]. Regardless of privacy settings, medical students must be cautious when deciding what to post on their SM profiles since the content can be leaked.

Among the students, 119 (37 with a public profile) uploaded recordings of lectures or workshops without obtaining the presenter's permission. This behavior is concerning, as comments and expressions made by educators or attendees may be taken out of context by worldwide viewers. Educators often tailor teaching material to their intended audience. They also ensure the cultural appropriateness of their expressions and

comments while observing the audience's social norms. If educators are aware that their work will be shared with a wider online audience, they may decide to change their appearance, behavior, and lecture content. They may also choose to avoid comments that may cause controversy among other groups. These fears have led many UK universities to implement lecture capture policies to manage the recording and dissemination of lecture content [24]. The policies address concerns related to intellectual property rights, emphasizing the need for the lecturer's consent before recording and sharing materials. Furthermore, in our study, 108 students (29 with a public profile) discussed online incidents that had occurred at their institutions. These incidents may have been unintentionally misrepresented by these students. Studies have proven that eyewitness accounts are not always accurate [25]. Additionally, these incidents may have been exaggerated online for comedic or dramatic purposes or unintentionally reveal confidential patient information. Unfortunately, public criticism of these online posts will be directed at the students' profession and medical school [26,27].

Our findings contribute to the growing body of literature on medical students' SM use by highlighting specific behaviors and awareness levels in the context of the Kingdom of Saudi Arabia. While many of our results align with previous studies, notable differences were also observed. For instance, similar to a French study, most of our students used YouTube for medical education [28]. Although, our numbers (314/340, 92%) are much higher than those in France (504/762, 66%). However, only 172/400 (43%) of our students use SM for education compared to 42/63 (67%) of Canadian students in 2015 [29]. Additionally, 60% - 92% of medical schools in the United States have also experienced unprofessional online behavior by medical students [10,14]. Most students in both regions reported using restrictive privacy settings, with only 20% - 37% of US students failing to do so [30,31]. However, unlike our American counterparts, our students are less likely to use a clear profile photo, with female students being less likely than male students to do so. By contrast, an American study found that 57% of medical students had a clear profile photo, with females being more likely to display one than males [31]. While in India, 80% of students used a clear profile photo [32]. This discrepancy may reflect cultural differences in SM use. In Saudi Arabia, where our study was conducted, cultural norms and societal expectations may influence their online behaviors. These findings emphasize the importance of contextualizing SM behaviors within cultural and geographical frameworks to develop targeted interventions that address both universal and region-specific challenges.

This study's findings are consistent with the results of other studies suggesting that medical school curricula should be regularly updated and adapted to the constantly changing clinical environment, which now includes the internet [4]. Developing guidelines alone would not be sufficient, as evidenced by the fact that 51% of US medical schools that reported incidents already had policies in place addressing online content [10]. Based on our findings, medical schools must integrate e-professionalism training into their curriculum. This refers to attitudes and behaviors that reflect traditional professionalism paradigms but are manifested through digital media [33]. These guidelines should not be restricted to patient privacy but must also emphasize respect and consideration for their professors, colleagues, and medical school. We recommend that medical schools (1) develop comprehensive e-professionalism guidelines, (2) implement mandatory training sessions on SM use, (3) regularly update curricula to reflect the evolving digital landscape and its impact on professional practice, (4) introduce regular audits and feedback sessions where students' SM activities are reviewed and constructive feedback is provided, and (5) develop an anonymous reporting system for unprofessional behavior, ensuring students can report concerns without fear of retribution.

The limitations of our study include the use of a voluntary questionnaire that depended on self-reporting. Additionally, the generalizability of the findings may be limited due to the single institution sample and cultural context. The potential impact of self-reporting bias must be acknowledged, as participants might underreport unprofessional behavior. Moreover, this study did not account for other possible confounding variables such as the influence of peers or external SM trends. Two of the researchers are associate professors and 3 of them are students at the institution which may have influenced their study design and interpretation of results. Furthermore, we did not examine the specific content of medical students' posts. We are, therefore, unaware if shared patient information followed Health Insurance Portability and Accountability Act guidelines and if posts positively or negatively depicted their school. Future studies should include content analysis of SM posts as that could provide deeper insights into the types of information shared and help identify specific areas for intervention. This analysis involves categorizing posts into themes such as educational content, patient confidentiality breaches, and professional interactions.

In conclusion, this study reveals significant gaps in medical students' online behavior that can affect their medical schools' image, patient care, and reputation. To foster students' understanding of these issues, e-professionalism must be included in training curricula and assessments. This curriculum should include workshops, regular seminars on e-professionalism, and integration of SM conduct into existing ethics courses. Now, more than ever, medical schools should ensure that students develop a sense of belonging and pride in their institution and care about how it is represented worldwide. Information-sharing guidelines should strive to strike a balance between clinical knowledge sharing, protecting patients' privacy, and reflecting an institution's values and public image.

Conflicts of Interest

None declared.

Multimedia Appendix 1
This is a copy of the survey.
[PDF File, 29 KB - mededu_v11i1e57812_app1.pdf]

References

1. Aslam S. Twitter by the numbers: stats, demographics & fun facts. Omnicore. 2024. URL: https://www.omnicoreagency.com/twitter-statistics/ [accessed 2025-03-07]
2. Asiri AK, Almetrek MA, Alsamghan AS, Mustafa O, Alshehri SF. Impact of Twitter and WhatsApp on sleep quality among medical students in King Khalid University, Saudi Arabia. Sleep Hypn 2018 Jan 26;20(4):247-252. [doi: 10.5350/Sleep.Hypn.2018.20.0158] [Medline: 28090183]
3. Alsuraihi AK, Almaqati AS, Abughanim SA, Jastaniah NA. Use of social media in education among medical students in Saudi Arabia. Korean J Med Educ 2016 Dec;28(4):343-354. [doi: 10.3946/kjme.2016.40] [Medline: 27907981]
4. Guckian J, Utukuri M, Asif A, et al. Social media in undergraduate medical education: a systematic review. Med Educ 2021 Nov;55(11):1227-1241. [doi: 10.1111/medu.14567] [Medline: 33988867]
5. Avcı K, Çelikden SG, Eren S, Aydenizöz D. Assessment of medical students' attitudes on social media use in medicine: a cross-sectional study. BMC Med Educ 2015 Feb 15;15(1):18. [doi: 10.1186/s12909-015-0300-y] [Medline: 25890252]
6. Rosen AO, Holmes AL, Balluerka N, et al. Is social media a new type of social support? Social media use in Spain during the COVID-19 pandemic: a mixed methods study. Int J Environ Res Public Health 2022 Mar 26;19(7):3952. [doi: 10.3390/ijerph19073952] [Medline: 35409634]
7. Judd T, Elliott K. Selection and use of online learning resources by first-year medical students: cross-sectional study. JMIR Med Educ 2017 Oct 2;3(2):e17. [doi: 10.2196/mededu.7382] [Medline: 28970187]
8. Jaffar AA. YouTube: an emerging tool in anatomy education. Anat Sci Educ 2012;5(3):158-164. [doi: 10.1002/ase.1268] [Medline: 22383096]
9. Scott K, Morris A, Marais B. Medical student use of digital learning resources. Clin Teach 2018 Feb;15(1):29-33. [doi: 10.1111/tct.12630] [Medline: 28300343]
10. Chretien KC, Greysen SR, Chretien JP, Kind T. Online posting of unprofessional content by medical students. JAMA 2009 Sep 23;302(12):1309-1315. [doi: 10.1001/jama.2009.1387] [Medline: 19773566]
11. Gibson M. Nursing students expelled for posting photo of a placenta on Facebook. TIME Magazine. 2011. URL: https://newsfeed.time.com/2011/01/04/nursing-students-expelled-for-posting-photo-of-a-placenta-on-facebook/ [accessed 2025-03-07]
12. Greysen SR, Chretien KC, Kind T, Young A, Gross CP. Physician violations of online professionalism and disciplinary actions: a national survey of state medical boards. JAMA 2012 Mar 21;307(11):1141-1142. [doi: 10.1001/jama.2012.330] [Medline: 22436951]
13. Barlow CJ, Morrison S, Stephens HON, Jenkins E, Bailey MJ, Pilcher D. Unprofessional behaviour on social media by medical students. Med J Aust 2015 Dec 14;203(11):439. [doi: 10.5694/mja15.00272] [Medline: 26654611]
14. Greysen SR, Kind T, Chretien KC. Online professionalism and the mirror of social media. J Gen Intern Med 2010 Nov;25(11):1227-1229. [doi: 10.1007/s11606-010-1447-1] [Medline: 20632121]
15. Cardiff university apology after students called "idiots. BBC News. 2021. URL: https://www.bbc.com/news/uk-wales-55633371 [accessed 2025-03-07]
16. Joseph‐Richard P, Jessop T, Okafor G, Almpanis T, Price D. Big brother or harbinger of best practice: can lecture capture actually improve teaching? British Educational Res J 2018 Jun;44(3):377-392 [FREE Full text] [doi: 10.1002/berj.3336]
17. MacKay JRD. Show and 'tool': how lecture recording transforms staff and student perspectives on lectures in higher education. Comput Educ 2019 Oct;140:103593. [doi: 10.1016/j.compedu.2019.05.019]
18. Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (HIPAA) privacy rule. US Department of Health and Human Services. 2022. URL: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html [accessed 2025-03-07]
19. Child swallows spongebob squarepants. CBS News. 2015. URL: https://www.cbsnews.com/news/child-swallows-spongebob-squarepants/ [accessed 2025-03-07]
20. Health ministry warns against negative practices in the media and social media. Ministry of Health Saudi Arabia. 2017. URL: https://www.moh.gov.sa/Ministry/MediaCenter/News/Pages/News-2017-07-09-001.aspx [accessed 2025-03-07]
21. Saudi guidelines for informed consent. Ministry of Health Saudi Arabia. 2019. URL: https://www.moh.gov.sa/en/Ministry/MediaCenter/Publications/Pages/Saudi-Guidelines-for-Informed-Consent.pdf [accessed 2025-03-07]
22. Li Q. Cyberbullying in schools: a research of gender differences. Sch Psychol Int 2006;27(2):157-170. [doi: 10.1177/0143034306064547]
23. Wang J, Iannotti RJ, Nansel TR. School bullying among adolescents in the United States: physical, verbal, relational, and cyber. J Adolesc Health 2009 Oct;45(4):368-375. [doi: 10.1016/j.jadohealth.2009.03.021] [Medline: 19766941]
24. Ibrahim Y, Howarth A, Stone I. Lecture capture policies: a survey of British universities. Postdigit Sci Educ 2021 Jan;3(1):144-161. [doi: 10.1007/s42438-020-00102-x]

25. Albright TD. Why eyewitnesses fail. Proc Natl Acad Sci U S A 2017 Jul 25;114(30):7758-7764. [doi: 10.1073/pnas.1706891114] [Medline: 28739937]
26. Reimann N. Report: 17 florida medical residents have coronavirus after throwing house party. Forbes. 2020 Jul 27. URL: https://www.forbes.com/sites/nicholasreimann/2020/07/27/report-17-florida-medical-residents-have-coronavirus-after-throwing-house-party/ [accessed 2025-03-07]
27. Frehse R, Compinoti MS. Ohio plastic surgeon who livestreamed patient operations on TikTok has state medical license revoked permanently. CNN Ohio. 2023 Jul 14. URL: https://edition.cnn.com/2023/07/13/us/ohio-doctor-tiktok-license-revoked/index.html [accessed 2025-03-07]
28. Clavier T, Chevalier E, Demailly Z, Veber B, Messaadi IA, Popoff B. Social media usage for medical education and smartphone addiction among medical students: national web-based survey. JMIR Med Educ 2024 Oct 22;10:e55149. [doi: 10.2196/55149] [Medline: 39437450]
29. El Bialy S, Jalali A. Go where the students are: a comparison of the use of social networking sites between medical students and medical educators. JMIR Med Educ 2015 Sep 8;1(2):e7. [doi: 10.2196/mededu.4908] [Medline: 27731847]
30. MacDonald J, Sohn S, Ellis P. Privacy, professionalism and Facebook: a dilemma for young doctors. Med Educ 2010 Aug;44(8):805-813. [doi: 10.1111/j.1365-2923.2010.03720.x] [Medline: 20633220]
31. Walton JM, White J, Ross S. What's on YOUR Facebook profile? Evaluation of an educational intervention to promote appropriate use of privacy settings by medical students on social networking sites. Med Educ Online 2015;20(1):26198434. [doi: 10.3402/meo.v20.28708] [Medline: 26198434]
32. Gupta S, Singh S, Dhaliwal U. Visible Facebook profiles and e-professionalism in undergraduate medical students in India. J Educ Eval Health Prof 2015;12:50. [doi: 10.3352/jeehp.2015.12.50] [Medline: 26582630]
33. Cain J, Romanelli F. E-professionalism: a new paradigm for a digital age. Curr Pharm Teach Learn 2009 Dec;1(2):66-70. [doi: 10.1016/j.cptl.2009.10.001]

Abbreviation

KAU: King AbdulAziz University
MFMER: Mayo Foundation for Medical Education and Research
OR: odds ratio
SM: social media

Edited by B Lesselroth; submitted 29.02.24; peer-reviewed by AM Pawar, JH Ye; revised version received 20.12.24; accepted 25.02.25; published 24.03.25.

Please cite as:
Farsi S, Sabbahi A, Sait D, Kabli R, Abduljabar G
Ethical Use of Social Media and Sharing of Patient Information by Medical Students at a University Hospital in Saudi Arabia: Cross-Sectional Survey
JMIR Med Educ 2025;11:e57812
URL: https://mededu.jmir.org/2025/1/e57812
doi:10.2196/57812

Original Paper

Instagram as a Tool to Improve Human Histology Learning in Medical Education: Descriptive Study

Alejandro Escamilla-Sanchez[1,2*], MD, PhD; Juan Antonio López-Villodres[1,2*], MD, PhD; Carmen Alba-Tercedor[1], PhD; María Victoria Ortega-Jiménez[1,2,3], MD, PhD; Francisca Rius-Díaz[4], MD, PhD; Raquel Sanchez-Varo[1,2,5], PhD; Diego Bermúdez[1], MD, PhD

[1]Department of Human Physiology, Human Histology, Anatomical Pathology and Physical and Sports Education, Faculty of Medicine, University of Malaga, Malaga, Spain

[2]IBIMA Bionand Platform Biomedical Research Institute, University of Malaga, Malaga, Spain

[3]Unit of Anatomical Pathology, University Hospital Virgen de la Victoria, Malaga, Spain

[4]Department of Public Health and Psychiatry, Faculty of Medicine, University of Malaga, Malaga, Spain

[5]Centre for Networked Biomedical Research in Neurodegenerative Diseases, Madrid, Spain

[*]these authors contributed equally

Corresponding Author:
Raquel Sanchez-Varo, PhD
Department of Human Physiology, Human Histology, Anatomical Pathology and Physical and Sports Education
Faculty of Medicine
University of Malaga
Bl Luis Pasteur 32, 1st floor
Malaga, 29071
Spain
Phone: 34 952131585
Email: raquelsv@uma.es

Abstract

Background: Student development is currently taking place in an environment governed by new technologies and social media. Some platforms, such as Instagram or X (previously known as “Twitter”), have been incorporated as additional tools for teaching and learning processes in higher education, especially in the framework of image-based applied disciplines, including radiology and pathology. Nevertheless, the role of social media in the teaching of core subjects such as histology has hardly been studied, and there are very few reports on this issue.

Objective: The aim of this work was to investigate the impact of implementing social media on the ability to learn human histology. For this purpose, a set of voluntary e-learning activities was shared on Instagram as a complement to traditional face-to-face teaching.

Methods: The proposal included questionnaires based on multiple-choice questions, descriptions of histological images, and schematic diagrams about the subject content. These activities were posted on an Instagram account only accessible by second-year medical students from the University of Malaga. In addition, students could share their own images taken during the laboratory practice and interact with their peers.

Results: Of the students enrolled in Human Histology 2, 85.6% (143/167) agreed to participate in the platform. Most of the students valued the initiative positively and considered it an adequate instrument to improve their final marks. Specifically, 68.5% (98/143) of the student body regarded the multiple-choice questions and image-based questions as the most useful activities. Interestingly, there were statistically significant differences between the marks on the final exam (without considering other evaluation activities) for students who participated in the activity compared with those who did not or barely participated in the activity (P<.001). There were no significant differences by degree of participation between the more active groups.

Conclusions: These results provide evidence that incorporating social media may be considered a useful, easy, and accessible tool to improve the learning of human histology in the context of medical degrees.

(JMIR Med Educ 2025;11:e55861) doi:10.2196/55861

KEYWORDS

medical education; medical students; histology; pathology; e-learning; computer-based; social media; Instagram; Meta; community-oriented; usability; utility; accessibility

Introduction

Social media platforms are web-based technologies particularly suited to facilitate the exchange of ideas through collaboration, interaction, and discussion. The accessibility and low cost of internet access, together with the high number of users of these platforms, make social media one of the easiest and most effective ways to disseminate information. In fact, 4.65 billion people, equivalent to 58.4% of the world's population, use social media [1]. In addition, most current medical students are far more knowledgeable and experienced with emerging technologies than preceding generations. Unlike traditional media (journals or television), social media emphasizes interactivity, motivation through social connections, and immediacy [2]. In this sense, the "social constructivism theory" states that interaction and socialization may help students learn and construct their knowledge and personal learning processes, supporting the use of social media for educational activities as a different tool for teaching and learning [3]. For all these reasons, social media platforms have progressively been incorporated into health care and medical education [4].

Technological advances have enabled rapid dissemination of medical updates through social networks such as Facebook, X (previously known as Twitter), or Instagram. Thus, students often have access to significant amounts of information, including content taught during traditional classes. Nevertheless, this information has not always been rigorously verified or is outdated, representing a formative disadvantage for medical students. Moreover, there is a lack of engagement and even dropout from classes because traditional education methodology is considered by the student body to be boring, unnecessary, or repetitive. Therefore, the faculty must adapt to meet their specific needs, changing traditional teaching styles and implementing new e-learning technologies [5].

Recently, the COVID-19 pandemic forced teaching staff to move further into a virtual education environment and highlighted the importance of communication between educators and learners through social media platforms. The rapid and efficient dissemination of information during the pandemic illustrated the significant influence of social media in the dissemination of medical literature and knowledge, not only among health care professionals but also among the student body [6,7]. For instance, Chan et al [8] demonstrated the benefits of using tools such as infographics posted on social media platforms (such as X and WeChat) to educate frontline health care workers about respiratory tract management and infection control in the setting of COVID-19. Thus, the pandemic prompted a paradigm shift in learning for students and medical residents by using different platforms (eg, YouTube, Zoom, Microsoft Teams) as e-learning tools under the new circumstances. Although face-to-face teaching is possible and desirable today, the use of social networks as educational instruments must continue with apps such as Instagram and aim to share image-based educational content to complement the classes.

Histology has long been an integral part of the medical curriculum [9] and continues to provide key information about biological tissues, physiology, and disease; it is therefore highly valued in clinical medicine and research. Furthermore, histopathology is a fundamental tool for diagnosis and prognosis. In addition, a thorough knowledge of histology is necessary for the surgical field and in general practice. However, histology and its nomenclature can be complex to understand for novice medical students, and consequently, it is often perceived as a secondary subject without clinical relevance [10]. From a pedagogical point of view, one of the main goals of histology courses is to ensure that students acquire the competencies necessary to understand histophysiology. For example, histology requires students to develop pattern recognition skills. Specifically, they must be able to identify what they are observing based on specific histologic features. Consequently, histology courses commonly include laboratory practices for students to train and develop these abilities. In this context, the study of histology through digital imaging might be a relevant alternative for the development of their curricula.

Instagram is a social networking service owned by Meta Platforms Inc that was launched in October 2010. Instagram allows photo and video sharing accompanied by text. The information can be shared either publicly or privately. Followers can archive shared posts, and the account's owner can track the number of people reached and give feedback to their followers. The literature shows that this social network is being used for educational purposes in medical schools, predominantly in imaging-related subjects such as radiology [11], ophthalmology [12], dermatology [13], anatomy [14], fertility [15], pathology [16], plastic surgery [17], dentistry [18], and (with very few proposals) in histology [19]. Understanding how students interact with these novel social media–based teaching environments and their approaches during e-learning processes is a matter of high relevance [20]. On the other hand, there is a lack of evidence on how the use of social networks impacts the learning and follow-up of Spanish medical students in the first years of their formation in the field of histology. In the first courses, the curricula of a Spanish medical degree include a basic thematic area with fundamental core subjects to obtain the essential knowledge for the subsequent study of pathological alterations. Among these subjects, some necessarily require the use of images, such as anatomy, cytology, histology, and microbiology. For that purpose, an educational experience was carried out using the social network Instagram to make the subject more attractive to the students of the official degree of Medicine at the University of Malaga in Spain during the 2022-2023 academic year. Our main objective was to test whether the use of Instagram might facilitate knowledge acquisition and increase engagement with histology, leading to a positive impact on students' qualifications. Additionally, we aimed to elucidate which type of visual material was more useful

for medical students. Finally, we determined students' perceptions of the integration of this tool in medical education.

Methods

Content of Histology in the Degree of Medicine at the University of Malaga

According to the syllabus for the degree in Medicine at the University of Malaga, histology is divided into 2 subjects (Human Histology 1 and 2) that are taught during the first and second years, respectively. The different didactic content is distributed sequentially, progressively increasing the theoretical difficulty (Table 1). First-year students learn general histology along with some special histology topics (eg, the immune system). The remaining systems and organs are studied during the second school year in the subject Human Histology 2. Some of the potential skills to be developed during these courses are knowledge about the architecture, morphology, and function of the different tissues or systems; recognizing the morphology and structure of tissues by microscopy and imaging techniques; and how to handle basic laboratory equipment and methodology. In addition, our curricula include the acquisition of some transversal competencies such as the capacity for analysis and synthesis, problem-solving or critical reasoning, and analysis, together with other abilities and skills (autonomous work, information management, and oral or written communication skills).

Table 1. Curricula content of the human histology subjects in the degree in Medicine at Malaga University.

Content	Issues dedicated to each topic, n
Human Histology 1	
Tissues (epithelial, muscle, osseous, connective, nervous)	17
Stem cells	1
Blood and hematopoiesis	3
Circulatory system	1
Immunity and lymphoid tissues	5
Human Histology 2	
Digestive system	6
Respiratory system	2
Urinary system	2
Genital apparatus	6
Tegumentary system	1
Endocrine system	6
Nervous system and neurosensorial organs	12

Sample Size

This study was carried out with 167 students enrolled in the subject Human Histology 2 in the degree in Medicine at the University of Malaga during the 2022-2023 academic year. The final examination was performed by most of the students (153 students), of which 143 participated until the end of the Instagram experience. Thus, only 10 (6.5%) of the 153 involved students did not follow our account.

Design of the Instagram Profile

After downloading the free app on a smartphone, a private Instagram account (username: @histologiauma) was created for the subject Human Histology 2 at the University of Malaga. The Instagram profile was linked to an institutional email address created to receive questions and comments from the student body of this subject. Students were notified of the availability of this account and were informed about the procedure to participate. For instance, they had to register by giving their real first name and last name. Once we checked they belong to the subject, the students were accepted as followers of @histologiauma.

Virtual Microscope Images

The images published in @histologiauma belong to the image bank of the Histology Unit of the Department of Human Physiology, Human Histology, Anatomical Pathology and Physical-Sports Education of the Medical School at the University of Malaga. During the COVID-19 pandemic, we introduced a highly interactive, web-based digital microscope system to view histological images during online practical lessons, either from classroom or personal computers. This virtual microscope is currently based on the digitalization of 66 slides, providing the element of real-time dynamic microscopy and offering students a truly innovative experience at exceptionally high resolution. Interestingly, this virtual microscope offers the possibility to capture specific tissue areas and use these pictures to formulate specific questions.

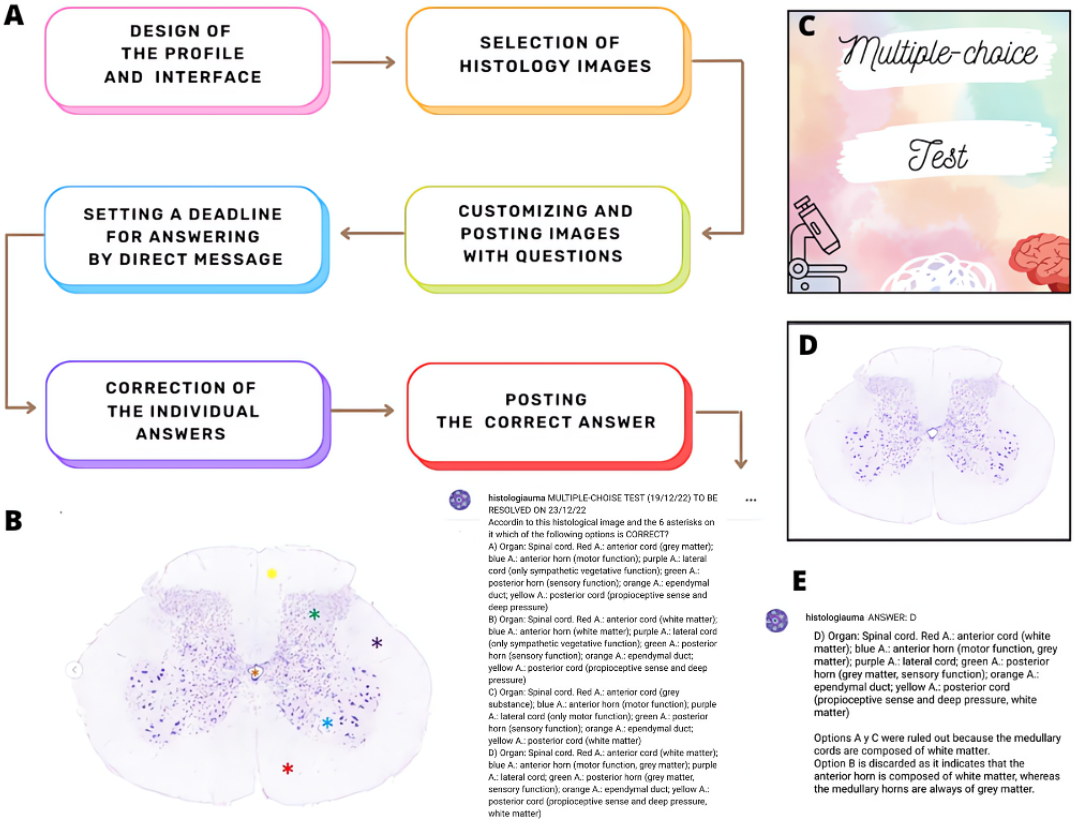
Account Content Feed

Two software applications were used to design the images: Canva and Microsoft Office PowerPoint. The free basic mode of Canva offers access to thousands of templates and 1 million free photos. Both applications allow drag-and-drop operations familiar to both average users and design professionals and

feature templates, photo filters, images, icons, and shapes useful for customizing histological images (eg, including different shapes to highlight structures or cells within a tissue, adding numbers or letters).

During a 3-month trial period, 35 posts were published, of which 5 were announcements about the account rules. The general process for uploading new content to @histologiauma account is summarized in Figure 1.

Figure 1. (A) Workflow for uploading a post in @histologiauma account, (B) screenshot of a post from @histologiauma account, (C) multiple-choice test interface, (D) image of a spinal cord transversal section with Klüver-Barrera staining, and (E) a student's direct message including a reasoned correct answer.



Ethical Considerations

The repository of digital images is composed of scanned slides with anonymized tissue remnants from Virgen de la Victoria University Hospital, whose patients provided signed informed consent for educational purposes.

The account @histologiauma was created as a private profile to be exclusively accessed by those second-year students of the degree in Medicine at the University of Malaga who voluntarily requested to participate. Images displaying captures from @histologiauma have been edited to make students' profiles unidentifiable. Moreover, all the surveys were anonymously filled out by students.

The manuscript is a retrospective case report that does not require ethics committee approval at our institution since no demographic nor clinical data from patients were used.

Type of Questions Posted on @histologiauma

The following sections were included in the Instagram platform for human histology education.

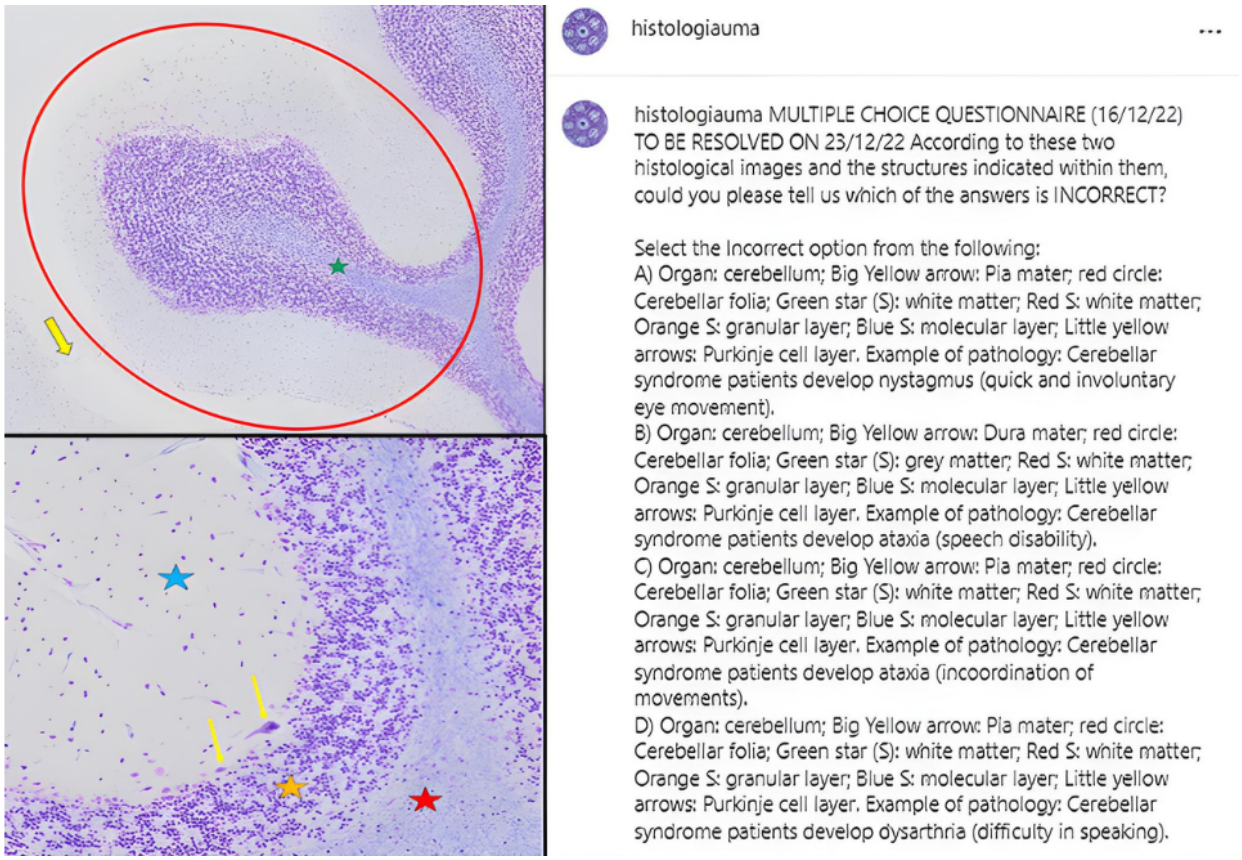
Image-Based Multiple-Choice Questions

There were 13 posts with image-based multiple-choice questions (Figure 2). Histological images of several organs studied during the academic year were posted. Different structures or cells were highlighted with arrows or other shapes (eg, stars, circles, asterisks). Multiple-choice questions with a single correct answer were posted, and 4 options were marked with labels A, B, C, and D in each question. For example, Figure 2 shows a post that asked students to select the incorrect option from the following answer choices: "A) Organ: cerebellum; Large yellow arrow: Pia mater; Red circle: Cerebellar folia; Green star: white matter; Red star: white matter; Orange star: granular layer; Blue star: molecular layer; Small yellow arrows: Purkinje cell layer. Example of pathology: Cerebellar syndrome, nystagmus as a quick and involuntary eye movement is included. B) Organ: cerebellum; Large yellow arrow: Dura mater; red circle: Cerebellar folia; Green star: gray matter; Red star: white matter; Orange star: granular layer; Blue star: molecular layer; Small yellow arrows: Purkinje cell layer. Example of pathology: Cerebellar syndrome, ataxia as a problem to speak is included. C) Organ: cerebellum; Large yellow arrow: Pia mater; red circle: Cerebellar folia; Green star: white matter; Red star: white matter; Orange star: granular layer; Blue Star: molecular layer; Small yellow arrows: Purkinje cell layer. Example of pathology: Cerebellar syndrome, ataxia or incoordination of movements is included. 4) Organ: cerebellum; Large yellow arrow: Pia mater; red circle: Cerebellar folia; Green star: white matter; Red star: white matter; Orange Star: granular layer; Blue star:

molecular layer; Small yellow arrows: Purkinje cell layer. Example of pathology: Cerebellar syndrome patients develop dysarthria (difficulty in speaking).” The correct answer is B.

No negative scores were given in the case of wrong answers. All questions had a single correct answer. Each student provided their answer, including a brief justification in the form of a private message on Instagram. They were then notified about their success or encouraged to try again in case of failure. The answers were made public 5 days later in a comment, accompanied by a summary of the most common mistakes.

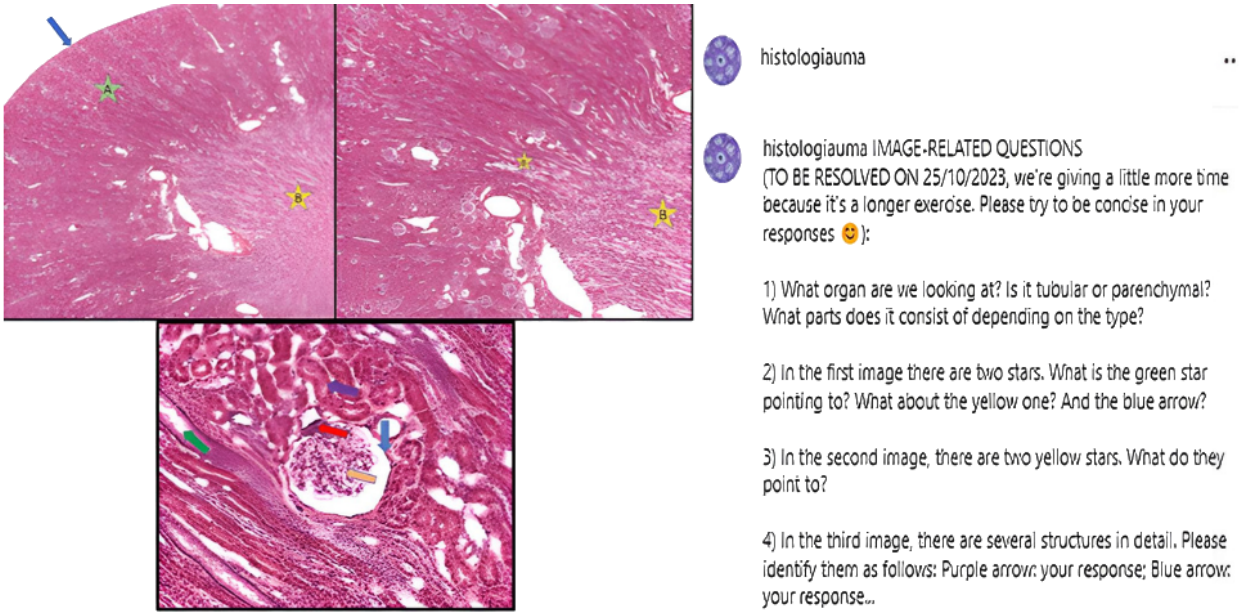
Figure 2. Screenshot of an image-based multiple-choice question about the cerebellum from the account @histologiauma.



Descriptions or Questions Associated With Histological Images

This section, consisting of 10 posts, showed histological images pointing out different structures and components to be identified by the students. For example, Figure 3 shows a post that included the following questions: “1) Identify the organ shown in the image. Is it a tubular or a parenchymatous organ? 2) What is the green star (A) pointing at? And the yellow star (B)? And the blue arrow?” Occasionally, comparisons between pathological and healthy tissues were posted, along with an introduction to clinical medicine. This section was conceived in accordance with the curricular competency entitled “From Histology to Medicine,” which aims to highlight the clinical aspects of human histology. The correct answer was published 5 days later as a comment on the post, and feedback was given to the students, as explained for the multiple-choice questions.

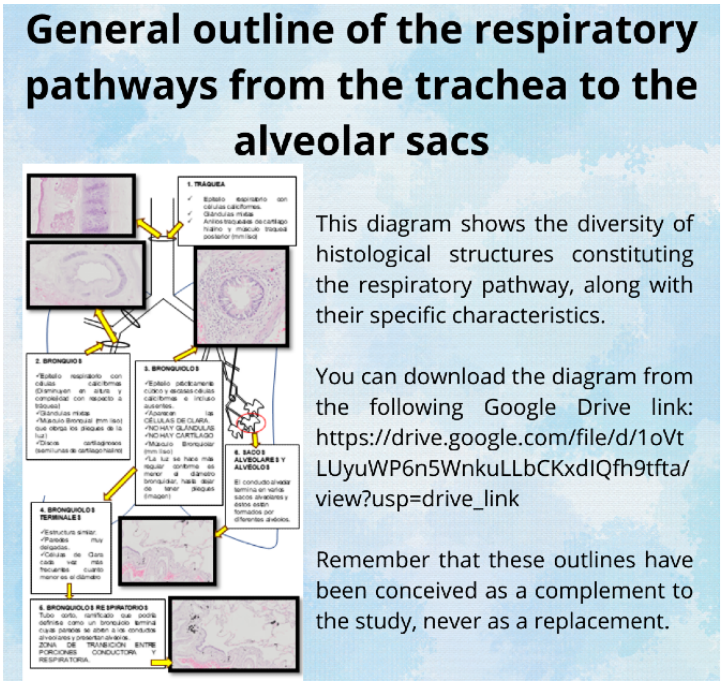
Figure 3. Screenshot of a histological image section (kidney) with the accompanying questions.



Didactic Schemes

This section included 7 posts based on student requests for visual or explanatory diagrams of the content they found most difficult. Teachers then prepared specific outlines based on these requests, avoiding the inclusion of new content. An example is shown in Figure 4. Diagrams were created using free-design and educational software, such as PowerPoint or Canva and stored in a shared Google Drive folder. The link to access the content was posted on the Instagram account and made available for 1 week. The content of these diagrams was derived from the theoretical material already provided to the students, as they were conceived as a complementary tool to the study. Students were also encouraged to make their own schemes to learn how to summarize concepts.

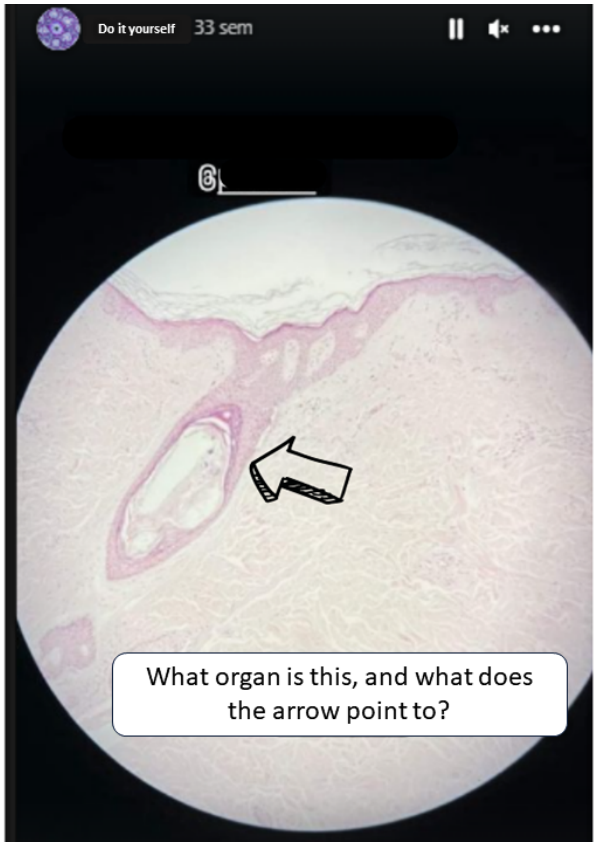
Figure 4. Screenshot of a scheme requested by students about the respiratory tract.



"Do It Yourself" Section

During the practical classes, students were encouraged to take images of histological slides through the eyepiece of the microscopes with their own smartphones. Later, they posted the images for 24 hours in the form of a story on @histologiauma accompanied by a specific question to be solved by their classmates. In total, 25 images were shared as stories, and an example is shown in Figure 5.

Figure 5. Screenshot of a "Do it yourself" section depicting a histological section of skin that was taken and elaborated on by one of the students.



Teaching Commitment

The teaching staff played a crucial role in providing feedback to the students, notifying them about their successes or mistakes through private messages and designing diagrams. Overall, 1 hour of work was required daily during the experience.

Rating

The activity was conceived as a voluntary pilot study. Students who actively participated could add a maximum of up to 0.50 points to their final mark, regardless of whether their interventions were successful. Thus, students earned points in proportion to their level of participation. A score of 10 was assigned to students who answered 100% of the questions. The student body was organized in 4 groups according to the rewards received on @histologiauma (group 1=0-4.4 points; group 2=4.5-6.5 points; group 3=6.5-8.5 points; group 4=8.5-10 points). Therefore, these points served as an indicator of participation.

Evaluation of the Activity as an Educational Innovation

To explore the influence of this innovative learning tool, data on student engagement and perceptions were collected during the lectures and at the end of the course through a final evaluation. The results were gathered throughout the academic semester with 3 anonymous surveys using a 5-point Likert scale. Each topic in the surveys covered a gradient of agreement with the statement presented (1=strongly disagree, 5=strongly agree). All questions were designed in Spanish by members of the teaching staff (AES, RSV, and DB) and later translated into English for publication.

At the beginning of this educational experience, the opinions of students about the inclusion of new technologies and the implementation of social media in our medical school were assessed through a first survey (called the pre-experience survey; Textbox 1).

This first survey also included specific questions regarding students' perceptions of the use of Instagram in the histology course (Textbox 2).

The second survey (middle experience; Textbox 3) was conducted 2 months after the start of the project. This questionnaire focused on the general operation of the account and their early perceptions of the experience. The third and final survey contained the same questions (Textbox 3) and was carried out during the last week of theory classes.

Textbox 1. General questions in the initial survey (pre-experience).

1. The university's educational systems are up to date and adapted to the times.
2. Most teachers use social networks as an educational resource.
3. Currently, the use of alternative educational tools is essential.
4. Bringing the basic subjects of a medical degree closer to the reality of professional practice is essential.
5. In general, teachers are concerned about updating educational tools.

Textbox 2. Specific questions for the initial survey (pre-experience).

1. Instagram facilitates access to educational content on histology.
2. Accessing Instagram allows me to consult the content anywhere at any time (eg, bus, train).
3. Test questions are a useful tool to review theoretical or practical content.
4. Downloading the subject outlines helps me to complete histology concepts.
5. I expect to improve my academic grades thanks to this academic experience.

Textbox 3. Specific questions for the 2nd (mid-experience) and 3rd (end of the experience) surveys.

1. I followed the @histologiauma updates daily.
2. I answered test questions.
3. I answered questions shared in the @histologiauma stories.
4. I shared some photographs in the “Do it yourself” section.
5. I answered the image-associated questions.
6. I used @histologiauma when traveling by public transport.
7. I used @histologiauma during class exchanges.
8. The test questions were adequate.
9. The schemes were useful.
10. We received highly personalized attention.
11. Circle the most useful sections of @histologiauma: “Do it yourself,” image-associated questions,” “multiple-choice questions,” “schemes.”
12. I will use @histologiuma to prepare for my final exams.

Statistical Analysis

Raw data from the survey responses were collected and analyzed using SPSS v.24 (IBM Corp). Descriptive statistics were used to characterize the data, with a frequency study carried out for each of the variables evaluated in the different surveys. Homoscedasticity (equality of variances) and normal distribution of the data were checked. Data corresponding to students’ marks (score range 0-10) are expressed as mean (SD). Mann-Whitney tests were conducted to compare overall grades from different cohorts (groups from different academic years or current cohort with the extra points versus the cohort without the extra points). The comparisons between the marks of the different groups (previously categorized into groups 1-4 according to the level of participation) and the degree of acquired knowledge evidenced by the final exam (global marks) were conducted using 1-way ANOVA followed by a post hoc Bonferroni test. Pearson correlation coefficients were calculated using the individual scores according to the students’ participation in @histologiauma and their final grades. An $r>1$ indicated a positive linear correlation between the 2 variables. The significance was set at 95% of confidence.

Results

Participant Demographics

From the very first post, 73 of 167 students enrolled in Human Histology 2 followed the account. At the end of the learning experience, there were 143 followers (143/167, 85.6%), of which 106 students had actively participated during the entire period. Analytics from these 143 students showed that 76.2% (109/143) were women and 23.8% (34/143) were men. Most (133/143, 93%) of them were 18 years to 22 years old and in their second year of the medical degree (139/143, 97.2%). There were only 4 repeaters of this subject, who were concomitantly in the third year of the degree. Full-time students represented 97.9% (140/143) of the respondents. The full demographic profile of the students is shown in Table 2.

Table 2. Student demographic data (N=143).

Characteristics	Results, n (%)
Gender	
Female	109 (76.2)
Male	34 (23.8)
Age (years)	
18-22	133 (93)
23-26	7 (4.9)
27-30	2 (1.4)
31-40	1 (0.7)
Academic year	
2	139 (97.2)
3	4 (2.8)
Enrollment	
Complete	140 (97.9)
Partial	3 (2.1)

Pre-Experience Test About Students' Perceptions of the Use of Social Media and New Technologies in Medical School Curricula

Of the participants, 83.9% (120/143) considered that the current educational system requires a significant update. Thus, 97.9% (139/143) of them strongly believed that the use of social networks should be significantly improved. Of the students, 99% (141/143) considered that using alternative educational tools is relevant, and 77.1% (110/143) of the students agreed that the use of social media such as Instagram facilitates access to the didactic content. Thereby, 95% (133/143) of the respondents found the subject more accessible thanks to @histologiauma, and 99.5% (142/143) believed they might improve their academic marks thanks to this experience.

Survey About the Students' Perceptions During the Experience

Multiple-choice questions (65/143, 45.7%) and image-based questions (33/143, 22.8%) were the students' favorite sections, with 96.5% (138/143) and 99.3% (142/143) of the students, respectively, considering them highly useful for learning the subject. In contrast, the schemes and "Do it yourself" sections were the favorite sections for 12.6% (18/143) and 0.8% (1/143), respectively, of the students. The remaining 18.1% (26/143) reported a stated preference for combinations of the different sections (multiple-choice questions + image-based questions: 17/143, 11.8%; multiple-choice questions + schemes: 5/143, 3.9%; image-based questions + schemes: 3/143, 2.4%). No other activities were included in the experience. Additionally, 96.7% (138/143) of the students felt well-supported and guided by the teaching staff throughout the experience. In addition, students showed no preference between public transport or the exchange of classes for visualizing the didactic content available on Instagram.

Impact of the Experience on Final Marks

The average grade obtained by the students from the Histology course during the academic year prior to the implementation of the experience (2021-2022: n=158 students) was 6.49 (SD 1.87) out of 10, whereas marks from the 2022-2023 cohort were significantly higher (mean 7.13, SD 1.68; $P<.002$; n=153), regardless of the extra points for participating in the experience. Once the earned points were included, the final outcome was not significantly different (2022/2023 without extra points: mean 7.29, SD 1.70; 2022/2023 with extra points: mean 7.13, SD 1.68; $P=.03$). Furthermore, the mean final grade from the 4 previous academic courses showed homogeneity in terms of having lower results (mean 6.12, SD 0.27; n=628 students) in comparison to our cohort. Overall, our data support that the use of social media produced a positive impact on students' performance, even without considering the points for participating in @histologiauma.

Interestingly, a positive linear correlation between individual participation scores and final marks (not including the extra reward points) was found ($r=0.439$, $P<.001$). Moreover, the ANOVA showed significant differences between students' marks according to their degree of participation ($P<.001$; Table 3). There was a trend of higher ratings according to the level of participation. The Bonferroni test showed that group 1 (the least engaged group with 0-4.4 points) achieved significantly lower global mean scores than the other 3 groups (all $P<.001$). Finally, there were no significant differences among groups 2 to 4 (all $P=.99$).

Table 3. Students' global marks (without the extra points) and rating obtained according to the degree of participation in @histologiauma.

Group	Global score, mean (SD)
Group 1 (0-4.4 points)	5.43 (2.65)
Group 2 (4.5-6.5 points)	7.44 (0.93)
Group 3 (6.5-8.5 points)	8.04 (0.51)
Group 4 (8.5-10 points)	8.27 (1.32)

Discussion

Background

Histology is one of the first morphological disciplines faced by medical students. Since it is necessary to integrate basic knowledge from other fields (eg, anatomy, cytology, biology, biochemistry) with spatial awareness, histology is perceived as a difficult subject by most learners. Moreover, students consider histology as irrelevant on their board examination (ie, "Spanish Specialized Health Training examination") and even for their future clinical practice [21,22]. Most current medical students use social networks daily and demand considerable effort from educators to make the subjects more attractive and dynamic. Creating a social media account is a free educational option that enables access to information and allows users to easily connect with others [23]. Thus, in this work, we analyzed the impact of using a specific Instagram account (@histologiauma) as a teaching resource during a histology course (2022-2023).

Principal Findings and Implications

Overall, our data demonstrate that medical students who followed and interacted with @histologiauma improved their exam scores compared with those who did not. In fact, a complete lack or a low level of participation generated significant differences in comparison with students who actively engaged with the activity. Most importantly, the enhancement of final grades compared with previous cohorts was not a direct consequence of the extra points awarded to the participants. Thus, improved test performance may serve as indirect and tangible evidence of better long-term knowledge acquisition [24]. These results are supported by the previous opinion of the majority of our students about the positive impact of this experience on academic results. In the first instance, the pre-experiment survey already showed that most of our students believed that social media is rarely used in educational contexts and considered that it may be relevant to include social media platforms as teaching tools, not only to increase accessibility to the content but also to improve their marks. In fact, the results demonstrated a positive disposition toward this innovative approach, since 99.5% of the participants believed they could improve their academic grades thanks to this experience even before participating in it.

Research on the strength or quality of motivation as a predictor of academic success has yielded both definitive and inconclusive findings. In this work, higher engagement and interaction with the content through the proposed interactive activities may have helped in the learning process, which was later reflected in the scores. Indeed, motivation is a determining factor not only for medical students but also for all students to develop sophisticated and successful learning strategies. A study on small group learning found that increased knowledge and understanding of subject matter increase students' motivation for studying and interest in the course content [24]. The "social constructivist theory" states that socialization can also help students during their personal learning processes [3]. In this sense, social media facilitates active interaction and collaboration by enabling instant communication and motivation [24,25].

Furthermore, our Instagram activities also served as additional virtual tests. Testing is no longer considered as only a tool for evaluation but also for learning [26,27]. Thus, using Instagram for educational purposes incorporates not only these phenomena in the process (as it could have been done through a virtual platform like Moodle) but also other factors that are particularly relevant for current young students: direct interaction with their classmates and immediacy, in addition to their own behavior and daily routine with smartphones and social media. We believe that all these factors increased the motivation and engagement of students with histology, leading to greater retention of the content that was finally reflected as higher scores.

Unfortunately, the information available on social media platforms might not be updated or subjected to peer review; thus, it may be invalid, incorrect, or even false. Conversely, @histologiauma is a platform controlled by our group of specialized teachers, prepared to guide learners toward appropriate knowledge according to the content of the subject. Therefore, the creation of a platform adapted by the teaching staff to the curricular content is ideal not only to boost interest but also to prevent students from accessing unreliable information [28].

Additionally, it is essential to comprehend the preferences of learners in order to create a quality digital learning environment [29,30]. During the experience, the image-based questions, multiple-choice questions, and histological descriptions were considered very useful by the students. Ultimately, this knowledge may help teachers to understand the strengths and weaknesses of the subject matter as well as its impact on adherence.

Comparison With the Literature

Numerous social media accounts disseminate information about many different types of pathologies to the general public. Although our work focuses on a course within the medical degree program, it is evident that Instagram serves as an optimal and cost-effective platform for capturing attention through passive learning in the field of histology and pathology [31]. In this sense, Nguyen et al [14] reported that 92.5% of students visit Instagram for educational purposes. Accordingly, 97.9%

of our respondents strongly believed that social networks should be implemented in higher education.

As far as we know, many accounts share educational content about pathology [16,32], but very few are specifically targeted at histology and assessing the impact of sharing this information on social media on medical students. Another novelty of our approach is that we used Instagram as an educational tool specifically tailored to our students, offering personalized content directly aligned with the course curriculum. Although many other studies have examined the use of social media in education, few have focused on how a targeted, image-based platform like Instagram can enhance engagement and learning outcomes in medical education, particularly in a subject highly reliant on visual materials.

For instance, Essig et al [19] from the School of Medicine at the University of North Carolina created an experience with the Instagram profile @InstaHisto in 2020, which is the most similar to our work in the existing literature. However, one of the main differences between these profiles could be summarized by the word "personalization." Our private account was created solely and exclusively for second-year students in the medicine degree program at the University of Malaga, in contrast with their public profile. Moreover, they examined the impact of the posts based on the number of views, not focusing on student interaction but rather on the general public. Instead, our work aimed to potentiate students' knowledge acquisition and to increase engagement with the subject. We also intended to understand students' perceptions about this educational tool. Similarly, the work by Essig et al [19] focused on the National Board of Medical Examiners final exams, reporting that 77% of their students found the histology content from @InstaHisto useful for passing the test. In this line, our survey data reflected a high degree of satisfaction with the utility in the educational environment of these virtual activities (96.5% and 99.3% with multiple-choice questions and image-based questions). More recently, Prabhu and Munawar [11] evaluated 49 Instagram profiles dedicated to the dissemination and teaching of radiology, concluding that it is indeed a better application of this image-based social media platform due to its easy accessibility and appeal to students. In this line, our data reflect that 95% of students believe that using Instagram would enhance their perception of the course and its appeal.

Limitations

In this work, rating grades increased after use of @histologiauma, even before adding the points awarded for students' participation in Instagram. It is noticeable that scarce involvement led to no or low improvement, and although there was no significant difference between the most active groups (G2 to G4), we did find a trend toward higher ratings according to the level of participation. In this sense, we cannot completely rule out the influence of other factors masking the impact of this experience, including other curricular or extracurricular activities performed during the school year or personal preferences regarding social media. Nevertheless, as far as we are concerned, standard students from the same academic course share identical academic schedules. All other activities performed during the histology course (such as problem-based learning or the section "From Histology to Medicine") were developed during theoretical classes or practice and were mandatory. However, we are aware that every academic group is different and repeating the experience during additional academic years would yield more reliable data. Relevant to this, the algorithm used by social media platforms like Instagram tends to favor posts from accounts with which users interact more frequently or with related content, creating information bias for customers [33]. Given this scenario, it is reasonable to interpret that students who interacted with @histologiauma above a threshold were later shown our account or similar profiles in their private feed more often than those who barely engaged or did not participate in the voluntary activity. This interaction likely results in students passively reviewing content each time they visit this platform, which finally positively impacts their acquisition and assimilation of knowledge and therefore their final results. However, this may also minimize the differences between the most active groups.

Another specific limitation of this work is that participating in these activities was not mandatory, which may have led to potential selection bias. We cannot rule out that students following the @histologiauma account were more eager to participate in additional activities than the general medical students. However, the high participation rate (85.6% of students enrolled) notably reduced the impact of this possibility. Even so, eliminating the optional nature of this activity would have yielded clearer data. In addition, it is possible that some of our students do not have an Instagram account because they do not find social networks attractive. Nevertheless, this was rarely the case, since we detected very few exceptions of students who performed the final exam without participating in Instagram (10 of 153 students). For future editions, we intend to propose @histologiauma as an educational instrument in a public mode, encouraging users to create their own hashtags and check the transcendence of the posts.

On the other hand, the use of social media platforms during the educational process of histology should be recommended only as a complement for regular teaching. Evidence that social media is not a panacea was provided in a separate analysis of 131 students who were using the microblog X during class with the aim of fostering student-faculty interaction on two campuses. Although it facilitated discussion, 71% of students found it distracting [34]. For this reason, it is important to find a balance between the usual lecture-based methodology and the inclusion of social media in higher education, not only to meet the curricular needs of students but also to ensure their engagement with their studies.

For future research, the sample size may be broadened to increase the validity and reliability of our findings and include cohorts from other courses or health science degrees such as podiatry, physiotherapy, or nursing. Moreover, specific tests about the content shown on the Instagram account could be implemented. The inclusion of a longitudinal study to track students' performance and engagement over multiple semesters would allow better understanding of the long-term impact of Instagram-based learning. Although, in general, we detected typical mistakes of pattern recognition of histological structures, a range of accuracy-based rewards could be incorporated into

the activity to avoid participation without true commitment. Finally, future experiences could explore the impact of different types of Instagram content (eg, video, live question-and-answer sessions, different quizzes).

Conclusions

Medical students consider there is inadequate use of social networks for teaching purposes, probably due to a lack of updated methodological approaches in the context of university subjects. Compared with the conventional educational system, social media platforms have a considerable impact on both teachers and students as they offer the possibility to easily connect and collaborate. In fact, one of the main objectives of medical education is to capitalize on the engaging nature of social media tools as part of an overall strategy to use a learner-centered approach. In addition, to increase student engagement during the first year of the degree in Medicine, it is desirable to use attractive didactic methods for learning histology. In this regard, the visual nature of histology is particularly appropriate for the introduction of new image-based tools. Thus, the aim of this study was to investigate an innovative online educational approach for histology based on an Instagram account specifically designed for medical students. In this work, we showed that the use of Instagram has great potential to improve not only the knowledge but also the scores of students of human histology. Our results provide evidence that this teaching strategy boosts students' learning motivation. In the near future, the classical practical lessons based on the physical microscope might not be enough to meet the needs of medical students. Therefore, Instagram may be considered as a relevant tool for current students to achieve their curricular objectives in a more dynamic, friendly, and enjoyable way under the supervision of the faculty.

Acknowledgments

Article processing charges were funded by Plan Propio de IBIMA-Plataforma Bionand 2024 and Project Oficina de Transferencia de Resultados de Investigacion (Office for Transference of Research Results) of the University of Malaga (reference: 806/86.6531 to AES). We thank the faculty of the Histology Unit from the University of Malaga for the histological images.

Conflicts of Interest

None declared.

References

1. Kemp S. Digital 2020: Global Overview Report. DATAREPORTAL. 2022 Jan 26. URL: https://datareportal.com/reports/digital-2022-global-overview-report [accessed 2025-01-20]
2. Wutoh R, Boren SA, Balas EA. eLearning: a review of Internet-based continuing medical education. J Contin Educ Health Prof 2004;24(1):20-30. [doi: 10.1002/chp.1340240105] [Medline: 15069909]
3. Kalasi R. The impact of social networking on new age teaching and learning: an overview. Journal of Education & Social Policy 2014;1(1):23-28 [FREE Full text]
4. Guckian J, Utukuri M, Asif A, Burton O, Adeyoju J, Oumeziane A, et al. Social media in undergraduate medical education: a systematic review. Med Educ 2021 Nov;55(11):1227-1241 [FREE Full text] [doi: 10.1111/medu.14567] [Medline: 33988867]
5. Kennedy G, Gray K, Tse J. 'Net Generation' medical students: technological experiences of pre-clinical and clinical students. Med Teach 2008 Feb;30(1):10-16 [FREE Full text] [doi: 10.1080/01421590701798737] [Medline: 18278643]
6. Katz M, Nandi N. Social media and medical education in the context of the COVID-19 pandemic: scoping review. JMIR Med Educ 2021 Apr 12;7(2):e25892 [FREE Full text] [doi: 10.2196/25892] [Medline: 33755578]
7. Villagaray-Pacheco N, Villacorta-Landeo P, Tejada-Llacsa PJ. [Social media and medical education in COVID-19 pandemic]. Rev Med Chil 2020 Aug;148(8):1220-1221 [FREE Full text] [doi: 10.4067/S0034-98872020000801220] [Medline: 33399791]
8. Chan AKM, Nickson CP, Rudolph JW, Lee A, Joynt GM. Social media for rapid knowledge dissemination: early experience from the COVID-19 pandemic. Anaesthesia 2020 Dec;75(12):1579-1582 [FREE Full text] [doi: 10.1111/anae.15057] [Medline: 32227594]
9. Chapman JA, Lee LMJ, Swailes NT. From scope to screen: the evolution of histology education. Adv Exp Med Biol 2020;1260:75-107. [doi: 10.1007/978-3-030-47483-6_5] [Medline: 33211308]
10. Tauber Z, Lacey H, Lichnovska R, Erdosova B, Zizka R, Sedy J, et al. Students´ preparedness, learning habits and the greatest difficulties in studying histology in the digital era: A comparison between students of general and dental schools. Eur J Dent Educ 2021 May 19;25(2):371-376. [doi: 10.1111/eje.12613] [Medline: 33012128]
11. Prabhu V, Munawar K. Radiology on Instagram: analysis of public accounts and identified areas for content creation. Acad Radiol 2022 Jan;29(1):77-83. [doi: 10.1016/j.acra.2020.08.024] [Medline: 32980242]
12. Huang AS, Abdullah AAN, Chen K, Zhu D. Ophthalmology and social media: an in-depth investigation of ophthalmologic content on Instagram. Clin Ophthalmol 2022;16:685-694 [FREE Full text] [doi: 10.2147/OPTH.S353417] [Medline: 35300033]

13. Chen JY, Gardner JM, Chen SC, McMichael JR. Instagram for dermatology education. J Am Acad Dermatol 2020 Oct;83(4):1175-1176. [doi: 10.1016/j.jaad.2020.02.001] [Medline: 32035941]
14. Nguyen VH, Lyden ER, Yoachim SD. Using Instagram as a tool to enhance anatomy learning at two US dental schools. J Dent Educ 2021 Sep 29;85(9):1525-1535 [FREE Full text] [doi: 10.1002/jdd.12631] [Medline: 33913160]
15. Peyser A, Goldstein L, Mullin C, Goldman RH. Fertility education: what's trending on Instagram. Fertil Res Pract 2021 Jan 18;7(1):3 [FREE Full text] [doi: 10.1186/s40738-021-00095-6] [Medline: 33461628]
16. Cutshall H, Hattaway R, Singh NP, Rais-Bahrami S, McCleskey B. The #Path2Path virtual landscape during the COVID-19 pandemic: preparing for the 2020 pathology residency recruitment season. Acad Pathol 2021;8:23742895211002783 [FREE Full text] [doi: 10.1177/23742895211002783] [Medline: 34192133]
17. Dorfman RG, Vaca EE, Mahmood E, Fine NA, Schierle CF. Plastic surgery-related hashtag utilization on Instagram: implications for education and marketing. Aesthet Surg J 2018 Feb 15;38(3):332-338. [doi: 10.1093/asj/sjx120] [Medline: 29040378]
18. Douglas NKM, Scholz M, Myers MA, Rae SM, Elmansouri A, Hall S, et al. Reviewing the role of Instagram in education: can a photo sharing application deliver benefits to medical and dental anatomy education? Med Sci Educ 2019 Dec;29(4):1117-1128 [FREE Full text] [doi: 10.1007/s40670-019-00767-5] [Medline: 34457591]
19. Essig J, Watts M, Beck Dallaghan GL, Gilliland KO. InstaHisto: utilizing Instagram as a medium for disseminating visual educational resources. Med Sci Educ 2020 Sep 17;30(3):1035-1042 [FREE Full text] [doi: 10.1007/s40670-020-01010-2] [Medline: 34457765]
20. Kebritchi M, Lipschuetz A, Santiague L. Issues and challenges for teaching successful online courses in higher education. Journal of Educational Technology Systems 2017 Aug 08;46(1):4-29 [FREE Full text] [doi: 10.1177/0047239516661713]
21. Johnson S, Purkiss J, Holaday L, Selvig D, Hortsch M. Learning histology - dental and medical students' study strategies. Eur J Dent Educ 2015 May;19(2):65-73 [FREE Full text] [doi: 10.1111/eje.12104] [Medline: 24809952]
22. Dennis JF. The HistoHustle: supplemental histology sessions to enrich student learning and self-efficacy. Med Sci Educ 2020 Dec;30(4):1725-1726 [FREE Full text] [doi: 10.1007/s40670-020-01060-6] [Medline: 34457834]
23. Prudencio J, Wongwiwatthananukit S, Lozano A, Xu Y. Instagram as a tool to enhance pharmacy student learning of ambulatory care pharmacy. Curr Pharm Teach Learn 2021 Feb;13(2):134-138. [doi: 10.1016/j.cptl.2020.09.007] [Medline: 33454069]
24. Kusurkar RA, Ten Cate TJ, van Asperen M, Croiset G. Motivation as an independent and a dependent variable in medical education: a review of the literature. Med Teach 2011;33(5):e242-e262. [doi: 10.3109/0142159X.2011.558539] [Medline: 21517676]
25. Hennessy C, Smith C. Digital and social media in anatomy education. Adv Exp Med Biol 2020;1260:109-122. [doi: 10.1007/978-3-030-47483-6_6] [Medline: 33211309]
26. Yang BW, Razo J, Persky AM. Using testing as a learning tool. Am J Pharm Educ 2019 Nov;83(9):7324 [FREE Full text] [doi: 10.5688/ajpe7324] [Medline: 31871352]
27. Murphy DH, Little JL, Bjork EL. The value of using tests in education as tools for learning—not just for assessment. Educ Psychol Rev 2023 Sep 08;35(3):1. [doi: 10.1007/s10648-023-09808-3]
28. Popoola-Samuel HAO, Bhuchakra HP, Tango T, Pandya ND, Narayan KL. Instagram and seizure: knowledge, access, and perception of circulating information on the internet. Cureus 2023 Jul;15(7):e41664 [FREE Full text] [doi: 10.7759/cureus.41664] [Medline: 37575724]
29. Clauson KA, Singh-Franco D, Sircar-Ramsewak F, Joseph S, Sandars J. Social media use and educational preferences among first-year pharmacy students. Teach Learn Med 2013 Apr;25(2):122-128 [FREE Full text] [doi: 10.1080/10401334.2013.770742] [Medline: 23530673]
30. Timothy PG, Jeffrey B, Kaitlyn L, Margarita VD. Delivery of educational content via Instagram. Med Educ 2016 Apr 13;50(5):575-576 [FREE Full text] [doi: 10.1111/MEDU.13009] [Medline: 27072461]
31. Gomaa B, Houghton RF, Crocker N, Walsh-Buhi ER. Skin cancer narratives on Instagram: content analysis. JMIR Infodemiology 2022;2(1):e34940 [FREE Full text] [doi: 10.2196/34940] [Medline: 37113805]
32. Schukow CP, Kilpatrick SE. Highlighting bone and soft tissue pathology on Instagram. Adv Anat Pathol 2023 Mar 06:1. [doi: 10.1097/PAP.0000000000000396] [Medline: 36882880]
33. Gurler D, Buyukceran I. Assessment of the medical reliability of videos on social media: detailed analysis of the quality and usability of four social media platforms (Facebook, Instagram, Twitter, and YouTube). Healthcare (Basel) 2022 Sep 22;10(10):1 [FREE Full text] [doi: 10.3390/healthcare10101836] [Medline: 36292284]
34. Fox BI, Varadarajan R. Use of Twitter to encourage interaction in a multi-campus pharmacy management course. Am J Pharm Educ 2011 Jun 10;75(5):88 [FREE Full text] [doi: 10.5688/ajpe75588] [Medline: 21829262]

Edited by B Lesselroth; submitted 04.01.24; peer-reviewed by F Tume, H Zou, J Amaro, M Garrosa, H Alshawaf; comments to author 14.08.24; revised version received 10.09.24; accepted 17.12.24; published 19.02.25.

Please cite as:
Escamilla-Sanchez A, López-Villodres JA, Alba-Tercedor C, Ortega-Jiménez MV, Rius-Díaz F, Sanchez-Varo R, Bermúdez D
Instagram as a Tool to Improve Human Histology Learning in Medical Education: Descriptive Study
JMIR Med Educ 2025;11:e55861
URL: https://mededu.jmir.org/2025/1/e55861
doi:10.2196/55861
PMID:

Virtual Standardized Patients for Improving Clinical Thinking Ability Training in Residents: Randomized Controlled Trial

Liyuan Xu, MMS; Qinrong Xu, BMS; Chunya Liu, MMS; Baozhen Chen, BMS; Chunxia Wang, MMS

Department of Endocrinology, Wenzhou Medical University Affiliated Quzhou Hospital, Quzhou People's Hospital, No. 100 Minjiang Avenue, Quzhou City, Quzhou, Zhejiang Province, China

Corresponding Author:
Chunxia Wang, MMS
Department of Endocrinology, Wenzhou Medical University Affiliated Quzhou Hospital, Quzhou People's Hospital, No. 100 Minjiang Avenue, Quzhou City, Quzhou, Zhejiang Province, China

Abstract

Background: Clinical internal medicine practice training traditionally relies on case-based teaching. This approach limits the development of students' clinical thinking skills. It also places significant pressure on instructors. Virtual standardized patients (VSPs) could offer an alternative solution. However, evidence on their feasibility and effectiveness remains limited.

Objective: This study aims to use the “VSPs in general practice” interactive diagnostic and teaching system, which uses VSPs to provide 3D virtual simulated patients and mimic virtual clinical scenarios. Medical students are trained through system-preset cases. This study aims to establish the clinical application of VSPs through a “VSPs in general practice” system and compare its effectiveness with traditional teaching in improving students’ clinical thinking ability.

Methods: A randomized controlled trial was conducted from October 20, 2022, to October 20, 2024. A total of 60 medical students interning at Quzhou People’s Hospital were enrolled and divided into 2 groups: the experimental group receiving VSP training (30/60, 50%) and the control group receiving traditional academic training (30/60, 50%). The teaching effectiveness was evaluated using basic knowledge assessments and virtual system scoring. After completing the course, students were surveyed with a questionnaire to assess their satisfaction with the course.

Results: All enrolled medical students completed the study. In the evaluation of training effectiveness, the experimental group showed significantly greater improvement in theoretical scores compared to the control group (mean 17.07, SD 4.24 vs mean 10.67, SD 4.91; $F_{1, 59}$=29.20; Cohen *d*=1.15; 95% CI 12.43-15.31; *P*<.001); the total score improvement in the virtual clinical thinking training system test was also significantly better in the experimental group than in the control group (mean 42.60, SD 9.56 vs mean 31.63, SD 7.24; $F_{1, 59}$=25.10; Cohen *d*=1.09; 95% CI 34.51-39.72; *P*<.001). Specifically, improvements in consultation skills (mean 8.76, SD 1.67 vs mean 7.66, SD 2.08; $F_{1, 59}$=31.09; Cohen *d*=0.55; 95% CI 7.70-8.70; *P*<.001), overall objective improvement (mean 11.97, SD 2.77 vs mean 8.15, SD 2.62; $F_{1, 59}$=30.08; Cohen *d*=1.16; 95% CI 9.21-10.91; *P*<.001), initial diagnostic ability (mean 8.74, SD 1.67 vs mean 7.66, SD 2.08; $F_{1, 59}$=4.91; Cohen *d*=0.55, 95% CI 7.70-8.70; *P*=.03), and ability to provide patient treatment (mean 7.23, SD 2.41 vs mean 5.72, SD 2.19; $F_{1, 59}$=6.42; Cohen *d*=0.63; 95% CI 5.85-7.01; *P*=.01) were significantly higher in the experimental group than in the control group. The questionnaire results indicated that 90% (27/30) of the students who participated in the VSPs’ training believed it could enhance their clinical thinking abilities.

Conclusions: VSPs reinforce the foundational knowledge of internal medicine among medical students and enhance their clinical thinking abilities, as well as improve their capacity for independent work. The VSP system is feasible, practical, and cost-effective, making it worthy of further promotion in clinical education.

(JMIR Med Educ 2025;11:e73196) doi:10.2196/73196

KEYWORDS

virtual standardized patients; clinical competence; clinical thinking ability; medical education; medical students

Introduction

Against the backdrop of the continuous expansion of higher medical education, clinical skills training is facing multiple developmental challenges. On one hand, medical schools are under real pressure to address the shortage of millions of health care professionals, urgently needing to scale up the training of high-quality clinical physicians [1]. On the other hand, they are constrained by the relative scarcity of high-quality teaching resources, leading to significant regional disparities in the quality of medical talent cultivation [2]. Clinical reasoning ability, as the core competency for medical students transitioning into qualified clinical physicians, essentially requires the integration of multidimensional clinical information such as patient history

collection, physical sign identification, laboratory data interpretation, and imaging analysis [3]. However, the current clinical thinking training system still has significant shortcomings, primarily relying on traditional teaching methods in ward practice. This approach not only lacks opportunities for students to independently construct clinical reasoning frameworks but also fails to stimulate the cultivation of critical thinking and clinical decision-making abilities [4].

Driven by the rapid development of information technology, virtual simulation technology is deeply reconstructing the medical education ecosystem. Virtual standardized patients (VSPs), defined as artificial intelligence (AI)–driven, interactive virtual humans that simulate real patient encounters in 3D environments, offer a scalable alternative. As a product of the deep integration of computer multimedia technology and clinical medicine, VSPs have become an important carrier for revolutionizing clinical thinking training paradigms by accurately simulating real diagnosis and treatment scenarios through intelligent interactive systems [5-7]. Unlike static case studies, VSPs dynamically respond to learners' actions, replicating nuanced clinical scenarios, from history-taking to diagnostic decision-making. With the rise of generative AI, modern VSP systems can now leverage natural language processing and machine learning to mimic diverse patient phenotypes, pathologies, and even emotional states, providing a near-lifelike training experience. In the field of international medical education, mature application systems like the DxR Clinician, developed by Southern Illinois University School of Medicine, have established benchmarks through AI-driven virtual patient simulations, dynamic clinical decision trees, and intelligent feedback mechanisms, making it a gold standard for clinical thinking training [8-10].

In contrast, VSP teaching in China is still in the exploratory stage of localization and adaptation and has not yet formed a large-scale application system. From the perspective of teaching implementation, the VSP system reconstructs the clinical competency development path through 3 core mechanisms [11]: first, constructing a full lifecycle disease spectrum database, covering typical and rare cases from newborns to the older population; second, creating a human-computer interactive consultation environment that requires learners to independently complete the entire process of medical history collection, physical examination operations, auxiliary examination selection, and treatment plan formulation; finally, relying on intelligent evaluation algorithms, millisecond-level feedback is provided to operational nodes, marking logical vulnerabilities and providing evidence-based medical treatment recommendations. This closed-loop system of "simulation practice, instant feedback, correction, and improvement" effectively breaks through the time and space limitations and ethical dilemmas of traditional bedside teaching.

Clinical empirical research shows that VSPs have unique advantages. In terms of diagnostic thinking, through the dynamic deduction of complex cases such as diabetic ketoacidosis, cultivate the ability to make timing decisions for differential diagnosis. In terms of treatment decision-making, individualized treatment plans based on drug metabolism characteristics are established through scenario simulations such as antibiotic tiered use [12]. In terms of humanistic literacy, a role-playing system for patients with depression is used to train their ability to analyze the social and psychological factors behind symptoms. This teaching paradigm of integrating reality and virtuality is reshaping the new landscape of clinical competence cultivation [13]. In recent years, the demand for medical safety in society has been constantly increasing, and the professional pressure on physicians has also increased accordingly. The long training period, heavy social responsibility, and high professional risks of physicians make it a key issue for medical education to improve skill levels in high-pressure environments. The virtual patient system provides a secure practice platform for medical students, allowing them to make mistakes in the virtual environment and learn from them, thereby reducing future medical errors in real clinical environments. This "trial and error learning" model not only helps cultivate high-level medical workers, but also provides new directions for the reform of medical education [14].

Traditional clinical internal medicine training heavily relies on case-based teaching, which often limits the development of students' clinical reasoning skills and places substantial demands on instructors [15]. While VSPs offer a promising alternative, empirical evidence regarding their feasibility and effectiveness in enhancing clinical thinking remains limited. This study aimed to evaluate whether VSP-based training outperforms traditional methods in improving medical students' clinical thinking abilities. Specifically, we addressed the following questions: (1) Does VSP training lead to greater improvements in theoretical knowledge and practical diagnostic skills compared to conventional case-based teaching? (2) Which components of clinical thinking are most influenced by VSPs? (3) How do learners perceive the utility of VSPs in their training? We hypothesized that the "VSPs in general practice" system, a 3D virtual platform simulating patient interactions, would significantly enhance clinical thinking metrics due to its immersive, repeatable scenarios. To test this, we conducted a randomized controlled trial (RCT) comparing VSP training with traditional teaching, assessing outcomes via standardized knowledge tests, virtual system scoring, and posttraining surveys. By rigorously evaluating VSPs, this study provides actionable insights into scalable, cost-effective alternatives to traditional clinical education, addressing gaps in evidence-based teaching innovations.

Methods

Trainee Recruitment

Based on the pilot data, we determined the sample size for this study to be 60 medical students. From a total pool of 85 eligible fifth-year undergraduate medical students interning at Quzhou People's Hospital (affiliated with Wenzhou Medical University and Zhejiang University of Traditional Chinese Medicine), we recruited 60 participants via consecutive sampling between October 20, 2022, and October 20, 2024 (.

Inclusion and Exclusion Criteria

We selected fifth-year undergraduate medical students from Wenzhou Medical University and Zhejiang University of Traditional Chinese Medicine who were interning at our

hospital. The exclusion criteria were (1) participants who had received VSP or standardized patient training, (2) participated in courses related to clinical thinking skills, and (3) failure to comply with the research plan or withdrawal from the study.

Randomization

Using computer-generated randomization, participants were randomly assigned to either the experimental group or the control group in a 1:1 ratio. Random grouping is conducted by individuals who have not had contact with the participants.

Training Curriculum and Setting

The control group received traditional clinical thinking training through a combined problem-based learning and real-patient exposure approach. Supervisors selected patients from routine admissions in the Department of Internal Medicine who met the criteria for typical clinical cases outlined in the teaching syllabus, prioritizing broad alignment with the VSP curriculum while preserving real-world clinical heterogeneity. No further screening for specific attributes was applied to maintain ecological validity. Students were divided into small groups (10 per group) to: (1) conduct standardized clinical workflows (history-taking, physical examinations, and interpretation of laboratory and auxiliary tests); (2) formulate preliminary diagnoses and treatment plans; (3) participate in structured problem-based learning discussions based on these real medical records. Supervisors provided systematic theoretical explanations and feedback throughout the process. All patient interactions adhered to hospital protocols, with explicit consent obtained for student involvement. While exact case matching with the VSP group was logistically unfeasible, supervisors ensured thematic consistency to mitigate exposure variability between groups.

The experimental group used the VSPs of the "VSPs in general practice" interactive diagnosis and treatment teaching system for teaching. The study used the HIWILL VSP System (version 2.1; Huawei Medical Technology), an interactive 3D clinical simulation platform designed for clinical thinking training. Each student in the experimental group completed 12 standardized VSP cases during the training period, with each case representing a different common clinical scenario in internal medicine. Students were allowed up to 3 attempts per case to achieve optimal performance, with immediate feedback provided after each attempt. The operation process of the VSP teaching system is shown in Figure 1.

Figure 1. Operation procedures of the virtual standardized patient system.

Selection and Improvement of VSPs

By using the VSPs included in the "VSPs in general practice" interactive diagnosis and treatment teaching system, the chief complaints, disease descriptions, consultations and physical examinations, laboratory and auxiliary examinations, as well as clinical cases required to be mastered in the treatment and teaching syllabus of VSPs are organically combined.

Application of Virtualization Teaching Software

Implement VSP teaching for experimental group students, and teachers provide guidance and training on software usage methods for students. Students form small groups of 10 and complete the entire clinical diagnosis and treatment process through interaction with VSPs. Obtain the patient's condition description, chief complaint, and abnormal signs in a virtual clinical environment. Based on the obtained consultation information, perform examinations on VSPs, including physical examination, laboratory examination, and, if necessary,

histopathological examination. Students conducted a preliminary diagnosis of VSPs, provided corresponding diagnostic criteria, and then treated the disease. Teachers provide guidance and systematic explanations and training on relevant professional theoretical knowledge. All teachers supervising VSP sessions completed a standardized 8-hour training program that included: technical operation of the VSP system, standardized facilitation techniques, methods for providing consistent feedback, and protocols for troubleshooting technical issues. Instructors were required to pass a competency assessment before supervising sessions (Multimedia Appendix 1).

Evaluation of Training Effectiveness

All students underwent medical record analysis exams and assessments of relevant professional theoretical knowledge upon enrollment. A 50-item multiple-choice exam covering core internal medicine topics was developed by a panel of 5 senior clinicians. And their clinical thinking abilities were evaluated through the intelligent module of the "VSPs in general practice" interactive diagnosis and treatment teaching system. The system evaluated performance across 4 domains: consultation skills (history taking and communication), diagnostic accuracy (initial and differential diagnoses), treatment planning (evidence-based interventions), and overall clinical reasoning (logical progression and efficiency). The score was recorded as the base score F_0, and they entered group teaching. After completing the teaching tasks, all students underwent medical record analysis exams and assessments of relevant professional theoretical knowledge again, and their clinical thinking abilities were reevaluated through the intelligent module of the "VSPs in general practice" interactive diagnosis and treatment teaching system. The score was recorded as F_1-score. Traditional case-based evaluations were structured to mirror the VSP domains but used paper-based scenarios. Both groups were assessed by blinded instructors using identical criteria.

The VSP system test primarily encompassed 4 components, with the medical interview section accounting for 25 points. This section evaluated the student's ability to gather relevant medical history through human-computer dialogue. The condition examination section was worth 30 points, comprising a physical examination worth 15 points and auxiliary examinations worth 15 points. It assessed the student's capability to perform necessary physical examinations based on the collected information and to select relevant tests and laboratory investigations required for diagnosis and differential, including histopathological examinations when necessary. The diagnosis and differential section carried 25 points, consisting of an initial diagnosis worth 16 points and a differential diagnosis worth 9 points. This section tested the student's ability to choose the correct diagnostic criteria based on known information to make an accurate diagnosis and to analyze inclusion and exclusion criteria to complete the differential diagnosis. The treatment plan section was worth 20 points, with medication management accounting for 17 points and nonpharmacological interventions for 3 points. It evaluated the student's competence in providing the correct treatment plan for the condition.

Reference Standard for VSP Scoring

The VSP scoring algorithm was cross-validated against blinded expert evaluations (3 senior clinicians) who reviewed 20% of randomly selected VSP sessions. Interrater reliability (Fleiss κ=0.81) confirmed consistency between automated and manual scoring. Discrepancies were resolved by consensus.

Feedback Questionnaire

Based on the characteristics of this study and reference literature, a self-designed learning effect questionnaire was used to investigate the effectiveness of prehospital emergency training among students. It mainly includes 5 questions, namely (1) beneficial for improving the ability to collect medical history, (2) it is conducive to improving the ability to analyze the condition, (3) beneficial for improving diagnostic and differential diagnostic capabilities, (4) beneficial for improving clinical treatment capabilities, and (5) it is conducive to improving clinical thinking ability. The evaluation criteria are divided into 4 levels: full agreement, agreement, partial agreement, and disagreement, which are filled in by students based on their true feelings. Those who fully agree or agree in the questionnaire will be considered as positive reviews, and the overall positive review rate and individual positive review rate will be comprehensively calculated. In addition, we conducted a questionnaire survey on 6 teachers responsible for clinical training to evaluate the potential impact of VSP courses on their work.

Statistical Analysis

SPSS 26.0 statistical software (IBM Corp) was used to perform statistical processing on the data obtained from this study, analyzing the differences in scores between each group of students before and after teaching, and analyzing the differences in F_0 and F_1-scores between the 2 groups of students, in order to evaluate the differences in clinical thinking ability between students before and after teaching and under 2 different teaching modes. Continuous variables are represented as mean (SD), while categorical variables are represented as frequency or percentage. All continuous outcomes were analyzed using independent *t* tests, reporting mean differences, 95% CIs, and Cohen *d* as effect size measures. For categorical outcomes, Chi-square tests with odds ratios (ORs) and 95% CIs were used.

Ethical Considerations

This study, which involved medical students as participants in an educational intervention evaluation, was granted a formal exemption from requiring full ethical review by the Quzhou People's Hospital Medical Ethics Review Committee (review number: 2025 - 004). This exemption is in full compliance with the institution's policy, as well as the institutional guidelines for educational research. Specifically, the study was deemed to fall under the category of low-risk educational interventions that are part of the normal evaluation of curriculum quality and teaching methods. The anonymity of the participants was ensured, and the data were collected and analyzed in an aggregated manner for research purposes only, posing minimal risk to the participants. All participants provided written informed consent before participating in the study and received

corresponding remuneration according to hospital regulations after the study ended.

Results

The Basic Characteristics of Participants

There were 30 students in the control group, including 18 females and 12 males, with a mean age of 21.73 (SD 0.85) years. The experimental group consisted of 30 students, including 16 females and 14 males (mean age 21.60, SD 0.73) years (Figure 2). The prehospital emergency training for both groups is 6 months, and the training teachers for both groups are the same 6 teachers (2 bishops and 4 assistants). The average credit scores of the 2 groups of students were 3.66 (SD 0.36) points and 3.58 (SD 0.32) points, respectively. There were no statistically significant differences between the control group and the experimental group in terms of gender, age, average credit scores, training time, and training teachers, as shown in Table 1.

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) flow diagram. SP: standardized patient training; VSP: virtual standardized patient.

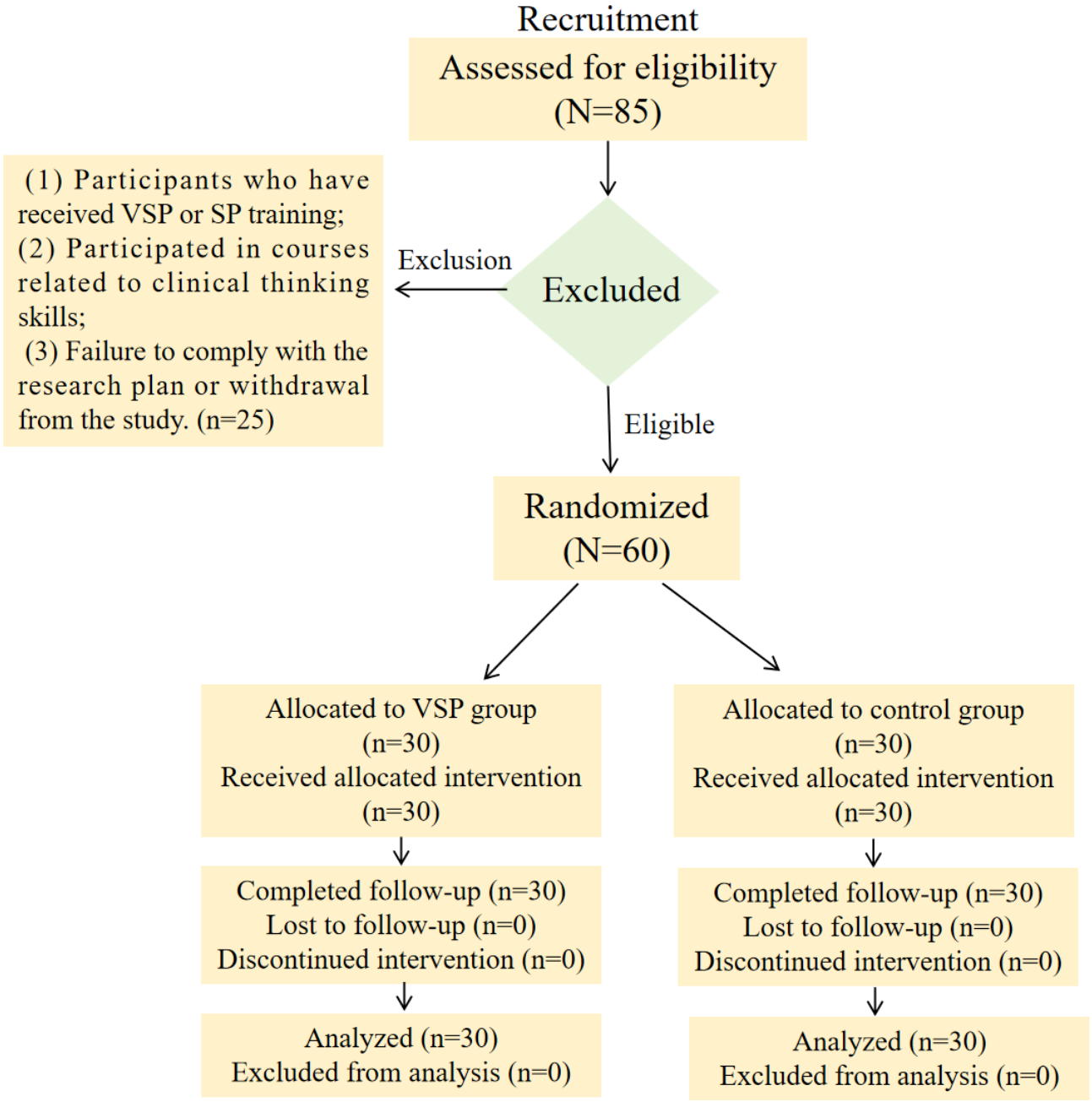


Table . General information of 2 groups of students.

	Control group (n=30)	Experimental group (n=30)	*t* test (*df*)	*P* value
Sex, n (%)			0.620 (59)	.28
Male	12 (40.0)	14 (46.7)		
Female	18 (60.0)	16 (53.3)		
Age (years), mean (SD)	21.73 (0.85)	21.60 (0.73)	3.454 (59)	.18
Grade point, mean (SD)	3.66 (0.36)	3.58 (0.32)	0.046 (59)	.75

Evaluation of Training Effectiveness

Systematic Knowledge Test

Figure 3A displays the theoretical performance results of 2 groups of trainees before and after training. Prior to clinical training, there was no significant difference in the theoretical performance between the experimental group and the control group (mean 65.83, SD 7.11 vs mean 66.57, SD 7.17; $F_{1, 59}$=0.16; Cohen *d*=0.10; 95% CI 64.36-68.03; *P*=.69). After the training, the theoretical performance of the experimental group was significantly better than that of the control group (mean 82.63, SD 5.76 vs mean 77.07, SD 5.89; $F_{1, 59}$=13.71; Cohen *d*=0.87; 95% CI 78.19-81.51; *P*<.001). Additionally, the improvement in theoretical scores was significantly greater in the experimental group compared to the control group (mean 17.07, SD 4.24 vs mean 10.67, SD 4.91; $F_{1, 59}$=29.20; Cohen *d*=1.15; 95% CI 12.43-15.31; *P*<.001).

Figure 3. Results of formative evaluation, including (A) theoretical score, (B) VSP training system improvement score, (C) medical interview improvement score, (D) score improvement for physical examination and auxiliary examination, (E) physical examination improvement score, (F) pathological examination improvement score, (G) improvement of identification and diagnostic abilities, (H) improvement of primary diagnosis ability, (I) improvement of differential diagnosis ability, (J) improvement of treatment plan ability, (K) enhancement of drug treatment plan ability, and (L) enhancement of nondrug treatment plan ability. VSP: virtual standardized patient. *P<.05, **P<.01.

Clinical Skill Test

The test results of the virtual clinical reasoning training system are shown in Figure 3B. Before the clinical training, there was no significant difference in the total scores of the virtual clinical reasoning training system test between the experimental group and the control group (mean 35.90, SD 8 vs mean 35.70, SD 9.55; $F_{1, 59}$=0; Cohen *d*=0.01; 95% CI 33.5-38; *P*=.96). After 4 weeks of training, the total scores of the virtual clinical reasoning training system test in the experimental group were significantly higher than those in the control group (mean 78.30, SD 7.98 vs mean 67.43, SD 8.74; $F_{1, 59}$=25.27; Cohen *d*=1.09; 95% CI 70.30-75.44; *P*＜.001). Additionally, the improvement in the total scores of the virtual clinical reasoning training system test was significantly greater in the experimental group compared to the control group (mean 42.60, SD 9.56 vs mean 31.63, SD 7.24; $F_{1, 59}$=25.10; Cohen *d*=1.09; 95% CI 34.51-39.72; *P*＜.001).

Medical Interview

Figure 3C illustrates the scores of medical history-taking abilities for both groups of medical students before and after the training. After 4 weeks of training, the medical history-taking ability scores of the medical students in the experimental group (mean 20, SD 2.33 vs mean 15.45, SD 5.70; $F_{1, 59}$=49.61; Cohen *d*=1.35; 95% CI 16.86-18.60; *P*<.001) and the improvement in medical history-taking ability scores (mean 12.30, SD 3.43 vs mean 7.77, SD 2.84; $F_{1, 59}$=31.09; Cohen *d*=1.17, 95% CI 9.04-11.04; *P*<.001) were significantly higher than those in the control group.

Physical Examination and Auxiliary Examination

Figure 3D-F displays the scores of the ability to conduct patient examinations for both groups of medical students before and after the training. After 4 weeks of training, the medical students in the experimental group showed significantly greater improvement in physical examination ability (mean 5.86, SD 1.44 vs mean 4, SD 1.34; $F_{1, 59}$=26.41; Cohen *d*=1.11; 95% CI 4.50-5.36; *P*<.001) and in auxiliary examination ability (mean 6.11, SD 1.74 vs mean 4.15, SD 1.63; $F_{1, 59}$=20.36; Cohen *d*=1.01; 95% CI 4.63-5.63; *P*<.001) compared to the control group. Additionally, the total improvement in objective examination scores was significantly higher in the experimental group than in the control group (mean 11.97, SD 2.77 vs mean 8.15, SD 2.62; $F_{1, 59}$=30.08; Cohen *d*=1.16; 95% CI 9.21-10.91; *P*<.001).

Diagnosis and Differentiation

Figure 3 depicts the scores of the ability to diagnose and perform differential diagnoses for both groups of medical students before and after the training. After 4 weeks of training, the medical students in the experimental group showed a significantly greater improvement in initial diagnosis scores (mean 8.74, SD 1.67 vs mean 7.66, SD 2.08; $F_{1, 59}$=4.21; Cohen *d*=0.55; 95% CI 7.70-8.70; *P*=.045) compared to the control group. However, there was no significant difference between the 2 groups in the improvement of differential diagnosis ability (mean 2.37, SD 1.24 vs mean 2.33, SD 1.09; $F_{1, 59}$=0.01; Cohen *d*=0.03; 95% CI 2.05-2.65; *P*=.91).

Treatment Plan

Figure 3J-L displays the scores of the ability to administer treatment to patients for both groups of medical students before and after the training. After 4 weeks of training, the medical students in the experimental group showed significantly greater improvement in medication management ability (mean 5.69, SD 2.27 vs mean 4.51, SD 2.14; $F_{1, 59}$=4.29; Cohen *d*=0.62; 95% CI 4.51-5.69; *P*=.04) and in nonpharmacological treatment ability (mean 1.54, SD 0.47 vs mean 1.21, SD 0.43; $F_{1, 59}$=7.85; Cohen *d*=0.68; 95% CI 1.25-1.50; *P*=.007) compared to the control group. Additionally, the improvement in the overall ability to provide treatment to patients was significantly higher in the experimental group than in the control group (mean 7.23, SD 2.41 vs mean 5.72, SD 2.19; $F_{1, 59}$=6.42; Cohen *d*=0.63; 95% CI 5.85-7.01; *P*=.01). The medium-large effect size (*d*=0.75) suggests that VSP training is not only statistically significant but also practically important in improving clinical reasoning. However, the wide CI (0.55-1.17) for Cohen *d* suggests that there is uncertainty about the true effect size and further validation is needed.

Questionnaire

After the training, 2 groups of clinical medicine students who participated in the training were invited to conduct a satisfaction survey. As shown in Table 2, 29 of 30 medical students (96.7%) in the experimental group believe that VSP training is beneficial for improving their ability to collect medical history and analyze medical conditions. The experimental group of 27 of 30 medical students (90%) believes that VSP training is beneficial for improving diagnostic and differential diagnostic abilities, clinical treatment abilities, and clinical thinking abilities.

Table . Results of the learning effectiveness questionnaire.

Questionnaire content	Control group, n	VSP[a] group, n	Chi-square (*df*)	*P* value
Beneficial for improving the ability to collect medical history			31.86 (59)	<.001
Strongly agree	5	22		
Agree	11	7		
Neutral	12	1		
Disagree	2	0		
Beneficial for improving the ability to analyze the condition			23.17 (59)	<.001
Strongly agree	12	27		
Agree	8	2		
Neutral	6	1		
Disagree	4	0		
Beneficial for improving diagnostic and differential diagnostic capabilities			22.35 (59)	＜.001
Strongly agree	7	23		
Agree	13	4		
Neutral	8	2		
Disagree	2	1		
Beneficial for improving clinical treatment capabilities			20.19 (59)	=.009
Strongly agree	10	19		
Agree	7	8		
Neutral	10	2		
Disagree	3	1		
Beneficial for improving clinical thinking ability			20.56 (59)	＜.001
Strongly agree	4	20		
Agree	11	7		
Neutral	9	2		
Disagree	6	1		

[a]VSP: virtual standardized patient.

Discussion

Principal Findings

This study found that VSPs can consolidate the basic knowledge of internal medicine among medical students, improve their clinical thinking ability, and enhance their ability to work independently. Clinical reasoning ability is one of the essential core competencies for medical students, requiring not only a solid theoretical foundation but also continuous accumulation through clinical practice by interacting with patients. Unlike clinical skills, which can be improved through repeated training on models, the VSPs in the “VSPs in general practice” interactive diagnostic and therapeutic teaching system are characterized by immersion, interactivity, and conceptualization [16]. The system establishes a simulated human model based on real clinical cases, incorporating various multimedia elements, such as images, sounds, and text to realistically recreate clinical environments. It allows real-time interaction through natural methods like vision and touch, simulating the clinical diagnosis and treatment processes. The system enables virtual examinations and treatments, including medical interviews, electrocardiogram examinations, physical examinations, laboratory tests, imaging studies, clinical interventions, medication administration, and device-based interventions [17]. All examinations performed by trainees on the virtual patients automatically provide feedback on the results. The VSPs present the patient’s condition and physical signs through lifelike 3D imaging, creating an immersive experience. Through an open diagnostic and therapeutic model, the system fully cultivates trainees’ ability to identify and solve problems, fostering clear, rigorous, and efficient thinking patterns. A comprehensive assessment system objectively evaluates the trainees’ clinical reasoning logic and mastery of basic skills. Under the premise of ensuring medical safety, it provides trainees with practice opportunities that closely resemble real clinical scenarios [15,18].

With the advancement of technology, VSPs have gradually become an important tool in medical education. VSPs are computer-based and AI-driven simulation tools designed to cultivate clinical reasoning abilities in medical students. By

simulating real clinical cases and providing an interactive learning environment, VSPs help medical students improve their diagnostic, therapeutic, and communication skills [7,19,20]. This study suggests that, compared to traditional clinical training, VSP-based training may enhance medical students' consolidation of theoretical knowledge and improve their abilities in clinical history-taking, physical and auxiliary examinations, diagnosis and differential diagnosis, as well as treatment planning. The findings indicate that virtual patient systems could contribute to reforming medical education curricula and support the goal of cultivating clinically competent professionals. VSPs offer several potential advantages: they can be reused indefinitely, allowing students to practice repeatedly in various scenarios without being influenced by the emotions or physical conditions of real patients, thus providing a safe practice environment. Additionally, VSPs may offer personalized feedback and guidance based on students' learning progress and performance. While the initial development costs are high, VSPs could, in the long term, reduce the overall costs of medical education.

This RCT compared the teaching effectiveness of virtual training and traditional academic training for more than 4 weeks. The results suggest that students experienced less stress in the simulated clinic environment, which may have allowed them to dedicate more time to learning. Through repeated practice in a safe setting, they appeared to become more familiar with diagnostic and therapeutic processes. By the midterm assessment, students in the VSP group showed improvement in comprehensive abilities such as medical interviewing and clinical judgment compared to the control group. However, further research is needed to validate these findings across diverse educational settings and larger cohorts. Summative evaluation results also indicated that the VSP simulation system might enhance students' theoretical knowledge, medical interview skills, syndrome differentiation, and treatment capabilities. Compared to traditional academic training, the use of standardized patient teaching appeared to improve students' interpersonal communication skills, potentially helping them establish more harmonious doctor-patient relationships in future practice. While these findings are promising, additional studies are required to assess the long-term impact of VSP training on clinical performance and patient outcomes. Despite some technical and methodological challenges, ongoing advancements in virtual simulation technology are expected to further refine VSPs, potentially making them a valuable tool in medical education. However, their full integration into curricula will depend on continued validation and adaptation to different learning environments.

Limitations

Content and Design Limitations

First, the VSP system primarily focuses on textbook-style disease presentations, potentially neglecting case variations and atypical scenarios encountered in real clinical practice. Additionally, the predefined decision-tree branches may restrict students' exploration of alternative diagnostic hypotheses, limiting the simulation's flexibility. Future research should enhance VSPs' realism and adaptability to better mimic clinical complexity.

Methodological Constraints

This study relied on self-reported data and lacked objective external evaluation, which may introduce instructor-dependent variability. While the sample size was adequate for detecting primary outcome differences, it limited subgroup analyses. Furthermore, the control group's nonstandardized patient pool, though intentionally designed to reflect real-world clinical training environments, could lead to confounding due to uneven case exposure. Multicenter studies with larger, standardized samples are needed to validate these findings.

Assessment and Long-Term Validity

The VSP-embedded assessment provided real-time metrics, but its alignment with the intervention might introduce measurement bias. To address this, we cross-validated results with standardized exams and faculty evaluations; however, future studies should incorporate third-party assessment tools. Moreover, while the study demonstrated significant short-term improvements in clinical knowledge and skills, the lack of long-term follow-up and direct validation with real patient interactions leaves the clinical relevance of these gains uncertain. Structured evaluations in real-world settings are warranted to confirm the predictive validity of VSP training.

Conclusions

This RCT demonstrates the feasibility of VSPs in improving the clinical competence of medical students. VSPs deserve more attention and promotion. Therefore, future research aims to improve the VSPs system to more effectively enhance the clinical thinking ability of medical students. Future studies should use a hybrid evaluation approach, where trainees first complete VSP training and are then assessed with standardized real patient encounters, to establish the translational validity of virtual training to clinical practice.

Funding

This study was funded by the Teaching Reform Project JG2022153 of Wenzhou Medical University, Quzhou Science and Technology Plan Project (2022015), and research projects of Quzhou People's Hospital in 2022 (YNB21 and YNB19).

Authors' Contributions

LX had full access to all the data in the study and took accountability for data integrity and the accuracy of data analysis. LX, QX, and CL conceptualized and designed the study. BC obtained and analyzed data. LX drafted and revised the article. CW obtained funding.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Visual abstract.
[PNG File, 526 KB - mededu_v11i1e73196_app1.png]

Checklist 1
CONSORT checklist.
[DOCX File, 21 KB - mededu_v11i1e73196_app2.docx]

References

1. Duvivier RJ, Boulet JR, Opalek A, van Zanten M, Norcini J. Overview of the world's medical schools: an update. Med Educ 2014 Sep;48(9):860-869. [doi: 10.1111/medu.12499] [Medline: 25113113]
2. Huang M, Yang H, Guo J, et al. Faculty standardized patients versus traditional teaching method to improve clinical competence among traditional Chinese medicine students: a prospective randomized controlled trial. BMC Med Educ 2024 Jul 24;24(1):793. [doi: 10.1186/s12909-024-05779-3] [Medline: 39049066]
3. Prince K, van Eijs P, Boshuizen HPA, van der Vleuten CPM, Scherpbier A. General competencies of problem-based learning (PBL) and non-PBL graduates. Med Educ 2005 Apr;39(4):394-401. [doi: 10.1111/j.1365-2929.2005.02107.x] [Medline: 15813762]
4. Dolmans D, De Grave W, Wolfhagen I, van der Vleuten CPM. Problem-based learning: future challenges for educational practice and research. Med Educ 2005 Jul;39(7):732-741. [doi: 10.1111/j.1365-2929.2005.02205.x] [Medline: 15960794]
5. Talbot T, Rizzo AS. Virtual human standardized patients for clinical training. In: Virtual Reality for Psychological and Neurocognitive Interventions: Springer; 2019:387-405. [doi: 10.1007/978-1-4939-9482-3_17]
6. Reger GM, Norr AM, Gramlich MA, Buchman JM. Virtual standardized patients for mental health education. Curr Psychiatry Rep 2021 Jul 15;23(9):57. [doi: 10.1007/s11920-021-01273-5] [Medline: 34268633]
7. Yang H, Xiao X, Wu X, et al. Virtual standardized patients versus traditional academic training for improving clinical competence among traditional chinese medicine students: prospective randomized controlled trial. J Med Internet Res 2023 Sep 20;25:e43763. [doi: 10.2196/43763] [Medline: 37728989]
8. Maicher KR, Zimmerman L, Wilcox B, et al. Using virtual standardized patients to accurately assess information gathering skills in medical students. Med Teach 2019 Sep;41(9):1053-1059. [doi: 10.1080/0142159X.2019.1616683] [Medline: 31230496]
9. Maicher KR, Stiff A, Scholl M, et al. Artificial intelligence in virtual standardized patients: combining natural language understanding and rule based dialogue management to improve conversational fidelity. Med Teach 2022 Nov 8;45(3):1-7. [doi: 10.1080/0142159X.2022.2130216] [Medline: 36346810]
10. Malik TG, Mahboob U, Khan RA, Alam R. Virtual patients versus standardized patients for improving clinical reasoning skills in ophthalmology residents. a randomized controlled trial. BMC Med Educ 2024 Apr 22;24(1):429. [doi: 10.1186/s12909-024-05241-4] [Medline: 38649884]
11. Kononowicz AA, Zary N, Edelbring S, Corral J, Hege I. Virtual patients--what are we talking about? A framework to classify the meanings of the term in healthcare education. BMC Med Educ 2015 Feb 1;15(1):11. [doi: 10.1186/s12909-015-0296-3] [Medline: 25638167]
12. Thistlethwaite JE, Davies D, Ekeocha S, et al. The effectiveness of case-based learning in health professional education. A BEME systematic review: BEME guide no. 23. Med Teach 2012;34(6):e421-e444. [doi: 10.3109/0142159X.2012.680939] [Medline: 22578051]
13. Cook DA, Erwin PJ, Triola MM. Computerized virtual patients in health professions education: a systematic review and meta-analysis. Acad Med 2010 Oct;85(10):1589-1602. [doi: 10.1097/ACM.0b013e3181edfe13] [Medline: 20703150]
14. Samson J, Gilbey M, Taylor N, Kneafsey R. Virtual simulated placements in health care education: scoping review. JMIR Med Educ 2025 Jun 10;11:e58794. [doi: 10.2196/58794] [Medline: 40548423]
15. Wengenroth L, Hege I, Förderreuther K, et al. Promoting occupational health in secondary schools through virtual patients. Comput Educ 2010 Dec;55(4):1443-1448. [doi: 10.1016/j.compedu.2010.06.007]
16. Zary N, Johnson G, Boberg J, Fors UGH. Development, implementation and pilot evaluation of a web-based virtual patient case simulation environment--Web-SP. BMC Med Educ 2006 Feb 21;6(1):10. [doi: 10.1186/1472-6920-6-10] [Medline: 16504041]

17. Kononowicz AA, Zary N, Edelbring S, Corral J, Hege I. Virtual patients--what are we talking about? A framework to classify the meanings of the term in healthcare education. BMC Med Educ 2015 Feb 1;15:11. [doi: 10.1186/s12909-015-0296-3] [Medline: 25638167]
18. Lewis KL, Bohnert CA, Gammon WL, et al. The association of standardized patient educators (ASPE) standards of best practice (SOBP). Adv Simul (Lond) 2017;2:10. [doi: 10.1186/s41077-017-0043-4] [Medline: 29450011]
19. Bond WF, Mischler MJ, Lynch TJ, et al. The use of virtual standardized patients for practice in high value care. Simul Healthc 2023 Jun 1;18(3):147-154. [doi: 10.1097/SIH.0000000000000659] [Medline: 35322798]
20. Consorti F, Mancuso R, Nocioni M, Piccolo A. Efficacy of virtual patients in medical education: a meta-analysis of randomized studies. Comput Educ 2012 Nov;59(3):1001-1008. [doi: 10.1016/j.compedu.2012.04.017]

Abbreviations

AI: artificial intelligence
CONSORT: Consolidated Standards of Reporting Trials
RCT: randomized controlled trial
VSP: virtual standardized patient

Edited by B Lesselroth; submitted 27.02.25; peer-reviewed by A Rubio-López, N Misra, R Singh, S Ajayi; revised version received 21.07.25; accepted 25.09.25; published 08.12.25.

Please cite as:
Xu L, Xu Q, Liu C, Chen B, Wang C
Virtual Standardized Patients for Improving Clinical Thinking Ability Training in Residents: Randomized Controlled Trial
JMIR Med Educ 2025;11:e73196
URL: https://mededu.jmir.org/2025/1/e73196
doi:10.2196/73196

Exploring the Role of Immersive Virtual Reality Simulation in Health Professions Education: Thematic Analysis

Jordan Talan, MHPE, MD; Molly Forster, MD; Leian Joseph, MD; Deepak Pradhan, MHPE, MD

Division of Pulmonary, Critical Care, & Sleep Medicine, Department of Medicine, NYU Grossman School of Medicine, 550 First Avenue, 15th Floor, Medical ICU, New York, NY, United States

Corresponding Author:
Jordan Talan, MHPE, MD
Division of Pulmonary, Critical Care, & Sleep Medicine, Department of Medicine, NYU Grossman School of Medicine, 550 First Avenue, 15th Floor, Medical ICU, New York, NY, United States

Abstract

Background: Although technology is rapidly advancing in immersive virtual reality (VR) simulation, there is a paucity of literature to guide its implementation into health professions education, and there are no described best practices for the development of this evolving technology.

Objective: We conducted a qualitative study using semistructured interviews with early adopters of immersive VR simulation technology to investigate use and motivations behind using this technology in educational practice, and to identify the educational needs that this technology can address.

Methods: We conducted 16 interviews with VR early adopters. Data were analyzed via directed content analysis through the lens of the Unified Theory of Acceptance and Use of Technology.

Results: The main themes that emerged included focus on cognitive skills, access to education, resource investment, and balancing immersion. These findings help to clarify the intended role of VR simulation in health professions education. Based on our data, we synthesized a set of research questions that may help define best practices for future VR development and implementation.

Conclusions: Immersive VR simulation technology primarily serves to teach cognitive skills, expand access to educational experiences, act as a collaborative repository of widely relevant and diverse simulation scenarios, and foster learning through deep immersion. By applying the Unified Theory of Acceptance and Use of Technology theoretical framework to the context of VR simulation, we not only collected validation evidence for this established theory, but also proposed several modifications to better explain use behavior in this specific setting.

(JMIR Med Educ 2025;11:e62803) doi:10.2196/62803

KEYWORDS

virtual reality; medical education; virtual reality simulation; extended reality; simulation; VR; health professions education; health education; thematic analysis; evolving technology; qualitative study; qualitative; semistructured interviews; educational experiences; theoretical framework

Introduction

Background

As technology rapidly advances in immersive virtual reality (VR) simulation, there is a growing interest among educators to develop VR simulation curricula for health professions education. However, there is a paucity of literature to guide these efforts, and there are no accepted best practices for the development or implementation of this technology. While experts anticipate the potential for VR to transform medical education [1], without a better understanding of the role VR will play in our training programs, these statements may amount to nothing more than vague future promises. Therefore, characterization of the early use of VR is imperative to clarify its evolving role and gain insights that will allow us to implement this technology to its fullest potential.

VR Simulation Technology

Immersive VR creates a simulated environment, allowing users to “step inside” a computer-generated world and engage authentically with their surroundings [1]. VR offers several potential benefits for health professions education, including facilitating distance learning and providing training that is difficult to deliver via traditional simulation [2]. In addition, VR shows comparable educational outcomes to high-fidelity mannequin simulation with more cost-effectiveness [3-7]. Many institutions are enthusiastic about VR simulation and are already piloting or studying VR curricula [1,8] However, there is still

much to learn in order to best guide the development and implementation of these curricula.

While prior research has concentrated on individual VR usage-scenarios or software evaluations [9,10], effective educational interventions require a broader understanding of the context of our learners [11]. Therefore, we must study VR user needs across a wider spectrum to guide development that aligns with the context of health professions training. By analyzing current VR educational practices, we can better identify the gaps that this technology can bridge, and move toward a consensus about how best to use VR simulation in the future. Without a better understanding of these gaps, we risk pouring resources into technology for technology's sake—a solution looking for a problem [12].

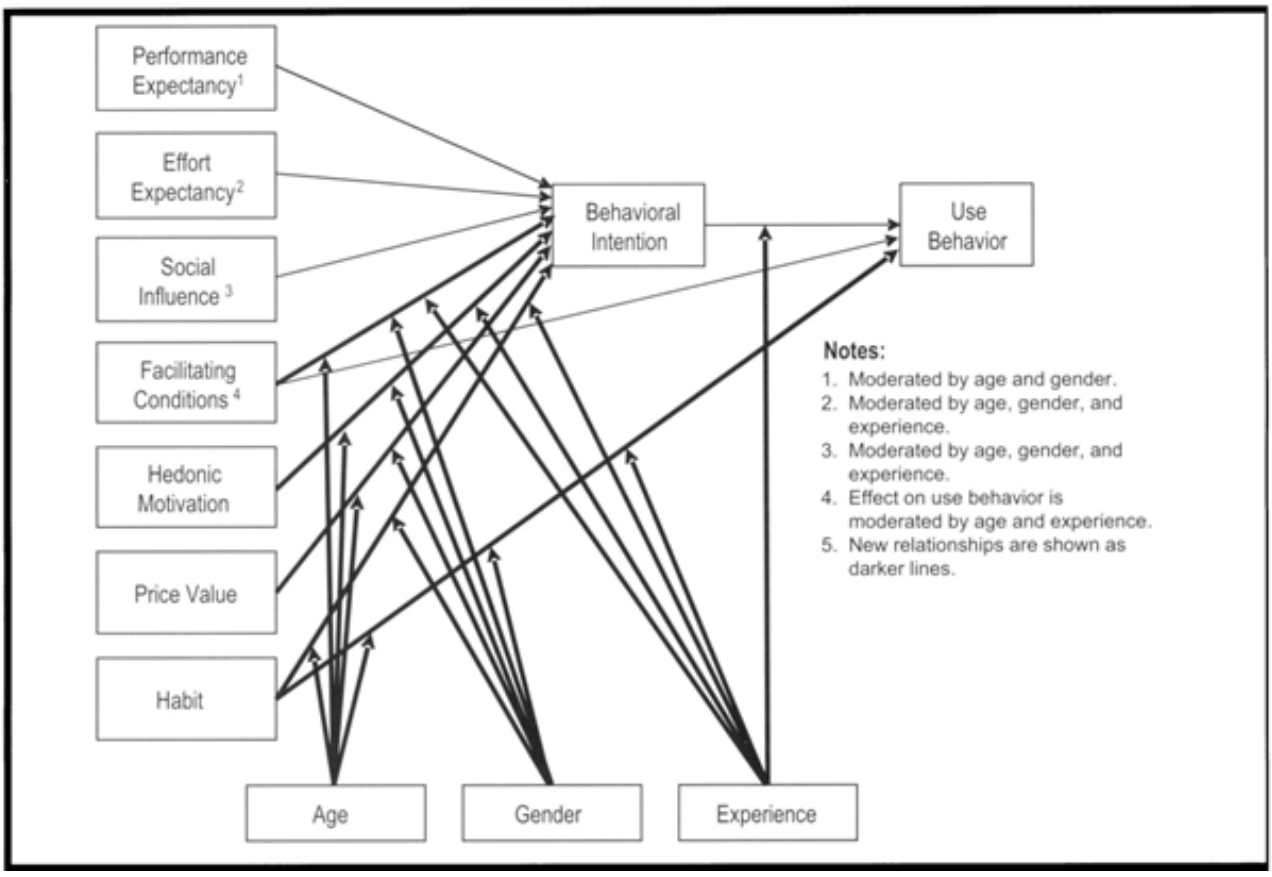
Study of Early Adopters

The technology adoption life cycle categorizes users into 5 groups based on their likelihood to adopt new technology: innovators, early adopters, early majority, late majority, and laggards [13]. Our study focuses on early adopters, as they represent educational stakeholders pioneering the implementation of VR within authentic educational environments and collaborating with VR innovators to adapt the technology to their needs. They therefore have expertise evaluating VR technology, but unlike the innovators, their experience is more practical than theoretical.

Unified Theory of Acceptance and Use of Technology

The Unified Theory of Acceptance and Use of Technology (UTAUT) explains factors that affect the adoption of new technologies and predicts future technology use [14]. UTAUT provides a robust theoretical framework for understanding the drivers incentivizing early adopters to embrace VR as an educational strategy. The original theory described 4 constructs as direct determinants of technology usage behavior (Figure 1): performance expectancy (user expectation that the technology improves performance), effort expectancy (ease associated with using the technology), social influence (user perception that others believe they should be using the technology), and facilitating conditions (organizational and technological infrastructure for technology implementation). These determinants are modified to varying degrees by user gender, age, or experience [14]. Extensively applied across multiple fields for assessing new technologies [15], the UTAUT was expanded to the Unified Theory of Acceptance and Use of Technology 2 (UTAUT2) with 3 additional constructs: hedonic motivation (pleasure derived from using the technology), price value (perceived cost of the new technology), and habit (degree of automatic use of the technology) [16].

Figure 1. The UTAUT2 from Venkatesh et al [16], used with permission. UTAUT2: Unified Theory of Acceptance and Use of Technology 2.



With validity evidence across multiple fields since 2003, UTAUT has become one of the most developed and intensive models to test new technologies [15]. In addition to being well-validated in multiple settings, UTAUT is an ideal theoretical framework to explore and better understand thoughts and behaviors associated with the use of VR simulation technology. While other theoretical frameworks have been used to approach VR simulation research, few will allow the isolation of factors associated with VR specifically rather than those that apply to simulation in general. Even fewer might facilitate the

prediction of its use in the future. For example, constructivist learning theory has been applied to VR simulation because learners can manipulate a problem and construct learning from active participation in an engaging experience [17]. Experiential learning has also been used to contextualize VR simulation because it provides a safe and forgiving training environment that facilitates learning by doing [3]. However, these theoretical frameworks serve better to characterize simulation in general rather than to focus on the specific experience provided by VR technology. The use of UTAUT2 on the other hand provides a structured framework by which we can distinguish the features of VR technology from other modes of simulation, and by which we can attempt to predict its future use.

Prior VR research using UTAUT2 focuses primarily on understanding learner experience and learner acceptance [18,19]. While these concepts are critical for the successful adoption of evolving technological innovations [20], we must progress further by investigating how VR simulation can address specific educational needs and gathering validation evidence for its most effective future role in the evolving landscape of health professions education.

Study Aims

To fill this gap in our understanding, this study interviews early adopters of VR simulation in health professions education, with the following aims: (1) characterize how early adopters are adapting VR to meet their educational needs, (2) define the educational problems or gaps that early adopters are trying to address with VR, and (3) explore factors influencing the ability of early adopters to meet their needs with VR.

Methods

Ethical Considerations

This study was approved by the New York University Langone Institutional Review Board (22‐01346). Informed consent was obtained from all study participants, and participants had the ability to opt out or withdraw from the study at any time. Interview transcripts were deidentified for confidentiality. Participants were not compensated.

Study Design

This is a qualitative study using thematic analysis of semistructured interviews. The research methodology is directed content analysis, starting with a limited code book of 7 a priori codes defined through the lens of the UTAUT2 theoretical framework, followed by an exploratory coding phase [21,22]. The research paradigm is postpositivist. Reporting was completed following the Standards for Reporting Qualitative Research guidelines [23].

Semistructured Interview Guide

We iteratively developed a semistructured interview guide based on our research questions and grounded in the UTAUT2 theoretical framework [24,25] (Multimedia Appendix 1). The interview guide was piloted with local stakeholders to ensure capture of meaningful data within the 45-minute interview timeframe.

Recruitment and Sampling

We recruited educational stakeholders who were identified as "early adopters" of immersive VR simulation technology. Inclusion criteria were experience educating, implementing, or researching with VR. Exclusion criteria included technology developers without educational practice experience, and participants with experience limited to 360° video, augmented reality, or nonsimulation immersive learning.

The first 3 participants were recruited as a convenience sample, as they were known to our research team based on their work with the American College of Chest Physicians to develop and pilot an immersive VR simulation program teaching endotracheal intubation. These participants were recruited as an entry into the community of VR early adopters, with subsequent recruitment by snowball sampling. We sought to map the terrain of VR use-cases by recruiting for maximal diversity. We asked if participants could identify additional early adopters who had different experiences (ie, worked with a different company, in a different learner setting, at a different institution, or who had differing perspectives on VR technology). We estimated a sample size of 12‐18 interviews. Data were iteratively analyzed for thematic saturation, and recruitment was terminated upon achieving saturation of meaning [26].

Interviews and Data Analysis

Each participant completed a 45-minute semistructured interview via Zoom videoconferencing. Interviews were audio-recorded and transcribed verbatim into a written document via Speechmatics software with manual verification. Transcripts were imported to ATLAS.ti (ATLAS.ti Scientific Software Development GmbH) web, which was used for iterative qualitative data coding and analysis. First-round coding was performed via an a priori coding template corresponding to the UTAUT2 domains. Any additional codes used process coding and descriptive coding. All codes were approved by 2 independent reviewers (JT and DP) with deliberation over any discrepancies. Second-round coding then checked all codes against the initial coding template, collapsing as necessary to capture any new domains not described by the UTAUT2 framework. Field notes and memos were maintained by both reviewers. Themes were identified and their interrelationships characterized [27]. Themes were then shared with study participants via member checking to ensure the accuracy of our analyses.

Reflexivity

JT and DP are Pulmonary/Critical Care Medicine physicians. JT has worked with technology companies and educational technologists researching immersive VR simulation, but is relatively suspicious of new technology unless it fulfills a specific need. DP is also an early adopter, who is a self-described "gamer" and owns a VR headset for recreational use. JT, DP, and MF are simulation educators at New York University. All authors kept memos to practice reflexivity throughout this study's period.

Results

Overview

We completed 16 semistructured interviews. Coding saturation occurred after 11 interviews and thematic saturation after 12 interviews. Four additional interviews were completed to ensure saturation of meaning [26]. Participant demographics are described in Table 1. Our study population included early adopters from diverse health professions whose educational interventions targeted the following groups of learners: physician trainees (premedical students, medical students, residents, and fellows), advanced practice providers (nurse practitioners and physician assistants), nurses and nursing students, respiratory therapists, pharmacists, and emergency service members (emergency medical technician students and paramedical students).

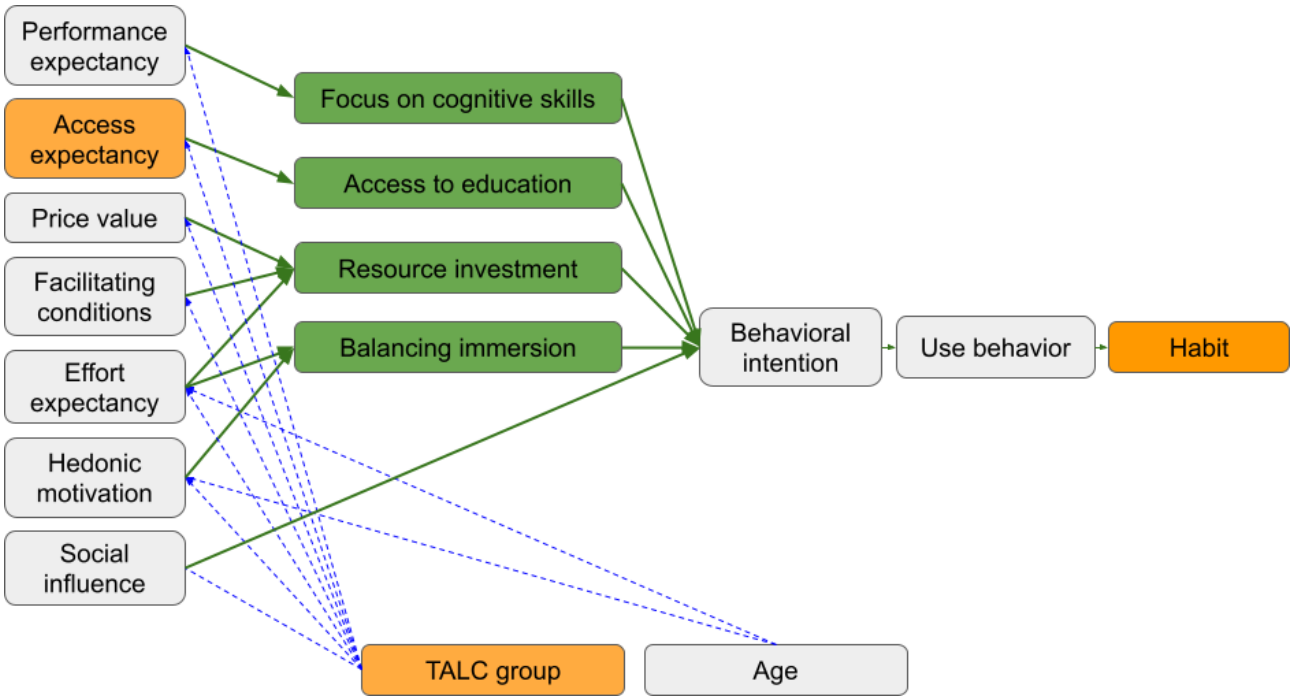
Table . Demographics of interview participants (N=16).

Demographics of interview participants	Values (n)
Gender	
Male	10
Female	6
Age (years)	
31 - 50	8
51 - 65	3
>65	3
Unknown	2
Technology adoption life cycle group	
Innovator	6
Early adopter	6
Early majority	3
Unknown	1
Geography	
Northeast (United States)	3
Midwest (United States)	5
South (United States)	4
West (United States)	3
Canada	1
Setting	
Urban	12
Suburban	4
Rural	0
Health profession	
Advanced practice provider (nurse practitioner or physician assistant)	1
Emergency medical service (paramedics or emergency medical technicians)	1
Health care education technologist	1
Nurse	3
Physician	8
Anesthesiology	2
Cardiology (pediatrics)	1
Emergency medicine (adult)	2
Armed forces	1
Emergency medicine (pediatrics)	1
Internal medicine	1
Pulmonary and critical care medicine	2
Respiratory therapist	1

Coding and Themes

First-round coding generated 38 unique codes: 7 from the a priori coding template corresponding to UTAUT domains, and 31 new codes via process and descriptive coding. Four themes were identified and examined for their interrelationships. The resulting synthesis and validation evidence for the UTAUT2 framework are depicted in a thematic map (Figure 2).

Figure 2. Thematic map of results. The thematic map illustrates our results and changes we have made to the theoretical framework. Themes (green boxes) are superimposed between each UTAUT construct and the resulting behavioral intention. The new addition of access expectancy is highlighted in orange. Green arrows illustrate which theme most strongly relates to which UTAUT construct. Blue arrows represent effect modifying relationships. TALC: technology adoption lifecycle; UTAUT: Unified Theory of Acceptance and Use of Technology.



Theme 1: Focus on Cognitive Skills

Study participants focused on VR simulation for the development of cognitive skills, including communication, teamwork, clinical reasoning, situational awareness, and interdisciplinary skills, occasionally referred to as "soft skills."

> *It's really just about talking to each other, right? And sharing that mental model…I think that's where you can really benefit from VR because it's not really about the tasks you're doing. It's how are you communicating…I think if you look at a lot of the sentinel events or the near misses that happen, it's based on communication.* [Participant #4]

For procedural skills, the role of VR simulation was limited to building procedural knowledge or situational awareness.

> *It does help you remember the different steps. You know, don't forget the suction at the head of the bed or, you know, [we will] have the patient vomit…and maybe they won't forget it now.* [Participant #4]

However, participants found the teaching of fine psychomotor skills, such as laryngoscopy or peripheral intravenous placement, to be limited in VR.

> *We knew that you cannot teach the fine motor skills of intubating a patient in virtual reality. It is very difficult to do with the kind of tools that are available right now. And so it was more about the thinking…I truly believe that the mental process of approaching an airway is just as important, if not more important, than the fine tuning of technical skills.* [Participant #1]

The most common barrier to teaching psychomotor skills in the VR environment was haptic technology, noted unanimously by every interview participant.

> *What it doesn't do well? One is teaching people how to do things that require them to use their hands and fine motor skills, even just like how to use tools. It's really challenging to teach somebody how to hold a laryngoscope, how to hold the endotracheal tube…I think the big limiting factor is the fact that you have to use controllers because the controllers only work a certain way. You must hold it this way. These are the few buttons that you have. You're not using your hands the way you would in real life.* [Participant #1]

In addition to limitations in simulating authentic tools, participants noted limitations in simulating the weight or feel of human anatomy necessary to learn fine psychomotor skills.

> *You put the [laryngoscope] blade in their mouth, the vocal cords show up on the screen and you just drop the tube and it just clicks right in. But right now, we don't have…that feedback where you feel your scope in your hand or you feel the weight of the jaw when you're going to lift up.* [Participant #8]

Many participants discussed the development of haptic gloves which can allow users to simulate touch and contact experiences. However, most found current solutions either cost-prohibitive or inadequate.

> *I don't see that type of fine motor feedback, you know, where I know how to put the needle into an arm for an IV, for example. That's not going to happen for - I would say that's decades away. Easily. I don't think there are good solutions right now in even the most*

> *expensive labs with experimental haptics for that.* [Participant #16]

Therefore, to teach psychomotor and procedural skills, participants turned to other forms of simulation technology such as mixed reality or mannequin simulation. Several participants offered learners a blended experience, using VR to create an immersive scenario, followed by a task trainer to simulate any necessary fine motor tasks.

Theme 2: Access to Education

The ability of VR simulation to facilitate distance learning was seen as a significant driver of use behavior for most participants.

> *We wanted to break down the barriers of requiring learners to physically come to a place to get this type of education. We want this education to be deliverable over long distances to people in other parts of the world.* [Participant #1]

In areas without the resources for a high-fidelity mannequin-based simulation laboratory, VR was seen as expanding access to high-quality simulation learning from expert educators.

> *Places that don't have simulation labs and all of those resources…available at academic medical centers, at [professional society] headquarters…But outside of large centers or hospitals that have access to a sim lab - and I think probably the majority of hospitals in the country do not - those hospitals don't have access to that type of education.* [Participant #1]

VR simulation was also used as a solution to reduce the cost, inconvenience, and sometimes danger associated with traveling to simulation centers.

> *The ability to do remote simulation at a much lower cost than requiring travel, that's a huge benefit…If you've got employees spread across the country, even across the state - and I'll use Wyoming, for example… everything's 8 hours away. It's icy half the year. So if…you've got students all over the state that are part of your paramedic program, and you have these guys driving throughout the winter to come to your simulation center. Like, what are the chances of something bad happening there? Pretty high to be honest.* [Participant #8]

Even in centers with existing high-fidelity simulation laboratories, participants found a role for VR in facilitating collaboration and standard-setting at a national or international level.

> *As people adopt these headsets…somebody from [University] could do the same ACLS training as somebody at [University] and it would be the same across institutions. And there's crosstalk - so different learning points and different perspectives and shared information and shared values in terms of education.* [Participant #6]

Finally, VR simulation also expanded access to education during the COVID-19 pandemic, responding to the need for social distancing, and significantly accelerating VR adoption.

> *If heaven forbid there was another pandemic, now we are set up that. If our students were at home on lockdown, as long as they had a headset, their learning would not be interrupted.* [Participant #13]

Overall, VR afforded a distance learning advantage, providing equitable access to high-quality simulation education to centers in diverse settings and learners in adverse scenarios.

Theme 3: Resource Investment

Implementation of a VR simulation curriculum required extensive resources, particularly upfront costs (funding and time commitments).

> *There is a capital purchase that has to be made just for the equipment itself. But…how do you develop that program in a manner that somebody is not spending tons and tons of time to bring one little educational module to fruition.* [Participant #3]

These upfront costs also related to the process of cocreation with technology companies.

> *There was generally some frustration during the build process because we're all clinicians and we're like, 'yeah, this thing needs to be this way'. And you're trying to communicate that with someone who has no medical experience and is a software programmer… We speak one language and they speak a different language, and there was some inability to communicate that effectively.* [Participant #3]

Once the programming was complete, participants also described an ongoing cost to maintain the software through updates or licensing.

> *Keeping these things alive is really…the cost to maintain software…for servers and engineers and updates and things like that. So without some sort of continued funding from somewhere, it will become a useless pile of code as soon as the next [operating system] update hits.* [Participant #9]

The investment required to develop novel VR programming was frequently more resource-intensive than anticipated, and the risk of failed investment was wasted time and money.

> *I've seen many cases of projects that are developed and they're just abandoned…I would walk into my office every morning and I had a stack of 16 boxes of headsets we didn't use.* [Participant #10]

Therefore, participants wanted more opportunities for creative collaboration and sharing of software programs. However, some felt limited by the current state of technology.

> *There's no great way to share content yet. So a lot of stuff is just getting reinvented over and over again, which is a really expensive way to do things.* [Participant #12]

Others felt restricted by the current incentives within the VR marketplace, concerned by compatibility between different software or hardware companies.

> *There needs to be an ability for me to use multiple vendors within my one headset without having to pay*

> *millions of dollars to do so…I don't know about everybody's budget. On my budget, I cannot afford to pay four different guys for completely different programs.* [Participant #8]

Generally, participants desired to use pre-existing software that was universally relevant for multiple institutions and multiple users, and compatible with a variety of hardware.

Theme 4: Balancing Immersion

The immersiveness of VR was a powerful experience associated with learner enjoyment.

> *Being in virtual reality is an immersive experience, and it's just hard to describe in words until you try it. But when people try it, it's like seeing a new color.* [Participant #1]

At its best, immersion increased learner presence, stimulated intellectual curiosity, and accelerated learning.

> *When you go in and you see an environment in 3D that looks exactly like your cardiac ICU…you immediately have a 'wow' thing. And what I love about that is immediately when I start this scenario, I never really hear like, 'Wait, what do you want me to do? Are we starting now? Is the patient supposed to have pulses?'…It's so immersive that people immediately feel like they're in a football game and it's kickoff.* [Participant #7]

Participants also valued VR immersion for minimizing distractions more than other simulation technologies.

> *The thing that's nice about virtual reality is you put the headset on and that's what you're doing, right? So you're not looking at your phone or checking your email while someone's trying to teach you.* [Participant #3]

However, immersion could also create extraneous cognitive load, detracting from learning. Participants described unnecessary environmental elements that distracted from learning objectives, along with some tasks that were frustrating to simulate in VR.

> *The picking up of items in the ICU was difficult always…With the limited controller toggles, it was not always intuitive how to pick something up. And even when they told you what to do, it still sometimes fell on the floor and stuff like that.* [Participant #2]

Sometimes immersive scenarios became more about navigating the VR environment than mastering intended learning objectives.

> *You're never going to drop some instrument on the tray 6 times…Like is the goal to learn to pick the scope up, or to [learn the procedure]?…I think a lot of people try and make the virtual world exactly like the real world…but I think you have to simplify the haptics…If it's just so frustrating because you're an intensivist and you can't pick up the needle drivers, then forget it. There shouldn't be a five minute learning curve on how to pick up needle drivers, right?* [Participant #12]

Participants found an ideal immersive balance when the virtual world accomplished the intended learning objectives, but was not overly complex to create frustration in navigating the environment. In this way, there was constructive alignment between the intended learning outcomes and the virtual learning activities.

Validation Evidence for the UTAUT2 Framework

Codes were confirmed for each previously described construct within the UTAUT2 framework [16]. Performance expectancy was the most frequently coded driver for use intention with VR technology. We also found age to modify the effect of certain constructs, with the younger generation more easily adapting to VR technology (effort expectancy) and demonstrating greater VR learning enjoyment (hedonic motivation).

> *I think the current generation of learners is…becoming more and more comfortable with virtual reality. So I think the buy-in of our new generation of learners is going to be really quick…And so I think they're going to help drive the need for this type of education.* [Participant #5]

The UTAUT2 theoretical framework was able to explain patterns in use behavior and intention related to VR simulation, and provided conclusions relevant to educational practice. However, we modified the UTAUT2 model, most notably adding "access expectancy" as an independent driver of behavioral intention. The importance of distance learning, expanded access, and equity in educational experience was sufficient to qualify as an independent construct. To illustrate its relative importance, there were more instances of coding for access expectancy than hedonic motivation, habit, or social influence. This may reflect that the UTAUT was initially described in the individual consumer marketplace while "access expectancy" applies more to the context of the educational technology marketplace. Further research would be necessary to explore this hypothesis.

The other notable change was seen in the UTAUT modifier "gender." There were no instances of coding applicable to gender by either independent reviewer, and themes identified did not differ by participant gender. We found no signal for gender as a modifier of any UTAUT2 construct. We suggest that this is a reflection of both time and context. The initial publication of the UTAUT was in 2003, wherein it was discussed that effort expectancies may be more salient to women than men, and that women may be more sensitive to others' opinions than men [16]. We believe this contextualization of gender roles and social norms to be antiquated and due for revision. Furthermore, in this cohort of career medical educators, we found no differences in motivating factors between men and women related to their intention or use behavior with simulation technology. Therefore, we eliminated "gender" from our thematic map.

Discussion

Principal Findings

Participating early adopters have adapted their use of VR to meet specific educational needs. Whether it be the need for

distance learning during the pandemic, the need to bridge geographical or institutional divides, or the need for wide dissemination of teaching to address gaps in knowledge or skills, early adopters are implementing VR as a method to expand access to high-yield educational interventions. In terms of the role that VR served among the studied population, it was used primarily to teach cognitive skills as opposed to psychomotor or procedural skills. The most common factors that affected how successful any given implementation of VR would be is related to how educators managed their resources (funding, time, and design effort) and the degree to which they were able to foster learning through deep immersion.

The results of this study establish drivers of use behavior, providing practical insights into the educational gaps that VR might address in the future. Furthermore, this study contributes validity evidence for the UTAUT framework in studying the evolving role of immersive VR simulation in health professions education. This study represents an advancement to the literature in this field as it encompasses a wider variety of VR use cases than prior work, and it uses a well-validated theoretical framework to reflect on the perspectives of a diverse population within the health professions education community.

Implications for Future Research

Based on our findings, we synthesize a set of research questions that may help define best practices for future VR development and implementation (Table 2). We also list example study ideas corresponding to each research question in order to provide additional context and encourage reflection. These examples are not meant to be comprehensive or prescriptive, but rather to demonstrate how researchers might approach these questions with a variety of different methodologies and paradigms that could advance the literature in this field.

Table . Suggested research questions for immersive VR[a] simulation technology.

Themes from this study and suggested research questions		Example future research study
Focus on cognitive skills		
	How can we best implement immersive VR simulation given its strength in teaching cognitive skills (eg, communication, teamwork, clinical reasoning, situational awareness, and interdisciplinary skills)?	Multi-institutional study comparing 2 different VR implementation methods and using a validated assessment for cognitive skills
	What innovations can improve the teaching of fine psychomotor tasks in the VR environment?	Validation study using novel haptic gloves for VR simulation and assessing learning outcomes
Access to education		
	How can VR be most effectively leveraged to provide distance learning?	Mixed methods (quantitative or qualitative) needs assessment for distance simulation learning in post-COVID health professions education
	How can VR be used as a tool to create equity of educational experience?	Comparative study of learner outcomes at highly resourced centers versus resource-limited training programs for VR simulation
	How can VR facilitate collaboration on a larger scale (eg, national or international)?	Descriptive study demonstrating feasibility of an international VR curriculum offered by a professional society
Resource investment		
	What are the upfront investments and preparation processes necessary to start a new VR simulation program?	Focus group study of early adopters with concentration on preparation and upfront costs for establishing a VR simulation program
	How can we increase availability and decrease barriers for using pre-existing VR software programs?	Thematic analysis of focus groups after piloting a VR multi-case library targeting undergraduate medical education learners
	What processes facilitate the creation of novel VR software that is relevant to external users, institutions, and learner groups?	Systematic review and subsequent guideline development project to describe best practices in creation of VR curricula
Balancing immersion		
	How can we achieve sufficient immersion to accomplish intended learning objectives without creating extraneous cognitive load and frustrating part-tasks?	Comparative study of learning outcomes in a high-fidelity versus low-fidelity VR environment

[a]VR: virtual reality.

Much ongoing VR simulation research focuses on demonstrating that VR is equally or more effective than traditional simulation modalities [3-6,28]. While this is an important question, it risks overshadowing other questions that are raised by early adopters

in this study: how might we improve the ability of VR technology to teach psychomotor skills? How can we use VR simulation to create equity between learner populations? What solutions exist for shared and collaborative creation of VR software? How can we leverage the incentives of the marketplace for VR technology companies? These questions could significantly impact the future use of this technology in health professions education.

Study Limitations

This study has several limitations. First, early adopters tend to be optimistic about the advantages of new technologies, sometimes underemphasizing associated challenges. To account for this bias, we designed our interview guide with prompts targeted equally toward the advantages and challenges of VR technology, and we practiced reflexivity among this study's team to appreciate the effects of any personal biases. Future studies should examine perspectives from early majority, late majority, and laggards, but these groups are not yet readily identifiable.

Second, this study used nonprobability sampling, which harbors potential for bias toward participants with similar experiences and perspectives. However, the small size of the VR educator community limits the feasibility of random sampling. We therefore attempted to compensate by seeking participants with diverse experiences, working with different software applications, different VR companies, or different learner populations. To further assure accuracy and freedom from bias, future studies should attempt triangulation of this data, for example via data source triangulation using focus groups or via theory triangulation, analyzing this data through a different theoretical lens [29].

Third, while we targeted diversity, all health professions were not represented. Our sample included only individuals from the United States and Canada, and participants from rural workplace settings were underrepresented. These considerations may be important, particularly if geography is found to independently affect use behavior.

Finally, regarding the UTAUT modifier "experience," our sample size was inadequate to analyze its role as a modifier of use behavior. Therefore, we did not include experience in our thematic map and further research will be necessary to explore how experience may affect use behavior with VR simulation.

Conclusion

We used the UTAUT2 framework in a directed content analysis using semistructured interviews to investigate the role of immersive VR simulation in health professions education. We identified 4 key themes elucidating use behavior related to VR simulation, suggesting its optimal applications include teaching cognitive skills, expanding access to educational experiences, offering a collaborative repository of relevant simulation scenarios, and enhancing immersion for intended learning objectives. These themes may help to inform best practices for the future development and implementation of immersive VR simulation programs.

As immersive VR simulation technology continues to evolve in health professions education, the VR educator community will continue to grow alongside the rapid technological advancements. Therefore, defining best practices for integrating this technology into training programs is critical. Future research should focus on leveraging VR simulation's unique capabilities as compared to traditional simulation modalities.

Acknowledgments

The publication of this study was supported by a Fellowship Grant from the Stony Wold-Herbert Fund, Inc.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Semistructured interview guide.
[DOCX File, 22 KB - mededu_v11i1e62803_app1.docx]

References

1. Pottle J. Virtual reality and the transformation of medical education. Future Healthc J 2019 Oct;6(3):181-185. [doi: 10.7861/fhj.2019-0036] [Medline: 31660522]
2. Haowen J, Vimalesvaran S, Kyaw BM, Car LT. Virtual reality in medical students' education: a scoping review protocol. BMJ Open 2021 May 26;11(5):e046986. [doi: 10.1136/bmjopen-2020-046986] [Medline: 34039577]
3. Abulfaraj MM, Jeffers JM, Tackett S, Chang T. Virtual reality vs. high-fidelity mannequin-based simulation: a pilot randomized trial evaluating learner performance. Cureus 2021 Aug;13(8):e17091. [doi: 10.7759/cureus.17091] [Medline: 34527478]
4. Nassar AK, Al-Manaseer F, Knowlton LM, Tuma F. Virtual reality (VR) as a simulation modality for technical skills acquisition. Ann Med Surg (Lond) 2021 Nov;71:102945. [doi: 10.1016/j.amsu.2021.102945] [Medline: 34840738]
5. Khan R, Plahouras J, Johnston BC, Scaffidi MA, Grover SC, Walsh CM. Virtual reality simulation training in endoscopy: a Cochrane review and meta-analysis. Endoscopy 2019 Jul;51(7):653-664. [doi: 10.1055/a-0894-4400] [Medline: 31071757]

6. Berg H, Steinsbekk A. The effect of self-practicing systematic clinical observations in a multiplayer, immersive, interactive virtual reality application versus physical equipment: a randomized controlled trial. Adv Health Sci Educ Theory Pract 2021 May;26(2):667-682. [doi: 10.1007/s10459-020-10019-6] [Medline: 33511505]
7. Katz D, Shah R, Kim E, et al. Utilization of a voice-based virtual reality advanced cardiac life support team leader refresher: prospective observational study. J Med Internet Res 2020 Mar 12;22(3):e17425. [doi: 10.2196/17425] [Medline: 32163038]
8. Kyaw BM, Saxena N, Posadzki P, et al. Virtual reality for health professions education: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019 Jan 22;21(1):e12959. [doi: 10.2196/12959] [Medline: 30668519]
9. Saab MM, Hegarty J, Murphy D, Landers M. Incorporating virtual reality in nurse education: a qualitative study of nursing students' perspectives. Nurse Educ Today 2021 Oct;105:105045. [doi: 10.1016/j.nedt.2021.105045] [Medline: 34245956]
10. Wong M, Chue S, Jong M, Benny HWK, Zary N. Clinical instructors' perceptions of virtual reality in health professionals' cardiopulmonary resuscitation education. SAGE Open Med 2018;6:2050312118799602. [doi: 10.1177/2050312118799602] [Medline: 30245815]
11. Snell L, Son D, Onishi H. Instructional design. In: Swanwick T, Forrest K, O'Brien BC, editors. Understanding Medical Education: Evidence, Theory, and Practice: The Association for the Study of Medical Education (ASME); 2018:89-100. [doi: 10.1002/9781119373780]
12. Scalea JR. Technology for technology's sake no longer. Ann Surg 2019 Feb;269(2):e24. [doi: 10.1097/SLA.0000000000003132] [Medline: 30614843]
13. Lau KHV, Greer DM. Using technology adoption theories to maximize the uptake of e-learning in medical education. Med Sci Educ 2022 Apr;32(2):545-552. [doi: 10.1007/s40670-022-01528-7] [Medline: 35261814]
14. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q 2003;27(3):425. [doi: 10.2307/30036540]
15. Momani A. The unified theory of acceptance and use of technology: a new approach in technology acceptance. Int J Soc Knowled Develop 2020 Jul;12:79-98. [doi: 10.4018/IJSKD.2020070105]
16. Venkatesh V, Thong JYL, Xu X. Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. MIS Q 2012;36(1):157. [doi: 10.2307/41410412]
17. Chen CJ. Theoretical bases for using virtual reality in education. Theme Sci Techn Educ 2009;2(2):71-90 [FREE Full text]
18. Lange AK, Koch J, Beck A, et al. Learning with virtual reality in nursing education: qualitative interview study among nursing students using the unified theory of acceptance and use of technology model. JMIR Nurs 2020;3(1):e20249. [doi: 10.2196/20249] [Medline: 34345791]
19. Bracq MS, Michinov E, Arnaldi B, et al. Learning procedural skills with a virtual reality simulator: an acceptability study. Nurse Educ Today 2019 Aug;79:153-160. [doi: 10.1016/j.nedt.2019.05.026] [Medline: 31132727]
20. Baniasadi T, Ayyoubzadeh SM, Mohammadzadeh N. Challenges and practical considerations in applying virtual reality in medical education and treatment. Oman Med J 2020 May;35(3):e125. [doi: 10.5001/omj.2020.43] [Medline: 32489677]
21. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
22. Brooks J, McCluskey S, Turley E, King N. The utility of template analysis in qualitative psychology research. Qual Res Psychol 2015 Apr 3;12(2):202-222. [doi: 10.1080/14780887.2014.955224] [Medline: 27499705]
23. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
24. DeJonckheere M, Vaughn LM. Semistructured interviewing in primary care research: a balance of relationship and rigour. Fam Med Community Health 2019;7(2):e000057. [doi: 10.1136/fmch-2018-000057] [Medline: 32148704]
25. Kallio H, Pietilä AM, Johnson M, Kangasniemi M. Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. J Adv Nurs 2016 Dec;72(12):2954-2965. [doi: 10.1111/jan.13031] [Medline: 27221824]
26. Hennink MM, Kaiser BN, Marconi VC. Code saturation versus meaning saturation: how many interviews are enough? Qual Health Res 2017 Mar;27(4):591-608. [doi: 10.1177/1049732316665344] [Medline: 27670770]
27. Kiger ME, Varpio L. Thematic analysis of qualitative data: AMEE Guide No. 131. Med Teach 2020 Aug;42(8):846-854. [doi: 10.1080/0142159X.2020.1755030] [Medline: 32356468]
28. Regehr G. It's not rocket science: rethinking our metaphors for research in health professions education. Med Educ 2010 Jan;44(1):31-39. [doi: 10.1111/j.1365-2923.2009.03418.x] [Medline: 20078754]
29. Carter N, Bryant-Lukosius D, DiCenso A, Blythe J, Neville AJ. The use of triangulation in qualitative research. Oncol Nurs Forum 2014 Sep;41(5):545-547. [doi: 10.1188/14.ONF.545-547] [Medline: 25158659]

Abbreviations:

UTAUT: Unified Theory of Acceptance and Use of Technology
UTAUT2: Unified Theory of Acceptance and Use of Technology 2
VR: virtual reality

Edited by LT Car; submitted 07.06.24; peer-reviewed by K Doo, SH Stige; revised version received 16.01.25; accepted 24.01.25; published 12.03.25.

Please cite as:
Talan J, Forster M, Joseph L, Pradhan D
Exploring the Role of Immersive Virtual Reality Simulation in Health Professions Education: Thematic Analysis
JMIR Med Educ 2025;11:e62803
URL: https://mededu.jmir.org/2025/1/e62803
doi:10.2196/62803

Original Paper

Knowledge Mapping and Global Trends in Simulation in Medical Education: Bibliometric and Visual Analysis

Hongjun Ba[1], MD; Lili Zhang[2], BA; Xiufang He[2], BA; Shujuan Li[2], MD

[1]Department of Pediatrics, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China

[2]Department of Pediatric Cardiology, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China

Corresponding Author:
Hongjun Ba, MD
Department of Pediatrics
The First Affiliated Hospital
Sun Yat-sen University
58 Zhongshan Road 2
Guangzhou, 510080
China
Phone: 86 15920109625
Email: bahj3@mail.sysu.edu.cn

Abstract

Background: With the increasing recognition of the importance of simulation-based teaching in medical education, research in this field has developed rapidly. To comprehensively understand the research dynamics and trends in this area, we conducted an analysis of knowledge mapping and global trends.

Objective: This study aims to reveal the research hotspots and development trends in the field of simulation-based teaching in medical education from 2004 to 2024 through bibliometric and visualization analyses.

Methods: Using CiteSpace and VOSviewer, we conducted bibliometric and visualization analyses of 6743 articles related to simulation-based teaching in medical education, published in core journals from 2004 to 2024. The analysis included publication trends, contributions by countries and institutions, author contributions, keyword co-occurrence and clustering, and keyword bursts.

Results: From 2004 to 2008, the number of articles published annually did not exceed 100. However, starting from 2009, the number increased year by year, reaching a peak of 850 articles in 2024, indicating rapid development in this research field. The United States, Canada, the United Kingdom, Australia, and China published the most articles. Harvard University emerged as a research hub with 1799 collaborative links, although the overall collaboration density was low. Among the 6743 core journal articles, a total of 858 authors were involved, with Lars Konge and Adam Dubrowski being the most prolific. However, collaboration density was low, and the collaboration network was relatively dispersed. A total of 812 common keywords were identified, forming 4189 links. The keywords “medical education,” “education,” and “simulation” had the highest frequency of occurrence. Cluster analysis indicated that “cardiopulmonary resuscitation” and “surgical education” were major research hotspots. From 2004 to 2024, a total of 20 burst keywords were identified, among which “patient simulation,” “randomized controlled trial,” “clinical competence,” and “deliberate practice” had high burst strength. In recent years, “application of simulation in medical education,” “3D printing,” “augmented reality,” and “simulation training” have become research frontiers.

Conclusions: Research on the application of simulation-based teaching in medical education has become a hotspot, with expanding research areas and hotspots. Future research should strengthen interinstitutional collaboration and focus on the application of emerging technologies in simulation-based teaching.

(JMIR Med Educ 2025;11:e71844) doi:10.2196/71844

KEYWORDS

medical education; simulation-based teaching; bibliometrics; visualization analysis; knowledge mapping

Introduction

In the rapidly evolving landscape of medical education, the integration of simulation-based training has emerged as a pivotal innovation. Simulation in medical education encompasses a broad spectrum of methodologies, including high-fidelity mannequins, virtual reality, standardized patients, and computer-based simulations [1,2]. These techniques aim to enhance clinical skills, decision-making, and teamwork among medical professionals without the direct involvement of real patients.

The adoption of simulation in medical training addresses several critical challenges [3,4]. First, it provides a safe and controlled environment where learners can practice and refine their skills. This is particularly crucial in high-stakes scenarios such as emergency medicine, surgery, and critical care, where errors can have severe consequences [5,6]. In addition, simulation allows for repetitive practice and immediate feedback, facilitating a deeper understanding of complex procedures and concepts.

Over the past few decades, there has been a significant increase in research focused on the effectiveness and impact of simulation-based education in the medical field [7,8]. This growing body of literature reflects the widespread recognition of simulation as a valuable educational tool. However, the rapid expansion of this field necessitates a comprehensive review and analysis to understand its development, trends, and future directions.

Several bibliometric analyses have been conducted on simulation in medical education [9,10], highlighting its growing importance and impact. However, these studies often focus on specific aspects of simulation, such as surgical training or virtual reality. Our study complements this body of research by providing a comprehensive overview of the entire field, including emerging technologies like 3D printing and augmented reality (AR), and by analyzing collaborative networks and thematic trends over a 20-year period.

A bibliometric analysis provides an ideal approach to systematically evaluate the literature on simulation in medical education. By using quantitative methods to analyze publication patterns, citation networks, and research themes, bibliometric studies can offer valuable insights into the evolution of this field. Such an analysis can identify key contributors, influential publications, and emerging trends, thereby guiding future research and practice.

This study aims to conduct a bibliometric analysis of the literature on simulation in medical education. By examining the scope, growth, and impact of research in this area, we seek to elucidate the current state of the field and identify potential gaps and opportunities for further investigation. Specifically, this analysis will focus on the following objectives:

1. To map the overall publication trends and growth in simulation-based medical education research.
2. To identify the most influential journals, articles, and authors contributing to this field.
3. To explore the thematic evolution and emerging trends within the literature.
4. To assess the collaborative networks and geographical distribution of research activities.

Through this comprehensive bibliometric analysis, we hope to provide a clearer understanding of the trajectory and impact of simulation in medical education, ultimately contributing to the enhancement of educational practices and outcomes in the medical field.

Methods

Data Acquisition and Search Strategy

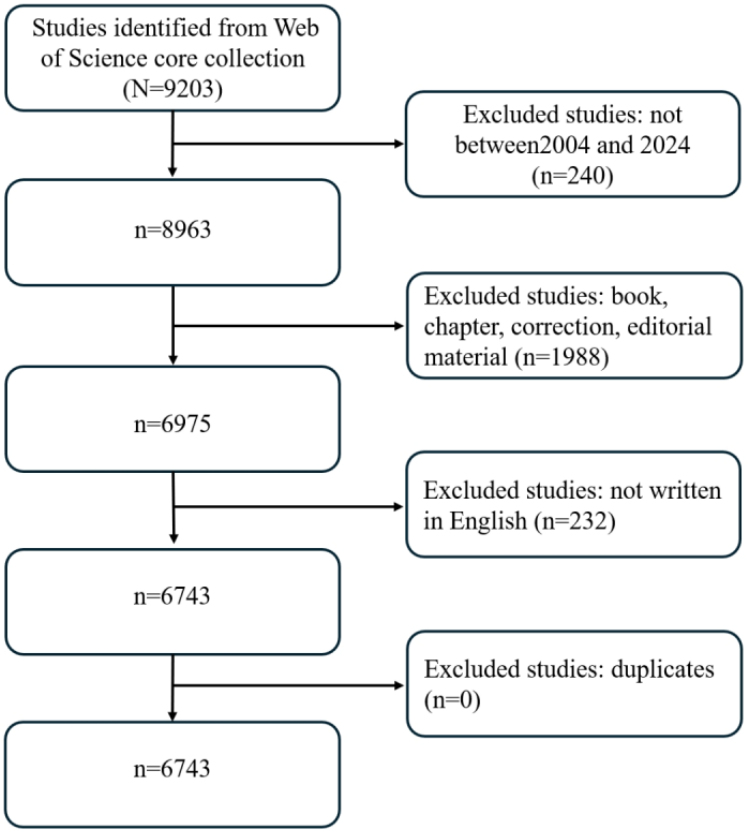
The search was conducted in the Web of Science Core Collection (WoSCC) database, which is widely recognized for its comprehensive coverage of high-quality, peer-reviewed literature [11,12]. While we acknowledge that including additional databases such as PubMed or Scopus could provide a more comprehensive dataset, the WoSCC was chosen for its superior bibliographic accuracy and extensive coverage of medical education research. Therefore, we opted to perform our search within this database. We conducted a search in the Web of Science for all relevant papers published between January 1, 2004, and December 31, 2024. The time frame from January 1, 2004, to December 31, 2024, was selected because it marks the period when simulation-based medical education began to gain significant traction in the literature, reflecting the growing recognition of its importance in medical training. In medical education, we define “simulation” as a teaching and training method that encompasses high-fidelity mannequins, virtual reality, standardized patients, and computer-based simulations.

The search formula “TS=(Medical education) AND TS=(Simulation)” was used. The inclusion criteria were as follows: (1) full-text publications related to simulation in medical education, including original research articles and review articles; (2) articles written in English; and (3) papers published between January 1, 2004, and December 31, 2024. We excluded conference abstracts, theses, dissertations, and nonpeer-reviewed articles to ensure the quality and relevance of the data. The exclusion criteria were (1) topics not related to simulation in medical education and (2) papers in the form of conference abstracts, theses, dissertations, and non–peer-reviewed articles to ensure the quality and relevance of the data. A plain text version of the papers was exported.

General Data

Between January 1, 2004, and December 31, 2024, the WoSCC database recorded a total of 6743 publications concerning simulation in medical education. This body of literature included contributions from 121 countries and regions, 510 institutions, and 858 authors. Figure 1 shows the process of literature searching and bibliometric analysis.

Figure 1. The workflow of data collection and bibliometric analysis.



Data Analysis

We used GraphPad Prism (version 8.0.2; Dotmatics) to illustrate annual publication trends. The methodological approach was validated through the use of CiteSpace and VOSviewer, both of which are widely recognized and extensively used in bibliometric research [13,14]. These tools have been shown to provide robust and reliable analyses of large-scale bibliometric data.

VOSviewer, a Java-based software developed by van Eck and Waltman in 2009, facilitates the construction of various types of network maps, such as bibliographic coupling, cocitation, and coauthorship networks. CiteSpace, developed by Professor Chaomei Chen, provides a dynamic platform for identifying and visualizing patterns and trends in scientific literature, enabling the exploration of knowledge domains and predictive analysis of research trajectories [14]. Our methodological approach involved setting specific parameters for network density (eg, keyword co-occurrence density of 0.0127), node inclusion thresholds (eg, minimum occurrence frequency of keywords), and time-slicing techniques to analyze temporal changes. The references corresponding to the software applications were verified against our citation list to ensure accuracy [13,14]. When using VOSviewer and CiteSpace for bibliometric analysis, we established standards for defining international collaboration. This was done by examining the authorship of papers, specifically the first and corresponding authors, to ensure a comprehensive capture of collaborative efforts from researchers across different countries.

Burst detection in CiteSpace is based on the Kleinberg algorithm, which models document streams using infinite-state automata to extract meaningful structures [15]. These analyses can reveal rapidly growing topics over extended periods as well as short-term themes.

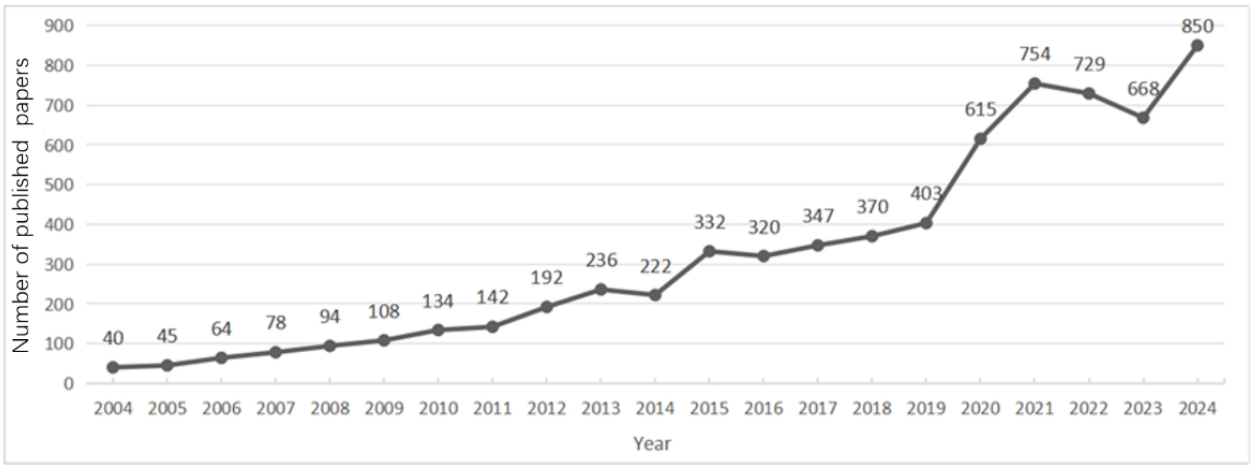
The rationale for selecting these techniques lies in their widespread application and effectiveness in bibliometric research. They provide robust and complementary insights into productivity, impact, and collaboration patterns within research fields.

Results

Publication Trend

Figure 2 shows that from 2004 to 2008, the annual number of publications on simulation teaching in medical education did not exceed 100 articles, indicating that research in this field was still in its nascent stage. Since 2009, the number of publications in this field has steadily increased, showing a trend of fluctuating growth. Specifically, the number of publications in 2015 surpassed 300 for the first time, and by 2020, this number had exceeded 500. This significant increase marks the growing attention and interest of scholars and researchers in the field of simulation teaching in medical education. Since 2020, the annual number of publications in this field has consistently remained above 500, reaching a peak of 850 articles in 2024. This further highlights the vigorous development and extensive influence of research in the field of simulation teaching in medical education.

Figure 2. Trend chart of publications in the past 20 years.



Country or Region and Institution Contributions

According to Figure 3A, the connections between circular nodes representing different countries to some extent reflect the existence of relationships and collaborations between these countries. Furthermore, the density of these connections in the network can serve as an important indicator of the closeness of collaborative relationships between countries. Among them, the countries with the highest number of publications are the United States (3083 articles), Canada (776 articles), England (510 articles), Australia (381 articles), and China (375 articles) (Table 1). In addition, countries such as Italy, the Netherlands, Sweden, and Belgium have numerous connections, indicating a complex network of relationships, which suggests that these countries have relatively close research collaborations with other regions.

Using the CiteSpace software, an institutional collaboration network diagram was obtained, as shown in Figure 3B. Upon statistical analysis, it was found that there are a total of 510 research institutions forming 1799 connections, with Harvard University being the central hub. The diagram reveals that the network density is 0.0139, indicating relatively weak collaborative relationships between research institutions, with a significant portion of them operating in a relatively independent research state. In terms of research output, the top-10 institutions by the number of publications are Harvard University, the University of Toronto, the University of California System, the University System of Ohio, Harvard Medical School, Mayo Clinic, Northwestern University, Feinberg School of Medicine, the University of Copenhagen, and Pennsylvania Commonwealth System of Higher Education (Table 2).

Figure 3. Network graph of national and institutional collaborations. (A) Network graph of national collaborations. (B) Network graph of institutional collaborations. The bubble size represents the number of publications.

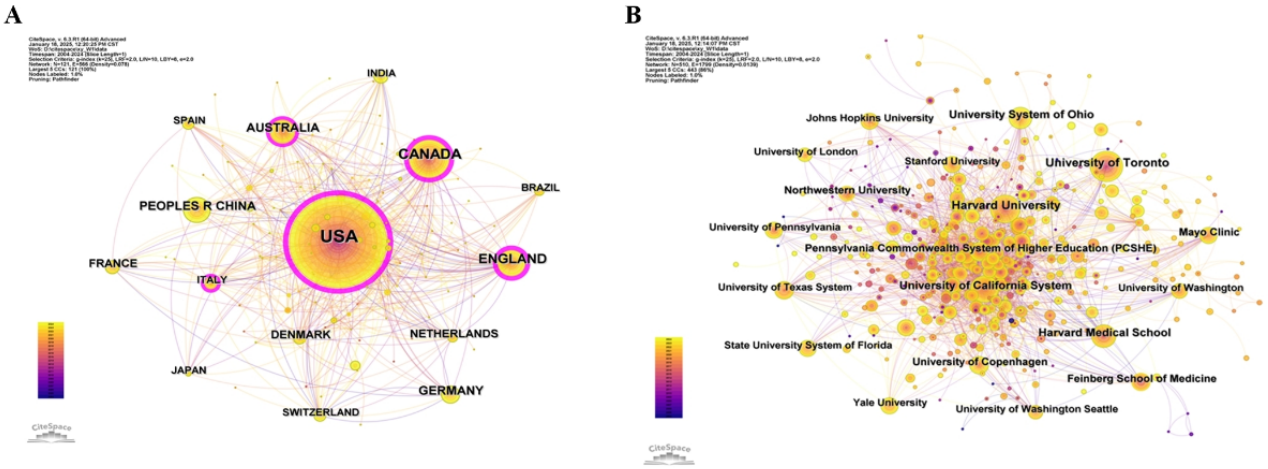


Table 1. Top-10 most productive countries or regions.

Rank	Country or region	Articles, n	Centrality	Percentage	Half-life
1	United States	3083	0.51	45.71%	14.5
2	Canada	776	0.22	11.52%	13.5
3	England	510	0.25	7.57%	15.5
4	Australia	381	0.12	5.64%	15.5
5	China	375	0.02	5.56%	15.5
6	Germany	340	0.05	5.04%	16.5
7	France	189	0.03	2.8%	15.5
8	Denmark	184	0.05	2.73%	15.5
9	The Netherlands	164	0.1	2.43%	15.5
10	Switzerland	142	0.05	2.1%	16.5

Table 2. Top-10 most productive institutions.

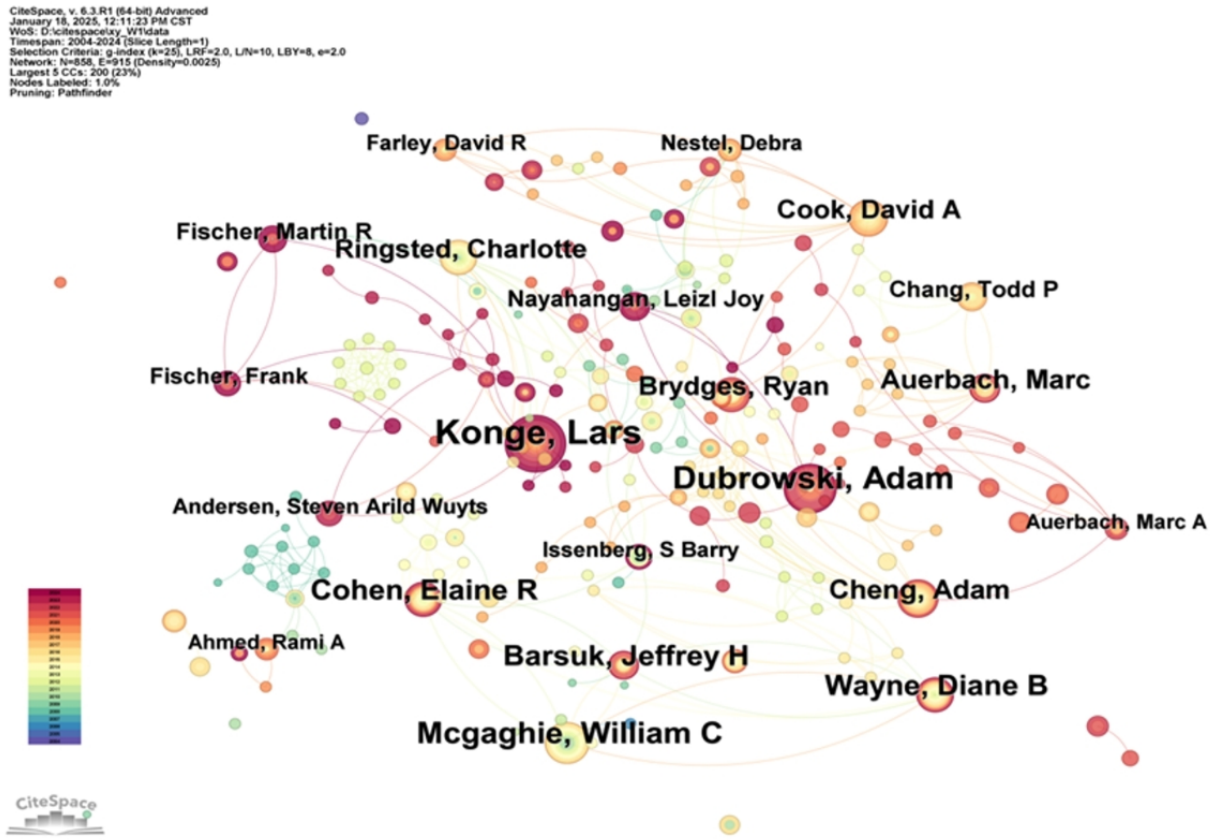
Rank	Institution	Country	Studies, n	Centrality	Half-life
1	Harvard University	United States	167	0.06	13.5
2	University of Toronto	Canada	153	0.04	11.5
3	University of California System	United States	125	0.01	12.5
4	University System of Ohio	United States	118	0.04	11.5
5	Harvard Medical School	United States	82	0.09	13.5
6	Mayo Clinic	United States	68	0.03	12.5
7	Northwestern University	United States	66	0.01	10.5
8	Feinberg School of Medicine	United States	65	0.01	10.5
9	University of Copenhagen	Denmark	61	0.01	14.5
10	Pennsylvania Commonwealth System of Higher Education	United States	60	0.02	13.5

Author Collaborations

The sample data were processed using CiteSpace, and the resulting author co-occurrence map is shown in Figure 4. In this map, each node represents a different author. The size of the node indicates the author's publication frequency, meaning the larger the node, the more publications the author has. When nodes are presented in the form of annual rings, the bandwidth of the color band corresponding to a particular year represents the number of papers published by the author that year, with a wider ring indicating more publications. The lines between nodes represent collaborative relationships between organizations or authors, with the thickness of the lines indicating the degree of collaboration.

Among the 6743 core journal articles, a total of 858 authors were involved. The top-10 authors by publication volume are Konge, Lars (69 papers); Dubrowski, Adam (50 papers); McGaghie, William C (37 papers); Wayne, Diane B (33 papers); Cohen, Elaine R (33 papers); Barsuk, Jeffrey H (30 papers); Auerbach, Marc (26 papers); Cheng, Adam (24 papers); Cook, David A (23 papers); and Ringsted, Charlotte (22 papers). Authors with 7 or more publications, a total of 44 individuals, were classified as the core author group, which accounts for only 5.1% of the total authors. In addition, there are 915 collaboration lines among the authors on the map, with a collaboration density of 0.0025, indicating a low-density level. The number of lines is relatively sparse, and the collaboration network map shows a relatively dispersed pattern. The largest collaboration network system is formed by the research team centered around Dubrowski, Adam; Nayahangan, Leizl Joy; Cheng, Adam; Auerbach, Marc A; and Cook, David A. The scale of collaboration is mainly presented in the form of individual or small-scale research teams, indicating that the core research team in this field has yet to be fully established.

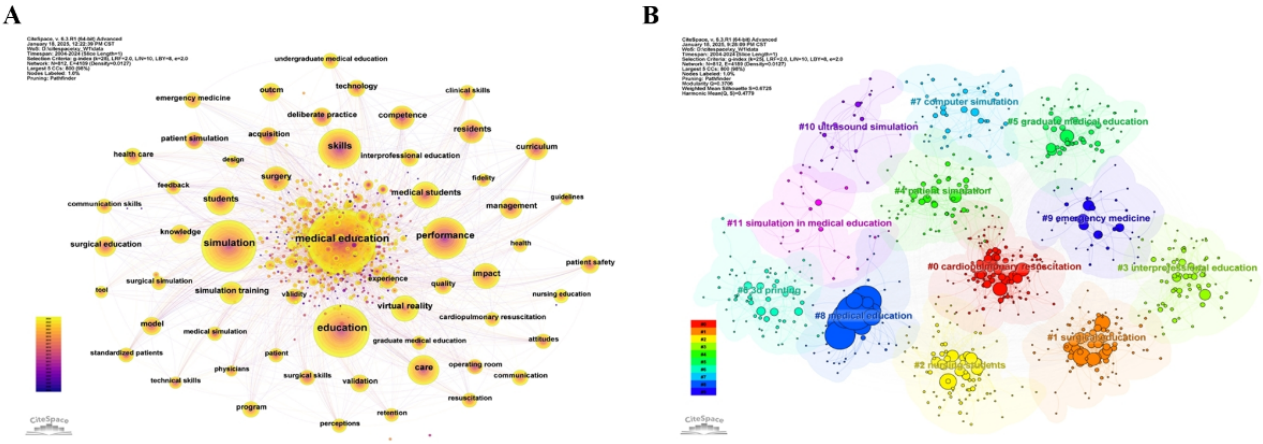
Figure 4. Network diagram of author collaborations. The bubble size represents the number of publications.



Keyword Co-Occurrence and Cluster

Using CiteSpace software to conduct a keyword co-occurrence analysis on the sample, the constructed keyword co-occurrence map is shown in Figure 5A. From the keyword co-occurrence analysis, a total of 812 common keywords were identified, forming 4189 connections, with a network density of 0.0127. The most frequently occurring keyword is “medical education,” accounting for 7.9%. This is followed by “education” and “simulation,” which account for 5.48% and 5.14%, respectively. The keywords “performance” and “skills” account for 3.49% and 3.29%, respectively. These keywords represent the current research hotspots and status in the field of simulation teaching in medical education.

Figure 5. Keyword co-occurrence and keyword clustering map. (A) Keyword co-occurrence map. (B) Keyword clustering map. The bubble size represents the number of publications.



Based on the keyword co-occurrence map, the log-likelihood ratio algorithm was used to cluster the keywords, resulting in a keyword clustering co-occurrence map. The Q value is 0.3707 (>0.3), and the S value is 0.6725 (>0.5), indicating a significant clustering structure and a high degree of clustering match. The map displays a total of 10 clustering areas, among which “cardiopulmonary resuscitation,” “surgical education,” “nursing students,” “interprofessional education,” and “patient simulation” are the five largest clusters (Figure 5B). Specifically, medical simulation teaching has become an important

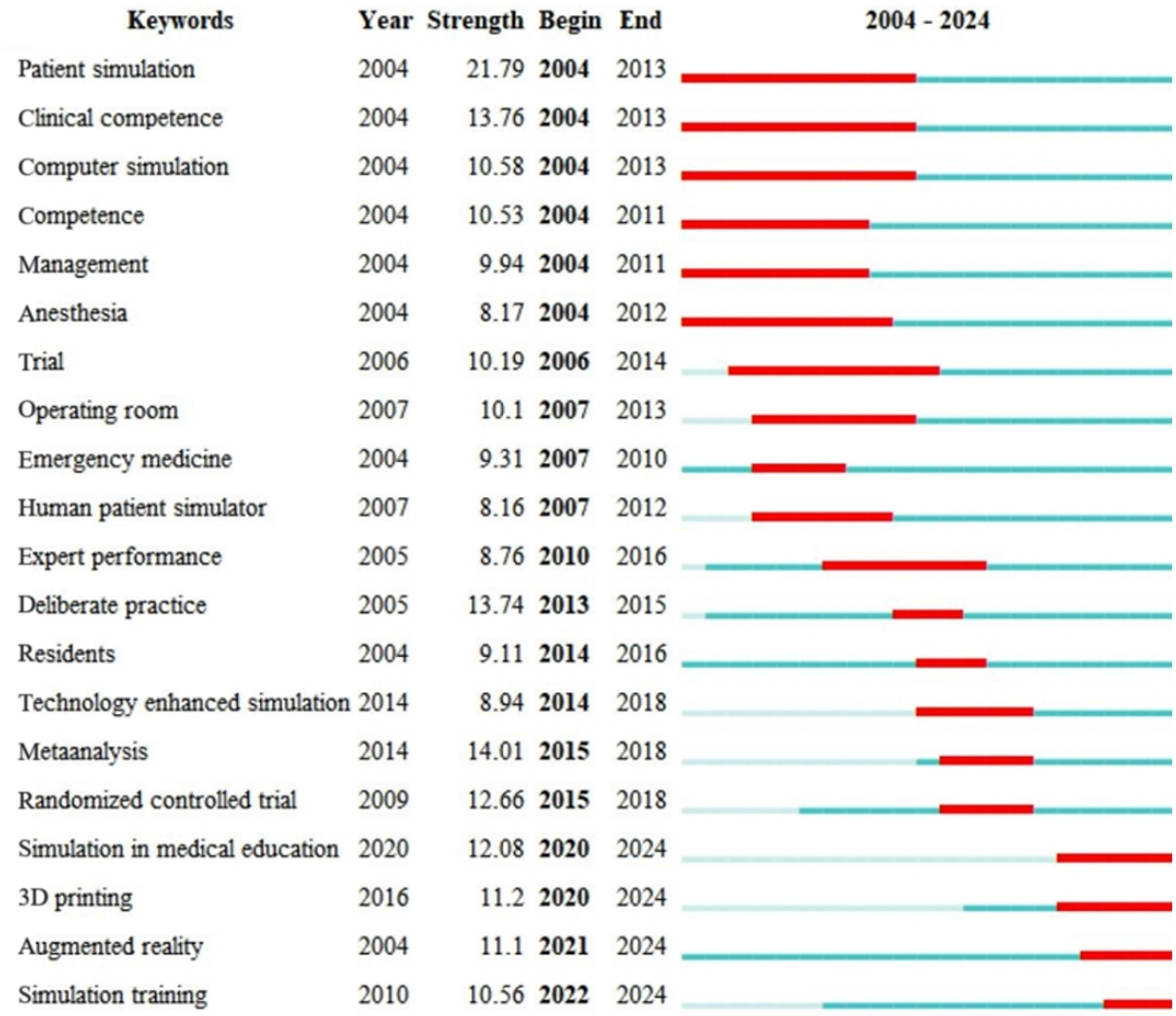
component of medical education, widely applied in various fields including cardiopulmonary resuscitation, surgical education, and nursing student training.

Keyword Citation Bursts

The keyword burst visualization analysis identified a total of 20 keywords in the field of simulation teaching in medical education from 2004 to 2024, along with their emergence intensity and start-end years. The relevant literature on keyword emergence is shown in Figure 6.

It can be observed that the keywords with high emergence intensity include "patient simulation," "randomized controlled trial," "clinical competence," and "deliberate practice." Meanwhile, the research area began to focus on "patient simulation" as early as 2004, which, along with "clinical competence" and "computer simulation," became one of the keywords with the longest duration of emergence. In recent years, researchers have increasingly focused on themes such as "trial" and "expert performance." The keywords that are still emerging represent the current research frontiers and trends, which include "simulation in medical education," "3D printing," "augmented reality," and "simulation training." The emergence of "3D printing" reflects the growing interest in using patient-specific anatomical models for surgical planning and training, offering a more personalized and immersive learning experience. Similarly, "augmented reality" signifies the integration of advanced technologies to create interactive and realistic training environments, enhancing the acquisition of clinical skills. These emerging trends highlight the transformative potential of technology in medical education, paving the way for more innovative and effective teaching methodologies.

Figure 6. Keyword burst graph (sorted by the beginning year of the burst). The blue bars denote the reference has been published; the red bars denote citation burstiness.



Discussion

Principal Findings

The results of our bibliometric analysis provide a comprehensive overview of the evolution, collaboration patterns, and thematic focus of simulation-based education research in the medical field. Key trends include a steady increase in publications from 2004 to 2024, particularly a surge after 2009, indicating a growing recognition of simulation's importance in medical education. By 2024, the publication count had peaked at 850, highlighting a transition of simulation from a novel approach to a staple in medical education. In addition, the United States emerged as the leading contributor with 3083 articles, reflecting substantial investment in education and research. Harvard University is a central hub for simulation-based medical education, despite a fragmented institutional landscape. Prominent authors like Lars Konge, Adam Dubrowski, and William C McGaghie drive the field, though the low density of collaborative networks suggests room for enhanced inter-institutional teamwork. Keyword analysis underscores the focus on competency-based education and practical skill acquisition, with emerging technologies like 3D printing and AR shaping future directions.

Comparison to Literature

Our findings are consistent with existing literature [16,17], which also highlights the increasing role of simulation in medical education over the past two decades. Previous studies have documented the rise in publications and the central role of the United States and key institutions like Harvard in advancing this field. However, our analysis provides a more granular look at the collaborative networks and thematic focuses, revealing a fragmented institutional landscape and the emergence of cutting-edge technologies that are less emphasized in earlier reviews.

Implications of Findings

The implications of these findings are multifaceted. The robust growth in simulation-based medical education research indicates a broad acceptance of its efficacy in improving medical training. The strong international collaboration suggests that best practices and innovative methodologies are being shared globally, potentially standardizing and enhancing simulation protocols. The emergence of new technologies like 3D printing and AR points to a future where simulation-based education will be more immersive and technologically advanced [18-20], which could significantly enhance learning outcomes and patient care. The integration of 3D printing and AR into simulation-based training can significantly improve clinical outcomes. 3D-printed anatomical models enable patient-specific simulations, allowing surgeons to practice complex procedures before operating on real patients, thus enhancing precision and reducing errors [21]. Similarly, AR creates immersive training environments, providing real-time feedback and interactive learning to enhance clinical skill acquisition [22]. However, challenges such as the high cost of equipment and the need for specialized training for educators and learners may limit their widespread adoption. Future research should explore cost-effective solutions to overcome these barriers and ensure broader access to these technologies in medical institutions.

The emergence of "cardiopulmonary resuscitation" as a major research hotspot reflects its critical importance in medical education and clinical practice. Cardiopulmonary resuscitation is a high-stakes procedure where errors can have severe consequences, making it an ideal candidate for simulation-based training [23]. Simulation allows learners to practice cardiopulmonary resuscitation in a controlled environment, receive immediate feedback, and refine their skills through repetitive practice [24]. This not only enhances individual competence but also improves team dynamics and communication during real-life emergencies.

Similarly, the focus on "surgical education" underscores the need for advanced training methods to prepare surgeons for complex procedures. Simulation-based training in surgical education has been shown to improve technical skills, reduce operative time, and enhance patient safety [25]. These findings highlight the transformative potential of simulation in addressing critical gaps in medical education and improving clinical outcomes.

Limitations

While the bibliometric analysis provides valuable insights, it has several limitations. First, the data might not capture all relevant publications, particularly those in non-English languages or those in less accessible databases, which could introduce selection bias. Second, the analysis relies on citation metrics, which may not fully reflect the quality or practical impact of the research. For instance, highly cited articles may not always represent the most impactful studies in terms of educational outcomes. Third, the low density of collaborative networks suggests that our findings might underrepresent the potential for interinstitutional synergy and innovation. Finally, a limitation of this study is the reliance on a single database (WoSCC), which may not capture all relevant publications. Future studies could expand the search to include additional databases such as PubMed and Scopus to enhance the robustness of the findings.

Suggestions

To address the identified limitations and enhance the impact of simulation-based education research, we suggest the following:

1. Increasing efforts to include diverse and international publications in future analyses.
2. Encouraging more interinstitutional collaborations to create a more cohesive research landscape.
3. Fostering larger, integrated research teams to deepen the scope of studies and drive innovation.
4. Embracing and further investigating emerging technologies to stay at the forefront of educational advancements.

Conclusions

In conclusion, the bibliometric analysis of simulation in medical education research reveals a dynamic field characterized by rapid growth, strong international collaboration, and evolving thematic focuses. The increasing trend in publications, significant contributions from leading countries and institutions,

and the integration of new technologies underscore the impactful nature of this research area. Moving forward, enhancing collaboration among institutions and expanding the core author network will be crucial. Future research should focus on integrating emerging technologies, such as 3D printing and AR, into medical education. For instance, studies could explore how 3D-printed anatomical models can enhance surgical training by providing realistic, patient-specific simulations.

Data Availability

All datasets generated for this study were included in the manuscript.

Authors' Contributions

HB conceived and designed the ideas for the manuscript. HB, LZ, XH, and SL participated in all data collection and processing. HB was the major contributor in organizing records and drafting the manuscript. All authors proofread and approved the manuscript.

Conflicts of Interest

None declared.

References

1. Motola I, Devine LA, Chung HS, Sullivan JE, Issenberg SB. Simulation in healthcare education: a best evidence practical guide. AMEE Guide No. 82. Med Teach 2013;35(10):e1511-e1530. [doi: 10.3109/0142159X.2013.818632] [Medline: 23941678]
2. So HY, Chen PP, Wong GKC, Chan TTN. Simulation in medical education. J R Coll Physicians Edinb 2019;49(1):52-57. [doi: 10.4997/JRCPE.2019.112] [Medline: 30838994]
3. Guze PA. Using technology to meet the challenges of medical education. Trans Am Clin Climatol Assoc 2015;126:260-270 [FREE Full text] [Medline: 26330687]
4. Kennedy CC, Cannon EK, Warner DO, Cook DA. Advanced airway management simulation training in medical education: a systematic review and meta-analysis. Crit Care Med 2014;42(1):169-178. [doi: 10.1097/CCM.0b013e31829a721f] [Medline: 24220691]
5. Hammond J. Simulation in critical care and trauma education and training. Curr Opin Crit Care 2004;10(5):325-329. [doi: 10.1097/01.ccx.0000140950.47361.c9] [Medline: 15385746]
6. Martín Parra JI, Manuel Palazuelos JC, Gómez Fleitas M. Pursuing quality in simulation-based surgical education. Cir Esp 2013;91(10):623-624. [doi: 10.1016/j.ciresp.2013.06.013] [Medline: 24143942]
7. Hepps JH, Yu CE, Calaman S. Simulation in medical education for the hospitalist: moving beyond the mock code. Pediatr Clin North Am 2019;66(4):855-866. [doi: 10.1016/j.pcl.2019.03.014] [Medline: 31230627]
8. Lu J, Cuff RF, Mansour MA. Simulation in surgical education. Am J Surg 2021;221(3):509-514. [doi: 10.1016/j.amjsurg.2020.12.016] [Medline: 33358139]
9. Walsh C, Lydon S, Byrne D, Madden C, Fox S, O Connor P. The 100 most cited articles on healthcare simulation: a bibliometric review. Simul Healthc 2018;13(3):211-220. [doi: 10.1097/SIH.0000000000000293] [Medline: 29613918]
10. Yao S, Tang Y, Yi C, Xiao Y. Research hotspots and trend exploration on the clinical translational outcome of simulation-based medical education: a 10-year scientific bibliometric analysis from 2011 to 2021. Front Med (Lausanne) 2021;8:801277 [FREE Full text] [doi: 10.3389/fmed.2021.801277] [Medline: 35198570]
11. Wu H, Li Y, Tong L, Wang Y, Sun Z. Worldwide research tendency and hotspots on hip fracture: a 20-year bibliometric analysis. Arch Osteoporos 2021;16(1):73. [doi: 10.1007/s11657-021-00929-2] [Medline: 33866438]
12. Vargas JS, Livinski AA, Karagu A, Cira MK, Maina M, Lu Y, et al. A bibliometric analysis of cancer research funders and collaborators in Kenya: 2007-2017. J Cancer Policy 2022;33:100331 [FREE Full text] [doi: 10.1016/j.jcpo.2022.100331] [Medline: 35792397]
13. van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 2010;84(2):523-538 [FREE Full text] [doi: 10.1007/s11192-009-0146-3] [Medline: 20585380]
14. Chen C. CiteSpace: A Practical Guide for Mapping Scientific Literature. New York, NY: Nova Science Publishers; 2016.
15. Kleinberg J. Bursty and hierarchical structure in streams. Data Min Knowl Discov 2003;7:373-397. [doi: 10.1023/A:1024940629314]
16. Elendu C, Amaechi DC, Okatta AU, Amaechi EC, Elendu TC, Ezeh CP, et al. The impact of simulation-based training in medical education: a review. Medicine (Baltimore) 2024;103(27):e38813 [FREE Full text] [doi: 10.1097/MD.0000000000038813] [Medline: 38968472]
17. Higgins M, Madan C, Patel R. Development and decay of procedural skills in surgery: a systematic review of the effectiveness of simulation-based medical education interventions. Surgeon 2021;19(4):e67-e77. [doi: 10.1016/j.surge.2020.07.013] [Medline: 32868158]

18. Childs BS, Manganiello MD, Korets R. Novel education and simulation tools in urologic training. Curr Urol Rep 2019;20(12):81. [doi: 10.1007/s11934-019-0947-8] [Medline: 31782033]
19. Smith B, Dasgupta P. 3D printing technology and its role in urological training. World J Urol 2020;38(10):2385-2391. [doi: 10.1007/s00345-019-02995-1] [Medline: 31676911]
20. Sun Z, Wong YH, Yeong CH. Patient-specific 3D-printed low-cost models in medical education and clinical practice. Micromachines (Basel) 2023;14(2):464 [FREE Full text] [doi: 10.3390/mi14020464] [Medline: 36838164]
21. Ravi P, Chepelev LL, Stichweh GV, Jones BS, Rybicki FJ. Medical 3D printing dimensional accuracy for multi-pathological anatomical models 3D printed using material extrusion. J Digit Imaging 2022;35(3):613-622 [FREE Full text] [doi: 10.1007/s10278-022-00614-x] [Medline: 35237891]
22. Nagayo Y, Saito T, Oyama H. Augmented reality self-training system for suturing in open surgery: a randomized controlled trial. Int J Surg 2022;102:106650 [FREE Full text] [doi: 10.1016/j.ijsu.2022.106650] [Medline: 35525415]
23. Demirtas A, Guvenc G, Aslan, Unver V, Basak T, Kaya C. Effectiveness of simulation-based cardiopulmonary resuscitation training programs on fourth-year nursing students. Australas Emerg Care 2021;24(1):4-10. [doi: 10.1016/j.auec.2020.08.005] [Medline: 32933888]
24. Laco RB, Stuart W. Simulation-based training program to improve cardiopulmonary resuscitation and teamwork skills for the urgent care clinic staff. Mil Med 2022;187(5-6):e764-e769. [doi: 10.1093/milmed/usab198] [Medline: 34050365]
25. Zubair U, Zubair Z. Surgical resident training in Pakistan and benefits of simulation based training. J Pak Med Assoc 2020;70(5):904-908. [doi: 10.5455/JPMA.282116] [Medline: 32400750]

Abbreviations

AR: augmented reality
WoSCC: Web of Science Core Collection

Edited by L Tudor Car; submitted 27.01.25; peer-reviewed by G Huang, W Huang; comments to author 27.02.25; revised version received 06.03.25; accepted 07.03.25; published 26.03.25.

Please cite as:
Ba H, Zhang L, He X, Li S
Knowledge Mapping and Global Trends in Simulation in Medical Education: Bibliometric and Visual Analysis
JMIR Med Educ 2025;11:e71844
URL: https://mededu.jmir.org/2025/1/e71844
doi:10.2196/71844
PMID:

Global Disparities in Simulation-Based Learning Performance: Serial Cross-Sectional Mixed Methods Study

Kashish Malhotra[1,2*], MBBS; Harshin Balakrishnan[2*], MBBS; Emily Warmington[3], MBBS; Vina Soran[3], MBBS; Francesca Crowe[2], PhD; Dengyi Zhou[4], MBChB; SIMBA AND CoMICs Team[3]; Punith Kempegowda[2,5], PhD

[1]Department of Surgery, Dayanand Medical College, Punjab, India

[2]Department of Applied Health Sciences, College of Medical and Dental Sciences, University of Birmingham, Edgbaston, Birmingham, United Kingdom

[3]College of Medical and Dental Sciences, University of Birmingham, Birmingham, United Kingdom

[4]Northwick Park Hospital, London North West University Healthcare NHS Trust, London, United Kingdom

[5]Queen Elizabeth Hospital, University Hospitals Birmingham NHS Foundation Trust, Birmingham, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Punith Kempegowda, PhD
Department of Applied Health Sciences, College of Medical and Dental Sciences, University of Birmingham, Edgbaston, Birmingham, United Kingdom

Abstract

Background: Simulated programs provide health care professionals (HCPs) with a learning opportunity to develop clinical competencies and improve patient outcomes in a safe and controlled environment. While the benefits of simulation training are well established, there is a paucity of research assessing its differential impact, if any. SIMBA (Simulation via Instant Messaging for Bedside Application) provides simulation-based learning through WhatsApp and Zoom (Zoom Video Communications, Inc) to increase HCPs' confidence in managing various medical conditions.

Objectives: This study aims to explore whether there are differences in the clinical performance of HCPs participating in SIMBA sessions based on gender, country of work, and training grade.

Methods: This study assessed participants in 17 SIMBA sessions from May 2020 to June 2022. WhatsApp chats containing participants' approach to the simulated scenarios were graded using an adapted version of the Global Rating Scale consisting of 6 domains: eliciting history; physical examination; investigations, diagnostic tests, and imaging; interpretation of investigations and imaging; clinical judgment; and management and follow-up or discharge plan. These domains were rated using a Likert-type scale of 1 (not done) to 5 (excellent) prior to the session based on expert inputs. All WhatsApp transcripts were evaluated against the scale postsimulation session. Unadjusted and adjusted means and 95% CIs of the scores for the 6 performance variables were calculated using multiple linear regression models. The *P* value for heterogeneity between the mean performance scores was calculated using likelihood ratio tests by using an analysis of variance.

Results: A total of 289 participants across 49 countries who completed pre-SIMBA and post-SIMBA surveys in the 17 simulation sessions were included in the analysis. Participants from high-income countries scored higher in all categories of the Global Rating Scale (GRS) except the physical examination and interpretation score. Junior-grade participants scored significantly higher in history taking (junior=4.2, middle=3.7, and senior=3.7; *P*=.003) and physical examination (junior=4.0, middle=3.7, and senior=3.5; *P*=.068), but this was not significantly different. There were no statistically significant differences in GRS scores between male and female participants.

Conclusions: The significant differences in clinical performance scores between low- and middle-income countries and high-income countries highlight the need for better medical education resources to bridge existing gaps in health care globally. The decrease in some clinical competency scores following career progression could be addressed by simulation-based training to maintain the same quality of history taking and physical examination skills. These outcomes, including no gendered differences in simulation-based learning, hold profound implications for tailoring medical education strategies, fostering equitable training, and elevating patient care standards on a global scale. The need for targeted interventions and capacity-building efforts via context-specific training and tailored approaches to health care education is emphasized.

(JMIR Med Educ 2025;11:e52332) doi:10.2196/52332

KEYWORDS

simulation; medical education; SIMBA; Global Rating Scale; low- and middle- income countries; high-income countries; global medical education; clinical training; gender equity; global disparity; disparity; discrepancy; inequality; health care education; clinical; cross-sectional study; linear regression model; physical examination; simulation-based learning

Introduction

Understanding the multifaceted global health care disparities is crucial for ensuring access to high-quality medical education [1]. The most common barriers to high-quality medical education are due to financial, resource, and accessibility inequities [2]. There is a shortage in the training resources to recruit and retain health care professionals (HCPs) in low- and middle-income countries (LMICs), leading to disparities in clinical learning opportunities compared with HCPs in high-income countries (HICs) [3]. This was further exacerbated following the COVID-19 pandemic [4,5].

E-learning and distance virtual simulation sessions can provide case-based learning and support the training of medical professionals in LMICs [6]. Specifically, initiatives using accessible technology platforms to develop high-quality educational programs can enhance postgraduate health care training [7,8]. Simulated programs provide HCPs with a learning opportunity to develop clinical competencies and improve patient outcomes in a safe and controlled environment [9,10]. However, there is an unequal distribution of simulation-based learning across LMICs, due to misconceptions regarding the cost and availability of programs [11]. Also, while the benefits of simulation training are well established, there is a paucity of research studying the differences, if any, in the clinical performance of HCPs from HICs and LMICs in simulation-based case scenarios. Understanding the similarities and differences in clinical performance can help identify areas for improvement in health care education and resource allocation in LMICs.

SIMBA (Simulation via Instant Messaging for Bedside Application) is a simulation-based learning model that uses WhatsApp and Zoom (Zoom Video Communications, Inc) platforms to improve HCPs' confidence in managing various medical conditions [12,13]. The model is built on the principles of Kolb's experiential learning theory and simulation gaming theory [14]. Kolb's experiential learning theory emphasizes the importance of learning through reflection on doing, where learners actively engage in a cycle of concrete experience, reflective observation, abstract conceptualization, and active experimentation. This framework guided the design of our SIMBA sessions, ensuring that participants were exposed to real-world scenarios and encouraged to reflect on their actions and apply new knowledge in future simulations. Simulation gaming theory further shaped our curriculum by integrating gamelike elements to create an immersive and interactive learning environment. Using realistic clinical cases, role-playing, and feedback loops in our sessions aligns with the theory's emphasis on learning through interaction and problem-solving within a simulated context. This approach fosters engagement and enhances skill development, making the learning process dynamic and participatory. It helped deliver and sustain good quality medical education during the COVID-19 pandemic [15]. It also provided alternative work experiences for premedical students [16] and a platform for medical students and junior doctors to improve teamwork and leadership skills [17]. Our previous work described how SIMBA provided equitable and high-quality postgraduate medical education to health care physicians in the LMICs and HICs [18].

As health care systems strive to provide safe and effective medical services, understanding the nuances of clinical competence among different training grades becomes a pivotal aspect of medical education and professional development. Exploring the clinical performance variations across various stages of HCP training offers valuable insights into how educational strategies and experiential learning impact the acquisition and application of clinical competencies. The need for such exploration is underscored by the potential ramifications these findings may have on patient care. An HCP's ability to accurately diagnose, interpret diagnostic tests, make informed clinical judgments, and formulate appropriate management plans can vary across different training grades. These variations can impact patient safety, the quality of health care services, and the overall effectiveness of medical teams [19-21]. In addition, as health care systems increasingly seek to bridge global disparities in medical education and patient outcomes, it is imperative to identify how training grade–related differences influence clinical performance across diverse settings.

In this study, we explore the differences in the performance and outcomes of HCPs participating in SIMBA sessions based on gender, country of work, and training grade. This study addresses this critical knowledge gap by examining the disparities among training grades within a single context and how the geographic context of HICs and LMICs influences these disparities. The inclusion of gender as a parameter is aligned with the aspiration to tailor medical education to the diverse needs of HCPs, while also acknowledging the broader societal implications of gender-sensitive health care delivery. The study's findings hold the potential to influence medical training practices, curriculum development, and professional mentorship in ways that are more attuned to the nuances of gender-specific experiences and perspectives.

Methods

Conducting SIMBA Sessions

The preparation and delivery of SIMBA sessions, including detailed flowcharts and examples, have been described in detail in our previous publication [12]. Each SIMBA session featured 4-6 clinical case scenarios representing a range of medical presentations commonly encountered in secondary care. These sessions were promoted via social media, junior doctor bulletins, and supporting organizations' websites.

The case scenarios were adapted from real-life cases with all patient-identifiable information removed. They included

comprehensive details such as presenting complaints, medical history, examination findings, clinical observations, investigation results (eg, blood tests and imaging), differential diagnoses, management strategies, and follow-up plans. Experts chairing the sessions rigorously reviewed and approved the case transcripts to ensure scientific accuracy and alignment with current medical guidelines.

Participation in the sessions was voluntary and free of charge. Participants registered in advance by providing their email addresses and WhatsApp numbers. Emails were used for presession communication, while WhatsApp facilitated real-time interaction during the simulations, providing access to all necessary session details and resources.

Before the simulation, participants completed a presimulation survey, which included informed consent, sociodemographic data collection, and self-assessment of their confidence in managing various clinical scenarios. This survey established a baseline measure of their confidence levels, covering both the session scenarios and the similar medical presentations.

During the simulation, HCPs from HICs and LMICs participated concurrently via WhatsApp. Moderators guided participants through the scenarios, encouraging them to interact as they would during real-life patient encounters. WhatsApp was also used to share key resources, such as links to surveys, investigations, and meeting invitations.

After completing all simulated cases, participants joined a Zoom meeting led by an expert. This interactive session provided an opportunity to discuss the cases, ask questions, and gain additional insights. Experts facilitated peer-to-peer discussions in the Zoom chat, fostering collaborative learning and ensuring a thorough exploration of each case.

After the simulation, participants completed a postsimulation survey, which included the confidence-rating questions from the presurvey to assess changes in confidence levels. The survey also collected feedback on their experience, highlighting successful aspects of the session and identifying areas for improvement.

WhatsApp interactions were recorded and later analyzed. Participants' performance was objectively assessed using a GRS [22], validated by experts prior to the session. The scale evaluated 6 domains: eliciting history, physical examination, investigations and diagnostic tests, interpretation of findings, clinical judgment, and management and follow-up plans. Each domain was scored on a Likert-type scale from 1 (not done) to 5 (excellent) [23], with total scores ranging from 6 to 30. This assessment measured participants' knowledge, accuracy, and thoroughness in handling the cases.

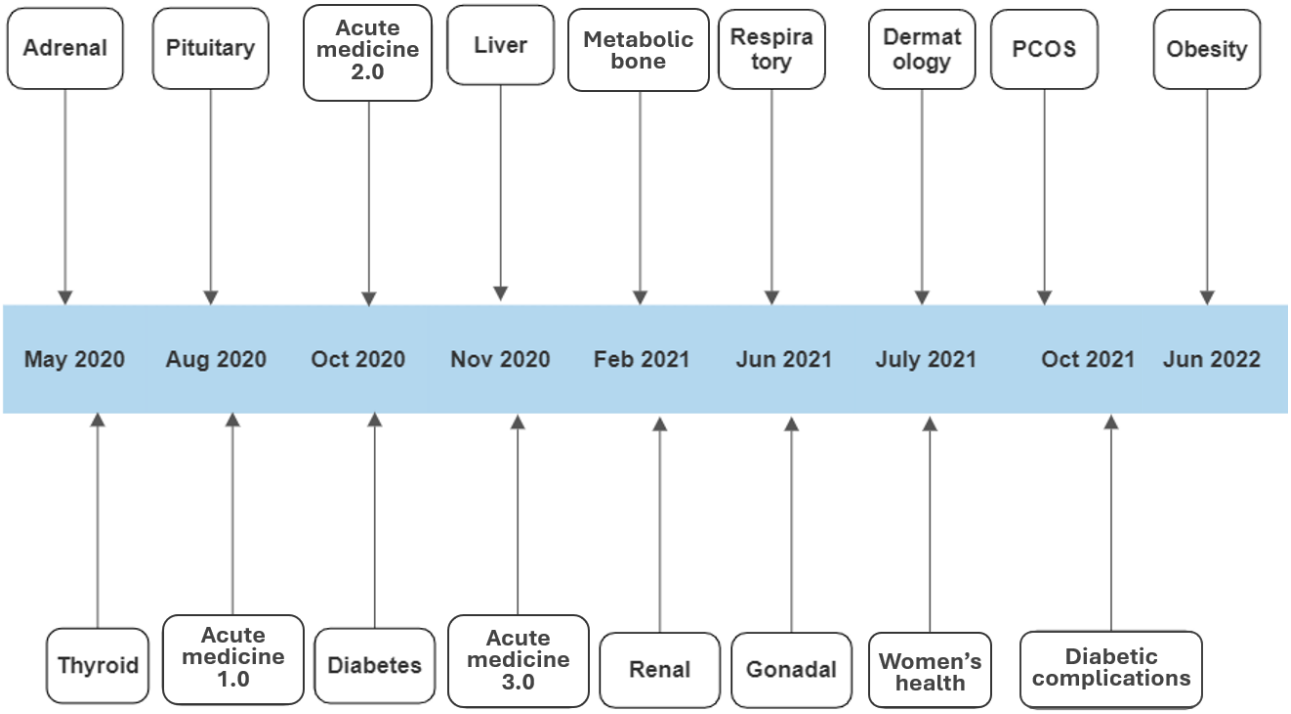
In addition, moderators provided personalized feedback using Pendleton's feedback model [24,25]. Feedback highlighted participants' strengths and offered constructive guidance for improvement, aiming to enhance their skills and confidence in managing similar scenarios in the future.

Data Collection

This study assessed participants in 17 SIMBA sessions from May 2020 to June 2022. The topics simulated included a wide range of medical specialties, as seen in Figure 1. All participants were invited to complete both the pre-SIMBA and post-SIMBA surveys voluntarily. The pre-SIMBA survey, distributed just before the session began, gathered basic sociodemographic information and self-reported confidence levels in managing simulated cases. The post-SIMBA survey, shared immediately after the expert case discussion, included similar questions on self-reported confidence in managing these cases. It also sought participants' feedback on their experience in the session. Each survey was administered at a single time point rather than as part of a test-retest design.

In addition, following each SIMBA session, WhatsApp chats containing the participants' approach to the simulated scenarios were graded using an adapted version of the GRS [22], which is reviewed by experts to confirm their appropriateness for the simulated case. Our assessment tool consisted of 6 domains (eliciting history; physical examination; investigations, diagnostic tests, and imaging; interpretation of investigations and imaging; clinical judgment; and management and follow-up or discharge plan), which are rated using a Likert-type scale of 1 (not done) to 5 (excellent) [23]. Moderators calculated the participant's score between a minimum of 6 and a maximum of 30, measuring the participant's knowledge, accuracy, and completeness in approaching the simulated case and assessing the patient. In addition to a numerical score, moderators provided participants with written feedback based on Pendleton's feedback model [24,25]. Pendleton's model is a structured approach to feedback that aims to create a positive and constructive learning environment. The model emphasizes a balanced feedback process by first highlighting the learner's strengths, then identifying areas for improvement, and finally engaging the learner in self-reflection. This structured feedback process encourages active learning and helps participants gain a deeper understanding of their performance, making it an effective tool for clinical education. GRS scores and feedback were emailed to participants 24 - 48 hours after the session. In the case of any queries, participants were provided with the contact details of the SIMBA team. To ensure uniformity and fairness across scoring, all moderators attended training sessions where they were taught how to use our assessment tool by experienced moderators. Any uncertainties that arised during the scoring process were resolved by discussion with experienced moderators.

Figure 1. Timeline of SIMBA (Simulation via Instant Messaging for Bedside Application) sessions held from 2020 to 2022. PCOS: polycystic ovary syndrome.



Statistical Analysis

Participants who completed both pre-SIMBA and post-SIMBA surveys were included in the analysis. This pre-post test design with presurvey and postsurvey specifically focused on assessing changes over time rather than the consistency of responses. Participants self-reported their country of residence during surveys, and the data (countries) were grouped into HICs and LMICs according to their country of residence based on the 2022 World Bank report [26]. Participants were classified as junior, middle, and senior grades based on their work experience. Junior-, middle-, and senior-grade HCPs had 0 - 2 years, 2 - 5 years, and >5 years of work experience, respectively.

Unadjusted and adjusted means and 95% CIs of the scores for the 6 performance variables (history taking, physical examination, investigation, interpretation, judgment, and management) by economic classification (HIC or LMIC), level of training (junior, middle, or senior grade), and gender (male or female) were calculated using multiple linear regression models, adjusting for gender, country, training, and number of WhatsApp messages (continuous), as appropriate. The mean (95% CI) difference of the pairwise comparisons between each performance score was calculated using paired *t* tests. The *P* value for heterogeneity between the mean performance scores was calculated using likelihood ratio tests by using an ANOVA. Two-sided *P* values of <.05 were considered statistically significant. We did not adjust the *P* value for multiple comparisons. All statistical analyses were performed using the Stata statistical software package (version 27.0; StataCorp).

Ethical Considerations

SIMBA was approved as an educational modality for postgraduate medical education by the Health Education West Midlands Diabetes and Endocrinology Specialist Training Committee. The study was approved by the Science, Technology, Engineering and Mathematics Ethics Committee at the University of Birmingham (ERN_2023 - 0495 and ERN_2023 - 0544). The study complied with all relevant ethical guidelines and institutional regulations throughout its design, implementation, analysis, and dissemination. Informed consent was taken from each participant and participation was voluntary. Participants were provided with an information sheet outlining the purpose of the study, the estimated duration of their involvement, the nature of the data collected, the identity of the investigators, and the storage and retention details. All data were anonymized at the point of analysis and stored on a secure, restricted access online drive. No identifiable patient data were used. No financial or material compensation was offered for participation.

Results

Overview

A total of 289 participants completed both pre-SIMBA and post-SIMBA surveys in the 17 simulation sessions. A total of 281 participants provided their country of residence. These participants worked in 49 countries across 5 continents, of which 36.7% (18/49) were classified as HICs, whereas 63.2% (31/49) were LMICs. Most participants worked in HICs (HICs: 191/281, 67.9%; LMICs: 90/281, 32.0%). Overall, the mean (SD) scores for the cohort were history taking: 3.6 (SD 3.9), physical examination: 3.5 (SD 3.7), investigation: 3.3 (SD 3.5), interpretation: 2.6 (SD 2.9), judgment: 3.1 (SD 3.3), and management: 2.6 (SD 2.8) (Table 1).

The values with the footnotes are the mean (95% CIs) for that particular score. Other values and corresponding *P* values are the mean difference (95% CIs) for the pairwise comparisons between scores taken from a paired *t* test for the 264 participants

who had a value for all 6 outcome scores. The difference is for the performance measure on the column compared with the performance measure on the corresponding row.

Table . Mean and 95% CIs difference between various performance measures using the global rating score for those participating in Simulation via Instant Messaging for Bedside Application (SIMBA).

	History	*P* value	Physical examination	*P* value	Investigation	*P* value	Interpretation	*P* value	Judgment	*P* value	Management
History	3.7 (3.6 - 3.9)[a]										
Physical examination	0.1 (0.0 - 0.2)	.02	3.6 (3.5 - 3.7)[a]								
Investigation	0.3 (0.2 - 0.4)	<.01	0.2 (0.1 - 0.3)	<.01	3.4 (3.3 - 3.5)[a]						
Interpretation	1.0 (0.8 - 1.2)	<.01	0.9 (0.7 - 1.1)	<.01	0.7 (0.6 - 0.8)	<.01	2.7 (2.6 - 2.9)[a]				
Judgment	0.5 (0.4 - 0.6)	<.01	0.4 (0.3 - 0.5)	<.01	0.2 (0.1 - 0.3)	<.01	−0.5 (−0.6 to 0.3)	<.01	3.2 (3.1 - 3.3)[a]		
Management	1.0 (0.9 - 1.2)	<.01	0.9 (0.8 - 1.1)	<.01	0.7 (0.6 - 0.8)	<.01	0.0 (−0.1 to 0.2)	.52	0.5 (0.4 - 0.6)	<.01	2.7 (2.6 - 2.8)[a]

[a]These values are the mean (95% CIs) for that particular score.

Comparing the Performance Between Participants From HIC and LMIC

After adjusting for sex, training, and number of WhatsApp messages, there were statistically significant differences in performance that were identified in the categories of history taking (HIC vs LMIC: 3.8 [3.7 - 3.9] vs 3.5 [3.3 - 3.7]; *P*<.01), investigations (3.6 [3.5 - 3.7] vs 3.1 [2.9 - 3.2]; *P*<.01), clinical judgment (3.4 [3.3 - 3.5] vs 2.9 [2.7 - 3.1]; *P*<.01), and management (2.9 [2.7 - 3.0] vs 2.3 [2.1 - 2.5]; *P*<.01) (Table 2 and Multimedia Appendix 1). A slight difference, although not statistically significant, was also found in physical examination (3.7 [3.5 - 3.8] vs 3.5 [3.2 - 3.7]; *P*=.13) and interpretation (2.8 [2.6 - 3.0] vs 2.6 [2.3 - 2.9]; *P*=.23).

Table . Mean and 95% CIs of various performance measures using the global rating score for low- or middle- and high-income countries for those participating in the SIMBA (Simulation via Instant Messaging for Bedside Application).

Performance measure	Unadjusted and adjusted	Low and middle income	High income	*P* value[a]
History taking	Unadjusted	3.5 (3.2 - 3.7)	3.8 (3.7 - 4.0)	<.01
	Adjusted[a]	3.5 (3.3 - 3.7)	3.8 (3.7 - 3.9)	<.01
Physical examination	Unadjusted	3.4 (3.2 - 3.6)	3.7 (3.5 - 3.8)	.05
	Adjusted[a]	3.5 (3.2 - 3.7)	3.7 (3.5 - 3.8)	.13
Investigation	Unadjusted	3.1 (2.9 - 3.2)	3.6 (3.5 - 3.7)	<.01
	Adjusted[a]	3.1 (2.9 - 3.2)	3.6 (3.5 - 3.7)	<.01
Interpretation	Unadjusted	2.6 (2.3 - 2.9)	2.8 (2.6 - 2.9)	.44
	Adjusted[a]	2.6 (2.3 - 2.9)	2.8 (2.6 - 3.0)	.23
Judgment	Unadjusted	2.9 (2.7 - 3.1)	3.4 (3.3 - 3.5)	<.01
	Adjusted[a]	2.9 (2.7 - 3.1)	3.4 (3.3 - 3.5)	<.01

[a]Adjusting for country, level of training, and the number of WhatsApp messages

Comparing the Performance Between Participants Based on Their Training Grade

A total of 213 participants provided their training grade in pre-SIMBA and post-SIMBA surveys, of which we classified 28.2% (60/213) as junior grade, 55.9% (119/213) as middle grade, and 16% (34/213) as senior grade. Junior-grade participants scored significantly higher in history taking (junior vs middle vs senior: 4.2 [4.0 - 4.5] vs 3.7 [3.5 - 3.9] vs 3.7 [3.4 - 4.0]; *P*<.01). Otherwise, there was no significant difference across the training grade for the rest of the domains (physical examination: 4.0 [3.7 - 4.2] vs 3.7 [3.5 - 3.9] vs 3.5

[3.1 - 3.8]; *P*=.07; investigation: 3.5 [3.3 - 3.7] vs 3.4 [3.3 - 3.6] vs 3.6 [3.3 - 3.8]; *P*=.53; clinical interpretation: 2.6 [2.3 - 2.9] vs 2.5 [2.3 - 2.8] vs 3.1 [2.7 - 3.6]; *P*=.08; clinical judgment: 3.4 [3.1 - 3.6] vs 3.3 [3.1 - 3.4] vs 3.4 [3.1 - 3.7]; *P*=.63; and management: 2.8 [2.5 - 3.0] vs 2.7 [2.6 - 2.9] vs 2.8 [2.5 - 3.1]; *P*=.94) (Table 3 and Multimedia Appendix 2).

Table . Mean and 95% CIs of various performance measures using the global rating score by level of training for those participating in SIMBA (Simulation via Instant Messaging for Bedside Application).

Performance measure	Unadjusted and adjusted	Junior grade	Middle grade	Senior grade	*P* value[a]
History taking	Unadjusted	4.2 (4.0 - 4.5)	3.7 (3.6 - 3.9)	3.6 (3.3 - 3.9)	<.01
	Adjusted[a]	4.2 (4.0 - 4.5)	3.7 (3.5 - 3.9)	3.7 (3.4 - 4.0)	<.01
Physical examination	Unadjusted	3.9 (3.7 - 4.2)	3.7 (3.5 - 3.9)	3.5 (3.1 - 3.8)	.09
	Adjusted[a]	4.0 (3.7 - 4.2)	3.7 (3.5 - 3.9)	3.5 (3.1 - 3.8)	.07
Investigation	Unadjusted	3.5 (3.3 - 3.7)	3.5 (3.3 - 3.6)	3.5 (3.2 - 3.8)	.97
	Adjusted[a]	3.5 (3.3 - 3.7)	3.4 (3.3 - 3.6)	3.6 (3.3 - 3.8)	.53
Interpretation	Unadjusted	2.6 (2.3 - 3.0)	2.5 (2.3 - 2.8)	3.1 (2.7 - 3.5)	.06
	Adjusted[a]	2.6 (2.3 - 2.9)	2.5 (2.3 - 2.8)	3.1 (2.7 - 3.6)	.08
Judgment	Unadjusted	3.3 (3.1 - 3.6)	3.3 (3.2 - 3.5)	3.4 (3.1 - 3.7)	.96
	Adjusted[a]	3.4 (3.1 - 3.6)	3.3 (3.1 - 3.4)	3.4 (3.1 - 3.7)	.63
Management	Unadjusted	2.7 (2.5 - 3.0)	2.8 (2.6 - 3.0)	2.7 (2.4 - 3.0)	.87
	Adjusted[a]	2.8 (2.5 - 3.0)	2.7 (2.6 - 2.9)	2.8 (2.5 - 3.1)	.94

[a]Adjusting for sex, country, and the number of WhatsApp messages.

Comparing the Performance Between Participants Based on Their Gender

A total of 199 participants provided their gender identity in pre-SIMBA and post-SIMBA surveys (male: 83/199, 41.7%; female: 116/199, 58.3%). There were no statistically significant differences in GRS scores between male and female participants (male vs female; history taking: 3.7 [3.5 - 3.9] vs 3.8 [3.6 - 3.9]; *P*=.53; physical examination: 3.7 [3.5 - 3.9] vs 3.6 [3.4 - 3.7]; *P*=.30; investigation: 3.4 [3.3 - 3.6] vs 3.4 [3.3 - 3.6]; *P*=.93; clinical interpretation: 2.9 [2.7 - 3.2] vs 2.7 [2.5 - 2.9]; *P*=.12; clinical judgment: 3.3 [3.1 - 3.5] vs 3.3 [3.1 - 3.4]; *P*=.75; and management: 2.8 [2.6 - 3.0] vs 2.7 [2.5 - 2.8]; *P*=.42) (Table 4).

Table . Mean and 95% CIs of various performance measures using the global rating score by number of WhatsApp messages for those participating in SIMBA (Simulation via Instant Messaging for Bedside Application).

Performance measure	Unadjusted and adjusted	Men	Women	*P* value[a]
History taking	Unadjusted	3.7 (3.4 - 3.9)	3.8 (3.6 - 4.0)	.52
	Adjusted[a]	3.7 (3.5 - 3.9)	3.8 (3.6 - 3.9)	.53
Physical examination	Unadjusted	3.7 (3.5 - 3.9)	3.6 (3.4 - 3.8)	.35
	Adjusted[a]	3.7 (3.5 - 3.9)	3.6 (3.4 - 3.7)	.30
Investigation	Unadjusted	3.4 (3.3 - 3.6)	3.4 (3.3 - 3.6)	.89
	Adjusted[a]	3.4 (3.3 - 3.6)	3.4 (3.3 - 3.6)	.93
Interpretation	Unadjusted	3.0 (2.7 - 3.2)	2.7 (2.4 - 2.9)	.09
	Adjusted[a]	2.9 (2.7 - 3.2)	2.7 (2.5 - 2.9)	.12
Judgment	Unadjusted	3.3 (3.1 - 3.5)	3.3 (3.1 - 3.4)	.79
	Adjusted[a]	3.3 (3.1 - 3.5)	3.3 (3.1 - 3.4)	.75
Management	Unadjusted	2.8 (2.5 - 3.0)	2.7 (2.5 - 2.8)	.51
	Adjusted[a]	2.8 (2.6 - 3.0)	2.7 (2.5 - 2.8)	.42

[a]Adjusting for country, level of training, and the number of WhatsApp messages.

Discussion

Principal Findings

Overall, participants scored higher in history taking and physical examination skills, while their scores were lower in interpretation and management skills. This information is valuable for fine-tuning future simulation programs to focus on improving interpretation and management abilities.

GRS is a validated tool for assessing clinical competencies across 6 critical domains: eliciting history, physical examination, investigations and diagnostic tests, interpretation of findings, clinical judgment, and management and follow-up plans. Although the GRS was not used to evaluate participants' abilities before the curriculum, it served as an objective measure during the simulation to assess performance and track growth. Combined with presimulation and postsimulation surveys, it allowed us to analyze changes in confidence levels and skill application.

The Pendleton model was adopted to provide structured feedback for the GRS output. This approach starts with positive reinforcement by highlighting what participants did well and then constructive guidance on areas for improvement. This approach ensures supportive feedback and facilitates skill enhancement.

Demographic data were collected during the presimulation survey to contextualize participants' backgrounds and identify trends in performance or confidence changes across demographic groups. While these characteristics are static, integrating them into a pre-post framework enabled us to correlate them with dynamic variables such as confidence and skill development.

The significant differences in clinical performance scores between LMICs and HICs highlight the need for better medical education resources to bridge existing gaps in health care across the globe. A recent analysis by the General Medical Council has identified a substantial attainment gap between international medical graduates from LMIC and candidates from the United Kingdom working in the National Health Service [27]. Subsequent recommendations advise the need for tailored approaches to tackle inequality and ensure consistent quality of medical education to trainees. Free and accessible programs, such as SIMBA, can facilitate this.

The observed differences in clinical performance between HCPs from HICs and LMICs can be attributed to various factors. HCPs from HICs often benefit from better infrastructure, advanced technology, and comprehensive training programs. In contrast, LMIC professionals face limited access to resources, inadequate training opportunities, and contextual factors that may impact their clinical performance. The comparative analysis highlights the need for context-specific training and tailored approaches to health care education in LMICs. Of note, a commonly identified issue with technology-enhanced simulation is the cost of simulators and limited understanding of the use and limitations of artificial intelligence or embedded computer algorithms [28,29]. SIMBA bypasses these hurdles as it has minimal costs, and the sessions can be conducted virtually.

Whilst all training grades performed similarly across investigation, clinical judgment, and management domains, significant differences were noted in history taking and physical examination scores. The decrease in these scores following career progression suggests the need for simulation-based training for senior trainees to maintain the high quality of history taking and physical examination skills. These insights hold implications for refining medical education approaches and optimizing patient care delivery. In an era of global health care parity, comprehending training grade–related performance disparities is instrumental in fostering contextually apt and universally equitable medical training standards. A study validating the use of smartphone accelerometers in neurosurgical simulations found higher acceleration scores among junior doctors than among senior registrars and consultants, suggesting improved technical performance in junior HCPs [30]. However, our findings also contrast with findings from a study of the American Board of Internal Medicine patient satisfaction questionnaire, which found a systematic improvement in the quality of consultations in directly observed clinical skills in senior attending physicians compared with junior residents [31]. The decline in history and examination skills may stem from increased workload, encouraging reliance on diagnostic investigations over thorough assessments. Deskilling—reduced use of certain skills in practice—also contributes to a decline in history and examination skills. The potential reason for contrast with findings from various national and international educational boards likely arises from differences in settings; formal assessments occur in controlled environments, encouraging conscious performance. In contrast, SIMBA sessions reflect real-life scenarios without direct observation, emphasizing authentic application of skills, which may reveal different trends.

Recommendations by Health Education England highlight the growing role of simulation-based education in core medical training, including procedural and emergency presentation learning sessions [32]. Its widespread use and implementation of a clear framework across all National Health Service trusts may allow physicians to further develop their skills and knowledge in line with up-to-date guidelines throughout their careers.

We did not find a significant difference in GRS between male and female participants. This is consistent with previous findings, which showed no significant gender-related differences in simulation assessment scores among emergency medicine residents assessed against 2 simulation cases related to the emergency medicine curriculum [33]. Similarly, simulation-based learning has been shown to eliminate gender differences in a virtual surgical training course in endoscopic proficiency [34]. However, while gender differences in the simulation were investigated, the translation of simulated skill to physical performance in the endoscopy suite was not investigated. Consequently, further prospective studies are required to assess the impact of gender on clinical performance. Previous studies have further demonstrated the role of simulation-based education in developing interpersonal communication skills and teamwork among medical students and HCPs [35,36]. Therefore, simulation-based learning may

provide objective and standardized scenarios to help remove gender-based differences in clinical outcome scores.

While mobile simulation-based learning appears to be widely accessible, reliable web access and cost of access continue to pose a barrier to virtual learning in many LMICs [37]. This is a particular issue in remote areas of the world, which may be subject to unscheduled power cuts [38]. Although communication platforms such as WhatsApp serve as low-cost initiatives for delivering medical education, emphasis must be placed on data security and confidentiality [39]. Consequently, standardized guidelines concerning the use of instant messaging in health care education are required to ensure its sustainability [40].

Limitations

We did not assess the change in perception toward virtual education over time. This may have influenced the findings of experience and difficulties that the participants encountered with web-based education from both LMICs and HICs. Future longitudinal studies must determine whether subsequent clinical decision-making is improved in real-life scenarios. Brain drain, that is, migration of doctors usually from LMICs to HICs for better opportunities and other varied reasons, is another important factor that needs to be considered as this phenomenon may have skewed the results. As SIMBA sessions are currently conducted only in English, it is a potential language barrier. As all the data collected were self-reported rather than objectively verified or tested, it is an inherent limitation of the study.

Regarding the physical examination, participants requested specific types of examinations, such as general physical examination, abdominal examination, or limb examination, and the corresponding findings were provided to them. The virtual clinical physical examination was not performed by the participants themselves, which is a limitation of our study. As we collected minimal identifying data, we do not have further information on the career breaks taken by a doctor for any reason that may have led to inaccurate classification as a junior, middle, or senior-grade doctor. Moreover, given the growing role of virtual learning in medical education, future work should focus on user perception of the feasibility of remote simulation-based training in LMICs.

Conclusions

Future simulation programs should enhance interpretation and management skills by incorporating customized content, feedback mechanisms, and real-world scenarios for HCPs. The significant differences in clinical performance scores between LMICs and HICs highlight the need for better medical education resources to bridge existing gaps in health care globally. The decrease in some clinical competency scores following career progression could be addressed by simulation-based training to maintain the same quality of history taking and physical examination skills. These outcomes, including no gendered differences in simulation-based learning, hold profound implications for tailoring medical education strategies, fostering equitable training, and elevating patient care standards on a global scale. The need for targeted interventions and capacity-building efforts via context-specific training and tailored approaches to health care education is emphasized.

Acknowledgments

The authors thank all the health care professionals who participated in this study. They thank all the moderators, session chairs, and endorsing organizations for their contribution and continued support. During the preparation of this work the authors used ChatGPT to improve language and readability. After using this tool or service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Data Availability

Data will be made available upon request.

Authors' Contributions

KM and HB are the joint first authors, having contributed to all aspects of the study. FC analyzed and interpreted the data analysis. PK conceptualized and supervised the delivery of all study aspects and critically reviewed the manuscript. EW, VS, and DZ contributed substantially to delivering the SIMBA sessions, developing the study and writing the manuscript. SIMBA and CoMICs team contributed toward all stages of this manuscript. All authors contributed substantially to drafting and approving the final draft of the manuscript. The final version has been reviewed and approved by all the authors.

The *SIMBA and CoMICs Team* members are Carina Pan, Pavithra Sakthivel, Anisah Ali, Isabel Allison, Tamzin Ogiliev, Haaziq Sheikh, Sung Yat Ng, Zahra Olateju, Maiar Elhariry, Eka Melson, Sangamithra Ravi, Abby Radcliffe, Rachel Nirmal, Aditya Swaminathan, Shams Ali Baig, Dwi Delson, Soon Chee Yap, Vardhan Venkatesh, and Fazna Rahman.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Global Rating Scale score of HIC versus LMIC participants (statistically significant results in history taking, investigations, clinical judgment, and management). HIC: high-income country; LMIC: low- and middle-income country.
[PNG File, 30 KB - mededu_v11i1e52332_app1.png]

Multimedia Appendix 2
Global Rating Scale score of participants by training grade (statistically significant results in history taking).
[PNG File, 37 KB - mededu_v11i1e52332_app2.png]

References

1. Morton S, Pencheon D, Squires N. Sustainable development goals (SDGs), and their implementation: a national global framework for health, development and equity needs a systems approach at every level. Br Med Bull 2017 Dec 1;124(1):81-90. [doi: 10.1093/bmb/ldx031] [Medline: 29069332]
2. Shah R, Ahluwalia S. The challenges of understanding differential attainment in postgraduate medical education. Br J Gen Pract 2019 Sep;69(686):426-427. [doi: 10.3399/bjgp19X705161] [Medline: 31467003]
3. Kinfu Y, Dal Poz MR, Mercer H, Evans DB. The health worker shortage in Africa: are enough physicians and nurses being trained? Bull World Health Organ 2009 Mar;87(3):225-230. [doi: 10.2471/blt.08.051599] [Medline: 19377719]
4. Papapanou M, Routsi E, Tsamakis K, et al. Medical education challenges and innovations during COVID-19 pandemic. Postgrad Med J 2022 May;98(1159):321-327. [doi: 10.1136/postgradmedj-2021-140032] [Medline: 33782202]
5. Alsoufi A, Alsuyihili A, Msherghi A, et al. Impact of the COVID-19 pandemic on medical education: medical students' knowledge, attitudes, and practices regarding electronic learning. PLoS ONE 2020;15(11):e0242905. [doi: 10.1371/journal.pone.0242905] [Medline: 33237962]
6. Frehywot S, Vovides Y, Talib Z, et al. E-learning in medical education in resource constrained low- and middle-income countries. Hum Resour Health 2013 Feb 4;11:4. [doi: 10.1186/1478-4491-11-4] [Medline: 23379467]
7. Ahmady S, Kallestrup P, Sadoughi MM, et al. Distance learning strategies in medical education during COVID-19: a systematic review. J Educ Health Promot 2021;10:421. [doi: 10.4103/jehp.jehp_318_21] [Medline: 35071627]
8. Stojan J, Haas M, Thammasitboon S, et al. Online learning developments in undergraduate medical education in response to the COVID-19 pandemic: a BEME systematic review: BEME Guide No. 69. Med Teach 2022 Feb;44(2):109-129. [doi: 10.1080/0142159X.2021.1992373] [Medline: 34709949]
9. Al-Elq AH. Simulation-based medical teaching and learning. J Family Community Med 2010 Jan;17(1):35-40. [doi: 10.4103/1319-1683.68787] [Medline: 22022669]
10. Melson E, Chen W, Zhou D, et al. Adaptation and use of media in an innovative simulation-based clinician training programme. BMJ Simul Technol Enhanc Learn 2021;7(6):650-652. [doi: 10.1136/bmjstel-2020-000808] [Medline: 35520955]
11. Rosen KR. The history of medical simulation. J Crit Care 2008 Jun;23(2):157-166. [doi: 10.1016/j.jcrc.2007.12.004] [Medline: 18538206]
12. Melson E, Davitadze M, Aftab M, et al. Simulation via instant messaging—Birmingham advance (SIMBA) model helped improve clinicians' confidence to manage cases in diabetes and endocrinology. BMC Med Educ 2020 Aug 18;20(1):274. [doi: 10.1186/s12909-020-02190-6] [Medline: 32811488]
13. Wallett L, Chen W, Thomas L, et al. Developing a simulation-based learning model for acute medical education during COVID-19 pandemic with Simulation via Instant Messaging - Birmingham Advance (SIMBA). BMJ Open Qual 2022 Apr;11(2):e001565. [doi: 10.1136/bmjoq-2021-001565] [Medline: 35396253]
14. Davitadze M, Ooi E, Ng CY, et al. SIMBA: using Kolb's learning theory in simulation-based learning to improve participants' confidence. BMC Med Educ 2022 Feb 22;22(1):116. [doi: 10.1186/s12909-022-03176-2] [Medline: 35193557]
15. Morgan G, Melson E, Davitadze M, et al. Utility of Simulation via Instant Messaging—Birmingham Advance (SIMBA) in medical education during COVID-19 pandemic. J R Coll Physicians Edinb 2021 Jun;51(2):168-172. [doi: 10.4997/JRCPE.2021.218] [Medline: 34131679]
16. Evans N, Davitadze M, Narendran A, et al. SIMBA as an alternative and/or an adjunct to pre-medical work experience during the COVID-19 pandemic. Future Healthc J 2021 Mar;8(1):e142-e145. [doi: 10.7861/fhj.2020-0219] [Medline: 33791494]
17. Ng CY, Allison I, Ooi E, Davitadze M, Melson E, Kempegowda P. Medical students' and junior doctors' leadership and teamwork skills improved after involvement with Simulation via Instant Messaging—Birmingham Advance (SIMBA). BMJ Lead 2022 Sep;6(3):233-236. [doi: 10.1136/leader-2021-000486] [Medline: 36170479]
18. Malhotra K, Ali A, Soran V, et al. Levelling the learning ground for healthcare professionals across the world through SIMBA: a mixed-methods study. BMJ Open 2023 Jul 10;13(7):e069109. [doi: 10.1136/bmjopen-2022-069109] [Medline: 37429686]
19. Adams LV, Wagner CM, Nutt CT, Binagwaho A. The future of global health education: training for equity in global health. BMC Med Educ 2016 Nov 21;16(1):296. [doi: 10.1186/s12909-016-0820-0] [Medline: 27871276]
20. Williams JS, Walker RJ, Egede LE. Achieving equity in an evolving healthcare system: opportunities and challenges. Am J Med Sci 2016 Jan;351(1):33-43. [doi: 10.1016/j.amjms.2015.10.012] [Medline: 26802756]

21. Kruk ME, Gage AD, Arsenault C, et al. High-quality health systems in the Sustainable Development Goals era: time for a revolution. Lancet Glob Health 2018 Nov;6(11):e1196-e1252. [doi: 10.1016/S2214-109X(18)30386-3] [Medline: 30196093]
22. Gerard JM, Kessler DO, Braun C, Mehta R, Scalzo AJ, Auerbach M. Validation of global rating scale and checklist instruments for the infant lumbar puncture procedure. Simul Healthc 2013 Jun;8(3):148-154. [doi: 10.1097/SIH.0b013e3182802d34] [Medline: 23388627]
23. Joshi A, Kale S, Chandel S, Pal DK. Likert scale: explored and explained. BJAST 2015;7(4):396-403. [doi: 10.9734/BJAST/2015/14975]
24. Burgess A, van Diggele C, Roberts C, Mellis C. Feedback in the clinical setting. BMC Med Educ 2020 Dec 3;20(Suppl 2):460. [doi: 10.1186/s12909-020-02280-5] [Medline: 33272265]
25. Pendleton D. The Consultation: An Approach to Learning and Teaching: No6: Oxford University Press; 1984.
26. World bank country and lending groups. World Bank Data Help Desk. 2022 Mar. URL: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519 [accessed 2022-09-04]
27. Tackling disadvantage in medical education. Health Education England. 2023 Mar. URL: https://www.gmc-uk.org/-/media/documents/96887270_tackling-disadvantage-in-medical-education-020323.pdf [accessed 2023-04-24]
28. Kansal R, Bawa A, Bansal A, et al. Differences in knowledge and perspectives on the usage of artificial intelligence among doctors and medical students of a developing country: a cross-sectional study. Cureus 2022 Jan;14(1):e21434. [doi: 10.7759/cureus.21434] [Medline: 35223222]
29. Barteit S, Guzek D, Jahn A, Bärnighausen T, Jorge MM, Neuhann F. Evaluation of e-learning for medical education in low- and middle-income countries: a systematic review. Comput Educ 2020 Feb;145(103726):103726. [doi: 10.1016/j.compedu.2019.103726] [Medline: 32565611]
30. Jensen Ang WJ, Hopkins ME, Partridge R, et al. Validating the use of smartphone-based accelerometers for performance assessment in a simulated neurosurgical task. Neurosurgery 2014 Mar;10 Suppl 1(57–64):57-64. [doi: 10.1227/NEU.0000000000000010] [Medline: 23756748]
31. Boudreau D, Tamblyn R, Dufresne L. Evaluation of consultative skills in respiratory medicine using a structured medical consultation. Am J Respir Crit Care Med 1994 Nov;150(5 Pt 1):1298-1304. [doi: 10.1164/ajrccm.150.5.7952556] [Medline: 7952556]
32. Enhancing UK Core Medical Training through simulation-based education: an evidence-based approach. Joint JRCPTB/HEE Expert Group. 2018 Oct 4. URL: https://www.jrcptb.org.uk/sites/default/files/HEE_Report_FINAL.pdf [accessed 2023-08-23]
33. Siegelman JN, Lall M, Lee L, Moran TP, Wallenstein J, Shah B. Gender bias in simulation-based assessments of emergency medicine residents. J Grad Med Educ 2018 Aug;10(4):411-415. [doi: 10.4300/JGME-D-18-00059.1] [Medline: 30154972]
34. Ritter EM, Lineberry M, Hashimoto DA, et al. Simulation-based mastery learning significantly reduces gender differences on the fundamentals of endoscopic surgery performance exam. Surg Endosc 2018 Dec;32(12):5006-5011. [doi: 10.1007/s00464-018-6313-y] [Medline: 30014324]
35. Blackmore A, Kasfiki EV, Purva M. Simulation-based education to improve communication skills: a systematic review and identification of current best practice. BMJ Simul Technol Enhanc Learn 2018;4(4):159-164. [doi: 10.1136/bmjstel-2017-000220] [Medline: 35519010]
36. Amsalem D, Gothelf D, Soul O, Dorman A, Ziv A, Gross R. Single-day simulation-based training improves communication and psychiatric skills of medical students. Front Psychiatry 2020;11:221. [doi: 10.3389/fpsyt.2020.00221] [Medline: 32265762]
37. Nyemike Simeon A, Abdulmujeeb Babatunde A, Lukman Abiodun N, Omogbadegun Olu R, Ido Emem A. Uptake, barriers, and determinants of e-learning among university students in selected low income countries in Sub-Saharan Africa amidst the COVID-19 disruption: an online survey. Adv Med Educ Pract 2022;13(609–17):609-617. [doi: 10.2147/AMEP.S357677] [Medline: 35707204]
38. Ukrani RD, Shaikh AN, Martins RS, Fatima SS, Naseem HA, Baig MA. Low-cost peer-taught virtual research workshops for medical students in Pakistan: a creative, scalable, and sustainable solution for student research. BMC Med Educ 2021 Nov 1;21(1):557. [doi: 10.1186/s12909-021-02996-y] [Medline: 34724950]
39. Mars M, Morris C, Scott RE. WhatsApp guidelines—what guidelines? A literature review. J Telemed Telecare 2019 Oct;25(9):524-529. [doi: 10.1177/1357633X19873233] [Medline: 31631763]
40. Luna D, Almerares A, Mayan JC 3rd, González Bernaldo de Quirós F, Otero C. Health informatics in developing countries: going beyond pilot practices to sustainable implementations: a review of the current challenges. Healthc Inform Res 2014 Jan;20(1):3-10. [doi: 10.4258/hir.2014.20.1.3] [Medline: 24627813]

Abbreviations

HCPs: health care professionals
HICs: high-income countries
LMICs: low- and middle-income countries
SIMBA: Simulation via Instant Messaging for Bedside Application

Edited by B Lesselroth; submitted 31.08.23; peer-reviewed by C Manlhiot, E Ogut, F Ismail, M Davitadze; revised version received 09.12.24; accepted 25.02.25; published 11.08.25.

Please cite as:
Malhotra K, Balakrishnan H, Warmington E, Soran V, Crowe F, Zhou D, SIMBA AND CoMICs Team , Kempegowda P
Global Disparities in Simulation-Based Learning Performance: Serial Cross-Sectional Mixed Methods Study
JMIR Med Educ 2025;11:e52332
URL: https://mededu.jmir.org/2025/1/e52332
doi:10.2196/52332

Health Care Professionals' Knowledge, Attitude, Practice, and Infrastructure Accessibility for e-Learning in Ethiopia: Cross-Sectional Study

Sophie Sarah Rossner[1]; Muluken Gizaw[1,2,3], PhD; Sefonias Getachew[1,2,3], PhD; Eyerusalem Getachew[1,2,3]; Alemnew Destaw[1,2,3]; Sarah Negash[1]; Lena Bauer[1], MD; Eva Susanne Marion Hermann[1]; Abel Shita[1]; Susanne Unverzagt[1,4], PhD; Pablo Sandro Carvalho Santos[1], PhD; Eva Johanna Kantelhardt[1,5], Prof Dr Med; Eric Sven Kroeber[1], MD

[1]Global and Planetary Health Working Group, Institute of Medical Epidemiology, Biometrics, and Informatics, Center of Health Sciences, Medical Faculty of the Martin Luther University Halle-Wittenberg, Magdeburger Str. 8, Halle (Saale), Germany

[2]Department of Epidemiology and Biostatistics, School of Public Health, Addis Ababa University, Addis Ababa, Ethiopia

[3]NCD working group, School of Public Health, Addis Ababa University, Addis Ababa, Ethiopia

[4]Institute of General Practice and Family Medicine, Center for Health Sciences, Martin-Luther-University Halle-Wittenberg, Halle (Saale), Germany

[5]Department of Gynecology, Martin-Luther-University Halle-Wittenberg, Halle (Saale), Germany

Corresponding Author:
Eric Sven Kroeber, MD
Global and Planetary Health Working Group, Institute of Medical Epidemiology, Biometrics, and Informatics, Center of Health Sciences, Medical Faculty of the Martin Luther University Halle-Wittenberg, Magdeburger Str. 8, Halle (Saale), Germany

Abstract

Background: Training of health care professionals and their participation in continuous medical education are crucial to ensure quality health care. Low-resource countries in Sub-Saharan Africa struggle with health care disparities between urban and rural areas concerning access to educational resources. While e-learning can facilitate a wide distribution of educational content, it depends on learners' engagement and infrastructure.

Objective: This study aims to assess knowledge, attitude, practice, and access to infrastructure related to e-learning among health care professionals in primary health care settings in Ethiopia.

Methods: In April 2023, we carried out a quantitative, questionnaire-based cross-sectional study guided by the knowledge, attitudes, and practice framework, including additional items on available infrastructure. The scores in each category are defined as "high" and "low" based on the median, followed by the application of logistic regression on selected sociodemographic factors. We included health care professionals working in general and primary hospitals, health centers, and health posts.

Results: Of 398 participants (response rate 94.5%), more than half (n=207, 52%) reported feeling confident about their understanding of e-learning and conducting online searches, both for general (n=247, 62.1%) and medical-related content (n=251, 63.1%). Higher levels of education were associated with better knowledge (adjusted odds ratio [AOR] 2.32, 95% CI 1.45-3.68). Regardless of financial and personal efforts, we observed a generally positive attitude. Almost half of the participants (n=172, 43.2%) reported using the internet daily, compared to 16.8% (n=67) of participants who never used the internet. Higher education (AOR 2.56, 95% CI 1.57-4.16) and income levels (AOR 1.31, 95% CI 1.06-1.62) were associated with higher practice scores of e-learning-related activities. Women, however, exhibited lower practice scores (AOR 0.44, 95% CI 0.27-0.71). Regular access to an internet-enabled device was reported by 43.5% (n=173) of the participants. Smartphones were the primarily used device (268/393, 67.3%). Common barriers to internet access were limited internet availability (142/437, 32.5%) and costs (n=190, 43.5%). Higher education (AOR 1.56, 95% CI 0.98, 2.46) and income (AOR 1.50; 95% CI 1.21-1.85) were associated with increased access to infrastructure, while it was decreased for women (AOR 0.48, 95% CI 0.30-0.77).

Conclusions: Although Ethiopian health care professionals report mixed levels of knowledge, they have a positive attitude toward e-learning in medical education. While internet use is common, especially via smartphone, the access to devices and reliable internet is limited. To improve accessibility, investments in the digital infrastructure and individual digital education programs are necessary, especially targeting women and those with lower income. Due to their widespread availability, e-learning programs should be optimized for smartphones.

(JMIR Med Educ 2025;11:e65598) doi:10.2196/65598

KEYWORDS

e-learning; electronic learning; health care professional; education; medical education; medical learning; physician; doctor; primary care; primary health care; digital; digital health; digital technology; digital intervention; Ethiopia; Ethiopian; cross-sectional study

Introduction

Background

Importance of Continuous Professional Development

Medical education is a fundamental part of a country's health care system [1]. Continuous professional development (CPD) and continuous medical education are used to regularly update and enhance health care professionals' knowledge and skills [2,3]. While CPD programs are seen as beneficial in low-and middle-income countries (LMICs) [4], their availability is often limited due to financial and personnel resources. This shortage of training opportunities for health care professionals affects the quality of medical care and sets patients' health at risk [5].

e-Learning as a Globally Endorsed Strategy

The World Health Organization (WHO) endorses e-learning as a key component of their learning strategy, with the goal of reaching 3 million health care professionals by 2028 through approximately 260 courses. This offer leverages advanced multilingual technology and strategic partnerships to ensure accessible, culturally relevant, and impactful health education worldwide [6]. The WHO has established the "WHO Virtual Campus," which offers tailored training programs, emphasizing the importance of conducting thorough needs assessments involving local health authorities and stakeholders prior to implementing e-learning initiatives [7]. The United Nations Online Learning Framework emphasizes the importance of understanding learners' background, experiences, motivations, and infrastructural conditions before implementing online learning [8].

Advantages of e-Learning

The advantages of e-learning include improving the access to education [9], particularly in geographically or economically isolated areas [10], and being cost-effective, easy to update while improving the user's performance and knowledge [11,12]. In addition, as a CPD strategy, e-learning has proven effective in enhancing both knowledge and procedural skills. Studies indicate that medical, surgical, and pharmaceutical (partly blended) e-learning courses received high levels of satisfaction and were considered useful by participants, even in low-resource settings [13-15]. e-Learning approaches have advanced health care quality and accessibility on a global scale [16].

Barriers in LMICs and Africa

Nevertheless, e-learning approaches are not universally suitable for all learners due to the diverse and individual nature of learning styles [17]. Their implementation often faces challenges in many LMICs with unique aspects in Africa compared to other parts of the world [18]. One of the main hurdles to implementing e-Learning projects on the continent is the low internet coverage of 43.2%, compared to 66% worldwide (2021) [18-20]. In Ethiopia, the internet penetration rate has slightly increased in recent years, reaching 19.4% in 2024 [6]. There is also a strong disparity in access to grid electricity between rural and urban areas. While almost all urban residents (95.7%) have access, only 43.6% of residents in rural areas are covered (2023 [21]). The lack of technical devices further disadvantages rural populations compared to urban residents [22], and as educational strategies shift toward distance learning, individuals in rural areas face even greater challenges in accessing education [23,24]. Further common barriers are costs, a lack of digital skills and knowledge, and computer fear [11,25-28]. The literature notes that Ethiopian educational institutions and students are insufficiently prepared for e-learning and have only used these platforms to a limited extent. This is mainly due to their focus on traditional education systems, which has resulted in a lack of efficient e-learning approaches. As a consequence, the development and adoption of modern digital learning methods remain restricted [21,22]. A systematic review (published in 2023) of challenges associated with e-learning in Sub-Saharan Africa noted that existing problems remained similar since 2016 without notable structural improvements [29].

Implementation Efforts in Africa

Several studies report on piloting and partly adopting e-learning in medical education. In Uganda, despite relevant efforts, nationwide implementation remains limited, with progress largely dependent on the commitment of individual institutions and educators, leading to uneven advancement across the country [30]. Many e-learning pilot projects are insufficiently adapted to local environments in LMICs [1], for example, addressing unreliable access and limited technological infrastructure [31]. This includes the usage of appropriate digital platforms and providing user support for a successful learning experience. Despite this, recruiting skilled personnel capable of managing and using e-learning effectively requires additional financial investment [9]. A baseline and needs assessment of the target group and area, as well as pretests, can identify strengths and limitations of e-learning tools and enable their adjustment to specific settings. Since e-learning is anticipated to be used increasingly in medical education worldwide, efforts are needed to be worthwhile to understand the target groups [32].

Rationale

The knowledge, attitudes, and practices (KAP) framework is a widely used approach to assess target groups in the context of public health [33]. It can be used to identify educational gaps and inform the design of targeted interventions, such as e-learning strategies. KAP studies provide a valuable foundation for tailoring digital learning tools to the specific needs, perceptions, and practices of diverse user groups. In our literature review, we did not find publications from LMICs or Sub-Saharan Africa specifically concerning the perspective of health care professionals regarding e-learning, highlighting the

need to address this gap [34]. The existing literature mainly focuses on students enrolled in universities and colleges during the COVID-19 pandemic [28,35,36], as well as individuals among the general population [37].

Objective

Our study aims to assess knowledge, attitude, practice, and access to infrastructure (KAP-Infrastructure) related to e-learning courses among health care professionals working at the primary health care level in Central and Southern Ethiopia. Our study provides an opportunity to offer insights and address the existing gaps and missing information in this area and setting.

Methods

Overview

We conducted a data collector–supported self-administered cross-sectional survey among health care professionals working as general practitioners, health officers, nurses, midwives, or health extension workers in 45 health facilities between April 3 and 28, 2023. The study was set in the Oromia region and the former Southern Nations, Nationalities, and Peoples Region (SNNPR), which was reorganized in August 2023 after the data collection. The study regions are now part of the South Ethiopia Regional State and the Central Ethiopia Regional State. More than 85% of residents in Oromia and 89% of residents in the former SNNPR live in rural areas [38,39]. The selection of the health care facilities (2 general hospitals, 3 primary hospitals, 14 health centers, and 26 health posts) was based on partner settings in a collaborative cervical and breast cancer care project between the Addis Ababa University School of Public Health, Ethiopia and the Martin Luther University Halle-Wittenberg, Germany. Ethiopia has a 3-tier health care system with specific responsibilities on each level to enable continuous care in the country (Multimedia Appendix 1). At the primary health care level, health posts, health centers, and primary hospitals deliver essential health care services, particularly in rural areas [23], where 77.4% of the Ethiopian population is resided. General and referral hospitals, as part of the secondary health care level, provide a wider range of medical services and treatments. Specialized hospitals form the tertiary level and provide extensive services for comprehensive health care [40]. We included 2 general hospitals as part of the secondary health care level to perform a sensitivity analysis comparison between health care levels (see Multimedia Appendix 1). Our study area is located approximately 150 km to the south and southwest of the capital, Addis Ababa, in a radius of about 100 km.

Outcome Variables

We investigated the 4 KAP-Infrastructure constructs, building items based on the following definitions (1) “Knowledge” assessed the health care professionals’ self-perceived practical abilities related to e-learning tasks. (2) “Attitude” represented participants’ motivation and expectations toward e-learning to support their professional education. (3) “Practice” concerned the conduct of activities related to e-learning. Finally, (4) “Access to infrastructure” covered individual and work-related infrastructural conditions of the health care professionals.

Questionnaire

The questionnaire contained questions on sociodemographic characteristics as well as the participants’ perspectives on e-learning based on the KAP model [33]. The KAP items had 5-step Likert scale answering options [41].

Besides using the United Nations Online Learning Framework as an orientation, we covered relevant aspects of e-learning by using items described in a systematic review outlining enablers and barriers affecting e-learning in health sciences [11]. We added a category “access to infrastructure” (KAP-Infrastructure) as a relevant aspect for e-learning conduct. We based this section on the European Union’s “Community survey on ICT usage in households and by individuals (2020)” [42], with additional items adapted from a community survey on information and communication technology [35,37] and a KAP study on online learning by college students in India [35].

The questionnaire contained 10 sociodemographic, 10 knowledge, 13 attitude, 8 practice, and 7 infrastructure items. It was initially developed in English and then translated into the Amharic and Afan Oromo languages. A blind back-translation into English ensured accuracy. Discrepancies were addressed and discussed. The main translation process involved 5 native speakers of Amharic and Afan Oromo, all of whom were proficient in English and had a medical background.

Due to the lack of a topically fitting validated research instrument, the questionnaire was developed by an interprofessional team of health scientists from Martin Luther University Halle-Wittenberg and Addis Ababa University based on comprehensive initial literature research. We performed pretests according to the Concurrent Think Aloud Method and adapted the questionnaire accordingly, ensuring its comprehensibility, suitability, and validity. Two rounds of pretests were conducted in Addis Ababa in March 2023, including 38 health care professionals in 2 health centers. These participants represented approximately 9% of the overall sample, closely meeting the targeted 5% per health care center. The items were tested for understandability and clarity in the Amharic and Afan Oromo languages. The received feedback was then reviewed and discussed by interdisciplinary teams in both Germany and Ethiopia, resulting in the exclusion or rephrasing of certain questions and answer choices, as well as modifications to the consent section. The process was guided by “Surveys and Questionnaires in Health Research” by Schofield and colleagues [43,44].

Sample Size and Sampling Technique

Before collecting the data, the sample size was calculated using the single population proportion formula. The level of KAP of health care professionals on e-learning in Ethiopia has not been studied yet. Therefore, the sample size was calculated with the some assumptions, including 50% proportion of the outcome among health care professionals, 5% marginal error, and 95% CI. Under these circumstances, the sample size is 384, and after adding a 10 % nonresponse rate, the final sample size was 423. For comparison, we also calculated the sample size to estimate a 2-sided CI for one proportion for the target variable “internet access” as a representation for the objective “infrastructure.”

With this frame we calculated a sample size of 395, including a 10% nonresponse rate. Since the calculated sample size of infrastructure was smaller than the sample size of KAP, we decided to achieve a larger sample size of 423. We selected participating health care professionals working in the affiliated health facilities using the convenience sample method and reached a final sample size of 398 participants (response rate 94.1%). All health care professionals, including general practitioners, health officers, nurses, midwives, and health extension workers present during data collection, were included in the study. However, those who were unavailable at that time and any professionals not listed above were excluded from the study. Ensuring a balanced participation between all occupations, we considered a proportional allocation for each health facility.

Data Analysis

First, we conducted basic descriptive statistics. After building sum scores for each KAP-Infrastructure category, the median was used to define "low" and "high" scores in each category, which were considered as binary response variables in downstream analyses (Multimedia Appendix 2). In order to detect potential monotonic relationships (collinearity) among predictor variables, we used Spearman rank correlation coefficient (ρ) [45] (refer Multimedia Appendix 3). This preliminary diagnostic analysis yielded a high correlation between the variables "health facility" and "occupation" (ρ=0.982). As a ρ above 0.9 is generally considered very high, we proceeded to remove the variable "occupation" from all multivariate logistic regressions in order to avoid redundancy and ensure convergence of statistical modeling. The remaining 7 explanatory variables were kept in all multivariate logistic regressions. We divided the participants into 5 similar-sized income groups (monthly income less than US $94, US $94 - US $115, US $115 - US $130, US $130 - US $151.2, and more than US $151.2 USD), according to the World Bank's country income classifications [46]. We applied a more precise breakdown that is often used in household or individual-level analyses by adding middle income. The income thresholds were established to produce comparable income ranges across different income-level groups. We defined regular usage of the internet and internet access as an available internet service in terms of messaging, browsing, downloading, and purchasing in the last 3 months for at least once a week [47-49]. Data imputation was carried out based on Spearman correlation indices among the variables. It was done for 0.7% of data points in the "religion" and "sex" variables and for 3.8% of data points in the "income" variable. The correlation indices among the components of the dependent variable (KAP-Infrastructure scores) were used to allow imputation as well, which led to 1.1% of data points in the dependent variables being imputed. We then investigated the relationship between KAP-Infrastructure score and multiple explanatory factors (age, sex, education level, income, and type of health facility) by calculating odds ratios (ORs) and adjusted ORs (AORs) and the corresponding 95% CI with a logistic regression, using SPSS version 28 (IBM Corp) [50]. An OR greater than 1 indicates that the factor is associated with higher KAP-Infrastructure scores. We tested each category for internal consistency by calculating Cronbach α (knowledge 0.897; attitude: 0.688; practice: 0.899; access to infrastructure: 0.810) [51]. The results indicated acceptable or high internal consistency. Furthermore, we assessed the fit of a logistic regression model using the Hosmer-Lemeshow test [52].

Ethical Considerations

We received approval from the Addis Ababa University Research Ethics Committee in Addis Ababa, Ethiopia (Prev: 423), and the ethics committee of the Martin Luther University Halle-Wittenberg-Medical Faculty (processing number 2023 - 003). Participant information was collected anonymously, and written consent was obtained in the questionnaire. Participants did not receive any compensation.

Results

Overview

We included a total of 398 participants (response rate, 94.1%). Overall, 25 health care professionals declined due to workload, unwillingness, or undisclosed reasons. Male (195/394, 49.5%) and female (199/394, 50.5%) participants were equally represented. The mean age was 31.0 (SD 7.0) years. Half of the health care professionals were nurses (200/398, 50.3%). In total, 86.7% (345/398) of the participants chose the Amharic version of the questionnaire. The mean income per month was US $138 (SD $57). Among the participants, 4 out of 5 worked in primary facilities (316/398, 79.4%; Table 1).

Table . Sociodemographic information.

Characteristic and category	Value
Region (N=398), n (%)	
SNNPR[a]	252 (63.3)
Oromia region	146 (36.7)
Type of health facility (N=398), n (%)	
Health post	31 (7.8)
Health center	158 (39.7)
Primary hospital	127 (31.9)
General hospital	82 (20.6)
Occupation (N=398), n (%)	
Nurse	200 (50.3)
Midwife	68 (17.1)
Health extension worker	30 (7.5)
Health officer	61 (15.3)
General practitioner	39 (9.8)
Age in years (N=390) mean (SD); range	31.0 (7.0); 20.0-64.0
Sex (N=394), n (%)	
Female	199 (50.5)
Male	195 (49.5)
Education level (N=396), n (%)	
Diploma	151 (38.1)
Bachelor of Science	208 (52.5)
Master of Science and above	37 (9.3)
Income group in US ($; per month; N=383)[b], n (%)	
<94	77 (19.3)
94-115	77 (19.3)
115-130	73 (18.3)
130-151.2	70 (17.6)
>151.2	86 (21.6)
Chosen language of the questionnaire (N=398), n (%)	
Amharic	345 (86.7)
Afan Oromo	53 (13.3)

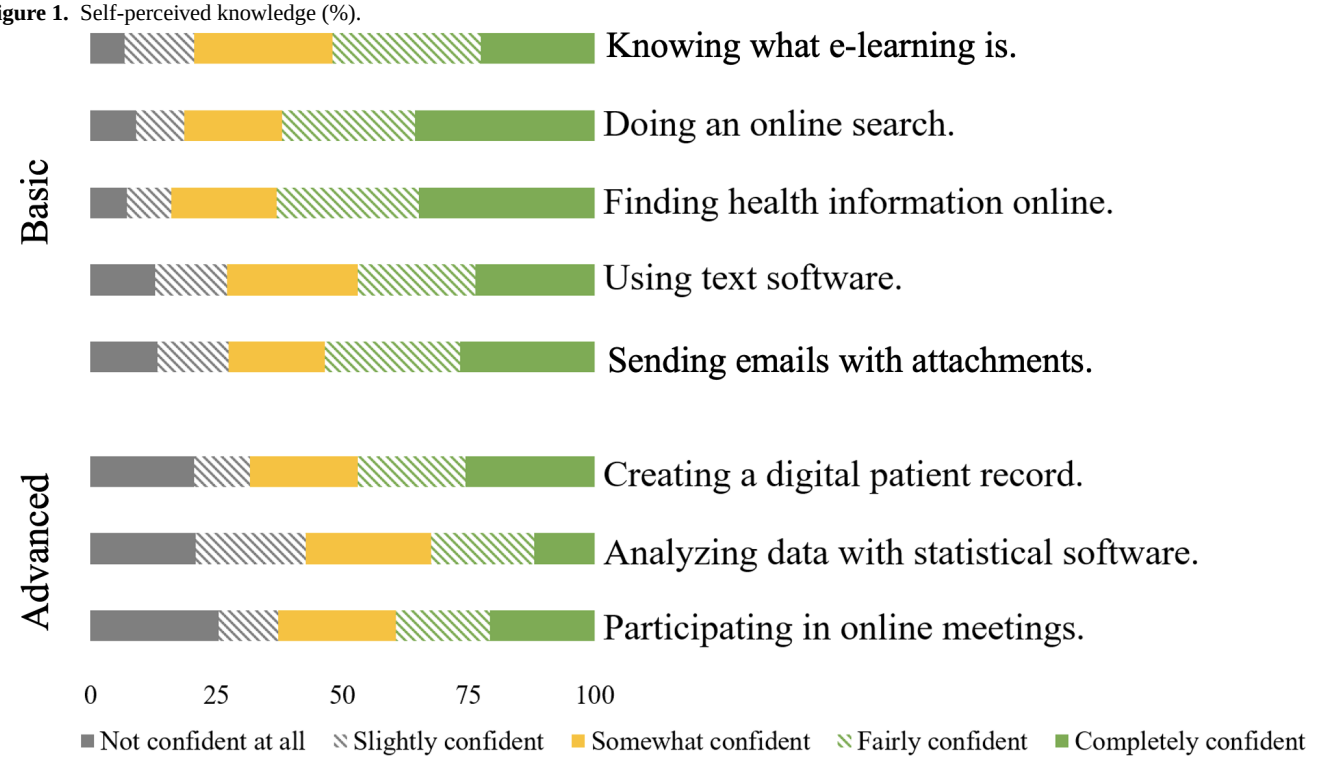
[a]Former Southern Nations Nationalities and Peoples Region.

[b]Exchange rate of 1 Ethiopian Birr to US $0.0167, as of April 16, 2023.

Knowledge

About half of the health care professionals were completely (90/398, 22.6%) or fairly confident (117/398, 29.4%) in knowing what e-learning is. Almost two thirds of them described a self-perceived competence in doing general (247/398, 62.1%) and medical-related (251/398, 63.1%) online searches. Over one-third were not or slightly confident in their ability to participate in online meetings (157/398, 37.2%; Figure 1).

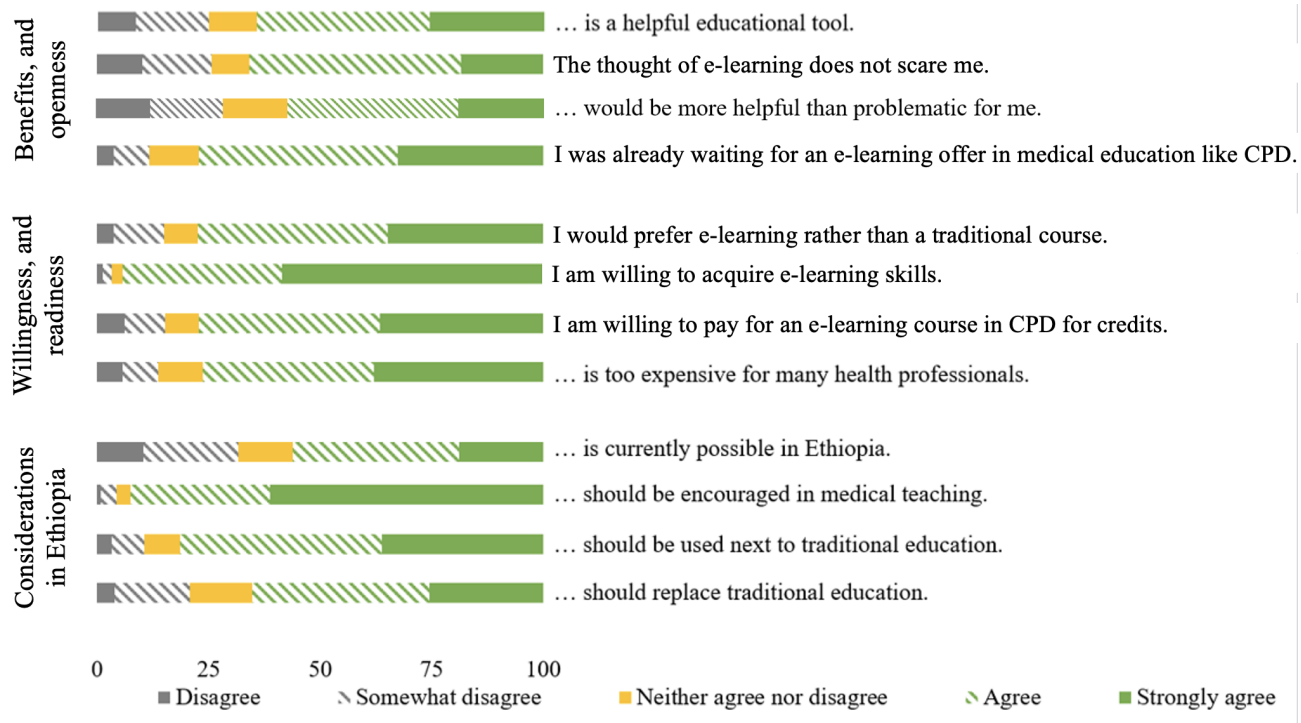
Figure 1. Self-perceived knowledge (%).



Attitude

Health care professionals predominantly had a positive attitude toward e-learning, with 94.2% (375/398) willing to put effort into acquiring e-learning skills. More than half (224/398, 56.2%) believed that e-learning is possible in rural parts of Ethiopia. Almost all health care professionals were in favor of integrating e-learning in medical teaching institutions (368/398, 92.5%). Some participants expressed fear (102/398, 25.7%) or perceived e-learning as more problematic than helpful (113/398; 28.4%; Figure 2).

Figure 2. Health care professionals' attitude toward e-learning (%). CPD: continuous professional development.

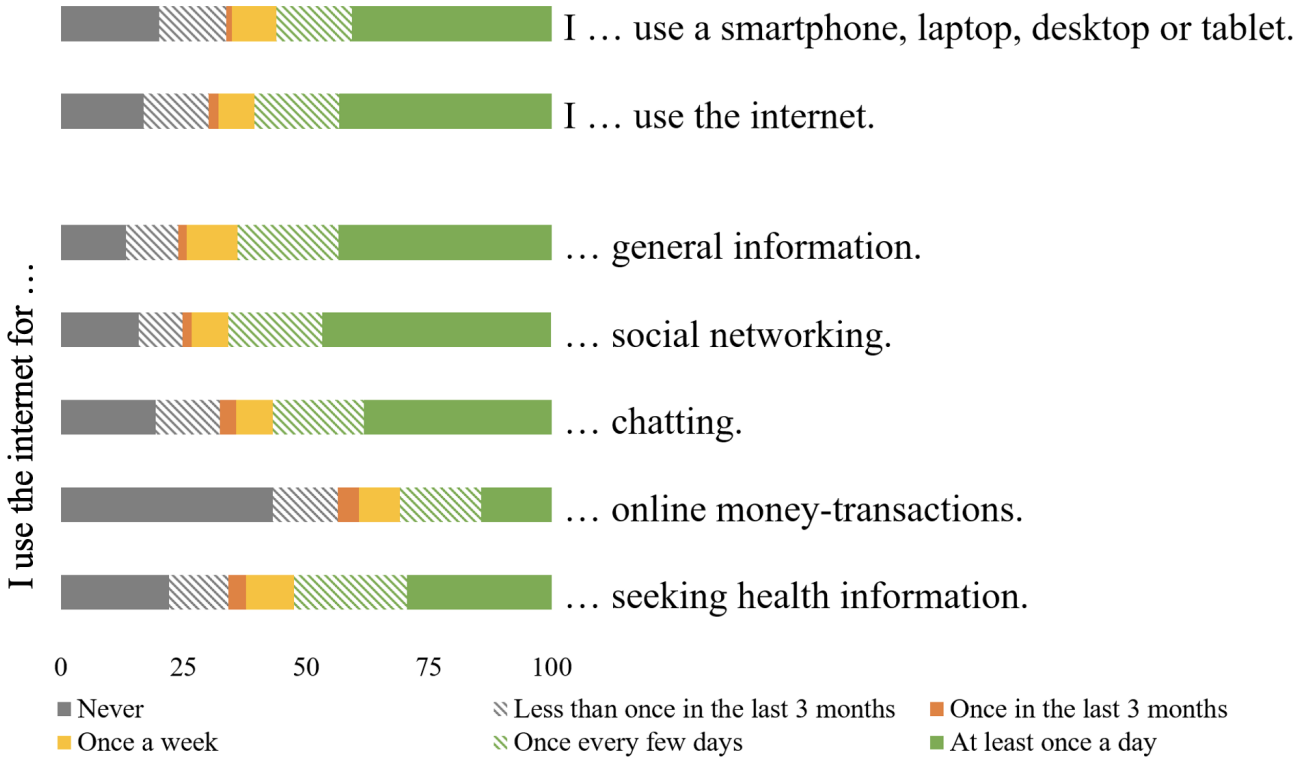


Practice

Daily internet usage was reported by 172/398 (43.2%) participants and 67/398 (16.8%) participants never used it. More than half used the internet for general (255/398, 64.1%) and medical-related searches (209/398, 52.5%), social networking (262/398, 65.8%), and chatting (226/398, 56.8%). In total, 2 out of 3 health care professionals stated a regular usage (at least

once a week) of an internet device (259/398, 65.1%) and pointed out smartphones as the most used device (268/393, 67.3%; Multimedia Appendix 4 and Figure 3).

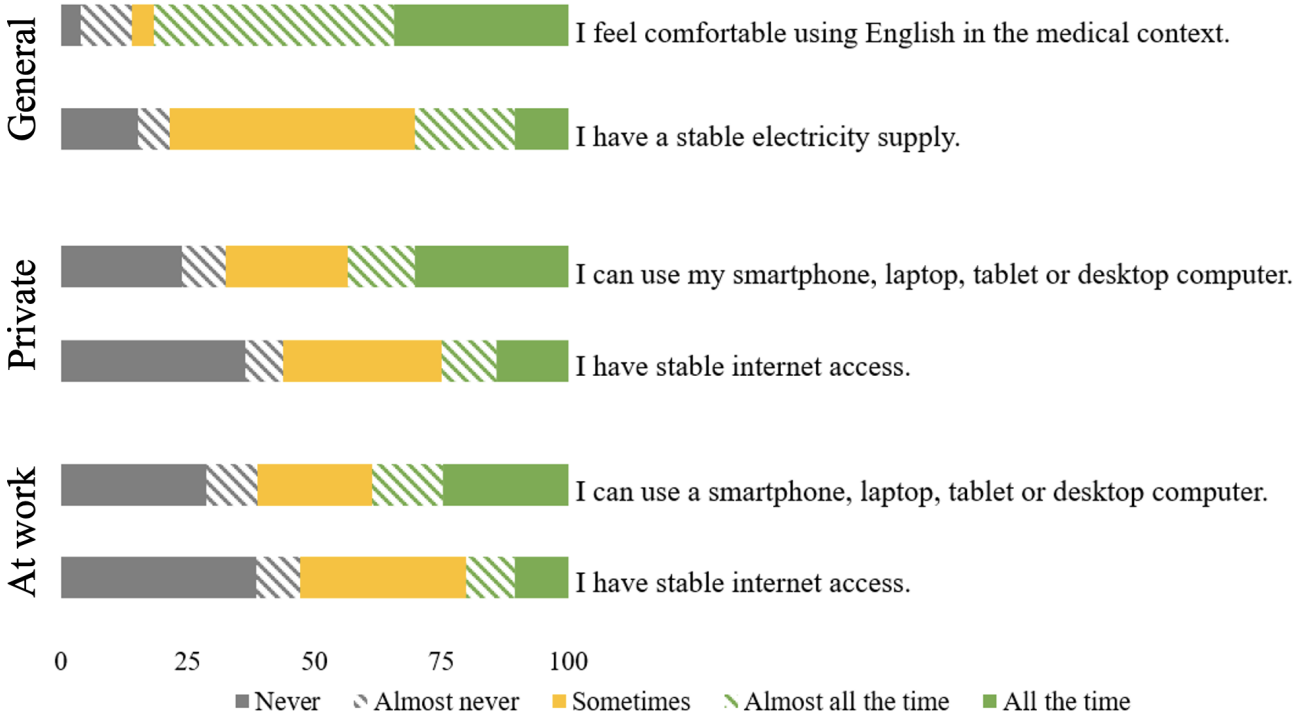
Figure 3. Practice of e-learning (%).



Access to Infrastructure

Access to digital devices (154/398, 38.7%) and stable internet (80/398, 20.1%) at the participants' workplace was less common than at home (access to private devices [173/398, 43.5%] and stable internet [99/398, 24.9%]). Common barriers included problems with internet availability (142/437, 32.5%) and internet costs (190/437; 43.5%; Multimedia Appendix 5 and Figure 4).

Figure 4. Health care professionals' access to infrastructure (in %).



Factors Associated With Knowledge, Attitude, Practice, and Infrastructure

Higher levels of education were associated with increased knowledge (AOR 2.31, 95% CI 1.45-3.68), whereas higher age was associated with lower knowledge (AOR 0.94, 95% CI 0.91-0.97). Health care professionals with a higher education level (AOR 2.56, 95% CI 1.57-4.16) and income group (AOR 1.31, 95% CI 1.06-1.62) were more likely to have higher practice scores. Female (AOR 0.44, 95% CI 0.6-0.71) and older participants (AOR 0.91, 95 % CI 0.87-0.95) were likely to report lower practice scores. Participants from higher income groups (AOR 1.50, 95% CI 1.21-1.85) had increased odds of having higher access to infrastructure. For detailed information see Table 2 and Multimedia Appendices 6 and 7.

Table . Factors associated with knowledge, attitude, practice–infrastructure.

Variable	Self-perceived knowledge, AOR[a] (95% CI)	Attitude, AOR (95% CI)	Practice, AOR (95% CI)	Access to infrastructure, AOR (95% CI)
Age (years)	0.94 (0.90-0.97)	1.01 (0.98-1.04)	0.91 (0.87-0.95)	0.94 (0.91-0.98)
Female sex[b]	0.69 (0.43-1.11)	1.01 (0.65-1.56)	0.44 (0.67-0.71)	0.48 (0.30-0.77)
Higher education level[c]	2.31 (1.45-3.68)	1.19 (0.78-1.82)	2.56 (1.57-4.16)	1.56 (0.98-2.46)
Higher income group	1.16 (0.95-1.41)	1.06 (0.88-1.28)	1.31 (1.06-1.62)	1.50 (1.21-1.85)
Health facility[d]				
Health post	2.00 (0.86-4.66)	1.48 (0.65-3.38)	0.70 (0.25-2.00)	0.31 (0.11-0.90)
Primary hospital	0.66 (0.39-1.09)	1.12 (0.69-1.78)	0.61 (0.36-1.04)	0.55 (0.33-0.91)
General hospital	1.97 (1.08-3.60)	0.60 (0.34-1.03)	1.01 (0.54-1.86)	1.08 (0.59-1.98)

[a]AOR: adjusted odds ratio.

[b]Male as reference.

[c] Diploma<Bachelor of Science<Mater of Science.

[d]Health center as reference.

Discussion

Principal Findings

The majority of the participants knew about e-learning and viewed themselves as competent in different e-learning-related activities. Still, we found difficulties in relevant knowledge aspects like participating in online meetings or searching for medical information. Moreover, a high proportion of workers had a positive attitude toward the use of e-learning. Almost half of the participants stated using the internet daily, whereas 16.8% (67/398) never use the internet. Infrastructure, such as technical devices and stable internet, was more often available at home than at work. The availability of stable internet was reported at below 25% (at home: 24.9% and at work: 20.1%).

Health care professionals with higher levels of education and income and those working in general hospitals had better KAP of e-learning and better access to infrastructure. Older health care professionals, women, and those working in health posts and primary hospitals reported reduced access to infrastructure.

Strengths and Limitations

First, the included health facilities show differences in terms of health service quality, with some having been rewarded international awards and not necessarily representing the typical peripheral hospital in Ethiopia. However, this diversity is part of the primary health care setting and should be represented in the study.

Second, during the data collection, we observed that some participants were confused about the definition of a smartphone, occasionally mistaking it for a tablet. These cases were clarified by the data collectors. Despite verification and testing of the translated questionnaire, we cannot completely rule out cases of misinterpretation. However, it is not expected to significantly affect the results.

Third, the health care facilities are exclusively located in Oromia and the former SNNPR region. Therefore, the results may not be generalizable for Ethiopia. While future research should investigate regional differences, it is reasonable to assume that our data from the context of primary health care settings are relevant and transferable to other parts of the country to a certain extent. Given the absence of studies in comparable settings, our data offer valuable insights for the development and implementation of effective e-learning courses.

Comparison With Prior Work

Participants report mixed levels of confidence regarding knowledge or e-learning–related tasks. More than 3 quarters of the health care professionals are at least somewhat confident in understanding the concept of e-learning (316/398, 79.4%), in performing basic digital tasks, such as general online searches (324/398, 81.4%), in finding professional health information on the internet (334/398, 83.9%), or in using writing programs, such as Microsoft Word (290/398, 72.9%). We did not find literature specifically targeting the knowledge of e-learning of health care professionals in LMICs, but the Ethiopian health care professionals were less confident in basic tasks, such as sending email attachments (53.5%) and conducting online searches (62.1%) compared to medical undergraduate students in India. Both studies identified creating patient records as

particularly challenging [36]. College students in India and Pakistan reported experiencing technical difficulties due to a lack of knowledge on e-learning. Nevertheless, they had an overall acceptable level of knowledge regarding media content and communication [28,35,53]. Possible reasons could include the higher internet penetration rate in India (52.4%) compared to Ethiopia (26.7%; 2024) [54,55]. This could be one of the reasons consequently leading to this difference.

Moreover, the lack of digital knowledge remains a significant challenge, further widening the digital divide between rural and urban areas [56,57]. Addressing these barriers is essential. Therefore, offering basic digital training programs can play a crucial role in bridging this gap [56]. Older health care professionals, those with lower education, and participants working outside general hospitals should be supported when implementing e-learning programs. A study from Kenya offers a further perspective on e-learning knowledge. Medical trainees faced difficulties in completing the courses due to limited technical skills and a lack of engagement from trainers and institutions responsible for developing the content. In addition to the extensive volume of e-learning materials and frequent online meetings, the courses contributed to feelings of being overwhelmed among the students. Approximately half of the students included in the study did not finish the course. Consideration should be given not only to the participants' knowledge but also to the content creators' [58].

Despite these mixed results, the overall attitude among Ethiopian health care professionals remains positive, and many even consider making financial and personal investments to support its implementation. In contrast, students from Uganda and Turkey showed an overall negative attitude toward e-learning. Participants' income, internet quality, ownership of technical devices, and previous use of academic sites were associated with attitude scores [59,60]. In Pakistan, similar issues did show, including technical and infrastructural challenges, limited technical skills, and the general disadvantages of e-learning, such as decreased connection and interaction with other participants, led to widespread dissatisfaction among the surveyed students. Consequently, 50% of them frequently missed e-learning classes. Nearly 80% considered e-learning less effective than traditional classroom instruction, particularly when it comes to acquiring practical skills [61]. Health care professionals in urban Taiwanese settings had a less positive attitude compared to those in rural settings but still valued the increase in knowledge, time saving, and diverse, and wide offer of information [62]. Medical students in Sudan favored a combination of electronic and traditional teaching methods [63]. Similarly, most of our participants (81.4%) are in favor of e-learning as an addition to traditional courses, even expressing a preference of e-learning over traditional workshops and lectures (77.4%). Overall, our results indicate that Ethiopian health care professionals are motivated to use e-learning for their professional development.

These positive attitudes can support the success of educative efforts [64], including e-learning [65]. e-Learning requires a higher level of self-motivation compared to other educational concepts [66], especially for participants with limited experience [67]. Incorporating feedback and evaluation is crucial in addressing these issues [11]. In addition, the rapidly evolving technical standards necessitate a positive attitude to continuously improve one's skills [68]. However, since a relevant share of health care professionals expressed fear, preparatory courses are advisable [1].

Regarding the practice of e-learning-related activities, almost half of the health care professionals (172/398, 43.2%) reported using the internet daily, which is less frequent than nursing students (68.1%) and comparable to patients from Gondar, Ethiopia (47%) [69,70]. Similar to our findings, being young, male, more educated, and with an urban residence is associated with increased internet use [70]. In rural communities in Sub-Saharan Africa, the internet is primarily used for browsing (13.9%) and email communication (13.2%), followed by information search (12.5%), chatting (10%), social networking (9.6%), and video conferencing (5.8%) [71]. Undergraduate students from Ethiopia with a rural background predominantly used the internet for social media [69]. Similarly, our participants most commonly used the internet for social networking and chatting, but also for general and medical searches. While a majority of primary health care professionals reported frequent internet use and engagement in e-learning–related digital activities, there is a relevant share of health care professionals (39.5%) who reported less frequent internet use, hinting at potential difficulties in handling online learning formats. To address this, targeted support for lower-income groups, women, and less-educated health care professionals is recommended.

Notably, smartphones emerged as the most commonly used device for accessing infrastructure, which aligns with findings from other Ethiopian settings [70]. Mobile phone ownership and usage has increased in countries of the global south, including rural communities in Sub-Saharan Africa, although access is not ubiquitous [70,72]. Mobile phones were among the least commonly used devices for use in our study, ranking lower than smartphones, tablets, laptops, and other devices. Smartphones are becoming increasingly important and prevalent as part of mobile health (mHealth) initiatives [73]. The WHO [25], along with other authors [19,26,74] recommends smartphones as a possible platform for e-learning to reach remote parts of countries with low resources. Health extension workers and midwives working in primary health care facilities in Ethiopia perceived the use of mobile devices in an mHealth program as easy and beneficial. The program was implemented for both the health care providers themselves as well as for the patients [75]. The Ethiopian Ministry of Health also introduced a community-based health extension program, which incorporates mHealth [76]. Nevertheless, rural communities in Sub-Saharan Africa used the mobile phone majorly for receiving (89.8%) and making calls (88.5%) and the least for internet browsing (27.3%) [71]. A study conducted in 4 primary hospitals in Ethiopia in 2018 revealed that 12.1% of health care professionals had internet access and 25.6% had access to a computer at work. Private computers were more often accessible at 33.3% [53]. Our study similarly showed higher accessibility to private internet and technical devices compared to work. Lacking bandwidth and stable internet access are common limitations, resulting in slow speed and low quality of e-learning

programs that may create difficulties for users to load digital content [9]. Power issues, shortage of skilled personnel for support, the cost of internet and necessary devices, and problems of viruses and malware can additionally hinder access to online education [77]. In addition, South Africa reports socioeconomic challenges that widen the digital divide and are thereby hindering the implementation of e-learning across various settings of the country [78]. For these reasons studies suggest e-learning courses with downloadable offline versions may help to establish accessible e-learning programs [79]. This increased accessibility and attempt to include students from different socioeconomic backgrounds, and in turn, positively influenced students' acceptance of e-learning in Amhara, Ethiopia, by enhancing their perceived ease of use [80].

Regarding possible gender differences, female health carecprofessionals had lower scores in knowledge, practice, and access to infrastructure while displaying a similar attitude. This is consistent with reports on the digital divide, both globally and in Ethiopia, which can in part be attributed to gender inequality [81]. For example, in Ethiopia, 14% of females used the internet, compared to 20% of males [82], and women faced lower computer literacy and access to infrastructure [19]. To reduce these disparities, supporting women, including the development of knowledge and skills, as well as addressing aspects of the costs of communication technology is necessary [83,84].

Conclusion

Health care professionals report mixed levels of knowledge in e-learning-related task while generally having a positive attitude toward e-learning in medical education. More than half use the internet regularly, especially via smartphones, while the access to digital infrastructure and reliable internet remains limited. Efforts should be made to improve access to internet access and technological infrastructure, as well as the adaptation of e-learning courses to the local needs. This could include the provision of introductory trainings, particularly for women and health care professionals with lower income and educationl; provision of downloadable offline learning options; and the adaptability of the course to smartphone use.

Acknowledgments

The study was also supported by a grant from Hospital Partnership through Deutsche Gesellschaft für Internationale Zusammenarbeit funded by the Ministry for Economic Cooperation and Development (ID 81281915). The project on which this publication is based was in part funded by the German Federal Ministry of Education and Research (01KA2220B). This study was also supported by a grant form the Else Kroener-Fresenius-Foundation (2018_HA31SP). This research was funded in part by Science for Africa Foundation to the Developing Excellence in Leadership, Training and Science in Africa (DELTAS Africa) program (Del-22 - 008) with support from Wellcome Trust and the UK Foreign, Commonwealth & Development Office and is part of the EDCPT2 program supported by the European Union.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Distribution of participants according to health care level, region, and type of health facility.
[PDF File, 108 KB - mededu_v11i1e65598_app1.pdf]

Multimedia Appendix 2
Cutoff points for categorization of high and low knowledge, attitude, practice, and infrastructure.
[PDF File, 79 KB - mededu_v11i1e65598_app2.pdf]

Multimedia Appendix 3
Correlation matrix.
[PDF File, 158 KB - mededu_v11i1e65598_app3.pdf]

Multimedia Appendix 4
Most often used device.
[PNG File, 13 KB - mededu_v11i1e65598_app4.png]

Multimedia Appendix 5
Reasons for restricted internet access.
[PDF File, 59 KB - mededu_v11i1e65598_app5.pdf]

Multimedia Appendix 6
Complete table factors associated to knowledge, attitude, practice, and access to infrastructure.

[PDF File, 119 KB - mededu_v11i1e65598_app6.pdf]

Multimedia Appendix 7
Distribution of the answered questionnaire.
[PDF File, 129 KB - mededu_v11i1e65598_app7.pdf]

References

1. Barteit S, Guzek D, Jahn A, Bärnighausen T, Jorge MM, Neuhann F. Evaluation of e-learning for medical education in low- and middle-income countries: a systematic review. Comput Educ 2020 Feb;145:103726. [doi: 10.1016/j.compedu.2019.103726] [Medline: 32565611]
2. CME content: definition and examples. Accreditation Council for Continuing Medical Education. 2023. URL: https://accme.org/accreditation-rules/policies/cme-content-definition-and-examples [accessed 2023-08-31]
3. Collin K, Van der Heijden B, Lewis P. Continuing professional development. Int J Training Development 2012 Sep;16(3):155-163. [doi: 10.1111/j.1468-2419.2012.00410.x]
4. Lilford RJ, Daniels B, McPake B, et al. Improving primary health-care services in LMIC cities. Lancet Glob Health 2025 May;13(5):e795-e796. [doi: 10.1016/S2214-109X(24)00537-0] [Medline: 40288387]
5. Hudspeth J, Curry CL, Sacks Z, Surena C. Continuing professional development in low-resource settings: Haiti as example. Ann Glob Health 2015;81(2):255-259. [doi: 10.1016/j.aogh.2015.03.004] [Medline: 26088091]
6. Digital 2024: ethiopia. DataReportal. 2025. URL: https://datareportal.com/reports/digital-2024-ethiopia [accessed 2025-06-07]
7. Bringing knowledge to practice. Virtual Campus for Public Health (VCPH/PAHO). 2025. URL: https://campus.paho.org/en [accessed 2025-05-12]
8. Online learning framework & toolkit. HR Portal. 2022. URL: https://hr.un.org/page/online-learning-framework-toolkit [accessed 2022-12-16]
9. Frehywot S, Vovides Y, Talib Z, et al. E-learning in medical education in resource constrained low- and middle-income countries. Hum Resour Health 2013 Feb 4;11:4 [FREE Full text] [doi: 10.1186/1478-4491-11-4] [Medline: 23379467]
10. Atlas of ehealth country profiles: the use of ehealth in support of universal health coverage. World Health Organization. 2023. URL: https://www.who.int/publications/i/item/9789241565219 [accessed 2025-09-19]
11. Regmi K, Jones L. A systematic review of the factors - enablers and barriers - affecting e-learning in health sciences education. BMC Med Educ 2020 Mar 30;20(1):91. [doi: 10.1186/s12909-020-02007-6] [Medline: 32228560]
12. Ruggeri K, Farrington C, Brayne C. A global model for effective use and evaluation of e-learning in health. Telemed J E Health 2013 Apr;19(4):312-321. [doi: 10.1089/tmj.2012.0175] [Medline: 23472702]
13. O'Flynn E, Ahmed A, Biswas A, Bempong-Ahun N, Perić I, Puyana JC. E-learning supporting surgical training in low-resource settings. Curr Surg Rep 2024;12(6):151-159. [doi: 10.1007/s40137-024-00399-8]
14. Schievano F, Mwamwitwa KW, Kisenge S, et al. Development, assessment and educational impact of a blended e-learning training program on pharmacovigilance implemented in four African countries. Front Med 2024;11:1347317. [doi: 10.3389/fmed.2024.1347317]
15. Williams E, Fernandes RD, Choi K, Fasola L, Zevin B. Learning outcomes and educational effectiveness of e-learning as a continuing professional development intervention for practicing surgeons and proceduralists: a systematic review. J Surg Educ 2023 Aug;80(8):1139-1149. [doi: 10.1016/j.jsurg.2023.05.017] [Medline: 37316431]
16. Aryee GFB, Amoadu M, Obeng P, et al. Effectiveness of eLearning programme for capacity building of healthcare professionals: a systematic review. Hum Resour Health 2024 Sep 2;22(1):60. [doi: 10.1186/s12960-024-00924-x] [Medline: 39223555]
17. Karkera S, Devendra N, Lakhani B, Manahan K, Geisler J. A review on modern teaching and learning techniques in medical education. EIKI Journal of Effective Teaching Methods 2025;2(1):63-80 [FREE Full text] [doi: 10.59652/jetm.v2i1.128]
18. Adeniyi IS, Hamad NMA, Adewusi OE, et al. E-learning platforms in higher education: a comparative review of the USA and Africa. Int J Sci Res Arch 2024;11(1):1686-1697 [FREE Full text] [doi: 10.30574/ijsra.2024.11.1.0283]
19. Bollinger R, Chang L, Jafari R, et al. Leveraging information technology to bridge the health workforce gap. Bull World Health Organ 2013 Nov 1;91(11):890-892. [doi: 10.2471/BLT.13.118737] [Medline: 24347719]
20. Internet penetration rate in Africa as of June 2022, compared to the global average. 2024. URL: https://www.statista.com/statistics/1176654/internet-penetration-rate-africa-compared-to-global-average/ [accessed 2024-02-22]
21. Access to electricity (% of population) - ethiopia. World Bank Group. 2025. URL: https://data.worldbank.org/indicator/EG.ELC.ACCS.ZS?locations=ET [accessed 2025-06-06]
22. Belay DG. COVID-19, distance learning and educational inequality in rural Ethiopia. Pedagogical Res 2020;5(4):em0082. [doi: 10.29333/pr/9133]
23. Continuous professional development. Ethiopian Ministry of Health. URL: https://ephi.gov.et/ntc/services/training/continuous-professional-developmental-training-cpd/ [accessed 2022-11-29]
24. Learning strategy. World Health Organization. 2023. URL: https://www.who.int/about/who-academy/learning-strategy [accessed 2023-09-04]

25. mHealth: new horizons for health through mobile technologies. World Health Organization. 2023. URL: https://www.afro.who.int/publications/mhealth-new-horizons-health-through-mobile-technologie [accessed 2023-08-03]
26. Mack HG, Golnik KC, Murray N, Filipe HP. Models for implementing continuing professional development programs in low-resource countries. MedEdPublish (2016) 2017;6:18 [FREE Full text] [doi: 10.15694/mep.2017.000018] [Medline: 38406456]
27. Childs S, Blenkinsopp E, Hall A, Walton G. Effective e-learning for health professionals and students--barriers and their solutions. A systematic review of the literature--findings from the HeXL project. Health Info Libr J 2005 Dec;22(Suppl 2):20-32. [doi: 10.1111/j.1470-3327.2005.00614.x] [Medline: 16279973]
28. Sarwar S, Akram D. KAP analysis of students regarding e-learning during Covid-19 in Universities of Lahore. J Prof Res Soc Sci 2021;8(2):24-35 [FREE Full text]
29. Van der Merwe TMD, Serote M, Maloma M. A systematic review of the challenges of e-learning implementation in sub-Saharan African countries: 2016-2022. Electron J e-Learn 2025;21(5):413-429 [FREE Full text] [doi: 10.34190/ejel.21.5.3075]
30. Oroma J, Oroma J, Wanga H, Wanga H, Fredrick N. Challenges of e-learning in developing countries: the ugandan experience. Presented at: 6th International Technology, Education and Development Conference; Mar 5-7, 2012; Valencia, Spain p. 3535-3543 URL: https://library.iated.org/view/OROMA2012CHA [accessed 2025-09-19]
31. Vyas R, Albright S, Walker D, Zachariah A, Lee MY. Clinical training at remote sites using mobile technology: an India–USA partnership. Distance Educ 2010 Aug;31(2):211-226. [doi: 10.1080/01587919.2010.498856]
32. Bankar MN, Bankar NJ, Singh BR, Bandre GR, Shelke YP. The role of e-content development in medical teaching: how far have we come? Cureus 2023 Aug;15(8):e43208. [doi: 10.7759/cureus.43208] [Medline: 37692742]
33. The KAP survey model (knowledge, attitudes, and practices). Spring-nutrition.org. 2014. URL: https://www.spring-nutrition.org/publications/tool-summaries/kap-survey-model-knowledge-attitudes-and-practices [accessed 2022-11-01]
34. Oluwadele D, Singh Y, Adeliyi TT. Trends and insights in e‐learning in medical education: a bibliometric analysis. Rev Educ 2023 Dec;11(3):e3431 [FREE Full text] [doi: 10.1002/rev3.3431]
35. Shekinah SI, Chinnasamy P, Deepsheka K, Singaram V. Impact of online education due to the pandemic among college students: knowledge, attitude and practices analysis with structural equation modeling. J Educ Health Promot 2022;11:189. [doi: 10.4103/jehp.jehp_995_21] [Medline: 36003250]
36. Visalam APK, Om A, Prakash PRK. Knowledge, Attitude and Practice towards E -Learning Among Medical Undergraduate Students. IOSR J Appl Phys 2021:1-4. [doi: 10.9790/4861-07430104]
37. Community survey on ICT usage in households and by individuals. 2017. URL: https://circabc.europa.eu/sd/a/81cbbefa-b48f-4ce6-a076-289e3f18daec/Questionnaire%20HH%202017%20v%200.11.pdf [accessed 2022-12-09]
38. SNNPR regional brief. UNICEF. URL: https://www.unicef.org/ethiopia/media/6516/file/SNNPR%20regional%20brief.pdf [accessed 2024-02-26]
39. Oromia regional brief. UNICEF. URL: https://www.unicef.org/ethiopia/media/6511/file/Oromia%20regional%20brief.pdf [accessed 2024-02-26]
40. Ethiopia rural population, percent. TheGlobalEconomy. 2022. URL: https://www.theglobaleconomy.com/Ethiopia/rural_population_percent/ [accessed 2024-02-26]
41. Bertram D. Likert scales: researchgate.net. 2007 URL: https://cspages.ucalgary.ca/~saul/wiki/uploads/CPSC681/topic-dane-likert.pdf
42. ICT usage in households and by individuals. 2020 URL: https://ec.europa.eu/eurostat/web/microdata/survey-ict-use-households-individuals
43. Schofield M, Forrester-Knauss C. Surveys and questionnaires in health research. In: Liamputtong P, editor. Research Methods in Health: Foundations for Evidence-Based Practice, 2nd Ed: Oxford University Press; 2013:198-218 URL: https://tinyurl.com/2e6kxscs [accessed 2025-09-19]
44. Porst R. Pretests for Evaluating the Questionnaire (Draft): A Workbook 4th Expanded Edition Geneva: International Labor Office; 1923, Vol. 9783658021177.
45. Myers L, Sirois MJ. Spearman correlation coefficients, differences between. In: Kotz S, Read CB, Balakrishnan N, et al, editors. Encyclopedia of Statistical Sciences: John Wiley & Sons; 2006:7901-7902. [doi: 10.1002/0471667196]
46. Hamadeh N, Van Rompaey C, Metreau E, Eapen SG. New world bank country classifications by income level: 2022-2023. WORLD BANK BLOGS. 2025. URL: https://blogs.worldbank.org/en/opendata/new-world-bank-country-classifications-income-level-2022-2023 [accessed 2025-05-16]
47. Glossary:internet use. Eurostat. 2023. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Internet_use [accessed 2023-08-04]
48. Chaudhuri A, Flamm KS, Horrigan J. An analysis of the determinants of internet access. Telecomm Policy 2005 Oct;29(9-10):731-755. [doi: 10.1016/j.telpol.2005.07.001]
49. Teo TSH. Demographic and motivation variables associated with Internet usage activities. Internet Research 2001 May 1;11(2):125-137. [doi: 10.1108/10662240110695089]
50. Downloading IBM SPSS statistics 28. IBM. URL: https://www.ibm.com/support/pages/downloading-ibm-spss-statistics-28 [accessed 2023-08-03]

51. Taber KS. The use of Cronbach's alpha when developing and reporting research instruments in science education. Res Sci Educ 2018 Dec;48(6):1273-1296. [doi: 10.1007/s11165-016-9602-2]
52. Hosmer-lemeshow test: definition. Statistics How To. 2016. URL: https://www.statisticshowto.com/hosmer-lemeshow-test/ [accessed 2024-01-06]
53. Awol SM, Birhanu AY, Mekonnen ZA, et al. Health professionals' readiness and its associated factors to implement electronic medical record system in four selected primary hospitals in Ethiopia. Adv Med Educ Pract 2020;11:147-154. [doi: 10.2147/AMEP.S233368] [Medline: 32110135]
54. India: internet penetration rate 2024. Statista. 2024. URL: https://www.statista.com/statistics/792074/india-internet-penetration-rate/ [accessed 2024-04-30]
55. Digital & connectivity indicators - ethiopia. Statista. 2024. URL: https://www.statista.com/outlook/co/digital-connectivity-indicators/ethiopia [accessed 2024-04-30]
56. Limaye RJ, Deka S, Ahmed N, Mwaikambo L. Designing eLearning courses to meet the digital literacy needs of healthcare workers in lower- and middle-income countries: experiences from the Knowledge for Health Project. Knowl Manag E-Learn 2015 Dec 15;7(4):601-615 [FREE Full text] [doi: 10.34105/j.kmel.2015.07.039]
57. van Dijk J, Hacker K. The digital divide as a complex and dynamic phenomenon. Inf Soc 2003 Sep;19(4):315-326. [doi: 10.1080/01972240309487]
58. Gachanja F, Mwangi N, Gicheru W. E-learning in medical education during COVID-19 pandemic: experiences of a research course at Kenya Medical Training College. BMC Med Educ 2021 Dec 11;21(1):612. [doi: 10.1186/s12909-021-03050-7] [Medline: 34893065]
59. Olum R, Atulinda L, Kigozi E, et al. Medical education and E-learning during COVID-19 pandemic: awareness, attitudes, preferences, and barriers among undergraduate medicine and nursing students at Makerere University, Uganda. J Med Educ Curric Dev 2020;7:2382120520973212. [doi: 10.1177/2382120520973212] [Medline: 33283049]
60. Güllü A, Kara M, Akgün Ş. Determining attitudes toward e-learning: what are the attitudes of health professional students? J Public Health (Berl) 2024 Jan;32(1):89-96. [doi: 10.1007/s10389-022-01791-3]
61. Abbas U, Parveen M, Sahito FS, Hussain N, Munir S. E-learning in medical education: a perspective of pre-clinical medical students from a lower-middle income country. BMC Med Educ 2024 Feb 20;24(1):162. [doi: 10.1186/s12909-024-05158-y] [Medline: 38378563]
62. Yu S, Yang KF. Attitudes toward Web-based distance learning among public health nurses in Taiwan: a questionnaire survey. Int J Nurs Stud 2006 Aug;43(6):767-774. [doi: 10.1016/j.ijnurstu.2005.09.005] [Medline: 16253261]
63. Alkanzi FK, Abd-algader AA, Ibrahim ZA, Krar AO, Osman MA, Karksawi NM. Knowledge, attitude and practice in electronic education among teaching staff and students in governmental medical faculties - Khartoum State. Sudan J Med Sci 2014;9(1):43-48 [FREE Full text]
64. Faridah I, Ratna Sari F, Wahyuningsih T, Putri Oganda F, Rahardja U. Effect digital learning on student motivation during covid-19. Presented at: 2020 8th International Conference on Cyber and IT Service Management (CITSM); Oct 23-24, 2020; Pangkal Pinang, Indonesia URL: https://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=9268688 [accessed 2025-09-19] [doi: 10.1109/CITSM50537.2020.9268843]
65. Samsudeen SN, Mohamed R. University students' intention to use e-learning systems. Interact Technol Smart Educ 2019 Sep 16;16(3):219-238. [doi: 10.1108/ITSE-11-2018-0092]
66. Nehme M. E-learning and students' motivation. Leg Educ Rev 2010;20(1):11. [doi: 10.53300/001c.6236]
67. Mailizar M, Burg D, Maulina S. Examining university students' behavioural intention to use e-learning during the COVID-19 pandemic: an extended TAM model. Educ Inf Technol 2021 Nov;26(6):7057-7077. [doi: 10.1007/s10639-021-10557-5]
68. Wilkinson A, While AE, Roberts J. Measurement of information and communication technology experience and attitudes to e-learning of students in the healthcare professions: integrative review. J Adv Nurs 2009 Apr;65(4):755-772. [doi: 10.1111/j.1365-2648.2008.04924.x] [Medline: 19228242]
69. Shiferaw KB, Mehari EA, Eshete T. eHealth literacy and internet use among undergraduate nursing students in a resource limited country: a cross-sectional study. Inform Med Unlocked 2020;18:100273. [doi: 10.1016/j.imu.2019.100273]
70. Shiferaw KB, Tilahun BC, Endehabtu BF, Gullslett MK, Mengiste SA. E-health literacy and associated factors among chronic patients in a low-income country: a cross-sectional survey. BMC Med Inform Decis Mak 2020 Aug 6;20(1):181. [doi: 10.1186/s12911-020-01202-1] [Medline: 32762745]
71. Balogun NA, Ehikhamenor FA, Mejabi OV, Oyekunle RA, Bello OW, Afolayan OT. Exploring information and communication technology among rural dwellers in sub-Saharan African communities. Afr J Sci Technol Innov Dev 2020 Jul 28;12(5):533-545. [doi: 10.1080/20421338.2019.1700668]
72. Marshall C, Lewis D, Whittaker M. Strengthening health systems in mhealth technologies in developing countries: a feasibility assessment and a proposed framework. : Health Information Systems Knowledge Hub, School of Population Health, The University of Queensland; 2013 URL: https://tinyurl.com/k84yzbux [accessed 2025-09-19]
73. Singh DP, Sachs PJD. 1 million community health workers in sub-Saharan Africa by 2015. Lancet 2013 Jul;382(9889):363-365 [FREE Full text] [doi: 10.1016/S0140-6736(12)62002-9]

74. Aranda-Jan CB, Mohutsiwa-Dibe N, Loukanova S. Systematic review on what works, what does not work and why of implementation of mobile health (mHealth) projects in Africa. BMC Public Health 2014 Feb 21;14(1):188. [doi: 10.1186/1471-2458-14-188] [Medline: 24555733]
75. Medhanyie AA, Little A, Yebyo H, et al. Health workers' experiences, barriers, preferences and motivating factors in using mHealth forms in Ethiopia. Hum Resour Health 2015 Jan 15;13(1):2 [FREE Full text] [doi: 10.1186/1478-4491-13-2] [Medline: 25588973]
76. Examining e-health. WorldCat.org. 2023. URL: https://www.worldcat.org/de/title/51211048?oclcNum=51211048 [accessed 2023-08-03]
77. Osei Asibey B, Agyemang S, Dankwah AB. The internet use for health information seeking among Ghanaian University students: a cross-sectional study. Int J Telemed Appl 2017;2017:1756473. [doi: 10.1155/2017/1756473] [Medline: 29225620]
78. Noorbhai H, Ojo TA. mHealth and e-Learning in health sciences curricula: a South African study of health sciences staff perspectives on utilisation, constraints and future possibilities. BMC Med Educ 2023 Mar 28;23(1):189. [doi: 10.1186/s12909-023-04132-4] [Medline: 36978117]
79. Barteit S, Neuhann F, Bärnighausen T, et al. Perspectives of nonphysician clinical students and medical lecturers on tablet-based health care practice support for medical education in Zambia, Africa: qualitative study. JMIR Mhealth Uhealth 2019 Jan 15;7(1):e12637. [doi: 10.2196/12637] [Medline: 30664475]
80. Belew A, Ketemaw A, Sitotaw G, et al. Acceptance of e-learning and associated factors among postgraduate medical and health science students at first generation universities in Amhara region, 2023: using modified technology acceptance model. BMC Med Educ 2024 Aug 5;24(1):838. [doi: 10.1186/s12909-024-05834-z] [Medline: 39103812]
81. Huyer S. ICTs, globalisation and poverty reduction: gender dimensions of the knowledge society part II. In: Gender Equality and Poverty Reduction in the Knowledge Society: Gender Advisory Board of the UN Commission on Science and Technology for Development; 2003. URL: http://gab.wisat.org/PartI.pdf [accessed 2025-09-19]
82. Digital development dashboard: ethiopia. International Telecommunication Union. URL: https://www.itu.int/en/ITU-D/Statistics/Documents/DDD/ddd_ETH.pdf [accessed 2023-09-04]
83. Ogato GS. The quest for gender responsive information communication technologies (ICTs) policy in least developed countries: policy and strategy implications for promoting gender equality and women's empowerment in Ethiopia. International Journal of Information Technology and Business Management ;15(1):23-44 [FREE Full text]
84. Gillwald A, Milek A, Christoph S. Gender assessment of ICT access and usage in africa. : Research ICT Africa URL: https://lirneasia.net/wp-content/uploads/2010/09/Gender_Paper_Sept_2010.pdf [accessed 2025-09-19]

Abbreviations

AOR: adjusted odds ratio
CPD: continuous professional development
KAP: knowledge, attitude, and practice
KAP-I: knowledge, attitude, practice, and infrastructure
LMIC: low- and middle-income country
mHealth: mobile health
OR: odds ratio
SNNPR: Southern Nations Nationalities and Peoples Region
WHO: World Health Organization

Edited by B Lesselroth, D Chartash; submitted 20.08.24; peer-reviewed by A Ivaturi, SM Sheikh Ghadzi; revised version received 25.06.25; accepted 25.06.25; published 25.09.25.

Please cite as:
Rossner SS, Gizaw M, Getachew S, Getachew E, Destaw A, Negash S, Bauer L, Hermann ESM, Shita A, Unverzagt S, Santos PSC, Kantelhardt EJ, Kroeber ES
Health Care Professionals' Knowledge, Attitude, Practice, and Infrastructure Accessibility for e-Learning in Ethiopia: Cross-Sectional Study
JMIR Med Educ 2025;11:e65598
URL: https://mededu.jmir.org/2025/1/e65598
doi:10.2196/65598

Original Paper

Effectiveness of an Interactive Web-Based Clinical Practice Monitoring System on Enhancing Motivation in Clinical Learning Among Undergraduate Nursing Students: Longitudinal Quasi-Experimental Study in Tanzania

Patricia Herman[1], MSc; Stephen M Kibusi[2], Prof Dr; Walter C Millanzi[3], PhD

[1]Department of Nursing, College of Health and Allied Sciences, Ruaha Catholic University, Iringa, United Republic of Tanzania

[2]Department of Public Health and Community Health Nursing, School of Nursing and Public Health, The University of Dodoma, Dodoma, United Republic of Tanzania

[3]Department of Nursing Management and Education, School of Nursing and Public Health, The University of Dodoma, Dodoma, United Republic of Tanzania

Corresponding Author:
Patricia Herman, MSc
Department of Nursing
College of Health and Allied Sciences
Ruaha Catholic University
Box 774
Iringa
United Republic of Tanzania
Phone: 255 788315184
Email: patriciaz1006@gmail.com

Abstract

Background: Nursing students' motivation in clinical learning is very important not only for their academic and professional achievement but also for making timely, informed, and appropriate decisions in providing quality and cost-effective care to people. However, the increased number of students and the scarcity of medical supplies, equipment, and patients, just to mention a few, have posed a challenge to educators in identifying and navigating the best approaches to motivate nursing students to learn during their clinical placements.

Objective: This study primarily used descriptive and analytical methods to examine undergraduate nursing students' desire for clinical learning both before and after participating in the program.

Methods: An uncontrolled longitudinal quasi-experimental study in a quantitative research approach was conducted from February to March 2021 among 589 undergraduate nursing students in Tanzania. Following a baseline evaluation, nursing students were enrolled in an interactive web-based clinical practice monitoring system by their program, institution, names, registration numbers, and emails via unique codes created by the lead investigator and trainers. The system recorded and generated feedback on attendance, clinical placement unit, selected or performed clinical nursing procedures, and in-between and end-of-shift feedback. The linear regression was used to assess the effect of the intervention (interactive web-based clinical practice monitoring system) controlled for other correlated factors on motivation in clinical learning (outcome) among nursing students. Nursing students' sociodemographic characteristics and levels of motivation in clinical learning were analyzed descriptively while a 2-tailed paired sample *t* test established a comparative mean difference in motivation in clinical learning between the pretest and the posttest. The association between variables was determined using regression analysis set at a 95% CI and 5% statistical significance.

Results: The mean age of study participants (N=589) was 23 (SD 2.69) years of which 383 (65.0%) were male. The estimated effect (β) of a 3-week intervention to improve nursing students' motivation in clinical learning was 3.041 (*P*=.03, 95% CI 1.022-7.732) when controlled for other co-related factors. The mean score for motivation in clinical learning increased significantly from the baseline (mean 9.31, SD 2.315) to the postintervention (mean 20.87, SD 5.504), and this improvement presented a large effect size of 2.743 (*P*<.001, 95% CI 1.011-4.107).

Conclusions: Findings suggest that an interactive web-based clinical practice monitoring system is viable and has the potential to improve undergraduate nursing students' motivation for clinical learning. One alternative clinical pedagogy that educators in

nursing education can use to facilitate clinical learning activities and develop motivated undergraduate nursing students is the integration of such technology throughout nursing curricula.

(JMIR Med Educ 2025;11:e45912) doi:10.2196/45912

KEYWORDS

clinical monitoring system; clinical practice; motivation in clinical learning; nursing students; smartphone; mobile phone; Ruaha Catholic University; web-based teaching

Introduction

Owing to changes in population, technology, and communicable and noncommunicable diseases, the uncertain nature of health care provision necessitates improvements in health care systems and the teaching and learning environment across health disciplines, including nursing [1]. Clinical nursing education, in particular, has become a fundamental aspect of the nursing profession that informs educators about the best and most innovative pedagogical strategies that are navigated into technology to increase nursing students' motivation in clinical learning and thus ensure consistent clinical attendance and passionate clinical learning around the world [2]. In clinical learning, motivation refers to an individual's inner drive toward an activity or behavior that defines his or her achievements, in this case, the attainment of skills and competence required in the nursing profession among nursing students [3]. Scholars have shown that motivated students can demonstrate metacognition, metacompetencies, and a sense of independence when providing high-quality, cost-effective health care to a diverse population [4-6].

Educators in nursing education are important individuals in both classroom and clinical teaching and learning activities to enhance nursing students' motivation in clinical learning during clinical placements for meta-competencies in diagnosing and making informed and appropriate decisions in providing quality and cost-effective care to people. Educators in nursing education are expected to be developed and empowered with pedagogical competencies to cope with advanced science and technology, increased rates of nursing student enrollments in middle and higher education institutions, scarcity of medical supplies and equipment, unimproved clinical teaching and learning environments, and availability of clients or patients as important individuals in clinical nursing education [7,8]. Empowering educators with knowledge and skills for facilitating clinical teaching and learning activities for nursing students may be intimately related to their pedagogical competencies in creating, supporting, mentoring, coaching, supervising, monitoring, and assessing nursing students [9,10].

Nursing students, developed by competent educators in nursing education, are believed to share their learning experiences with both educators and peers. They show motivation in their clinical learning activities and client or patient care provision as they demonstrate a sense of independence, a positive professional identity, and the attitude, skills, and values of lifelong learners and professionals [11]. Motivating nursing students to learn has risen to be prominent in clinical nursing education because it is regarded as a critical component in promoting consistent clinical attendance and learning [12]. Investing in the motivation of nursing students also ensures the development of competent nursing graduates who can demonstrate safe, ethical, and legal practices, which are foundational and essential aspects of clinical nursing education [6,13]. Contrary to what is expected of educators in nursing education in the twenty-first century, their current clinical nursing education practices are more traditional, with pedagogics such as bedside tutorials, lectures, demonstrations, discussions, case studies, portfolios, and nursing meetings, to name a few, being widely used [11,14,15]. However, due to the increased enrollment rate of nursing students with the unchanging pedagogical trend, a shortage of academic faculty, and a limited number of trained clinical instructors, the aforementioned clinical pedagogics are doubted to be able to enhance interactive communication between educators in nursing education and students.

Nevertheless, they are doubted on their abilities to establish teaching and learning feedback or experiences from students and nursing students' motivation in clinical learning [16]. Moreover, they demonstrate weaknesses in not developing them with clinical meta-competencies for providing quality and cost-effective care to people [17]. Scholars have linked the permanent implementation of conventional clinical nursing education pedagogics to a lack of interactive communication, teaching and learning feedback from students, and clinical absenteeism as remarkable signs of unmotivated nursing students worldwide [18]. The work by Rahman et al [12] has demonstrated that clinical absenteeism has been linked to an unsatisfactory clinical learning environment as well as a shortage of competent educators in nursing education. Nevertheless, the implementation of conventional clinical supervision, mentorships, support, monitoring, and evaluation measures such as registration books and follow-up books fail to manage a large group of nursing students in clinical settings [19]. Nursing students' avoidance and lack of enthusiasm in attending their daily practical activities during their clinical placements result in substandard care delivered to clients or patients alongside unethical professional conduct and poor customer care [20].

This work is based on the belief that understanding and implementing novel clinical pedagogical strategies will assist educators in nursing education in grasping and navigating the best ways to inspire unmotivated students to learn in a clinical context over conventional pedagogies. Adoption and integration of technology have been prioritized and have proven to be timely, quick, cost-efficient, long-term, and beneficial in enhancing students' motivation in their learning activities [21,22]. Authors of this work agree with other scholars that it appears to be timely for clinical nursing education to transform its pedagogics to technology-based ones to fill educational gaps demonstrated by conventional pedagogics for enhancing nursing

students' motivation in clinical learning, particularly in low- and middle-income countries such as Tanzania [23]. As it has worked elsewhere, web-based learning is currently becoming an increasingly vital instructional tool in nursing education, as it offers the potential to promote motivation to learn among students [24].

Scholars including Mico et al [25] have highlighted that the web-based learning approach, which has been endorsed as an essential educational tool, responds effectively and efficiently to nurses' needs and experiences in their clinical practices. Web-based technology in education is linked with a networking interactive model to promote communication with a feedback mechanism to enhance active learning between the users [26]. However, little about the integration of technology in clinical nursing education has been documented in Tanzania to mentor, supervise, support, monitor, and evaluate nursing students during their clinical placement [27]. If the situation remains unattended, nursing students will continue to be developed conventionally and continue to be unmotivated in their clinical learning, gaining little clinical competencies to work independently and confidently to deliver ethical, quality, and cost-effective care to people.

Therefore, this study intended to fill the gap by primarily using descriptive and analytical methods to examine undergraduate nursing students' desire for clinical learning both before and after participating in the program.

Methods

This study was conducted by taking into consideration international and national research standards and ethics. Moreover, it was informed by the institutional postgraduate guidelines and regulations of The University of Dodoma [28].

Study Location

The study was conducted in Dodoma Regional Referral Teaching Hospital in Tanzania, which accommodates a large number of nursing students from different nursing training institutions within Dodoma region. Some previous scholars [11,29,30] suggest that proximity to the research environment where the intervention is being piloted or implemented aids in receiving rapid input from the consulted experts, trainers, and participants, and ensuring periodic monitoring of its integrity. Aside from the availability of nursing students, the location was chosen because of the availability and accessibility of the consulted experts and trainers for their evaluation and appraisal of the web-based clinical monitoring system throughout the design and piloting procedures.

Study Design and Approach

As it has also been used by some previous scholarly works [31], this study used an uncontrolled longitudinal quasi-experimental design (pre-post tests) with a quantitative research approach among 589 randomly selected undergraduate nursing students in Tanzania from February 2021 to March 2021. The study began with a screening method to choose individuals who satisfied the inclusion criteria and were willing to engage in the study, followed by a baseline assessment to determine their initial level of motivation in clinical learning. Following the intervention, the same participants were exposed to the system for 3 weeks, with 1 week set apart for the end line evaluation (posttest) after the intervention as a follow-up assessment.

Study Population

To maximize the diversities of an intervention's effects, the study recruited undergraduate nursing students in diploma and bachelor's degree programs from 2 middle and 2 higher training institutions in the Dodoma region of Tanzania's central region.

Sample Size Determination

The minimum sample size of this study was determined based on the findings from previous studies [5]. Their findings revealed a baseline mean score of knowledge about how to plan learning activities of 53.10, whereas the end line mean score was 54.21. The following formula was used to determine the minimum sample size as suggested by previous studies [29,31] to be used when researchers wish to conduct an uncontrolled quasi-experimental study design:

where n=minimum sample size

$Z\alpha$: Tabulated Z value set at 95% (1.96) CI

$Z\beta$: Tabulated Z value set at 80% (0.84) power to demonstrate a statistical difference between pretest and posttest.

σ: Polled SD = 7.511399669835177

SD_1: SD 1 (from previous studies = 10.21)

SD_2: SD 2 (from previous = 11.02)

δ: Mean difference $(M_2 - M_1)^2 = 0.11$

M_1: Mean 1 (from previous studies = 53.10)

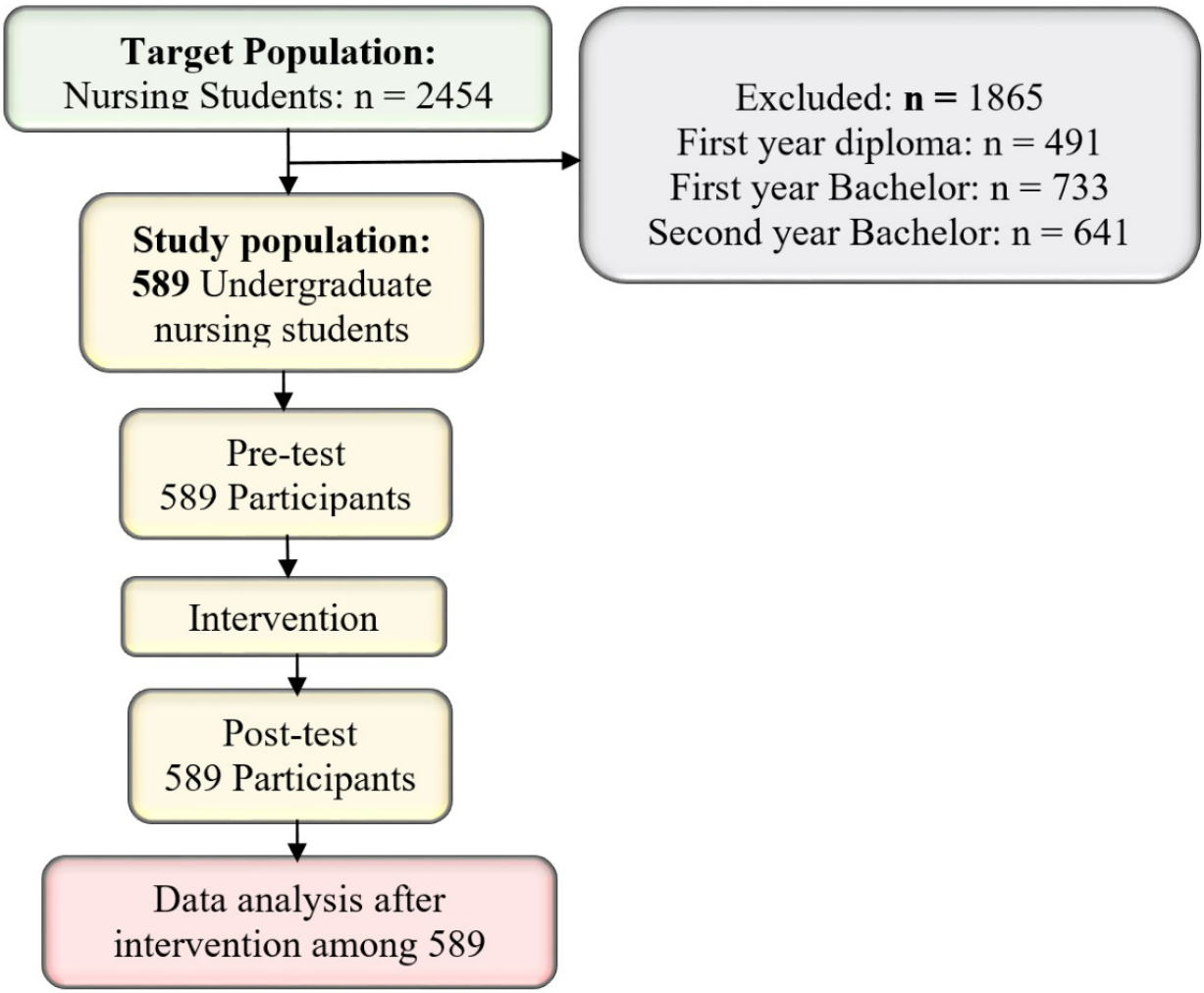
M_2: Mean 2 (from previous studies = 53.21)

With the addition of a 10% attrition adjustment of the calculated sample size (n=54) = (535+54) = 589. Therefore, the minimum sample size of this study was 589 nursing students.

Recruitment Procedures of the Study Participants

The framework of recruiting study participants has been benchmarked from some previous scholarly works [11,32-37]. As shown in Figure 1, 2454 nursing students were eligible to join the study. However, 589 nursing students met the inclusion criteria and participated in the study, assessed at baseline, at end line, and their data analyzed. Sums (1865/2454, 76%) were excluded due to various reasons including not having started clinical placements (first-year diploma in nursing [n=491], a first-year bachelor of science in nursing [n=733], and a second-year bachelor of science in nursing [n=641]). There was no loss to follow-up among nursing students who joined the study and thus the completion rate was 100%.

Figure 1. A flow pattern of sampling procedure among nursing students. From field data (2021).



Eligibility Criteria

Inclusion Criteria

Nursing students were recruited for this study based on their willingness after being informed about the purpose, benefits, and drawbacks of participating in this study. The study recruited second-, third-, and fourth-year undergraduate nursing students who were in clinical placements at the time of this study. In addition, undergraduate nursing students with Information and Communication Technology literacy and those with iPhones or iPads were recruited for the study.

Exclusion Criteria

Nursing students who reported being unwell and unable to converse or participate in the study were excluded from the study. Students without institutional registration numbers, those unable to use computers or smartphones, and nursing students recruited for other studies or projects were not eligible to participate in this study.

Sampling Procedures

As recommended by previous scholars [38-41], probability sampling techniques through multistage sampling methods were used to reach and study nursing students in this study. Stage 1: a simple random sampling technique by lottery method was used to select regions and districts. Stage 2: As shown in Tables 1 and 2, stratified sampling methods were used to select higher training institutions (institutions A and B). On the other hand, lower training institutions (institutions C and D) were randomly selected using a stratified sampling technique.

Nevertheless, a stratified random sampling technique was used to select the nursing program (diploma in nursing and bachelor of science in nursing) and classes (who are in their clinical placement). The training health facility was selected purposely because it is only a major regional referral hospital in the Dodoma region used by multiple training institutions for training nursing students in clinical settings. Stage 3: systematic sampling methods were used to select the study participants. As shown in Tables 1 and 2, a proportionate formula: $n=[P_i \times (n/p)]$ was used to determine a minimum sample size of nursing students per training institution and their program, respectively, whereas n_i is the proportional sample size, P_i is the total targeted population, n is the number of nursing students, and P is the minimum sample size of the study.

Table 1. Proportionate sample size by the sampled training institutions[a].

Name of training institution	Total population, n	Proportionate, n
Institution A	968	232
Institution B	998	239
Institution C	367	88
Institution D	121	30

[a]From Study plan (2021). Total number of nursing students among the sampled training institutions = 2454 ($P_i \times n/P$).

Table 2. Proportionate sample size by training nursing program[a].

Training institution	Nursing program	Total population, n	Proportionate ($P_i \times n/P$)
Institution A	Diploma in nursing	230	55
Institution A	Bachelor in nursing	738	177
Institution B	Bachelor in nursing	828	198
Institution B	Diploma in nursing	171	41
Institution C	Diploma in nursing	367	88
Institution D	Diploma in nursing	121	30

[a]From Study plan (2021).

Intervention

The intervention was completed in 3 weeks, with 1 week set aside for the final evaluation (posttest). It was carried out on a daily basis (6 hours a day), per the clinical rotation and placement schedules established by the nursing students' respective training institutions. Within the 3 weeks of the intervention, all students were expected to perform and complete the assigned clinical learning task or activities and models on a daily basis and within 6 hours of their duty shifts. The primary aim of the intervention was to descriptively and analytically measure a change in motivation in clinical learning among undergraduate nursing students before and after participating in the program. Trained research trainers who also had clinical nursing education competence implemented the intervention. As illustrated in Figure 2, the system included a welcoming window as well as several menus or nodes such as clinical instructors, academic staff, clinical nursing procedures, clinical attendance, evaluation forms, announcements, code generator, professional programs, clinical library, system feedbacks and reports, and students' node, which included students' profiles, clinical notes, attendance, evaluation forms, and announcement menus.

Referring to Figure 3, nursing students were required to arrive at their shift very early each day before the tuned time in the system, which was 7:30 AM (East African time), to be assigned codes of a specific day or date generated by the trainers. The generated codes allowed them to log in to the system and sign out in the presence of the trainer at the end of the shift, which was tuned at 1:30 PM (East African time). Any student who arrived late for his or her shift and who did not provide web-based feedback on his or her performance of the chosen clinical nursing procedure in the presence of clinical instructors or trainers in between shifts, or left the shift before the fixed time of signing out was considered absent in that particular shift. The system provided a daily shift attendance report that included the dates, students' names and registration numbers, department, ward or unit, duty shift, clinical nursing procedure executed, and evaluation score for each student.

Nonetheless, as shown in Figure 4, several nursing clinical procedures were imported into the system, including giving and receiving reports; patients admission; changing patients' positions on a bed; counseling; bed making; administration of intravenous, intermuscular, and oral medications; catheterization; kangaroo mother care; administration of oxygen to patients (oxygen therapy); cardiopulmonary resuscitation; management of patients with preeclampsia or eclampsia; blood transfusion; per-vagina examination; wound dressing; mouth care; management of postpartum hemorrhage; assessment of placenta; management of pregnant women in a first and second stage of labor; assessment of new-born babies; history taking during labor; and vital sign assessment. After selecting a clinical nursing procedure for a particular duty shift, nursing students had to adhere to the 6 stages of conducting it including identification of the patient alongside getting informed consent; preparation of the environment; one-self; equipment appropriate for the chosen procedure; and performing and finishing the procedure. Each clinical nursing procedure was featured in several activities, which nursing students were supposed to follow, adhere, and implement chronologically.

Figure 2. A welcoming window, menu window, and nursing students' node with various menus in the system. From field data (2021).

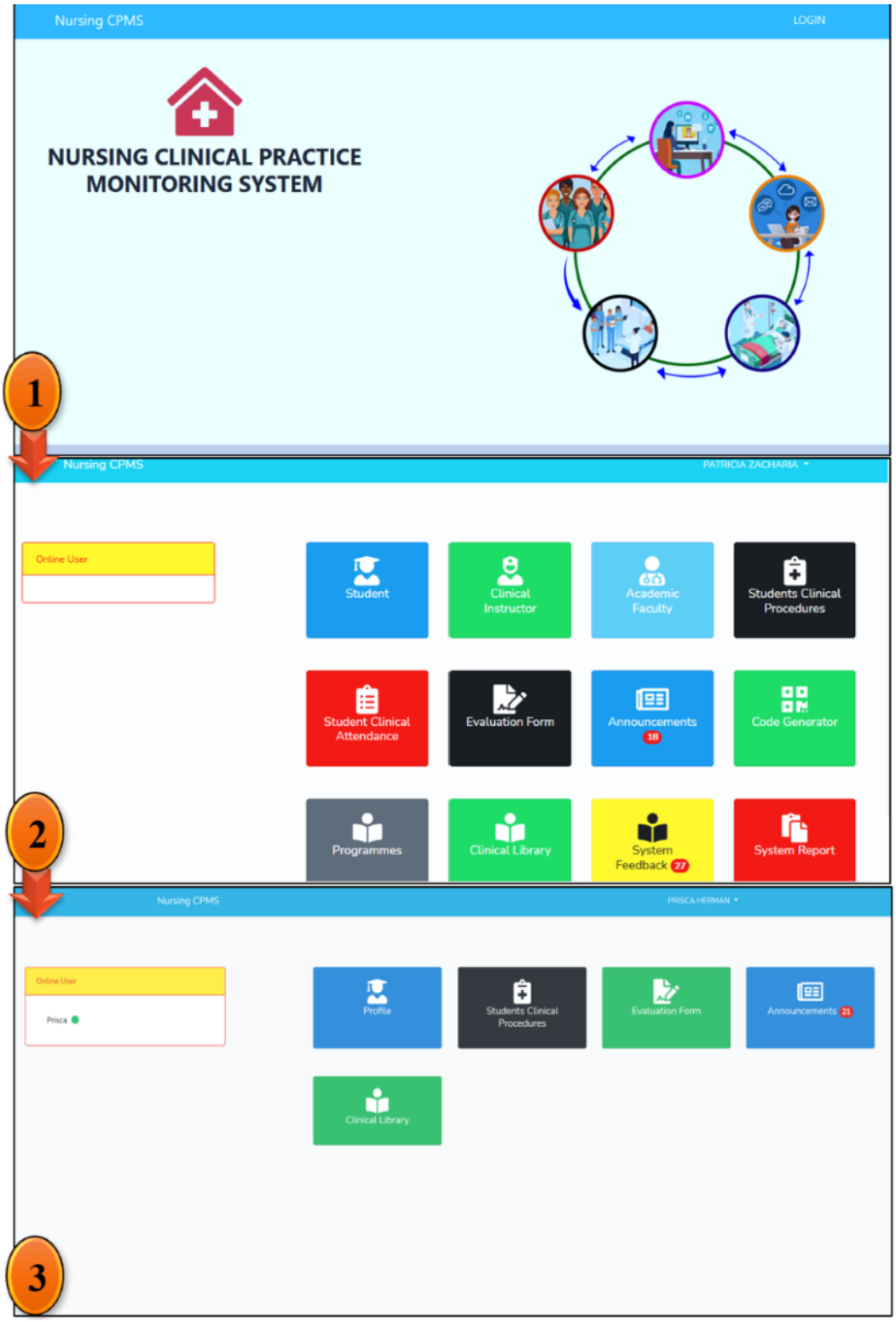


Figure 3. Code generator window, nursing students' login window, and an example of a system-generated report. From field data (2021).

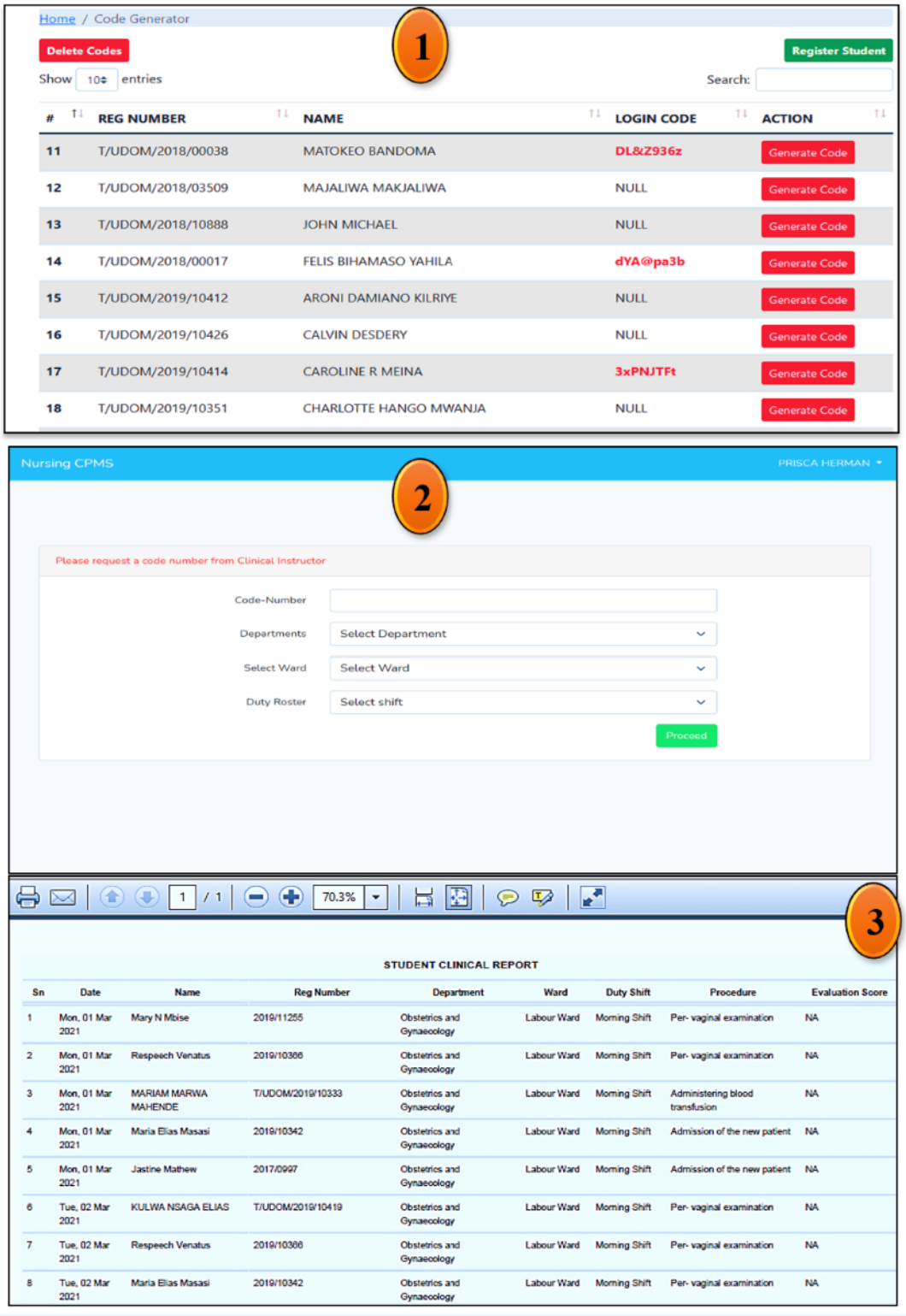
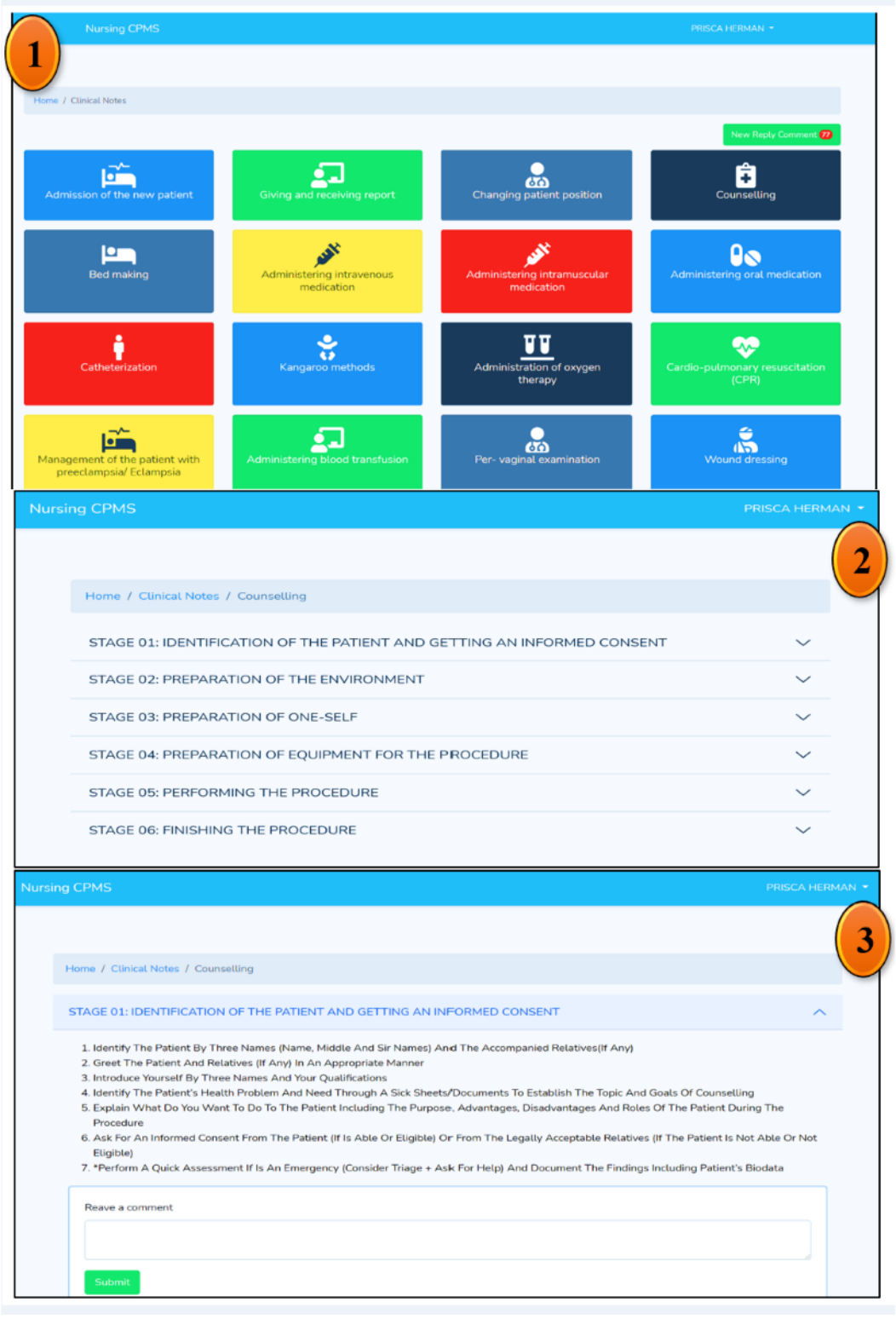


Figure 4. Variety of clinical nursing procedures, stages, and activities imported into the system for nursing students. From field data (2021).



Procedures for Getting Nursing Students Into the System, Supervising, Supporting, Monitoring, and Evaluating Them

Undergraduate nursing students were recruited based on their units of clinical placements including the emergency unit, labor and postnatal ward, pediatric ward, and male medical ward within Dodoma Regional Referral Hospital. Resources such as computers, smartphones or iPads, electricity, web connectivity, clinical nursing procedure checklists, duty rosters, notebooks, pens, and shift objectives were the requirements for the intervention. Nursing students' institutional registration numbers were used as username identities during the login procedures into the system.

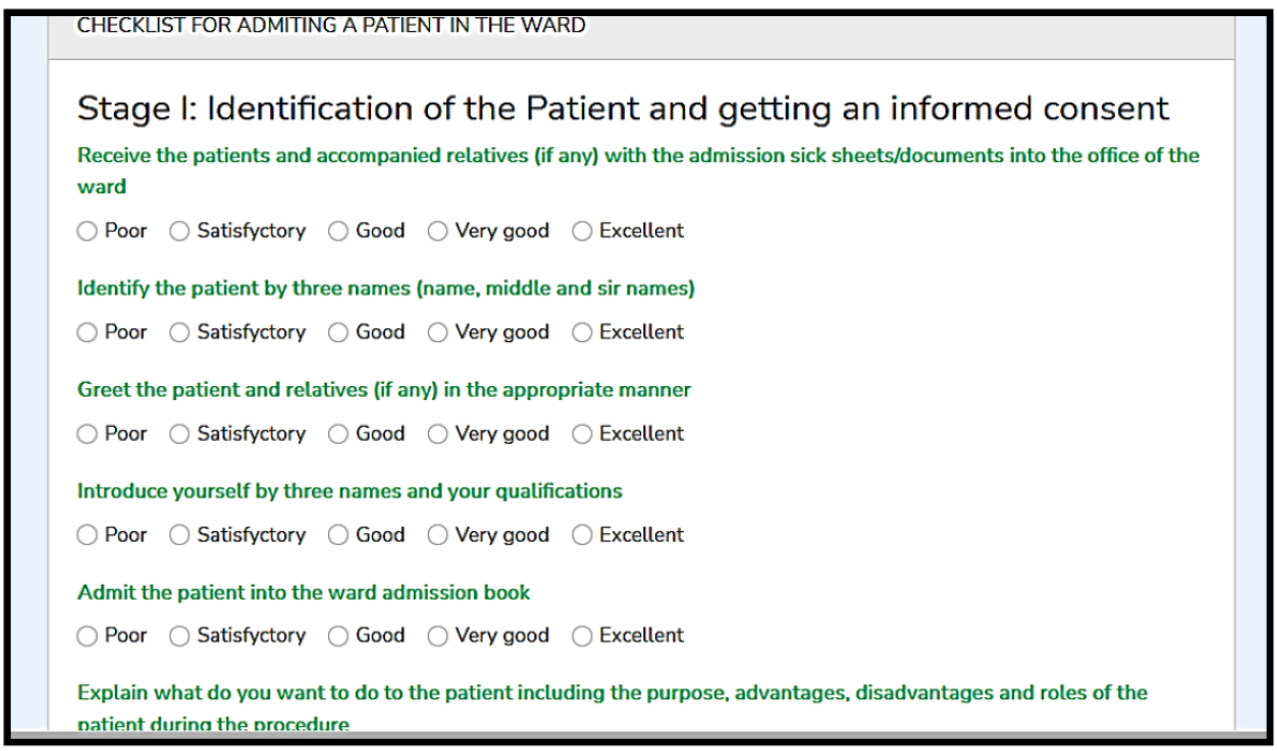
Clinical nursing procedures and evaluation checklists were imported into the system during the intervention to allow nursing students and clinical instructors to access them and have an interactive room to identify and discuss them before starting to

care for the patients on that particular day. Nursing students had to ask for a system-generated code in the presence of a (system implementers) clinical instructor to be able to log in to the system and counted to have reported timely in a particular duty shift as the mode of monitoring and keeping clinical attendance reports among them. Furthermore, after the successful registration of nursing students into the system, system implementers had to confirm the presence of nursing students throughout the specified duty shift.

A shift progress Min-reports (brief or summary report of what a student has done for a particular time of a shift) was posted by nursing students in the system including patients' progress after performing the procedure. On the other hand, trained research trainers had the role of periodically posting announcements into the system to enhance interactive and reciprocal communication among nursing students and academic faculty. They also had the role of summing up nursing students' clinical signs of progress and filling the evaluation forms at the end of a duty shift. All clinical teaching and learning activities performed by the trained research trainers and nursing students were tracked, monitored, and sometimes addressed by the principal investigator (PZH) through a system WebTop. All self-rated evaluations on the performed clinical procedures, day feedback, and experiences of the interactive web-based clinical practice monitoring system among nursing students and system implementers were captured through the system before they signed out at the end of the duty shift.

Only 1 session (morning duty shift) was preferred for the intervention as negotiated by hospital administration and system implementers. The intervention was implemented for 6 hours equivalent to 1 duty shift in a day with a duration ranging from 7:30 AM to 1:30 PM (East African time) among nursing students. Each student had to carry out and finish one of the aforementioned clinical nursing procedures every day for 3 weeks for learning to take place. This amounted to about 20 clinical nursing procedures in total. All students finished the same clinical nursing procedures during the intervention's 3-week duration including giving and receiving reports; patients admission; changing patients' positions on a bed; counseling; bed making; administration of intravenous, intermuscular, and oral medications; catheterization; kangaroo mother care; administration of oxygen to patients (oxygen therapy); cardiopulmonary resuscitation; management of patients with preeclampsia or eclampsia; blood transfusion; per-vagina examination; wound dressing; mouth care; management of postpartum hemorrhage; assessment of placenta; management of pregnant women in first and second stages of labor; assessment of newborn babies; history taking during labor; and vital sign assessment. Assessment of nursing students' motivation in clinical learning was performed using academic motivation questionnaires that were administered to them as pretests and posttests. As shown in Figure 5, the evaluation of the system was performed through nursing students' and system implementers' experiences feedback inventory that was posted in the system and filled out before signing out of the system.

Figure 5. An example of a self-rated evaluation form imported into the system for nursing students. From field data (2021).



Strategies Used to Maintain Adherence to the Intervention Among Nursing Students

This section describes the measures used to increase nursing students' adherence during intervention implementation in the field. Throughout the intervention, nursing students' registration numbers were used as identifiers during system deployment. Being unidentified and acknowledged by name would most likely make nursing students feel at ease and confident in their clinical learning privacy, allowing them to feel free and inspired

to follow the intervention regimen. Furthermore, with the assistance of academic faculty and clinical instructors, the study was carried out under the training institutional clinical rotation schedule, which most likely aided this study in ensuring that nursing students adhered to the intervention because they had to attend their shifts accordingly. The use of technology to assist clinical learning among nursing students appeared to impress them, particularly the ways interactive communications and physical contact were assured to them. Nursing students and system implementers have enhanced adherence to the intervention among nursing students by posting various announcements, peer teaching and learning, and self-evaluation forms.

Data Collection Procedures

The focus of the current effort was mostly on presenting the study's quantitative findings. Quantitative data were gathered at two-time intervals, baseline, and end line (directly postintervention), in different unoccupied rooms, as agreed upon by the administrative authority of the respective training institutes. The research assistants were chosen on purpose based on their eligibility criteria, which included at least 1 year of data-gathering experience. Nursing students were given a brief explanation of completing the questionnaire before having copies provided by research assistants. The researcher and assistants were present throughout the process to supervise and address the overstretched immediate and long-term problems before collecting all copies of completed surveys and protecting them as part of nursing students' confidentially. Before leaving the room, nursing students were informed of the timetable and method for intervention.

Research Tools and Instruments

The data collection tool for the quantitative part of the work consisted of 2 parts: the sociodemographic characteristics profiles and the part that measured nursing students' motivation in clinical learning. Gender, age, marital status, interest, motive for joining the nursing profession, accommodation, and marital status were all part of the sociodemographic profile. The study used a 5-point Likert scale to assess motivation in clinical learning using an Academic Motivation Scale composed of 28 items adopted from previous scholarly works [5,42-44]. The 5-point Likert scale ranged from 1=strongly disagree to 5=strongly agree. Its subscales included intrinsic motivation which was assessed by 12 items, extrinsic motivation (12 items), and amotivation (4 items) of which the findings were dichotomized into an "Agree" (assigned a value of "1") response that described the action to have been performed by the participant; otherwise, "Disagree" ("0" value) for unperformed action or behavior.

Before data collection, the principal component analysis was performed at a measure of Keiser-Meyer-Olkin and Bartlett test value of >0.5 and a significance level of ≤5% to measure the weight of each item. The findings revealed that all 28 items scored >0.3, and thus, they were all retained for data collection. Scoring the variable of motivation in clinical learning was adopted in a study conducted by Millanzi and Kibusi [6] that nursing students who scored 0-16 were categorized into low motivation in clinical, those who scored 17-24 had moderate learning motivation, and nursing students who scored 25 and above demonstrate high motivation in clinical learning.

Validity and Reliability of the Study

The validity and reliability tests were performed first before subjecting the tool to the actual field for data collection. The tool was shared with the subject matter and statistician for suitability of the items, reliability, and ambiguity to fit for knowledge to undergraduate nursing students. The pretest of research tools was conducted among 60 nursing students at a location that was different from the sampled study setting. The finding of a pretest was subjected to the exploratory factor analysis to determine the weight of each item to assess the outcome of interest and the reliability scale analysis to determine the internal consistency of the tool and presented using Cronbach α values as recommended by previous studies [45]. The findings of the scale analysis on motivation in clinical learning were found to be Cronbachα=0.840 (>0.7), which was statistically reliable for the actual data collection.

Data Analysis

An SPSS software program (version 23; IBM Corp) was used to analyze data. Before data analysis, cleaning was done to ensure the completeness, accuracy, and clarity of the information in the questionnaires. A normality test was performed to determine the distribution of data to opt for the mode of data analysis. Findings of the normality test revealed that data motivation in clinical learning was approximately normally distributed, and thus, parametric statistical measurements were adopted. Descriptive statistic was used to establish participants' sociodemographic characteristics profiles of the study participants such as age, sex, program, and year of study, just to mention a few, and the motivation in clinical learning. For inferential statistics, a 2-tailed paired *t* test was performed to determine a comparative mean score change and difference in motivation in clinical learning among nursing students between the pretest and the posttest.

The multiple linear regression analysis model by considering the control of other factors as independent variables such as sociodemographic characteristics profiles and the intervention (interactive web-based clinical practice monitoring system) was performed to establish the extent of association with the outcome variable of interest (motivation in clinical learning). The multiple linear regression was opted for because several factors were controlled during analysis and the outcome variable of interest was treated as a scale variable. The goal of this study was to determine the net effect of the intervention by taking into account and controlling for the sociodemographic characteristic profiles of nursing students, as it was hypothesized that these profiles would also have an impact on the outcome of interest (motivation in clinical learning) over the intervention. Findings of inferential statistics were presented in tabular forms by means, of SDs, *t* values, significance level (<5%), and 95% CI. The following logistic regression model was used:

where Y is dependent variable; a is intercept; b, c, d, and e are slopes; X_1, X_2, X_3, and X_4 are independent (explanatory) variables; and e is residual (error).

Ethical Considerations

This study conformed to The University of Dodoma institution's postgraduate guidelines and regulations after being approved and given an ethical clearance (number MA.84/261/02/81) by the institutional research ethics review committee. All participants provided written informed consent after being explained about the study and their freedom to participate in it. Data collection procedures were performed in separate and unoccupied venues that were available in the respective institution premises. Participants' names were not included in the data collection tools and their information was secured by the principal investigator (PZH) using folders with passwords. Given that participants were in their academic clinical rotations calendar for clinical learning activities, there were no compensations of either time or monetary incentives throughout the study.

Results

Proportional Distribution of Nursing Students by Their Sociodemographic Characteristics Profiles

The completion rate of the study was 100% of the studied participants. Findings in Table 3 indicate that 65.0% (383/589) of nursing students were males while 79.6% (469/589) of the sample were younger than 24 years, with a mean age of 23 (SD 2.689, range 19-50) years. Accommodated participants in their respective training institutions' hostels accounted for 71.5% (421/589) while 63.7% (375/589) of them were enrolled in bachelor of science in nursing and 33.6% (198/589) and 30.1% (177/589) of them were in their fourth and third year of studies, respectively. A majority of nursing students 69.4% (406/589) were not interested in joining nursing programs. However, those who were interested in joining nursing education were driven by a belief that it is a secure profession (567/589, 96.3%), caring to save peoples' lives (491/589, 83.4%), autonomy to practice (478/589, 81.2%), and generous salary and employment benefits (438/589, 74.4%). Other findings were found as shown in the table.

Table 3. Proportional distribution of nursing students by their sociodemographic characteristics (n=589)[a].

Variable	Values
Age (years), mean (SD; range)	23 (2.689; 19-50)
Age (years), n (%)	
<24	469 (79.6)
25-34	115 (19.5)
>35	5 (0.9)
Institution, n (%)	
Training institution A	232 (39.4)
Training institution B	239 (40.6)
Training institution C	88 (14.9)
Training institution D	30 (5.1)
Sex, n (%)	
Male	383 (65.0)
Female	206 (35.0)
Marital status, n (%)	
Single	543 (92.2)
Married	46 (7.8)
Accommodation, n (%)	
In-campus	421 (71.5)
Off-campus	168 (28.5)
Program of study, n (%)	
Diploma in nursing and midwifery	214 (36.3)
Bachelor of science in nursing	375 (63.7)
Year of study , n (%)	
Second-year diploma in nursing	89 (15.1)
Third-year diploma in nursing	125 (21,2)
Third-year bachelor of science in nursing	177 (30.1)
Fourth-year bachelor of science in nursing	198 (33.6)
Interested to join the nursing profession, n (%)	
No	409 (69.4)
Yes	180 (30.6)
Reason to join the nursing profession	
Generously salary and employment benefits, n (%)	
Yes	438 (74.4)
No	151 (25.6)
A secured profession, n (%)	
Yes	567 (96.3)
No	22 (3.7)
Autonomy to practice, n (%)	
No	478 (81.2)
Yes	111 (18.8)
Caring to save people’s lives, n (%)	
No	491 (83.4)

Variable	Values
Yes	98 (16.6)
Opportunity to travel worldwide, n (%)	
Yes	421 (71.5)
No	168 (28.5)
Job availability, n (%)	
Yes	398 (67.5)
No	191 (32.5)

[a]From field data (2021).

The Overall Distribution of the Level of Motivation in Clinical Learning and Their Domains Among Nursing Students

The findings from Table 4 revealed that there was no significant difference between the proportion of nursing students with low and moderate motivation in clinical learning (261/589, 44.3%; and 328/589, 55.7%, respectively). However, baseline findings indicated that none of the nursing students demonstrated high motivation in clinical learning 0.0% (n=0). On the other hand, baseline findings of the motivation domains indicated that 94.7% (558/589) of nursing students were not intrinsically motivated in clinical learning contrary to the end line findings, which indicated that 90.5% (533/589) demonstrated inner motive in clinical learning. Highly motivated nursing students to learn in clinical settings accounted for 67.7% (399/589) while only 4.9% (29/589) of them demonstrated lower motivation in clinical learning. Other findings were observed as shown in Table 4.

Table 4. Overall distribution of the level of motivation in clinical learning and their domains among nursing students in the Dodoma region (N=589)[a].

Variable	Pretest	Posttest
Overall motivation in clinical learning, n (%)		
High learning motivation	0 (0)	399 (67.7)
Moderate learning motivation	328 (55.7)	160 (27.2)
Low learning motivation	261 (44.3)	29 (4.9)
Motivation subscales		
Intrinsic motivation in clinical learning, n (%)		
No	558 (94.7)	56 (9.5)
Yes	31 (5.3)	533 (90.5)
Extrinsic motivation in clinical learning, n (%)		
No	506 (85.9)	106 (18.0)
Yes	83 (14.1)	482 (81.8)
Amotivation, n (%)		
Yes	475 (80.6)	313 (53.1)
No	114 (19.4)	276 (46.9)

[a]Field data (2021).

Overall Mean Score Change and Mean Difference in Motivation in Clinical Learning Between Pretest and Posttest Among Nursing Students

As shown in Table 5, there was a statistically significant increase in mean scores changes of motivation in clinical learning from mean 9.31 (SD 2.315) at baseline to mean 20.87 (SD 5.504) at the end line. A comparative analysis of motivation performance among nursing students between pretest and posttest was found to be statistically significant (mean 11.566, SD 5.667; t_{588}=49.496; P<.001; 95% CI 11.107-12.025). The findings suggest that nursing students scored high on motivation in clinical learning in the posttest as compared with the pretest. Moreover, the findings in Table 5 indicated that there was an increase in mean scores among nursing students per domain of motivation to clinical learning between the pretest (mean 3.74, SD 1.231) and the posttest (mean 9.53, SD 2.762).

Table 5. Overall mean score change and mean difference in motivation in clinical learning between pretest and posttest among nursing students (N=589)[a].

	Pretest, mean (SD)	Posttest, mean (SD)	Mean difference, mean (SD)	*t* test (*df*)	*P* value	95% CI	Effect size (Cohen *d*)	95% CI
Motivation in clinical learning	9.31 (2.315)	20.87 (5.504)	11.56 (5.667)	49.496 (588)	.001	11.107-12.025	2.743	1.011-4.107
Domains of motivation in clinical learning								
Intrinsic	3.74 (1.231)	9.53 (2.762)	5.800 (2.968)	47.421 (588)	.001	5.559-6.040	N/A[b]	N/A
Extrinsic	4.49 (1.42)	8.77 (3.325)	4.276 (3.474)	29.845 (588)	.001	3.994-4.557	N/A	N/A
Amotivation	2.57 (1.1871)	1.08 (1.392)	1.492 (1.644)	22.031 (588)	.001	1.359-1.625	N/A	N/A

[a]From field data (2021).

[b]N/A: not applicable.

Moreover, they also demonstrate higher scores in their extrinsic motivation to learning in clinical settings at the end line (mean 8.77, SD 3.325) than at baseline (mean 4.49, SD 1.42) while amotivation performance decreased from mean 2.57 (SD 1.187) at baseline to mean 1.08 (SD 1.392) at the end line. The effect size of the intervention on motivation in clinical learning among nursing students was computed using Cohen *d* formula (mean 2 minus mean 1 divided by a pooled SD). Findings showed that the intervention demonstrated an effect size of 2.74 (*P*<.001; 95% CI 1.011-4.107), which is a high effect size based on Cohens *d* classifications of effect sizes [46].

The Estimated Effect of an Intervention (Interactive Web-Based Clinical Practice Monitoring System) Controlled for Other Co-Related Factors on Motivation in Clinical Learning Among Nursing Students at Posttest

About 58.5% variation in motivation in clinical learning scores is explained by the explanatory variables included in the model. The overall model was statistically significant (*f*=25.6; *P*=.001). Findings in Table 6 indicate that reasons to join the nursing profession such as due to the opportunity to demonstrate autonomy (β=1.590; *P*=.02; 95% CI 0.279-3.901), the opportunity to travel around the world (β=1.648; *P*=.04; 95% CI 0.583-4.713), job availability (β=1.409; *P*=.001; 95% CI 1.046-5.772), and other corelated factors were statistically significantly associated with motivation in clinical learning among nursing students against their counterparts. The estimated effect (β) of a 3-week intervention to improve nursing students' motivation in clinical learning was 3.041 (*P*=.03, 95% CI 1.022-7.732) when controlled for other correlated factors.

Table 6. The estimated effect of an intervention (interactive web-based clinical practice monitoring system) controlled for other corelated factors on motivation in clinical learning among nursing students at posttest (N=589)[a,b].

Variable	Estimate (β)	SE	*P* value	95% CI
Intervention				
Pretest	1	N/A[c]	N/A	N/A
Posttest	3.041	0.308	.03[d]	1.022-7.732
Institutions	N/A	N/A	N/A	N/A
Institution A	1	N/A	N/A	N/A
Institution B	.729	0.883	.23	1.956-0.467
Institution C	.932	0.794	.65	6.052-3.753
Institution D	.312	1.117	.12	0.194-1.757
Programs				
Diploma	1	N/A	N/A	N/A
Bachelor	.281	0.483	.56	1.229-0.668
Year of study				
Second-year diploma	1	N/A	N/A	N/A
Third-year diploma	.593	0.809	.36	1.806-0.655
Third-year bachelor	.936	0.926	.71	5.315-3.611
Fourth-year bachelor	.744	0.617	.01[d]	0.431-3.417
Age group (years)				
<24	1	N/A	N/A	N/A
24-34	1.149	2.496	.49	0.669-1.407
>35	.782	0.497	.42	1.827-0.768
Sex				
Female	1	N/A	N/A	N/A
Male	.434	0.905	.58	0.851-1.512
Marital status				
Single	1	N/A	N/A	N/A
Married	.575	0.627	.39	1.184-3.007
Accommodation				
In-campus	1	N/A	N/A	N/A
Off-campus	.852	2.272	.73	1.981-1.385
Interested to join the nursing profession				
No	1	N/A	N/A	N/A
Yes	.250	0.043	.051[d]	0.335-0.165
Reason to join nursing				
Autonomy to practice				
Yes	1.590	0.667	.02[d]	0.279-3.901
No	1	N/A	N/A	N/A
Caring patients				
Yes	1.107	0.679	.10	0.226-2.441
No	1	N/A	N/A	N/A
Opportunity to travel around worldwide				

Variable	Estimate (β)	SE	*P* value	95% CI
Yes	1.648	1.512	.04[d]	0.583-4.713
No	1	N/A	N/A	N/A
It is the secured profession				
Yes	.563	0.551	.31	0.519-1.645
No	1	N/A	N/A	N/A
Reasonable payment				
Yes	.033	0.538	.95	1.089-1.023
No	1	N/A	N/A	N/A
Job availability				
Yes	1.409	0.694	.001[d]	1.046-5.772
No	1	N/A	N/A	N/A

[a]From field data (2021).

[b]R^2=0.865, f=116; P<.001, significant at P<.05, and significant at P<.001.

[c]N/A: not applicable.

[d]Variables that are significantly associated with the outcome variable.

Discussion

Principal Findings

In terms of the study's focus and objective, the implementation of an interactive web-based clinical practice monitoring system for nursing students' motivation in clinical learning was feasible and practical in a clinical setting with consistent electricity supply and web connectivity. Nursing students indicated moderate to high levels of motivation in clinical learning after 3 weeks of system implementation, compared with when they were not exposed to it. Nursing students demonstrated a capacity to plan, identify, and access academic resources and help, as well as participate in clinical practices with minimal support from clinical instructors, trainers, and academic faculty, according to the end line assessment. Nonetheless, contrary to the existing practices where nursing students are not allowed formally to use electronic devices in clinical settings, the use of electronic devices while students are in clinical settings such as smartphones, iPads, and computers would most likely extrinsically motivate nursing students to attend their daily duty shifts.

The posttest results show that nursing students who are mentored, supported, monitored, supervised, and evaluated using an interactive web-based clinical practice monitoring system are more efficient in terms of clinical attendance and completing clinical activities on time. The findings of this study indicated that nursing students showed a readiness to stick to their clinical duty roster, report on the clinical environment on time, receive and give reports, and complete their assigned responsibilities using an interactive web-based clinical practice monitoring system. Furthermore, students expressed a desire to ask questions and locate clinical resources to help them learn not just to win a prize or a grade but also to broaden their knowledge and skills. Despite a significant change in clinical learning motivation among nursing students, the system improved clinical attendance as an indicator that it motivated and enhanced their interest and willingness with a sense of being confident, independent, and autonomous to engage in clinical learning activities than when conventional clinical pedagogies such as attendance books, bed tutorials, or assignments were predominantly used.

Similarly to the findings from other previous scholars [2,21], the implementation of an interactive web-based clinical practice monitoring system would allow nursing students to interact with one another while performing clinical nursing procedures, as well as interact instantly and timely with trainers, clinical instructors, or academic faculty for any support or mentorship. Nonetheless, as argued by some previous scholars [5,47] that individuals' behaviors are sometimes shaped by their personalities, this study found that nursing students' motivation in clinical learning was partly attributed to what motivated them to join nursing education programs, such as challenges in job availability, opportunities to demonstrate autonomy in nursing, and opportunities to travel around the world when individuals enrolled in the nursing profession. Such corelated aspects to the intervention would most likely drive nursing students to study nursing programs extrinsically rather than intrinsically to match their academic and living goals.

Referring to the findings from some previous scholarly works on students' abilities to demonstrate an interest in their learning activities becomes a precursor for them to be motivated to identify and locate learning resources and thus engage in learning activities actively [48,49]. Similarly, scholars' [6,50,51] low motivation in clinical learning has been linked to uninteresting clinical teaching techniques, an uncomfortable learning environment, a lack of interest, and unclear clinical objectives, all of which contribute to clinical absenteeism among nursing students. This study's findings on motivation to learn are consistent with those found by Millanzi and Kibusi [6], for example, who argued that innovative pedagogy has the potential of improving learning motivation among nursing students.

Nevertheless, Allvin et al [52], claimed that while the clinical environment is important for nursing students' academic achievement, students' learning motivation is positively correlated with an adequate number of competent and qualified clinical instructors who use innovative clinical pedagogies that enhance their motivation in clinical learning, including the prescription and integration of interactive web-based clinical practice monitoring system in nursing curricula. In the same way, Aghajari et al [53] observed that the majority of nursing students lack academic motivation during clinical placement practices because the clinical learning environment, as well as clinical teaching and learning pedagogies, does not motivate them to learn and meet their clinical academic potential as lifelong learners. This study's and prior research findings indicate to underline that a favorable, learner-centered, and technology-based innovative clinical pedagogy may positively boost clinical learning motivation among nursing students in nursing education.

Limitations of the Study

The study did not involve a control group to maximize the validity of findings on the efficacy of interactive web-based clinical practice monitoring systems on the outcomes of interest. The use of a single group may obscure the interpretation of the effect size on the outcome variables because standard clinical teaching pedagogies would also produce effects on the outcome of interest. Therefore, findings on the effect of interactive web-based clinical practice monitoring systems need to be interpreted cautiously by considering this limitation. The study suffers from a methodological limitation, as it did not adopt a randomized controlled trial (a true intervention) to estimate the random effect of the interactive web-based clinical practice monitoring system on the outcome variables over the standard clinical teaching pedagogies. Therefore, findings need to be treated and interpreted cautiously by considering that with a quasi-experimental design random effect of the intervention on the outcome, variables would not be established without outweighing its effect over the standard clinical teaching pedagogy (control group) if it could be involved.

Conclusions

The findings of this study demonstrate that it is possible to teach, mentor, supervise, support, monitor, and assess nursing students throughout their clinical placements by adopting, implementing, and evaluating an interactive web-based clinical practice monitoring system. As a minimum exposure of at least 3 weeks of the interactive web-based clinical practice monitoring system, nursing students may demonstrate nursing students' attendance and motivation in clinical learning by their will than is now performed, where sanctions and other associated techniques are used to force them to attend their daily clinical duty shifts accordingly. Incorporating technology into clinical nursing education pedagogics nursing curricula can be an alternative educational technique for educators in nursing education to facilitate clinical learning activities to develop motivated and passionate undergraduate nursing students to engage and learn efficiently and effectively during their clinical placements. Nonetheless, managing a big group of nursing students was proven to be achievable with the use of an interactive web-based clinical practice monitoring system.

Furthermore, an interactive web-based clinical practice monitoring system was feasible to not only mentor, supervise, monitor, and support nursing students but also record students' clinical attendance and the type and number of clinical nursing procedures learned and practiced, as well as generate clinical formative evaluation reports. In other words, an interactive web-based clinical practice monitoring system can be used as an innovative clinical pedagogical approach in clinical teaching and learning to improve nursing students' motivation in clinical learning as a precursor to clinical competence for quality and cost-effective care to people.

Acknowledgments

This study would not have been possible without the support of The University of Dodoma and the Dodoma Regional Referral Training Hospital. On the other hand, the publication of this work would not be possible without support from JMIR Publications for granting a discount of article-processing charges for this work to be published. The work was privately sponsored.

Data Availability

The datasets generated during and/or analyzed during this study are available from the corresponding author on reasonable request via patriciaz1006@gmail.com.

Authors' Contributions

PH participated in the conceptualization, methods and materials, resources, investigation, writing the original draft, revision, and editing the draft of the work. SK contributed to the conceptualization, methodology, supervision, revision, and editing of the draft of the work. WM participated in the conceptualization, methods and materials, data curation, analysis, writing, and editing the original draft of the work. The authors have read and approved the manuscript.

Conflicts of Interest

None declared.

References

1. Oguguo BC, Ajuonuma JO, Azubuike R, Ene CU, Atta FO, Oko CJ. Influence of social media on students' academic achievement. Int J Eval Res Educ 2020;9(4):1000.
2. Ghasemi MR, Moonaghi HK, Heydari A. Strategies for sustaining and enhancing nursing students' engagement in academic and clinical settings: a narrative review. Korean J Med Educ 2020;32(2):103-117 [FREE Full text] [doi: 10.3946/kjme.2020.159] [Medline: 32486620]
3. Soroush A, Andaieshgar B, Vahdat A, Khatony A. The characteristics of an effective clinical instructor from the perspective of nursing students: a qualitative descriptive study in Iran. BMC Nurs 2021;20(1):36 [FREE Full text] [doi: 10.1186/s12912-021-00556-9] [Medline: 33663461]
4. Carey M, Kent B, Latour J. Experiences of undergraduate nursing students in peer assisted learning in clinical practice: a qualitative systematic review. JBI Database System Rev Implement Rep 2018;16(5):1190-1219 [FREE Full text] [doi: 10.11124/JBISRIR-2016-003295]
5. Millanzi W, Kibusi S. Exploring the effect of problem based facilitatory teaching approach on motivation to learn: a quasi-experimental study of nursing students in Tanzania. BMC Nurs 2021;20(1):3 [FREE Full text] [doi: 10.1186/s12912-020-00509-8] [Medline: 33397332]
6. Millanzi W, Kibusi S. Exploring the effect of problem-based facilitatory teaching approach on metacognition in nursing education: a quasi-experimental study of nurse students in Tanzania. Nurs Open 2020;7(5):1431-1445 [FREE Full text] [doi: 10.1002/nop2.514] [Medline: 32802363]
7. Forman T, Armor D, Miller A. A review of clinical informatics competencies in nursing to inform best practices in education and nurse faculty development. Nurs Educ Perspect 2020;41(1):E3-E7. [doi: 10.1097/01.NEP.0000000000000588] [Medline: 31860501]
8. Chewaka Gamtessa L. Correlation between academic and clinical practice performance of nursing students at a pediatrics and child health nursing course; Mizan-Tepi University, Ethiopia. Adv Med Educ Pract 2021;12:155-162 [FREE Full text] [doi: 10.2147/AMEP.S294650] [Medline: 33623467]
9. Kaphagawani N, Useh U. Clinical supervision and support: exploring pre-registration nursing students' clinical practice in Malawi. Ann Glob Health 2018;84(1):100-109 [FREE Full text] [doi: 10.29024/aogh.16] [Medline: 30873795]
10. National Association of School Nurses. Framework for 21st century school nursing practice™: clarifications and updated definitions. NASN Sch Nurse 2020;35(4):225-233. [doi: 10.1177/1942602X20928372] [Medline: 32491974]
11. Millanzi WC, Herman PZ, Hussein MR. The impact of facilitation in a problem-based pedagogy on self-directed learning readiness among nursing students: a quasi-experimental study in Tanzania. BMC Nurs 2021;20(1):242 [FREE Full text] [doi: 10.1186/s12912-021-00769-y] [Medline: 34872553]
12. Rahman T, Ghani DM, Kausar S. Factors contributing to absenteeism in undergraduates nursing students. Pakistan J Educ 2021;38(2).
13. Cusack L, Thornton K, Drioli-Phillips P, Cockburn T, Jones L, Whitehead M, et al. Are nurses recognised, prepared and supported to teach nursing students: mixed methods study. Nurse Educ Today 2020;90:104434. [doi: 10.1016/j.nedt.2020.104434] [Medline: 32315837]
14. Lenouvel E, Chivu C, Mattson J, Young JQ, Klöppel S, Pinilla S. Instructional design strategies for teaching the mental status examination and psychiatric interview: a scoping review. Acad Psychiatry 2022;46(6):750-758 [FREE Full text] [doi: 10.1007/s40596-022-01617-0] [Medline: 35318592]
15. Sommer I, Larsen K, Nielsen CM, Stenholt BV, Bjørk IT. Improving clinical nurses' development of supervision skills through an action learning approach. Nurs Res Pract 2020;2020:9483549 [FREE Full text] [doi: 10.1155/2020/9483549] [Medline: 32148957]
16. Tesfaye TS, Alemu W, Mekonen T. Perceived clinical practice competency and associated factors among undergraduate students of medicine and health science collage in Dilla University, SNNPR, Ethiopia. Adv Med Educ Pract 2020;11:131-137 [FREE Full text] [doi: 10.2147/AMEP.S235823] [Medline: 32110133]
17. Mashala YL. The impact of the implementation of free education policy on secondary education in Tanzania. Int J Acad Multidiscip Res Internet 2019;3(1):6-14 [FREE Full text]
18. Thirumoorthy G. Outcome based education (OBE) is need of the hour. Int J Res GRANTHAALAYAH 2021;9(4):571-582.
19. Mlaba ZP, Emmamally W. Describing the perceptions of student nurses regarding barriers and benefits of a peer-mentorship programme in a clinical setting in KwaZulu-Natal. Health SA 2019;24:1118 [FREE Full text] [doi: 10.4102/hsag.v24i0.1118] [Medline: 31934419]
20. Agarwal G, Gaber J, Richardson J, Mangin D, Ploeg J, Valaitis R, et al. Pilot randomized controlled trial of a complex intervention for diabetes self-management supported by volunteers, technology, and interprofessional primary health care teams. Pilot Feasibility Stud 2019;5:118 [FREE Full text] [doi: 10.1186/s40814-019-0504-8] [Medline: 31673398]
21. Kinnunen U, Heponiemi T, Rajalahti E, Ahonen O, Korhonen T, Hyppönen H. Factors related to health informatics competencies for nurses—results of a national electronic health record survey. Comput Inform Nurs 2019;37(8):420-429. [doi: 10.1097/CIN.0000000000000511] [Medline: 30741730]
22. Yuliawan D, Widyandana D, Nur Hidayah R. Utilization of Nursing Education Progressive Web Application (NEPWA) media in an education and health promotion course using Gagne's model of instructional design on nursing students:

quantitative research and development study. JMIR Nurs 2020;3(1):e19780 [FREE Full text] [doi: 10.2196/19780] [Medline: 34345790]
23. Gemuhay H, Kalolo A, Mirisho R, Chipwaza B, Nyangena E. Factors affecting performance in clinical practice among preservice diploma nursing students in Northern Tanzania. Nurs Res Pract 2019;2019:3453085 [FREE Full text] [doi: 10.1155/2019/3453085] [Medline: 30941212]
24. Anisenkov A, Zhadan D, Logashenko I. A Web-based control and monitoring system for DAQ applications. EPJ Web Conf 2019;214:01049. [doi: 10.1051/epjconf/201921401049]
25. Mico O, Santos P, Caldo R. Web-based smart farm data monitoring system?: a prototype. Int J Agric Environ Food Sci 2016;3(3):85-96.
26. Kotcherlakota S, Pelish P, Hoffman K, Kupzyk K, Rejda P. Augmented reality technology as a teaching strategy for learning pediatric asthma management: mixed methods study. JMIR Nurs 2020;3(1):e23963 [FREE Full text] [doi: 10.2196/23963] [Medline: 34406970]
27. El Idrissi WEM, Chemsi G, Kababi KE, Radid M. Assessment practices of student's clinical competences in nurse education. Open Nurs J 2021;15(1):47-54. [doi: 10.2174/1874434602115010047]
28. UDOM. Regulations and Guidelines for Postgraduate Programmes [Internet]. 3rd ed. In: The University of Dodoma. Dodoma: Office of the Deputy Vice-Chancellor; Academic, Research and Consultancy; 2019.
29. Salim MA, Gabrieli P, Millanzi WC. Enhancing pre-school teachers' competence in managing pediatric injuries in Pemba Island, Zanzibar. BMC Pediatr 2022;22(1):691 [FREE Full text] [doi: 10.1186/s12887-022-03765-6] [Medline: 36461011]
30. Cleophace W. The effectiveness of integrated reproductive health lesson materials in a problem-based pedagogy on knowledge, soft skills, and sexual behaviors among adolescents in Tanzania. In: The University of Dodoma. Dodoma, Tanzania: The University of Dodoma; 2021:1-538.
31. Haramba S, Millanzi W, Seif S. Enhancing nursing student presentation competences using Facilitatory Pecha Kucha presentation pedagogy: a quasi-experimental study protocol in Tanzania. BMC Med Educ 2023;23(1):628 [FREE Full text] [doi: 10.1186/s12909-023-04628-z] [Medline: 37661279]
32. Shitindi G, Millanzi W, Herman P. Perceived motivators, knowledge, attitude, self-reported and intentional practice of female condom use among female students in higher training institutions in Dodoma, Tanzania. Contracept Reprod Med 2023;8(1):16 [FREE Full text] [doi: 10.1186/s40834-022-00208-6] [Medline: 36750970]
33. Mwanja C, Herman P, Millanzi W. Prevalence, knowledge, attitude, motivators and intentional practice of female genital mutilation among women of reproductive age: a community-based analytical cross-sectional study in Tanzania. BMC Womens Health 2023;23(1):226 [FREE Full text] [doi: 10.1186/s12905-023-02356-6] [Medline: 37138247]
34. Millanzi W, Herman P, Mtangi S. Knowledge, attitude, and perceived practice of sanitary workers on healthcare waste management: a descriptive cross-sectional study in Dodoma region,Tanzania. SAGE Open Med 2023;11:20503121231174735 [FREE Full text] [doi: 10.1177/20503121231174735] [Medline: 37223674]
35. Millanzi WC, Herman PZ, Ambrose BA. Feeding practices, dietary adequacy, and dietary diversities among caregivers with under-five children: a descriptive cross-section study in Dodoma region, Tanzania. PLoS One 2023;18(3):e0283036 [FREE Full text] [doi: 10.1371/journal.pone.0283036] [Medline: 36947536]
36. Millanzi W, Osaki K, Kibusi S. Non-cognitive skills for safe sexual behavior: An exploration of baseline abstinence skills, condom use negotiation, self-esteem, and assertiveness skills from a controlled problem-based learning intervention among adolescents in tanzania non-cognitive skills for safe sexual behavior: an exploration of baseline abstinence skills, condom use negotiation, self-esteem, and assertiveness skills from a controlled problem-based learning intervention among adolescents in Tanzania. Glob J Med Res Internet 2020;20:43-50 [FREE Full text]
37. Millanzi W, Osaki K, Kibusi S. Attitude and prevalence of early sexual debut and associated risk sexual behavior among adolescents in Tanzania; evidence from baseline data in a Randomized Controlled Trial. BMC Public Health 2023;23(1):1758 [FREE Full text] [doi: 10.1186/s12889-023-16623-6] [Medline: 37689638]
38. Millanzi W, Kibusi SM, Osaki KM. Effect of integrated reproductive health lesson materials in a problem-based pedagogy on soft skills for safe sexual behaviour among adolescents: a school-based randomized controlled trial in Tanzania. PLoS One 2022;17(2):e0263431 [FREE Full text] [doi: 10.1371/journal.pone.0263431] [Medline: 35192640]
39. Millanzi W, Kibusi S, Osaki K. Effect of integrated reproductive health lesson materials in a problem-based pedagogy on soft skills for safe sexual behaviour among adolescents: a school-based randomized controlled trial in Tanzania. PLoS One 2022;17(2):e0263431 [FREE Full text] [doi: 10.1371/journal.pone.0263431] [Medline: 35192640]
40. Millanzi W, Osaki K, Kibusi S. The effect of educational intervention on shaping safe sexual behavior based on problem-based pedagogy in the field of sex education and reproductive health: clinical trial among adolescents in Tanzania. Health Psychol Behav Med 2022;10(1):262-290 [FREE Full text] [doi: 10.1080/21642850.2022.2046474] [Medline: 35251774]
41. Millanzi WC. Adolescents' World: Know One Tell One Against Unsafe Sexual Behaviours, Teenage Pregnancies and Sexually Transmitted Infections Including Chlamydia. In: IntechOpen. London, UK: IntechOpen; 2022:1541-1548.
42. Lucidi FF. The Academic Motivation Scale (AMS): factorial structure, invariance, and validity in the Italian context. TPM Test Psychom Methodol Appl Psychol 2008;15(4):211-220.
43. Utvær BK, Haugan G. The academic motivation scale: dimensionality, reliability, and construct validity among vocational students. J Vocat Educ Train 2016;6(2):17-45.

44. Taghipour H, Gilaninia S, Jalali M, Azizipour H, Javad S, Razaghi R. Standardizing of academic motivation scale. J Basic Appl Sci Res 2012;2(2):1186-1192.
45. Taber KS. The use of Cronbach's alpha when developing and reporting research instruments in science education. Res Sci Educ 2017;48(6):1273-1296 [FREE Full text] [doi: 10.1007/s11165-016-9602-2]
46. Brydges C. Effect size guidelines, sample size calculations, and statistical power in Gerontology. Innov Aging 2019;3(4):igz036 [FREE Full text] [doi: 10.1093/geroni/igz036] [Medline: 31528719]
47. Ibrahem S, Elsayed N, Mohamed H, Shereda A. Impact of clinical simulation on student's levels of motivation, satisfaction and self-confidence in learning psychiatric mental health nursing. Int J Spec Educ 2022;37(3):6214-6228.
48. Taylor I, Bing-Jonsson P, Wangensteen S, Finnbakk E, Sandvik L, McCormack B, et al. The self-assessment of clinical competence and the need for further training: a cross-sectional survey of advanced practice nursing students. J Clin Nurs 2020;29(3-4):545-555. [doi: 10.1111/jocn.15095] [Medline: 31714619]
49. Konttila J, Siira H, Kyngäs H, Lahtinen M, Elo S, Kääriäinen M, et al. Healthcare professionals' competence in digitalisation: a systematic review. J Clin Nurs 2019;28(5-6):745-761. [doi: 10.1111/jocn.14710] [Medline: 30376199]
50. Alaagib NA, Musa OA, Saeed AM. Comparison of the effectiveness of lectures based on problems and traditional lectures in physiology teaching in Sudan. BMC Med Educ 2019;19(1):365 [FREE Full text] [doi: 10.1186/s12909-019-1799-0] [Medline: 31547817]
51. Solomon Y. Comparison between problem-based learning and lecture-based learning: effect on nursing students' immediate knowledge retention [Response To Letter]. Adv Med Educ Pract 2021;12:163-164 [FREE Full text] [doi: 10.2147/AMEP.S305514] [Medline: 33623468]
52. Allvin R, Bisholt B, Blomberg K, Bååth C, Wangensteen S. Self-assessed competence and need for further training among registered nurses in somatic hospital wards in Sweden: a cross-sectional survey. BMC Nurs 2020;19:74 [FREE Full text] [doi: 10.1186/s12912-020-00466-2] [Medline: 32774153]
53. Aghajari Z, Loghmani L, Ilkhani M, Talebi A, Ashktorab T, Ahmadi M, et al. The relationship between quality of learning experiences and academic burnout among nursing students of Shahid Beheshti University of Medical Sciences in 2015. Electron J Gen Med 2018;15(6). [doi: 10.29333/ejgm/93470]

Edited by T de Azevedo Cardoso, T Leung; submitted 22.01.23; peer-reviewed by RJ Medina, L Nunes, MF Deschênes; comments to author 03.08.23; revised version received 11.10.23; accepted 15.03.24; published 23.04.25.

Please cite as:
Herman P, M Kibusi S, C Millanzi W
Effectiveness of an Interactive Web-Based Clinical Practice Monitoring System on Enhancing Motivation in Clinical Learning Among Undergraduate Nursing Students: Longitudinal Quasi-Experimental Study in Tanzania
JMIR Med Educ 2025;11:e45912
URL: https://mededu.jmir.org/2025/1/e45912
doi:10.2196/45912
PMID:

Deconstructing Participant Behaviors in Virtual Reality Simulation: Ethnographic Analysis

Daniel Loeb[1,2*], MD, MEd; Jamie Shoemaker[3*], RN, BSN; Kelly Ely[1,3*], RN, BSN; Matthew Zackoff[1,2,3], MD, MEd

[1]Division of Critical Care Medicine, Cincinnati Children's Hospital Medical Center, 3333 Burnett Avenue, Cincinnati, OH, United States

[2]Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, United States

[3]Center for Simulation and Research, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States

[*]these authors contributed equally

Corresponding Author:
Daniel Loeb, MD, MEd
Division of Critical Care Medicine, Cincinnati Children's Hospital Medical Center, 3333 Burnett Avenue, Cincinnati, OH, United States

Abstract

Background: Virtual reality (VR)–based simulation is an increasingly popular tool for simulation-based medical education, immersing participants in a realistic, 3D world where health care professionals can observe nuanced examination findings, such as subtle indicators of respiratory distress and skin perfusion. However, it remains unknown how the VR environment affects participant behavior and attention.

Objective: This study aimed to describe clinician attention and decision-making behaviors during interprofessional pediatric resuscitation simulations performed in VR. We used video-based focused ethnography to describe how participant attention and behavior are altered in the VR environment and reflect how these changes may affect the educational profile of VR simulation.

Methods: The research team analyzed scenarios with the question, "How does a completely virtual reality environment alter participant attention and behavior, and how might these changes impact educational goals?" Video-based focused ethnography consisting of data collection, analysis, and pattern explanation was conducted by experts in critical care, resuscitation, simulation, and medical education until data saturation was achieved.

Results: Fifteen interprofessional VR simulation sessions featuring the same scenario—a child with pneumonia and sepsis—were evaluated. Three major themes emerged: Source of Truth, Cognitive Focus, and Fidelity Breakers. Source of Truth explores how participants gather and synthesize information in a VR environment. Participants used the patient's physical examination over ancillary data sources, such as the cardiorespiratory monitor, returning to the monitor when the examination did not align with expectations. Cognitive Focus describes the interplay between thinking, communicating, and doing during a VR simulation. The VR setting imposed unique cognitive demands, requiring participants to process information from multiple sources, make rapid decisions, and execute tasks during the scenario. Participants experienced increased task burden when virtual tasks did not mirror real-world procedures, leading to delays and fixation on certain actions. Fidelity Breakers reflects how technical and environmental factors disrupted focus and hindered learning. Navigational challenges, such as unintended teleportation and difficulties interacting with the virtual patient and equipment, disrupted participant immersion. These challenges underscore the current limitations of VR in reproducing the tactile and procedural aspects of real clinical care.

Conclusions: Participants' focus on the physical examination findings in VR, as opposed to the cardiorespiratory monitor, potentially indicates simulation of an identical, more patient examination-centered approach to clinical data gathering. In addition, the multiple data sources allowed for participant cognitive load and task burden that may better mirror real-life clinical care. However, technical features that required straying from real-world task completion, as well as other navigational and interactional challenges in VR, led to breaks in fidelity and shifted focus away from the learning objectives. These findings underscore the need for continued research on how simulation modality, fidelity, and technical challenges may influence participant attention and behavior, to allow thoughtful alignment between desired learning objectives and mode of training.

(JMIR Med Educ 2025;11:e65886) doi:10.2196/65886

KEYWORDS

simulation; virtual reality; video review; resuscitation; CPR; immersion; VR; ethnographic analysis; tool; respiratory distress; pneumonia; sepsis; children; scenarios; ethnography; VR-based simulation; training; training method; development; cardiopulmonary resuscitation

Introduction

Simulation-based medical education (SBME) emerged as a way to practice skills without placing patients at risk [1]. Simulation has been shaped by the tools and technology available and has primarily consisted of ever-evolving computerized manikins designed to represent a patient. Computerized manikins have been well aligned with training toward and evaluation of specific goals and behaviors, such as practicing high-quality cardiopulmonary resuscitation [2], enhancing team dynamics [3], and establishing procedural competency [4]. Recently, technological advancements have allowed SBME to move beyond computerized manikins. Digital simulation environments can allow for a simulated patient to exist in 3D space as a virtual patient. These new approaches to simulation change the way participants perceive and interact with the patient during training scenarios [5-10].

Institutions have begun to explore the use of several digital-based simulation modalities, including augmented reality (AR) and virtual reality (VR)–based simulation. AR simulation places a 3D virtual patient in an actual physical space [11]. This virtual patient can either exist alone or as a superimposed image anchored onto a physical manikin [12]. AR passthrough technology obviates the need to replicate a physical environment in a virtual space, since the simulation occurs in the physical world. However, the technology needed to allow a seamless passthrough AR experience remains limited [13].

Immersive VR places participants in a fully digital 3D environment. This environment can be a digital twin of the actual clinical space used by the participants [12,14] or a completely novel training environment [8]. Achieving high-fidelity physical interactions and closely matching real-life clinical skills in VR remains challenging [15].

SBME is built upon the foundational presumption that the training environment, and the associated degree of realism and fidelity, impacts learning through the quantity and complexity of cognitive tasks required (ie, cognitive load theory) [16], as well as through allowing new knowledge construction through experiences and interactions (ie, constructivism) [17]. Augmented and VR, which significantly change the environment where participants engage, learn, and grow, could influence the attention (ie, directed cognitive engagement) and behaviors (ie, observable participant actions) exhibited by participants. In addition, while some of this influence may be purposeful (ie, realistic presentation of clinical findings), it is unknown whether there are additional influences that may be inherent to the training modality and independent of the clinical context and learning objectives. While these new and innovative tools present opportunities for training and assessment, an understanding of their influence on trainee attention and behavior is necessary to optimize our ability to use them to achieve ideal educational outcomes.

Our previous findings found that participant behaviors differed substantially between an AR-enhanced and a manikin-based simulation of an identical clinical scenario. The modality influenced the observed communication dynamics, participant behaviors, and fidelity breakpoints that influenced participant attention [5]. However, a similar description of attention and behavior in a VR clinical simulation currently does not exist. As these digital training modalities are becoming increasingly accessible, addressing this gap is crucial for supporting educators in developing and implementing precision training that is potentially more impactful through thoughtful alignment between learning objectives and chosen training modality. This study aimed to identify and categorize provider attention and behavior during VR SBME to aid in establishing which educational objectives are best addressed through this novel simulation modality.

Methods

Theoretical Framework

This study is anchored in the theoretical construct that the inherent characteristics of a simulation modality may influence participant attention and behavior. In addition, understanding that influence is key to ensuring that specific learning objectives are feasible to accomplish with the use of a given simulation modality. This aligns with the educational theory of constructivism, which posits that learners actively build knowledge by engaging in “hands-on” experiences, experimentation, and reflection within authentic contexts [17]. In a VR simulation, these authentic contexts are approximated by realistic clinical environments that include realistic digital patients and equipment to encourage exploration, role adoption, and collaborative decision-making. However, the same immersive features that make VR engaging (eg, richly detailed visuals, interactivity, and collaboration) may also impose additional cognitive demands (eg, technical work-arounds or deviations from reality) that could influence the construction of knowledge in unintended ways. The impact of those cognitive demands can best be understood through the lens of cognitive load theory, which highlights the limited capacity of working memory and its subsequent impact on knowledge or skill acquisition [16]. With VR simulations, expanded data sources in the form of virtual patients, equipment, environments, or other assets will influence the cognitive load burden on participants and potentially influence their behavior and subsequent learning. At the same time, novel controls and interface complexity may inadvertently add additional cognitive load, drawing learners’ mental resources away from essential clinical tasks and therefore potentially diminishing their bandwidth for the desired learning objectives. The integration of constructivism and cognitive load theory allows for a theoretical framework that aims to capture how participants effectively manage mental resources while constructing new knowledge through immersion, as well as identify when extraneous load hinders that process. This blended theoretical lens guided our data analysis and interpretation of observed learner behaviors in VR-based pediatric resuscitation simulations.

Study Design

We used video-based focused ethnography [18,19] to study a cohort of video-recorded VR simulations. Focused ethnography is a methodological adaptation of traditional ethnography, offering a targeted and time-efficient approach to studying

specific phenomena within specific contexts. This research strategy prioritizes depth of understanding over breadth, using concentrated data collection and a clear research question to describe particular social, cultural, or organizational processes. For this study, the goal was to understand how VR-based simulation affects participant attention and behavior. With this goal in mind, the research team was primed with the initial research question [19]: "How does a virtual reality simulation influence participant attention and behavior, and how might these changes impact educational goals?"

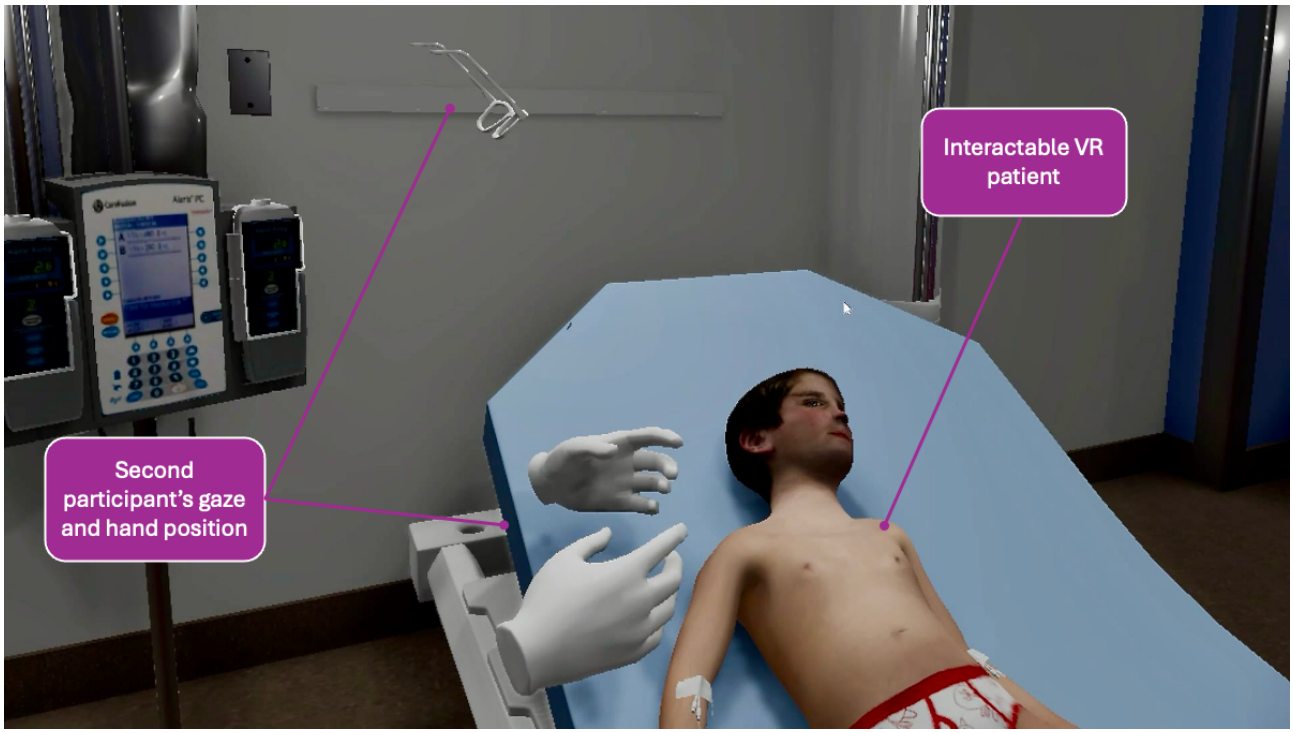
During this focused exploration, the team first identified and classified the data, then described and analyzed the specific behaviors, and finally moved to pattern explanation [18,19]. This study was approved by the Institutional Review Board at Cincinnati Children's Hospital Medical Center. It was funded, in part, by the Laerdal Foundation.

Data Corpus

Our research team, composed of an interdisciplinary group with combined expertise in critical care (DL, KE, and MZ), resuscitation (DL and MZ), simulation (DL, KE, and JS), and medical education (DL, JS, and MZ), reviewed a series of VR simulation recordings captured during a validation study of a specific VR curriculum [20]. The simulation took place in a virtual environment that mirrored a typical high-acuity room, as might exist in an emergency department or intensive care unit. Participants in the simulation included pediatric critical care and emergency medicine attendings and fellows, and pediatric intensive care unit and emergency department nurses. Each physician-nurse dyad had no previous exposure to the scenario and completed it only once.

The virtual patient was located on a hospital bed within a room that included a vital sign monitor, an intravenous pole with fluids, an oxygen delivery device, a stethoscope, a thermometer, a syringe for drawing blood samples, and a large fluid-filled syringe for triggering administration of a fluid bolus. A representative first-person perspective of the VR environment can be found in Figure 1. The virtual patient was an 8-year-old male with pneumonia and sepsis. The rationale for this scenario was that pneumonia frequently progresses to sepsis, and key stakeholders regarded this as a realistic scenario for clinical assessment and management training. The clinical status of the virtual patient changed throughout the simulation, conveying alterations in mental status (ranging from conversant to altered), perfusion (mottled skin that progressed to poor perfusion and cyanosis), and respiratory status (superimposed retractions and tachypnea).

Figure 1. First-person perspective of the virtual reality (VR) experience.



The scenario was coded in Unity (Unity Technologies) and accessed via an Oculus Rift headset connected to a VR-capable laptop, which allowed 2 participants within the same simulation to move freely and interact with the scenario and each other, including communicating regarding shared findings in real time. The purpose of the simulation was to assess the participants' ability to recognize, describe, and begin to treat a patient with developing shock. A detailed description of this VR simulation was previously described by Zackoff et al [20]. Participants received a 5-minute orientation to navigation in the VR environment and the functionality of the virtual patient and equipment from a trained VR simulation facilitator. They then received the case description and were instructed to begin their assessment and management of the patient as they would in real life. The VR simulation facilitator was present in the VR environment for the orientation and then exited the VR but remained present in real life to address questions and provide feedback throughout the remainder of the simulation session.

First-person audiovisual data recorded from the perspective of both users (a nurse and a physician) were stored in a password-protected, encrypted server, annotated via Vimeo (Vimeo, Inc), and coded in Microsoft Excel. Multiple audiovisual feeds, capturing the perspective of both participants within the virtual environment, allowed for data triangulation [21].

Research Team

Considering reflexivity [22] and the desire for analytic triangulation [21], the research team was composed of a heterogeneous group of experts in resuscitation team leadership and medical education (DL), resuscitation bedside nursing (KE), and SBME (JS). DL is a practicing pediatric critical care physician as well as a simulation educator. KE is a simulation educator and former critical care nurse. JS is a simulation educator and a former pediatric emergency department nurse. The data analysis was also overseen by MZ, a pediatric critical care physician and education scientist with expertise in designing, implementing, and evaluating SBME using innovative modalities such as VR and AR. MZ served as a senior consultant, providing guidance and expertise at scheduled intervals and during critical decision points where consensus within the research team was required.

Focused Research Question

The research team reviewed the data, primed with the following generative question [23], "How does a virtual reality simulation influence participant attention and behavior, and how might these influences impact achievable educational goals?"

Using a focused ethnographic approach is particularly suited for foundational research in which little is known about how subjects interact with a new environment (or, more specifically, a technology that places subjects in a new environment). Rather than measure changes against a specific baseline or control group, a focused ethnographic approach provides foundational descriptions of behaviors [18]. Instead of benchmarking against manikin-based or live simulations, our goal was to capture and interpret emergent behaviors, areas of challenge, and decision-making processes that may be uniquely influenced by VR-based training.

Phased Qualitative Analysis of Participant Attention and Behavior

To investigate participant attention and behavior within a virtual environment, we conducted a 3-phase analysis of VR simulations [18,24]. We followed a focused ethnographic strategy [18,19,24] incorporating inductive coding techniques [25]. This approach was chosen to (1) provide in-depth, context-specific insights into participant behavior in VR and (2) allow unanticipated themes to emerge from the data rather than imposing pre-existing hypotheses.

Phase 1: Identification and Classification

The initial phase involved exposing the research team to the VR scenarios. All participants were oriented to the technology, simulation, and virtual environment. The team then watched an initial group of scenarios to further familiarize themselves with the simulation and VR environment. For each study video, each researcher took independent field notes [26]—timestamped observations documenting participant statements, team dynamics, and points of interest [18]. These notes were stored, compiled, and subsequently treated as data and shared during collective analysis sessions.

The units of analysis for these notes were discrete, timestamped episodes of participant behavior. Team members independently categorized these notes into two broad types: (1) verbal events: any distinct verbalization by a participant (eg, "I'm going to give a bolus," or "He looks cyanotic") and (2) Interaction events: any observed action in the VR environment (eg, checking capillary refill, reaching for oxygen supplies, and teleporting to a different location). This unit of analysis helped capture how participants directed attention, interpreted clinical cues, and engaged with the virtual environment.

During the collective analysis sessions, the research team met together and used triangulation techniques to reconcile discrepancies in individual coding and ultimately reach consensus during the development of a comprehensive codebook. The team compared individual code lists, resolved discrepancies through discussion, and compiled a unified preliminary codebook. During this stage, our theoretical framework served as a sensitizing concept, alerting us to phenomena related to the active construction of knowledge and influence of cognitive load (eg, repeated attempts to use virtual equipment, indicating extraneous load on top of the active construction of clinical assessment knowledge).

This codebook served as a standardized tool for subsequent research phases (Phase 2: Description and Analysis and Phase 3: Pattern Explanation). Following the consolidation of the codebook, the research team independently applied it to each VR simulation session video recording. After coding each simulation session, the team reconvened to discuss and refine the codebook further as needed. This iterative process continued until data saturation [27] was achieved.

Phase 2: Description and Analysis

After achieving data saturation, the research team transitioned to the second phase—description and analysis. A rigorous review of the developed codebook was conducted to identify recurring patterns in participant attention and behavior across the reviewed simulated scenarios, guided by the focused research question.

The recurring findings were subsequently organized into distinct clusters to inform themes, providing a comprehensive overview of participant engagement during the VR simulations. To ensure the validity and consistency of these themes, an independent researcher (MZ) again triangulated the data.

Phase 3: Pattern Explanation

The final phase used the established themes to facilitate a descriptive analysis of the potential ramifications of using VR as an educational tool. By exploring the patterns of attention and behavior exhibited during VR-based simulations, we identified unique advantages and limitations associated with this modality. We reflected on how these insights may aid

educators in determining which educational objectives are best aligned with VR training.

Ensuring Code Validity

We used multiple strategies to enhance the credibility and dependability of our findings. First, we used analyst triangulation [21,28] by involving researchers with diverse backgrounds in critical care, nursing, and simulation education. Each analyst independently reviewed and coded selected video segments, then convened to discuss and reconcile any discrepancies. This iterative group process helped refine the codebook, ensuring that multiple perspectives were integrated and minimizing the influence of individual bias. We also maintained reflexive memos [22] to document personal assumptions or theoretical leanings; these memos were revisited throughout data analysis to encourage ongoing reflection on how our backgrounds might shape interpretations.

In addition, we created an audit trail [29] that outlined key decisions made during coding and theme development. This trail included records of coding updates, consensus discussions, and rationale for code merging or splitting, providing a transparent account of how the analysis evolved. By combining triangulation, reflexivity, and rigorous documentation, we sought to strengthen the trustworthiness of our findings and facilitate replicability of the study's analytic approach.

Ethical Considerations

The primary study and this secondary analysis received approval from the Cincinnati Children's Hospital Institutional Review Board, which granted a waiver of documentation of informed consent in accordance with 45 CFR 46.116(d). This waiver was permissible because the research posed no more than minimal risk to participants, did not compromise their rights and welfare, could not be feasibly conducted without the waiver, and would provide subjects with additional relevant information after participation, when appropriate. Given its educational focus and the absence of risk to participants, the study met these criteria. Participation in the simulations was voluntary, with no compensation provided. All data on participant performance were stored securely on a password-protected server.

Results

Overview

The ethnographic analysis was conducted on the recordings of 15 interprofessional teams of 2 users (a nurse and a physician) who completed the VR simulation. The simulations were recorded from both the perspective of the nurse and the team lead. This analysis revealed 3 major themes and associated subthemes that offer insights into the interplay between simulation modality and participant attention and behavior (Table 1).

Table . Ethnographic analysis results.

Themes, subthemes, and their examples	Illustrative descriptors and quotes
Source of Truth	
Initial reliance on virtual cues	
Observing movements	• The nurse and team lead observe the virtual patient swatting at the team with his hands and struggling to breathe. The nurse states, "Pending a [blood gas to confirm], we are probably looking at assisting ventilation one way or another." [N11]
Mental status	• "Quick cardiopulmonary assessment–His mental status is down; his airway is open. I need to take a listen to him, but I can see him breathing, cap refill is about 3 - 4 seconds, he looks like crap… I am thinking shock, likely sepsis." [P8] • "Timmy? HELLOO?" Meanwhile, the team lead Interacts with the virtual patient's foot and then performs a sternal rub. [P6]
Linking physical examination to clinical progression	• After evaluating the patient, the team lead states, "Airway seems to be ok. He seems to be a little more alert." Asks the virtual patient, "How you doing buddy?" The team lead taps on that patient's chest, and while sounding stressed, says, "He isn't responding." As the virtual patient starts to respond, the team lead sounds calmer. [P13] • Team lead says the virtual patient has bounding "femoral" and "jugular" pulses as he palpates them sequentially. [P2]
Dynamic reliance based on response	
Increasing reliance on the cardiopulmonary monitor when the examination does not change as expected	• "[After receiving fluid, the] blood pressure is still a little low, so that may be the cause of his tachycardia, but he also has fever, so…(provider does not finish sentence)." [P5] • Midway through the simulation, the team lead notes that the patient still has poor perfusion and is still working hard to breathe. He recognizes that about half of the fluid bolus is in, yet the patient has not improved. He then turns to the monitor, remarking on the tachycardia and hypoxemia. He asks for vancomycin to be added while staring at the monitor. [P10]
Cognitive Focus	
Task burden and scenario complexity	
Synthesizing large amounts of information quickly	• Team lead listens to the lungs, looks at the monitor, and states, "He looks kind of like ass... why is his face purple?" [P3] • After giving a fluid bolus, the team lead reports, "His heart rate is trending down nicely." [P5] • The nurse responds, "His perfusion in his hands is 3 - 4 seconds." [N5] • The team lead leans over the virtual patient and watches the cap refill animation.
Transitioning from interpretation to action	• "He looks sick, he is a little cyanotic...can you put oxygen on him?" [P13] • The team lead requests, "septic workup stuff." Quickly asks for several interventions, and then asks the nurse for suggestions. [P8]
Cognitive demands as an unexpected consequence of simulation modality	
Completing examination in VR[a]	• The team lead reaches for the stethoscope on the table so she can listen to the lungs. She struggles initially to grab it, but quickly figures it out. She then moves the stethoscope to the virtual patient but struggles to use it in the VR environment. Meanwhile, communication between the nurse and team lead stops while the team lead is task-burdened with learning how to perform an assessment of the lungs in the VR environment. [P4]

Themes, subthemes, and their examples	Illustrative descriptors and quotes
Completing interventions in VR	• The team lead asks the nurse to give the patient a fluid bolus. The nurse moves to follow the order and realizes that she does not know how to perform the action in the VR environment. The facilitator then proceeds to walk her through the process. [N10]
Fidelity Breakers	
Navigational challenges	
Unintended teleportation	• The team lead puts the headset on but finds himself in the ceiling of the virtual environment. He attempts to correct but ends up in the floor, spending several moments trying to get in the correct position. [P13]
Trouble walking	• The team lead accidentally teleports to a different location in the room, but is able to quickly get back to her original location without assistance. [N9]
Interaction challenges	
Applying interventions to patient	• The team has accidentally taken off the mask while attempting a sternal rub. The entire team is discussing how to get the mask back on and struggle to get it appropriately positioned. The facilitator interrupts, acknowledges the attempt to place the mask, and instructs the team to move on as if the mask has been applied. [P11]
Triggering animations	• The team is struggling to get the capillary refill animation to trigger. At first the nurse attempts and is unable, then the team lead then joins. Together, they spend 140 seconds trying to trigger the capillary refill animation in several locations—the foot, the hand, and the sternum. Meanwhile the virtual patient has globally poor perfusion, low blood pressure, and tachycardia, which is not addressed because the team is stuck on trying to trigger the animation. [P8, N8]

[a]VR: virtual reality.

Theme 1, Source of Truth, outlines the hierarchy of informational sources for participants analyzing clinical scenarios, including subthemes related to initial reliance on virtual cues and dynamic reliance based on responses. Theme 2, Cognitive Focus, describes the interplay between thinking, communicating, and doing, with subthemes of dynamic task burden based upon scenario complexity and the cognitive demands of the virtual environment. Theme 3, Fidelity Breakers, explores moments where the boundary between real life and simulation becomes most pronounced, leading participants to deviate from their typical behavior in real-life encounters. Subthemes within this theme relate to navigational and interactional challenges.

Theme 1: Source of Truth

Participants face novel challenges in a VR medical simulation that they may not have encountered in previous traditional simulation experiences. They are immersed in a completely virtual environment, with real-world objects replaced with digital replicas. This immersion requires participants to reorient themselves to the virtual room and their ability to move and interact within it. We gained insights into how participants navigate this new environment and identify reliable sources of information to guide their actions within the simulated medical case.

Subtheme 1.1: Initial Reliance on Virtual Cues

The VR environment offers a visually realistic virtual patient within a replicated clinical environment. This immersive environment replaces the real world with a digital replica, requiring participants to acclimate briefly before focusing on dynamic elements within the room. As participants acclimated to the virtual environment, their focus aligned to three major sources of information: (1) the virtual manikin, (2) the cardiorespiratory monitor, and (3) other participants in the same simulation.

After orientation to the VR environment, participants consistently prioritized soliciting data from the virtual patient. This early prioritization was evident in their frequent comments on the patient's movements, mental status, and perfusion. The recorded first-person perspective showed that participants spent the majority of their time looking at the patient.

Subtheme 1.2: Dynamic Reliance Based on Response

Though initially focused on the virtual patient, participants did use the cardiorespiratory monitor as a secondary source of information. They checked the monitor periodically for brief durations; however, this behavior changed when the patient's response to interventions deviated from expectations. For example, if the virtual patient's perfusion did not improve after administration of a fluid bolus, participants shifted to a heavier reliance on the monitor readings, potentially discounting some

qualitative physical examination cues in favor of objective data. In such cases, participants became more reliant on the monitor readings, devoting greater attention for extended periods. This shift suggested a transition from using the monitor for confirmation to using it as a primary source of information guiding decision-making.

When both the patient's presentation and the monitor readings deviated from expected responses to interventions (eg, no immediate improvement in perfusion or heart rate after an intravenous fluid bolus), participants would collaborate to review the case. When working together, participants made decisions based on a combination of group consensus, individual opinions, and real-life clinical algorithms applicable to the context of the simulation (ie, treatment of pneumonia or sepsis). Nonverbal communication was limited due to simulation constraints. Only the participants' virtual hands and glasses (corresponding to their headset orientation) were visible in the simulation, and thus participants could not assess eye contact or body language as markers of consensus or concern, which is a major deviation from real-life interactions.

If no alternative explanation existed for the perceived incongruity, the participant might seek clarification from the facilitator as the final source of objective data, ranging from asking for clarification on the VR simulation's functionality to attempting to draw the facilitator into the clinical scenario as an additional participant.

Theme 2: Cognitive Focus

Participant attention is finite, and the dynamic interplay between the VR environment and participants' cognitive load unfolded during the simulated scenario. By introducing an entirely virtual world and placing participants in it, VR introduces unique features that may impact the cognitive load and task burden placed upon the participant [30,31]. We observed how the virtual environment affected participants' ability to process information, make decisions, and execute actions within the simulation. These effects occurred both as an intentional part of the design and as unintended consequences of the VR simulation's technical limitations.

Subtheme 2.1: Task Burden and Scenario Complexity Dynamically Influenced Cognitive Demands

In the initial phase of the simulation, participants faced a high cognitive load as they synthesized the clinical situation while simultaneously familiarizing themselves with the complexities of the VR environment. Participants needed to integrate information from multiple sources: the virtual patient's visual presentation, physiological data displayed on the monitor, and communication with other team members. As noted in Theme 1, participants initially focused heavily on the virtual patient and the data it provided to anchor their navigation of the clinical scenario.

As the scenario progressed, a critical shift in cognitive demands occurred. The focus transitioned from understanding the patient's condition to taking decisive actions. Participants needed to execute physical assessments, such as auscultation and capillary refill checks, alongside tasks that included applying supplemental oxygen and initiating intravenous fluid resuscitation. This transition, in many cases, aligns with the natural progression of patient care, moving from initial assessment to formulating and implementing a treatment plan.

Subtheme 2.2: Cognitive Demands as an Unexpected Consequence of the Virtual Environment

Cognitive demands as an unintended consequence of the virtual environment (ie, additional demands due to unique VR functionality that differs from real life) impacted participant experience at times. When virtual tasks differed substantially from their real-world counterparts in terms of complexity or intuitiveness, participants appeared to experience increased task burden. We inferred this increased task burden primarily through observed behaviors, such as repeated attempts to perform a task, expressions of frustration, or prolonged fixation on a single action. Certain actions, such as administering a fluid bolus, were intentionally simplified for VR programming feasibility, as it was seen as outside the stated objectives of the training exercise. However, this simplification created a discrepancy from real-world procedures. Experienced nurses, accustomed to a specific bolus administration technique, appeared frustrated by the VR version. This frustration led to disproportionate task fixation on what would otherwise have been a rudimentary task for the team, resulting in excessive focus on technical skills that may have impeded uptake of the simulation's core learning objectives.

Theme 3: Fidelity Breakers

Factors intrinsic to the VR environment appeared to challenge participant immersion and disrupt their sense of "being there." These disruptions, termed Fidelity Breakers, appeared to break participants' engagement with the simulation and shift their focus from clinical reasoning to technical troubleshooting. The identified Fidelity Breakers fell into the 2 main categories of navigational challenges and interactional challenges.

Subtheme 3.1: Navigational Challenges and Learner Adaptations—Limitations in Movement Within the Virtual Space

Participants began each session positioned at the foot of the virtual patient's bed. Given their starting positioning relative to the patient and the equipment, participants universally tried to move around the room to complete a physical examination or use equipment. However, limitations in the VR system at times disrupted these efforts. Several instances of unintended teleportation were observed: rather than smoothly walking to their desired location via the hand-held controller, participants would inadvertently press a button that transported them to an unintended spot—sometimes on the opposite side of the bed or even partially on top of the patient (refer to Table 1, Fidelity Breakers). Such abrupt relocations caused noticeable confusion and delays while participants struggled to reposition themselves.

As a workaround, several participants elected to lean or reach instead of fully reteleporting to examine the patient more closely. Although this approach facilitated some aspect of examination, it also compromised realism by limiting participants' freedom to perform typical bedside maneuvers—such as placing a stethoscope on the chest or palpating a pulse—which would otherwise be straightforward

during an actual patient encounter. Conversely, some participants became completely fixated on these limitations and spent an outsized amount of time just trying to move around the room, rather than participating in the simulation.

Subtheme 3.2: Interaction Challenges—Difficulties Interacting With the Virtual Patient or Equipment

Interactions with the virtual patient and equipment at times appeared to challenge the authenticity of the simulation. Some key examination maneuvers were difficult for participants to perform in VR. For instance, some participants experienced collision detection problems when interacting with the patient. An example includes attempts to apply a supplemental oxygen mask to the virtual patient, which would sometimes accidentally pass through the face (clipping) or fail to trigger the animation altogether. This deviation from real-world actions hindered the intervention and assessment process. Reliance on VR controllers for interactions, as opposed to one's hands, additionally presented participants with a learning curve that occasionally led to repeated difficulties performing core physical examination skills. A prime example was the challenge of triggering the capillary refill animation, a vital step in evaluating the patient's hemodynamic status. In some cases, participants became fixated on manipulating the controllers to achieve this specific action, neglecting other aspects of the simulation for several minutes.

Discussion

Principal Findings

This study identified 3 major themes regarding the impact of a headset-based VR simulation on participant attention and behavior. VR was found to influence participants to rely primarily on the virtual patient for clinical data and decision-making; however, the participants' reliance on "Theme 1: Source of Truth" shifted to the cardiorespiratory monitor when the expected patient response to interventions did not occur. In addition, learner behaviors in the VR environment closely reflected the natural progression of patient care, moving from initial assessment to formulating and implementing a treatment plan, allowing closer approximation of the "Theme 2: Cognitive Focus" experienced during clinical care. Finally, participants faced navigational and procedural challenges that were "Theme 3: Fidelity Breakers," skewing participant attention to the simulation's technical details rather than the intended learning objectives.

Participant Reliance on Patient-Centered Data

Practicing medicine in a completely virtual environment is an unfamiliar experience for most medical staff and can be disorienting. In this constructed environment, participants primarily relied on the patient examination as their source of objective data, with focus shifting to the cardiorespiratory monitor only when the virtual patient did not respond as expected to interventions. This grounding of the patient as the center of the simulation, as opposed to a primary reliance on ancillary data, represents a key potential benefit of VR training—reinforcing the patient examination as central in the practice of medicine. Although these observations suggest that VR might help draw learners' attention to direct patient cues, this conclusion is tentative without a direct comparison to other simulation modalities. While VR could reinforce the patient examination as central in the practice of medicine, further research should verify whether this pattern is unique to VR or equally present in other simulation approaches.

Limitations of the VR Environment for Procedural Training

Though participants focused on the patient examination, this examination was at times limited. Due to the unique learning curves and inherent technological limitations for movement within the environment and physical manipulation of the patient and equipment, participants often relied primarily on audiovisual cues. In addition, technological glitches causing accidental teleporting and clipping resulted in participants electing to limit their movement around the room, favoring leaning and reaching over moving around the environment, as highlighted through the examples listed in Table 1, Fidelity Breakers. These challenges and resultant learner adaptations suggest that VR might have limitations for learning objectives focused on examination maneuvers or procedures requiring "hands-on" skill development. Until there is no need for technical workarounds or deviations from reality to perform skills in VR, as observed in this study, experience with these tasks in VR may not support skill acquisition that can be directly applied to real-life clinical care. In addition, the cognitive load added by navigating these workarounds may distract from or prevent the acquisition of other learning objectives. Alternatively, these skills may be better learned, practiced, or assessed with alternative modes of simulation, such as computerized manikin-based or AR-enhanced simulation, and thus specific studies examining specific skill acquisition in different modes of simulation are needed.

Task Fidelity and Cognitive Load Considerations

A virtual patient has the advantage of being able to portray a broad and realistic range of physical findings, as well as dynamic responses to therapy. However, inherent to VR is the requirement of creating digital content for the participant's entire field of view and expected actions—extending to the environment, equipment, and the interactions with said equipment. Though education and design teams can meticulously craft a virtual environment, any deviations from the real-world environment risk distracting the participant from the primary objectives of the training exercise. This dissonance extends from observable deviations from the real world to divergences in how to perform tasks. Although this VR simulation was designed to focus on recognizing and treating shock, participants were at times distracted by activities outside of that direct focus, such as needing to relearn how to administer a fluid bolus in VR. This intervention was intentionally simplified to reduce the technical learning curve of the simulation, but this simplification had the unintended effect of distracting the learners from the primary focus of the simulation. Future work might explore balancing necessary simplifications with realistic task fidelity to minimize extraneous cognitive load in VR.

Interface-Specific Training Requirements

VR-based simulation demands additional orientation and training that differ from what is typically provided in manikin-based scenarios. Because certain actions—such as administering a fluid bolus in this scenario (refer to Table 1, N10)—are not replicated 1:1 from real-life procedures, participants must allocate some cognitive load to learning the VR mechanics. This interface-based burden can divert attention from the core clinical objectives, underscoring the importance of structured, VR-specific training sessions before learners engage in complex clinical tasks, and considering these extraneous loads when targeting specific learning objectives.

Communication and Team Dynamics

Decision-making when treating critical illnesses must be quick to be effective. In VR simulations, team leads made decisions based on a combination of group consensus, individual opinions, and real-life clinical algorithms. Group discussions were frequent, and team communication dynamics appeared to significantly impact decision-making. The VR scenarios included only two participants (plus a facilitator), making decision-making conversational. However, nonverbal communication was limited for this VR simulation, as only participants' virtual hands and glasses were visualized, which limited assessment of body language and eliminated the ability for eye contact. For this specific VR simulation, the modality encouraged verbal communication around shared observations and findings but lacked feasible options for nonverbal communication. While this reliance on explicit verbal communication may facilitate clarity in certain respects, we cannot determine whether it is more frequent or effective than in traditional simulations without further comparative research. Finally, despite the presence of ongoing communication, individual parallel assessments continued, suggesting that participants processed information independently. A more complex environment or scenario, with potentially a larger team of clinicians and staff, might result in different communication dynamics than what we observed. Assessing the attention and behavior of larger teams in more complex clinical contexts represents key future work for our team.

Limitations

While our study sheds light on important aspects of VR training for pediatric resuscitation, several limitations warrant consideration. First, the data originated from a single institution over a limited timeframe, encompassing only 15 simulations. While this sample size is relatively small, the participants represent a significant portion of the pediatric resuscitation team at a major academic medical center, potentially fostering generalizability to similar institutions. In addition, data saturation was achieved after analyzing 15 scenarios, indicating that further analyses would not likely yield novel findings. Second, the focused ethnography approach involves inductive analysis, meaning the researchers' experience and expertise are inherently woven into the methodology. This strengthens the analysis by adding depth and richness to the conclusions drawn. Notably, the research team was purposefully assembled to leverage their expertise in simulation and resuscitation, allowing them to inject diverse perspectives into data analysis and enrich the interpretation of findings.

However, this methodological approach lacks a direct comparator or control group. Because our objective was to describe behaviors in a previously unexamined context, fully immersive VR, we did not measure change from a standard baseline (eg, manikin-based or live simulations). This limitation was partially mitigated by providing rich, context-specific observations from multiple angles, coupled with cross-disciplinary analyst triangulation. Future studies may build on these findings by including mixed method comparisons across different simulation platforms to quantify how VR might augment or hinder specific learning objectives. In addition, there is currently no literature that describes the focus of interdisciplinary teams during real-life clinical care—a key future direction for our team to enhance our ability to align participant behaviors during simulation with those expected during real life.

Finally, this study used a single VR scenario, and thus the conclusions drawn must be considered in relation to the specific technical design and functionality of that scenario. However, while there are a range of VR systems and approaches at various stages of development and availability, the foundational principles of navigating and interacting in a virtual world are likely similar across platforms. As VR technology advances, limitations in movement within the virtual space may be resolved. For example, wireless head-mounted displays allow for easier movement in the VR space by untethering users from stationary computers. Yet even at its best, VR can only approximate—rather than fully replicate—the fidelity of genuine hands-on, kinesthetic practice. Further research can help clarify how these core elements influence participants' clinical reasoning, teamwork, and skill acquisition, enabling educators to better tailor VR-based simulations to specific learning objectives.

Conclusions

The conclusions drawn from this study provide a comprehensive understanding of how a VR-based simulation may influence participant attention and behavior, revealing both strengths and limitations of this training modality. Our findings indicate that VR simulations create a unique environment where participants predominantly rely on virtual patient assessments and audiovisual cues. This setup prioritizes patient-centered data over ancillary information, fostering focused clinical decision-making. However, the constraints of VR, such as limited physical interaction and navigational challenges, can impact the authenticity of the training experience, suggesting that VR may be better suited for clinical assessment or cognitive objectives over physical skill–based objectives.

As the landscape of SBME evolves, it is crucial to recognize the potential and limitations of VR simulations compared to other modalities. The ability to create a fully immersive, yet controlled, environment offers significant training advantages for objectives that focus on rapid decision-making and patient assessment. Nevertheless, the technological constraints that limit physical interactions highlight the persistent need for complementary simulation modalities, such as AR and

computerized manikin-based simulations, which may better support hands-on skill development. Next steps include the delineation of how the growing armamentarium of simulation modalities and tools can be optimally leveraged to best train a broad range of clinical skills.

Acknowledgments

We would like to acknowledge the Laerdal Foundation for its funding contribution to this research. All the authors are responsible for this manuscript and have participated in the conceptualization, drafting, or revision of the manuscript and have seen and approved the final version submitted. This study was funded, in part, by project support from the Laerdal Foundation. Matthew Zackoff received previous project support through the Place Outcomes Research Award Program at Cincinnati Children's Hospital Medical Center to create the video dataset analyzed in this study.

Data Availability

The datasets generated and analyzed during the study are not publicly available due to the video consent, which stipulates that the collected videos cannot be used outside the current research study, and because individual participant identifiers are included in the field notes. The datasets are available from the corresponding author on reasonable request.

Authors' Contributions

DL contributed to the conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, supervision, visualization, and writing of the original draft. JS contributed to data curation, formal analysis, investigation, and writing—review and editing. KE contributed to data curation, formal analysis, investigation, and writing—review and editing. MZ contributed to conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, supervision, validation, visualization, and writing—review and editing.

Conflicts of Interest

None declared.

References

1. Grenvik A, Schaefer J. From Resusci-Anne to Sim-Man: the evolution of simulators in medicine. Crit Care Med 2004 Feb;32(2):S56-S57. [doi: 10.1097/00003246-200402001-00010] [Medline: 15043230]
2. Perkins GD, Boyle W, Bridgestock H, et al. Quality of CPR during advanced resuscitation training. Resuscitation 2008 Apr;77(1):69-74. [doi: 10.1016/j.resuscitation.2007.10.012] [Medline: 18083288]
3. Emani SS, Allan CK, Forster T, et al. Simulation training improves team dynamics and performance in a low-resource cardiac intensive care unit. Ann Pediatr Cardiol 2018;11(2):130-136. [doi: 10.4103/apc.APC_117_17] [Medline: 29922009]
4. Sanchez LD, DelaPena J, Kelly SP, Ban K, Pini R, Perna AM. Procedure lab used to improve confidence in the performance of rarely performed procedures. Eur J Emerg Med 2006 Feb;13(1):29-31. [doi: 10.1097/00063110-200602000-00007]
5. Loeb D, Shoemaker J, Parsons A, Schumacher D, Zackoff M. How augmenting reality changes the reality of simulation: ethnographic analysis. JMIR Med Educ 2023 Jun 30;9(1):e45538. [doi: 10.2196/45538] [Medline: 37389920]
6. Jensen L, Konradsen F. A review of the use of virtual reality head-mounted displays in education and training. Educ Inf Technol 2018 Jul;23(4):1515-1529 [FREE Full text] [doi: 10.1007/s10639-017-9676-0]
7. Plotzky C, Lindwedel U, Sorber M, et al. Virtual reality simulations in nurse education: a systematic mapping review. Nurse Educ Today 2021 Jun;101:104868. [doi: 10.1016/j.nedt.2021.104868] [Medline: 33798987]
8. de Cote E, Flores Herrera AM, Giorgi E, Cattaneo T. Augmented reality (AR) and virtual reality (VR) as tools to empower vulnerable communities: opportunities and challenges for designers. In: Giorgi E, Cattaneo T, Flores Herrera AM, Aceves Tarango V, editors. Des Vulnerable Communities Cham: Springer International Publishing; 2022:307-321. [doi: 10.1007/978-3-030-96866-3_16]
9. Ruthenbeck GS, Reynolds KJ. Virtual reality for medical training: the state-of-the-art. Journal of Simulation 2015 Feb;9(1):16-26. [doi: 10.1057/jos.2014.14]
10. Tursø-Finnich T, Jensen RO, Jensen LX, Konge L, Thinggaard E. Virtual reality head-mounted displays in medical education: a systematic review. Sim Healthcare 2023;18(1):42-50. [doi: 10.1097/SIH.0000000000000636]
11. Moro C, Birt J, Stromberga Z, et al. Virtual and augmented reality enhancements to medical and science student physiology and anatomy test performance: a systematic review and meta‐analysis. Anatomical Sciences Ed 2021 May;14(3):368-376 [FREE Full text] [doi: 10.1002/ase.2049]
12. Zackoff MW, Cruse B, Sahay RD, et al. Development and implementation of augmented reality enhanced high-fidelity simulation for recognition of patient decompensation. Simul Healthc 2021 Jun 1;16(3):221-230. [doi: 10.1097/SIH.0000000000000486] [Medline: 32910102]

13. Zhan T, Yin K, Xiong J, He Z, Wu ST. Augmented reality and virtual reality displays: perspectives and challenges. iScience 2020 Aug 21;23(8):101397. [doi: 10.1016/j.isci.2020.101397] [Medline: 32759057]
14. Dewan M, Vidrine R, Zackoff M, et al. Design, implementation, and validation of a pediatric ICU sepsis prediction tool as clinical decision support. Appl Clin Inform 2020 Mar;11(2):218-225. [doi: 10.1055/s-0040-1705107] [Medline: 32215893]
15. Kouijzer M, Kip H, Bouman YHA, Kelders SM. Implementation of virtual reality in healthcare: a scoping review on the implementation process of virtual reality in various healthcare settings. Implement Sci Commun 2023 Jun 16;4(1):67. [doi: 10.1186/s43058-023-00442-2] [Medline: 37328858]
16. Sweller J. Cognitive Load Theory: Springer; 2011. [doi: 10.1007/978-1-4419-8126-4]
17. Cobern WW. Constructivism. J Educ Psychol Consult 1993 Mar;4(1):105-112. [doi: 10.1207/s1532768xjepc0401_8]
18. Andreassen P, Christensen MK, Møller JE. Focused ethnography as an approach in medical education research. Med Educ 2020 Apr;54(4):296-302. [doi: 10.1111/medu.14045] [Medline: 31850537]
19. Higginbottom GM, Boadu N, Pillay J. Guidance on performing focused ethnographies with an emphasis on healthcare research. Qual Rep 2013;18(17):1-16. [doi: 10.46743/2160-3715/2013.1550]
20. Zackoff MW, Cruse B, Sahay RD, et al. Multiuser immersive virtual reality simulation for interprofessional sepsis recognition and management. J Hosp Med 2024 Mar;19(3):185-192. [doi: 10.1002/jhm.13274] [Medline: 38238875]
21. Flick U, von Kardorff E, Steinke I, editors. A Companion to Qualitative Research: SAGE Publications; 2004. URL: https://hu.kln.ac.lk/units/rc/media/attachments/2021/09/17/a_companion_to_qualitative_research.pdf [accessed 2025-10-22]
22. Dodgson JE. Reflexivity in qualitative research. J Hum Lact 2019 May;35(2):220-222. [doi: 10.1177/0890334419830990] [Medline: 30849272]
23. Agee J. Developing qualitative research questions: a reflective process. Int J Qual Stud Educ 2009 Jul;22(4):431-447. [doi: 10.1080/09518390902736512]
24. Knoblauch H, Schnettler B. Videography: analysing video data as a 'focused' ethnographic and hermeneutical exercise. Qual Res 2012 Jun;12(3):334-356. [doi: 10.1177/1468794111436147]
25. Thomas DR. A general inductive approach for qualitative data analysis. Am J Eval 2003;27(2). [doi: 10.1177/1098214005283748]
26. Phillippi J, Lauderdale J. A guide to field notes for qualitative research: context and conversation. Qual Health Res 2018 Feb;28(3):381-388. [doi: 10.1177/1049732317697102] [Medline: 29298584]
27. Fusch P, Ness L. Are we there yet? Data saturation in qualitative research. TQR 2015;9(20). [doi: 10.46743/2160-3715/2015.2281]
28. Carter N, Bryant-Lukosius D, DiCenso A, Blythe J, Neville AJ. The use of triangulation in qualitative research. Oncol Nurs Forum 2014 Sep;41(5):545-547. [doi: 10.1188/14.ONF.545-547] [Medline: 25158659]
29. Cutcliffe JR, McKenna HP. Expert qualitative researchers and the use of audit trails. J Adv Nurs 2004 Jan;45(2):126-133. [doi: 10.1046/j.1365-2648.2003.02874.x] [Medline: 14705996]
30. Han J, Zheng Q, Ding Y, Faculty of Education, Southwest University, Chongqing, China. Lost in virtual reality? Cognitive load in high immersive VR environments. JAIT 2021;12(4) [FREE Full text] [doi: 10.12720/jait.12.4.302-310]
31. Albus P, Vogt A, Seufert T. Signaling in virtual reality influences learning outcome and cognitive load. Comput Educ 2021 Jun;166:104154. [doi: 10.1016/j.compedu.2021.104154]

Abbreviations

AR: augmented reality
SBME: simulation-based medical education
VR: virtual reality

Edited by B Lesselroth; submitted 28.08.24; peer-reviewed by C Merritt, MD Lotbiniere-Bassett, O Meruvia-Pastor; revised version received 15.04.25; accepted 22.09.25; published 27.10.25.

Please cite as:
Loeb D, Shoemaker J, Ely K, Zackoff M
Deconstructing Participant Behaviors in Virtual Reality Simulation: Ethnographic Analysis
JMIR Med Educ 2025;11:e65886
URL: https://mededu.jmir.org/2025/1/e65886
doi:10.2196/65886

A Virtual Simulator to Improve Weight-Related Communication Skills for Health Care Professionals: Mixed Methods Pre-Post Pilot Feasibility Study

Fiona Quigley[1], PhD; Leona Ryan[2], PhD; Raymond Bond[3], PhD; Toni McAloon[4], PhD; Huiru Zheng[3], PhD; Anne Moorhead[1], PhD

[1]Institute of Nursing and Health Research, School of Communication and Media, Ulster University, 2-24 York Street, Belfast, United Kingdom

[2]School of Psychology, Ollscoil na Gaillimhe – University of Galway, Galway, Ireland

[3]School of Computing, Ulster University, Belfast, United Kingdom

[4]Institute of Nursing and Health Research, Ulster University, Belfast, United Kingdom

Corresponding Author:
Fiona Quigley, PhD
Institute of Nursing and Health Research, School of Communication and Media, Ulster University, 2-24 York Street, Belfast, United Kingdom

Abstract

Background: Discussing weight remains a sensitive and often avoided topic in health care, despite rising prevalence of obesity and calls for earlier, more compassionate interventions. Many health care professionals report inadequate training and low confidence to discuss weight, while patients often describe feeling stigmatized or dismissed. Digital simulation offers a promising route to build communication skills through supporting repeatable and reflective practice in a safe space. VITAL-COMS (Virtual Training and Assessment for Communication Skills) is a novel simulation tool designed to support health care professionals in navigating weight-related conversations with greater understanding and skill.

Objective: This study aimed to assess the potential of VITAL-COMS as a digital simulation training tool to improve weight-related communication skills among health care professionals.

Methods: A mixed-method feasibility study was conducted online via Zoom (Zoom Video Communications) between January to July 2021, with UK-based nurses, doctors, and dietitians. The intervention comprised educational videos and 2 simulated patient scenarios with real-time verbal interaction. Pre- and posttraining self-assessments of communication skills and conversation length were collected. Participants also completed a feasibility questionnaire. Descriptive statistics were used to analyze the feasibility questionnaire, and open-ended feedback was analyzed using content analysis. Paired-samples *t* tests were used to assess changes in communication skills and conversation length before and post training.

Results: In total, 31 participants completed the study. There was a statistically significant improvement in self-assessed communication skills following training (mean difference=3.9; 95% CI, 2.54 - 5.26; t_{30}=−5.76, *P*=.001, Cohen *d*=1.03). Mean conversation length increased significantly in both scenarios: in the female patient scenario, from 3.73 (SD 1.36) to 6.08 (SD 2.26) minutes, with a mean difference of 2.35 minutes (95% CI, 1.71 - 2.99; t_{30}=7.49, *P*=.001, Cohen *d*=1.34); and in the male scenario, from 3.61 (SD 1.12) to 5.65 (SD 1.76) minutes, a mean difference of 2.03 minutes (95% CI, 1.51 - 2.55; t_{30}=8.03, *P*=.001, Cohen *d*=1.44). Participants rated the simulation positively, with 97% (95% CI 90% - 100%) supporting wider use in health care and 84% (95% CI 71% - 97%) reporting emotional engagement. Content analysis of feedback generated two themes: (1) adapting to this form of learning and (2) recognizing the potential of simulation to support reflective, skills-based training. A minority, 13% (95% CI 1% - 25%) expressed a preference for alternative learning methods.

Conclusions: VITAL-COMS was feasible to implement and acceptable to a diverse group of health care professionals. Participants demonstrated significant improvements in self-assessed communication skills and patient-scenario engagement. The simulation was perceived as realistic, emotionally engaging, and well-suited for training in sensitive conversations. These findings support further development and integration of VITAL-COMS into health education programs. Next steps include the translation of the insights identified in this study to inform a tool supported by generative artificial intelligence.

(JMIR Med Educ 2025;11:e65949) doi:10.2196/65949

KEYWORDS

overweight; weight stigma; weight-related communication; virtual simulation; eLearning; medical education; healthcare communication; doctors; nurses; dietitians; simulator; professionals; healthcare; obesity; HCP; 3D; virtual; VITAL-COMS; tool; nutrition; mixed-method; digital training

Introduction

Obesity is a leading contributor to chronic disease and reduced quality of life [1,2]; yet, weight remains a sensitive and often avoided topic in clinical settings [3,4]. Despite national and global policy goals and clinical guidelines calling for earlier, more compassionate intervention [5-7], many health care professionals (HCPs) report inadequate training, lack of confidence, and fear of damaging the patient relationship when discussing weight [8-11]. Patients, in turn, often describe feeling stigmatized, misunderstood, or dismissed when seeking support for weight-related concerns [12-15] Addressing this training gap requires more than knowledge transfer—it requires practical, reflective skills development in how to talk about weight [16-18].

Digital or virtual simulation offers a flexible and scalable route to delivering communication skills training [19-21]. Simulation-based training enables learners to rehearse conversations in a safe environment and build confidence through repetition, feedback, and debrief [22]. When designed with behavioral science and communication theory, virtual simulations can foster the deeper understanding needed to engage patients with empathy and nuance [23-25]. For weight-related communication, where misalignment between HCP intention and patient perception is common [26-28], virtual simulation tools offer a safe space to surface and reflect on these tensions.

VITAL-COMS (Virtual Training and Assessment for Communication Skills) was developed to address the misalignments between HCPs and patients when discussing weight. The tool presents HCPs with simulated patient scenarios to practice sensitive, weight-related conversations, guided by principles from the COM-B (communication, opportunity, and motivation) model to support behavior change [23], Hargie's skilled interpersonal communication model [24], and the theory of deliberate practice [25]. The full development of VITAL-COMS is described elsewhere [26-32]. This study focuses on the preliminary findings of the feasibility study.

VITAL-COMS uses the Wizard of Oz (WoZ) method to control the simulated patient's responses. In this design technique, a human operator (FQ) manages the system's output to mimic intelligent responses that do not yet exist technologically [33,34]. WoZ allows designers to prototype complex interactions and gather authentic user feedback without the cost or limitations of fully developed artificial intelligence (AI) [35]. WoZ allows design ideas to be tested when the technology and AI do not yet exist or if it would take too much time or expense to explore a new idea. The virtual patient character in the simulator is operated by the primary author (FQ) using a WoZ controller designed for the simulation. In VITAL-COMS, this means tailoring verbal feedback from the virtual patient in real time to match the learner's communication, offering a high-fidelity, emotionally resonant training experience. More detail on the WoZ controller and how VITAL-COMS works can be found in Multimedia Appendices 1 and 2.

WoZ methods have been successfully used in health care education to simulate patient interactions, particularly where natural language, emotion, and timing are crucial [36]. This approach offers a flexible and iterative way to explore what kinds of digital communication training are most effective, while also revealing learners' needs and reactions [37]. As HCPs continue to navigate the challenges of supporting patients living with obesity, tools like VITAL-COMS provide a timely and evidence-based approach to addressing the training gaps.

Aim

The aim of this pilot feasibility study was to assess the potential of the VITAL-COMS digital simulation training tool to improve weight-related communication skills among HCPs.

Research Questions

This feasibility study addresses the following research questions Textbox 1:

Textbox 1. Primary and secondary questions.

Primary research question

- Does the VITAL-COMS (virtual training and assessment for communication skills) virtual simulator improve weight-related health communication skills?

Secondary research questions

- What is the perceived fidelity of the Wizard of Oz–based simulation from the participants' perspective?
- What is the perspective on VITAL-COMS leading to improvement in health care practice?

Methods

Design

A pre-post, mixed methods feasibility study was conducted to evaluate the viability of a virtual simulator designed to enhance weight-related health communication skills (VITAL-COMS). The study is reported following the CONSORT (Consolidated Standards of Reporting Trials) statement: extension to pilot and feasibility trials [38].

Technical Design

The VITAL-COMS virtual simulator was developed using the Unity 3D game engine (version 2019.3.5f). Development incorporated a suite of Unity's C# libraries to construct the user interface and implement speech simulation for the virtual patients. 3D character models were generated using Adobe Mixamo, and the Unity add-on components SALSA Lip-Sync, EmoteR, and Eyes were integrated to facilitate realistic speech articulation and emotional expression in the virtual patients. The tool comprised a pretraining stage of watching short educational videos on obesity science, weight-related communication, and weight stigma, then using the simulation tool (Multimedia Appendices 1 and 2).

Participants and Recruitment

This study used purposive and snowball sampling strategies to obtain a sample of 31 participants. This sample size aligns with recommendations for feasibility trials, which suggest total sample sizes ranging from N=24 to 50 [39]. The overall sample comprised both UK-based undergraduate and postgraduate allied health professionals and qualified doctors, nurses, and dietitians. Undergraduate and postgraduate students (n=2) were recruited via an email distributed by course directors at Ulster University. Targeted courses included undergraduate Nursing, Physiotherapy, Dietetics, and Occupational Therapy, as well as the MSc in Advanced Practice Nursing. Eligibility for participation was determined by enrollment in courses with curricula specifically addressing lifestyle factors related to the prevention, management, and recovery of ill health.

Qualified doctors, nurses, and dietitians (n=29) were recruited using purposive and snowball sampling. Eligibility for professional participants was contingent upon current practice within United Kingdom primary or secondary care settings and documented clinical experience in weight management counseling. Recruitment involved a social media advertisement posted on Ulster University's Twitter (currently known as X) and Facebook (Meta) accounts. The advertisement was subsequently disseminated through snowball sampling by tagging UK-based health care organizations and obesity charities. All participants provided informed consent before participation.

Procedure and Materials

Quantitative Data Collection

Upon obtaining informed consent, a participation date was arranged. Participants were subsequently provided with a secure, password-protected Zoom (Zoom Video Communications) meeting link via email before the scheduled session. Participants were invited to complete a brief demographic questionnaire, "About You." Participants also completed a self-reported knowledge assessment questionnaire on weight-related communication skills. The questionnaire was developed in relation to the existing literature and piloted with an advisory group (Multimedia Appendix 3).

The training intervention was delivered in 2 sequential parts, with participants offered the option of a brief break between sessions. In part 1, participants engaged in simulated consultations with 2 virtual patient scenarios, 1 male and 1 female. During these scenarios, participants interacted verbally with the virtual patient characters. In part 2, participants repeated the same scenarios, but with the addition of supplementary educational videos designed to enhance communication skills within each scenario (Multimedia Appendices 4 and 5). Following the completion of part 2, participants repeated the knowledge assessment questionnaire to evaluate potential changes in self-reported weight-related communication knowledge and skill levels. The temporal data associated with the primary author's activation of the WoZ controller buttons were recorded to facilitate an analysis of the author's controller operation during the simulated patient interactions.

The training concluded with a feasibility questionnaire, adapted from the technology acceptance model [40] and the digital simulation literature. Piloted and refined with the advisory group, it included yes or no items such as "I'd like to see this type of training used more in healthcare" and "I can see other uses for this type of training." Feasibility was assessed by calculating the percentage of "yes" responses.

Qualitative Data Collection

The feasibility questionnaire included one open-ended question, where participants were asked to add any other thoughts on the training.

Data Management and Analysis

Quantitative Data

Descriptive statistics were calculated for participant characteristics at baseline, including demographic data from the "About You" questionnaire, feasibility questionnaire responses, and WoZ controller button activation timestamps. All questionnaire data were collected online via the Qualtrics survey platform. To address the aims of the feasibility study, the primary focus was on descriptive data. However, exploratory significance testing using paired-samples *t* tests was conducted to evaluate within-group mean differences in weight-related communication skills between pre- and posttraining knowledge self-assessment responses. Data were analyzed using SPSS (version 28; IBM Corp, 2021).

Qualitative Data

Qualitative data from the open-ended feasibility questionnaire responses were analyzed using a systematic content analysis methodology [41]. The primary author initiated the analysis by familiarization with the textual data. A coding system was developed, informed by the initial patterns identified within the data. This coding scheme was then applied to code the data, facilitating the identification of recurrent themes and patterns.

The identified themes were reviewed with another author (LR) to ensure alignment with the raw data. Finally, these themes were interpreted to explain key participant perspectives and experiences.

Ethical Considerations

Ethical approval for this study was granted by the Ulster University Communications and Media Ethics Filter Committee in December 2020 (reference number: CMFC-20-012). Signed, written informed consent was obtained from all participants via email after they responded to study advertisements and confirmed eligibility. Participants completed the study in private settings of their choosing via Zoom. Data from the online Zoom meetings were transcribed from downloaded audio recordings and anonymized using participant identification numbers. All identifiable information was removed, and data were stored securely in password-protected files accessible only to the research team. No compensation was provided to participants.

Results

Overview

Participant attrition was observed during the feasibility study. Among student participants, a 33% dropout rate occurred between the consent and completion stages. For HCPs, a 68% attrition rate was noted from initial expression of interest via social media to the provision of informed consent. Among those HCPs who provided consent, 20% did not complete the study. Participant reports indicated that the COVID-19 pandemic significantly impacted their available time, contributing to attrition. In addition, some participants reported a perceived lack of content knowledge influencing their decision not to complete the study.

Participant Characteristics

A total of 31 participants, comprising a UK-wide cohort of male and female nurses, doctors, dietitians with varying levels of professional experience, and dietetic students from Ulster University, completed the feasibility study (Table 1). The sample was predominantly female (29/31, 94%) and exhibited a significant proportion of participants with over 10 years of professional practice (21/31, 68%).

Table . Characteristics of participants who completed the feasibility study.

Participant characteristics and professions		Participants, n (%)
Profession		
	Nurse	8 (26)
	Doctor	9 (29)
	Dietitian	12 (39)
	Student (dietitian)	2 (6)
Gender		
	Male	2 (6)
	Female	29 (94)
Years in practice		
	1 - 5	6 (20)
	6 - 10	2 (6)
	10+	21 (68)
	Student (dietitian)	2 (6)
Work area		
	Primary care	16 (52)
	Secondary care	9 (29)
	Other	4 (13)
Location		
	Northern Ireland	9 (29)
	Scotland	4 (13)
	Wales	2 (6)
	England	16 (52)
Previous training or experiences in weight management		
	Yes	19 (61)
	No	12 (39)
Previous experiences of eLearning simulations		
	Yes	8 (26)
	No	23 (74)

Descriptive Statistics

All dietitians (12/12, 100%) reported previous training in weight management. Few doctors (2/9, 23%) and 50% (4/8) of nurses reported training in weight management. Dietitians and nurses were more likely to have reported training in weight management than doctors. Students were less likely to participate in the study (2/31, 6%). Only 26% (8/31) of participants had previous experience with simulation-based learning. Most participants were from England (16/31, 52%), while the smallest proportion came from Wales (2/31, 6%).

Participants' usage time for the VITAL-COMS tool ranged from 18.1 to 41.4 minutes. The mean duration was 24.6 (SD 5.4, median 24.0) minutes. Most participants completed the tool within 20 to 30 minutes, though one participant took over 40 minutes, which may represent a mild outlier.

Inferential Statistics

A paired-samples *t* test revealed a statistically significant improvement in participants' self-assessed weight-related communication skills following training (t_{30}=−5.76, *P*=.001). Specifically, mean scores increased from 28.36 (SD 7.6) pretraining to 32.35 (SD 5.7) post-training, with a mean difference of 3.9 (SD 3.7; 95% CI 2.54 - 5.6). The effect size, as measured by Cohen *d*, was 1.03, indicating a strong effect. Doctors and nurses demonstrated the most substantial improvements, while dietitians showed the least.

Paired-samples *t* tests revealed statistically significant increases in conversation length following training for both the female (*P*=.001) and male virtual patient scenarios (*P*=.001). For the female scenario, the mean conversation length increased from 3.73 (SD 1.36) minutes pretraining to 6.08 (SD 2.26) minutes post training, representing a mean difference of 2.35 (SD 1.75, 95% CI 1.73 - 2.97) minutes. This increase was statistically

significant (t_{30}=7.49, P=.001) with a large effect size (Cohen d=1.34). Similarly, for the male scenario, the mean conversation length increased from 3.61 (SD 1.12) minutes pretraining to 5.65 (SD 1.76) minutes post training, with a mean difference of 2.03 (SD 1.41, 95% CI 1.53 - 2.53) minutes. This increase was also statistically significant (t_{30}=8.03, P=.001) and demonstrated a large effect size (Cohen d=1.44).

Fidelity of VITAL-COMS and the WoZ-Based Simulator

Following completion of the feasibility questionnaire, participants were invited to provide open-ended feedback on their experience with VITAL-COMS. Content analysis of the 84% (26/31) of participants who responded yielded two overarching themes: (1) "Getting used to this type of learning" and (2) "What this type of learning can do for people."

"Getting Used to This Type of Learning"

Participants generally reported positive reactions to the VITAL-COMS tool, though many noted an initial adjustment period. Some participants could immediately engage, while others found the lack of facial expression in the characters difficult to adjust to. Participants perceived a favorable balance between realism and learning stimulation, as illustrated by the comment

> *An enjoyable experience – enough verisimilitude to be meaningful and to stimulate learning, but not so real that it felt uncomfortable or threatening.* [Doctor, male]

This balance is referred to as effortful learning and is often a criticism of poorly designed eLearning content – that is, poorly designed digital learning does not challenge the user enough. Participants also reported initial feelings of awkwardness, stating, "It was very different to anything I have done before and I felt awkward, but if more used to it, I think it would be really useful" (Nurse, female). Despite this, the tool's overall effectiveness was consistently acknowledged.

"What This Training Does for People"

Participants reflected on who might benefit most from the tool, how the tool might help them to think and reflect, and the difference between this type of training and classroom-based training. Training for undergraduate students was mentioned frequently as a potential use for the tool. This seemed to be because it helped participants to practice sensitive conversations without the risk of failure in real life -

> *I think it is great to have this type of training as a "no risk" forum for practicing sensitive topics - the opportunity to do so has been limited to 'real life' where there is often more harm than good whilst people are upskilling.* [Nurse, female]

Mentions of reflection and making participants think were common from most participants. From the educational videos, the use of examples and good questions to ask was valued as a practical approach to engage with patients. The scenario-based learning approach and being able to try again seemed particularly helpful to prompt thinking and reflection,

> *Interesting to do this. Made me think about what I was saying and also already thinking about how I would respond differently again.* [Dietitian, female]

Comparisons between this type of virtual training and classroom training were drawn by many participants. Some felt that the addition of a facilitator for group discussions would add to the realism and ability to reflect. Participants suggested a combined or blended learning approach might be useful:

> *I think it would be good to have in combination with other training as we miss the emotional/ feelings and body language side which is important.* [Dietitian, female]

Participants' Perspective of the Feasibility Study

Poststudy feasibility questionnaire results indicated a predominantly positive reception among participants. Specifically, a substantial majority expressed a desire for increased usage of such tools within health care settings (97%, 95% CI 91% - 100%) and identified potential applications beyond the study context (97%, 95% CI 91% - 100%) Furthermore, 90% (95% CI 80% - 100%) of participants anticipated colleagues benefiting from the tool's implementation. Notably, 84% (95% CI 71% - 97%) reported experiencing emotional engagement during the simulation, suggesting a high degree of perceived fidelity in the learning experience. The use of simulated training for sensitive topics, such as weight management, was favorably received by 84% (95% CI 71% - 97%) of participants. While a majority (74%, 95% CI 59% - 90%) preferred scenario-based simulated learning for skills development, a minority, 13% (95% CI, 1% - 25%) expressed a general dislike for this pedagogical approach. In addition, 19% (95% CI 5% - 33%) indicated a preference for alternative learning modalities.

Discussion

Principal Findings

This study evaluated the feasibility and preliminary efficacy of VITAL-COMS, a previously developed and usability-tested weight-related communication training tool for HCPs. Participants demonstrated statistically significant improvements in self-assessed weight-related communication skills following training with VITAL-COMS. This was accompanied by a significant increase in the duration and engagement level of conversations with virtual characters. Specifically, doctors and nurses exhibited the greatest improvement, likely due to their comparatively limited previous formal training in this domain, while dietitians, who possess greater baseline expertise in weight-related communication, showed smaller, albeit still statistically significant, gains. These findings suggest VITAL-COMS is effective in enhancing both communication confidence and consultative engagement among health care professionals.

Most participants rated the training as realistic and appropriately challenging. Only 13% of participants reported disliking the training modality, and 19% expressed a preference for alternative learning approaches. Several participants suggested a blended model, combining simulator-based training with

subsequent peer or patient role-play. This highlights the importance of adaptable delivery formats to optimize engagement and satisfaction. The tool's real-time feedback functionality contributed to longer, more detailed conversations, demonstrating its potential for direct application in real-world clinical settings.

The observed lower levels of improvement among dietitians may reflect their pre-existing confidence and familiarity with weight-related communication. Consequently, future iterations of VITAL-COMS should incorporate more advanced content tailored to meet specific developmental needs of this professional group.

Low student participation, likely attributable to pandemic-related pressures, academic workloads, and hesitancy to engage with sensitive topics, was a notable limitation. This was corroborated by feedback from participating students. Future iterations could address this by integrating the training into course curricula, offering greater flexibility in scheduling, and providing novice learners with more comprehensive introductory guidance on navigating difficult conversations.

VITAL-COMS directly addresses common barriers to effective weight-related health care communication, particularly misalignment between HCPs' assumptions and patient needs. Through educational videos and real-time feedback, the tool facilitates the reduction of misunderstandings that can compromise communication quality. Posttraining feasibility questions and feedback interviews indicated that participants engaged in reflexivity, recalling previous patient interactions and identifying emotionally challenging encounters. This process is a recognized driver of behavioral change and quality improvement in health care [42,43].

Furthermore, the tool also addresses weight stigma by drawing participants' attention to the impact of weight stigma in the educational videos (Multimedia Appendix 5). Weight stigma is a well-documented barrier in clinical settings [4,7,16]. By foregrounding patient experiences and modeling empathetic communication, VITAL-COMS aims to mitigate patient anticipation of judgment or internalized weight stigma, both of which contribute to disengagement in weight-related care [44,45].

Implications of Findings

The VITAL-COMS training offers a precision-based approach and nuanced approach to weight-related communication training, moving beyond traditional weight management messaging such as "eat less, move more." By fostering the exploration of complex etiological factors contributing to weight gain, including biological, environmental, and psychosocial determinants, VITAL-COMS facilitates a more individualized and respectful patient-centered dialogue. This approach aligns with national and international health policy initiatives, such as the NHS (United Kingdom) health service long-term plan [46] and other international guidelines and calls to action [47-49], which emphasize the importance of enhanced nutrition and weight stigma education among health care professionals. VITAL-COMS supports these objectives by promoting in-depth patient discussions encompassing eating behaviors, appetite regulation, stress management, and sleep hygiene. Furthermore, the tool's inherent adaptability allows for future customization, including scenario diversification and integration into broader health care curricula.

Although most participants completed the VITAL-COMS tool within 20 - 30 minutes, one participant spent over 40 minutes using it. This longer duration likely reflects individual variation in communication style and clinical practice. In this case, the participant was an experienced dietitian working in secondary care, where consultations may be longer and more complex. Compared to others, the participant incorporated more patient-centered techniques, including taking a detailed weight history and using more open-ended questions to explore the patient's perspective. This suggests that while the tool is generally time-efficient, practitioners with more in-depth communication approaches or working in specialist settings may require additional time. It highlights the flexibility of VITAL-COMS to adapt to different clinical styles and patient needs, rather than being a rigid script.

The observed minority of participants who expressed negative feedback regarding the training modality mirrors findings from other studies evaluating digital simulation and virtual reality-based interventions, where user preferences and technological comfort levels exhibit heterogeneity. As demonstrated by Chang et al (2023) [50], perceived ease of use and digital confidence significantly influence the acceptance and usage of simulation-based training tools. This observation is consistent with the technology acceptance model [40], which posits that perceived usefulness and usability are critical determinants of technology adoption. Consequently, the implementation of hybrid learning models that integrate digital and interpersonal learning strategies may enhance engagement and satisfaction across diverse learner profiles.

Comparison With Previous Work

VITAL-COMS represents a novel, skills-based, communication-focused digital simulation tool specifically designed, tested, and evaluated for weight-related conversations. This innovation builds upon the growing body of research demonstrating the efficacy and acceptability of virtual character or virtual human interventions in health care communication training [51-53]. The scalable, immersive, and interactive nature of VITAL-COMS aligns with recent advancements in this area. Ryan et al (2024) [53] established a behavioral framework delineating essential training needs for health care professionals in weight-related communication and obesity science, many of which were also evident in the design of VITAL-COMS. However, VITAL-COMS distinguishes itself through its robust training needs analysis [30,31] and nuanced approach to modality (instructional videos, repetition, and guided feedback) and evaluation. This comprehensive methodology facilitates a more targeted and effective learning experience compared to recent progress.

Significant gaps remain in scalable weight-related communication skills training, particularly given the urgent need to address the rising prevalence of obesity [54,55]. This is shown through the current most available digital learning approaches as they often exhibit limitations in their scope and

delivery. For instance, Logue et al [56], 'Small talk, big difference' intervention, while grounded in the COM-B behavioral theory, provided only a single one-hour online session and lacked opportunities for practical skill development and reflective learning. Similarly, the INTERACT study in Germany [57], which incorporated eLearning and motivational interviewing content, failed to demonstrate sustained improvements in key patient outcomes. In contrast, VITAL-COMS emphasizes experiential learning through practice, repetition, and reflection. Developed using the Unity 3D simulation platform, VITAL-COMS facilitates ongoing development of nuanced communication strategies and allows for adaptability to a wider range of patient scenarios. This focus on experiential learning represents a significant advancement in equipping health care professionals to address weight stigma and engage in sensitive weight discussions with confidence. Comparative platforms such as the Royal College of GPs' obesity hub [58] and the global SCOPE training program [59] offer valuable educational material. However, these platforms typically lack immersive and interactive components. The inclusion of hands-on, realistic role-play with VITAL-COMS distinguishes it from these existing resources, thereby enhancing its potential for impactful translation into clinical practice.

Strengths and Limitations

The limited sample size, combined with recruitment challenges among students, likely exacerbated by the COVID-19 pandemic and the sensitive nature of the training topic, constrains the generalizability of the findings. Furthermore, while the communication self-assessment tool demonstrated preliminary promise, it requires further validation against established competency frameworks. The reliance on self-report measures, while common in communication skills training, introduces recognized limitations associated with subjective reporting bias. To mitigate potential inconsistencies arising from human facilitation, a recognized limitation of WoZ methodologies, the research team implemented a rigorous protocol to standardize instructions and ensure a consistent participant experience. This approach adhered to core principles of WoZ design experiments. Nevertheless, the potential for residual variability in human facilitation remains a consideration when interpreting the study's findings.

Future Work

VITAL-COMS, as an advanced prototype, used a WoZ methodology to simulate real-time natural language interaction with virtual patient characters, a functionality that was not readily available at the time of development. The findings of this feasibility study provide valuable insights for the potential integration of emerging technologies, such as generative AI chatbots, which now offer real-time natural language interaction and feedback capabilities. These advancements present a promising alternative to the WoZ approach, potentially enabling broader and more scalable training delivery. Furthermore, the feasibility of automated AI-driven assessment should be explored in future iterations of VITAL-COMS, addressing the inherent limitations associated with self-report assessment measures. Future development of VITAL-COMS includes expanding the range of characters and scenarios, adapting content for different health care settings and professional groups, incorporating advanced technologies, and ultimately, informing the development of the next iteration of the VITAL-COMS prototype driven by generative AI.

Conclusion

This study found that VITAL-COMS was acceptable to a diverse group of HCPs and it improved weight-related communication skills. While most participants were enthusiastic about the real-time simulation, some also recognized its potential value when used alongside more traditional learning methods. Further development is needed to tailor the training to different professional groups. The findings from this study will be used to inform the next iteration of the VITALS-COMS communication skills training tool driven by generative AI.

Acknowledgments

The authors wish to acknowledge the contribution of the study advisory group whose input was invaluable in the design of the training tool VITAL-COMS. We also wish to thank the study participants who provided their valuable time to take part in the feasibility study.

This study was funded as part of a PhD by the Department for the Economy Northern Ireland.

Authors' Contributions

Conceptualization: FQ (lead), RB (supporting)
Formal analysis: FQ (lead), RB (supporting)
Funding acquisition: AM
Investigation: FQ
Methodology: FQ (lead), AM (supporting), RB (equal), LR (equal)
Project administration: AM
Resources: AM (lead), RB (equal)
Supervision: AM (lead), RB (equal), TmcA (supporting), HZ (supporting)
Validation: RB (Lead), LR (supporting)
Visualization: FQ (lead), RB (supporting)
Writing – original draft: FQ (lead), AM (supporting), RB (equal), TmcA (equal), HZ (equal)

Writing – review & editing: FQ (lead), LR (equal)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Multimedia demonstration of VITAL-COMS (virtual training and assessment for communication skills), how the health care professional talks to the virtual character.
[MP4 File, 4411 KB - mededu_v11i1e65949_app1.mp4]

Multimedia Appendix 2
Presentation from the International Conference on Communication in Healthcare (2022), explaining how VITAL-COMS works and how Wizard of Oz is used as part of the design.
[PPTX File, 6707 KB - mededu_v11i1e65949_app2.pptx]

Multimedia Appendix 3
Self-reported knowledge assessment questionnaire on weight-related communication skills.
[PNG File, 86 KB - mededu_v11i1e65949_app3.png]

Multimedia Appendix 4
Example educational video 1 used in VITAL-COMS (virtual training and assessment for communication skills) to educate on obesity science and communication.
[MP4 File, 5611 KB - mededu_v11i1e65949_app4.mp4]

Multimedia Appendix 5
Example educational video 2 used in VITAL-COMS (virtual training and assessment for communication skills) to educate on weight stigma.
[MP4 File, 2363 KB - mededu_v11i1e65949_app5.mp4]

References

1. GBD 2015 Obesity Collaborators, Afshin A, Forouzanfar MH, et al. Health effects of overweight and obesity in 195 countries over 25 years. N Engl J Med 2017 Jul 6;377(1):13-27. [doi: 10.1056/NEJMoa1614362] [Medline: 28604169]
2. Adult BMI Collaborators. Global, regional, and national prevalence of adult overweight and obesity, 1990-2021, with forecasts to 2050: a forecasting study for the Global Burden of Disease Study 2021. Lancet 2021;405(10481):813-838. [doi: 10.1016/S0140-6736(25)00355-1]
3. Talbot A, et al. People with weight ‐ related long ‐ term conditions want support from GPs. Clin Obes 2021:e12471. [doi: 10.1111/cob.12471]
4. Ryan L, Coyne R, Heary C, et al. Weight stigma experienced by patients with obesity in healthcare settings: A qualitative evidence synthesis. Obes Rev 2023 Oct;24(10):e13606. [doi: 10.1111/obr.13606] [Medline: 37533183]
5. Wharton S, Lau DCW, Vallis M, et al. Obesity in adults: a clinical practice guideline. CMAJ 2020 Aug 4;192(31):E875-E891. [doi: 10.1503/cmaj.191707] [Medline: 32753461]
6. Ells LJ, Ashton M, Li R, et al. Can we deliver person-centred obesity care across the globe? Curr Obes Rep 2022 Dec;11(4):350-355. [doi: 10.1007/s13679-022-00489-7] [Medline: 36272056]
7. Cardel MI, Newsome FA, Pearl RL, et al. Patient-centered care for obesity: How health care providers can treat obesity while actively addressing weight stigma and eating disorder risk. J Acad Nutr Diet 2022 Jun;122(6):1089-1098. [doi: 10.1016/j.jand.2022.01.004] [Medline: 35033698]
8. O'Keeffe M, Flint SW, Watts K, Rubino F. Knowledge gaps and weight stigma shape attitudes toward obesity. Lancet Diabetes Endocrinol 2020 May;8(5):363-365. [doi: 10.1016/S2213-8587(20)30073-5] [Medline: 32142624]
9. Rubino F, et al. Attitudes about the treatment of obesity among healthcare providers. Obes Sci Pract 2021;7(1):1-11. [doi: 10.1002/osp4.518]
10. McHale CT, Laidlaw AH, Cecil JE. Predictors of weight discussion in primary care consultations: A multilevel modeling approach. Patient Educ Couns 2022 Mar;105(3):502-511. [doi: 10.1016/j.pec.2021.07.008] [Medline: 34253384]
11. Ryan L, O'Donoghue G, Crotty M, et al. Factors that influence general practitioners' obesity-related clinical practices and determinants of behavior to target to promote best practice in obesity care: A qualitative exploration. Obes Sci Pract 2024 Oct;10(5):e70012. [doi: 10.1002/osp4.70012] [Medline: 39345781]
12. Blackburn M, Stathi A. Moral discourse in general practitioners' accounts of obesity communication. Soc Sci Med 2019 Jun;230:166-173. [doi: 10.1016/j.socscimed.2019.03.032] [Medline: 31030008]

13. O'Donoghue G, Cunningham C, King M, O'Keefe C, Rofaeil A, McMahon S. A qualitative exploration of obesity bias and stigma in Irish healthcare; the patients' voice. PLoS ONE 2021;16(11):e0260075. [doi: 10.1371/journal.pone.0260075] [Medline: 34843517]
14. Crompvoets PI, Nieboer AP, van Rossum EFC, Cramm JM. Perceived weight stigma in healthcare settings among adults living with obesity: A cross-sectional investigation of the relationship with patient characteristics and person-centred care. Health Expect 2024 Feb;27(1):e13954. [doi: 10.1111/hex.13954] [Medline: 39102661]
15. Philip SR, Phelan SM, Standen EC, et al. Lessons learned from patients' weight-related medical encounters: Results from 34 interviews. Patient Educ Couns 2024 Oct;127:108336. [doi: 10.1016/j.pec.2024.108336]
16. Albury C, Strain WD, Brocq SL, et al. The importance of language in engagement between health-care professionals and people living with obesity: a joint consensus statement. Lancet Diabetes Endocrinol 2020 May;8(5):447-455. [doi: 10.1016/S2213-8587(20)30102-9] [Medline: 32333880]
17. Sagi-Dain L, Echar M, Paska-Davis N. How to talk with patients about weight? Viewpoints of 1697 individuals with overweight and obesity. Patient Educ Couns 2022 Mar;105(3):497-501. [doi: 10.1016/j.pec.2021.09.031] [Medline: 34620519]
18. Tremblett M, Webb H, Ziebland S, Stokoe E, Aveyard P, Albury C. Talking delicately: Providing opportunistic weight loss advice to people living with obesity. SSM Qual Res Health 2022 Dec;2(100162):ne. [doi: 10.1016/j.ssmqr.2022.100162] [Medline: 36531292]
19. Sim JJM, Rusli KDB, Seah B, et al. Virtual simulation to enhance clinical reasoning in nursing: a systematic review and meta-analysis. Clin Simul Nurs 2022;69:26-39. [doi: 10.1016/j.ecns.2022.05.006]
20. Verkuyl M, Lapum JL, St-Amant O, Hughes M, Romaniuk D. Curricular uptake of virtual gaming simulation in nursing education. Nurse Educ Pract 2021 Jan;50:102967. [doi: 10.1016/j.nepr.2021.102967] [Medline: 33465565]
21. Meyer K, James D, Amezaga B, White C. Simulation learning to train healthcare students in person-centered dementia care. Gerontol Geriatr Educ 2022;43(2):209-224. [doi: 10.1080/02701960.2020.1838503]
22. Li YY, Au ML, Tong LK, et al. High-fidelity simulation in undergraduate nursing education: a meta-analysis. Nurse Educ Today 2022;111:105291. [doi: 10.1016/j.nedt.2022.105291]
23. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6(1):42. [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
24. Hargie O. Skilled Interpersonal Communication, 7th edition: Routledge; 2021. [doi: 10.4324/9781003182269]
25. Ericsson A, Hoffman RR, Kozbelt A, Williams AM. The Cambridge Handbook of Expertise and Expert Performance, 2nd edition: Cambridge University Press; 2018. [doi: 10.1017/9781316480748]
26. Caterson ID, Alfadda AA, Auerbach P, et al. Gaps to bridge: Misalignment between perception, reality and actions in obesity. Diabetes Obes Metab 2019 Aug;21(8):1914-1924. [doi: 10.1111/dom.13752] [Medline: 31032548]
27. Chang JE, Lindenfeld Z, Chang VW. Obesity and Patient Activation: Confidence, Communication, and Information Seeking Behavior. J Prim Care Community Health 2022;13:21501319221129731. [doi: 10.1177/21501319221129731] [Medline: 36222682]
28. Standen EC, Rothman AJ, Mann T. Weight loss advice from a healthcare provider is motivating, but it is also stigmatizing: an experimental, scenario-based approach. Ann Behav Med 2025 Jan 4;59(1):kaaf018. [doi: 10.1093/abm/kaaf018] [Medline: 40165436]
29. Quigley F, Moorhead A, Bond R, Zheng H, McAloon T. A virtual reality training tool to improve weight-related communication across healthcare settings. 2019 Sep 10 Presented at: Proceedings of the 31st European Conference on Cognitive Ergonomics; BELFAST United Kingdom p. 19-22. [doi: 10.1145/3335082.3335121]
30. Quigley F, Moorhead A, Bond R, McAloon T, Zheng H. Identifying nutritional myths when healthcare professionals communicate about weight and obesity in healthcare settings. Proc Nutr Soc 2020;79(OCE3). [doi: 10.1017/S0029665120007612]
31. Quigley F, Moorhead A, Bond RR, McAloon T, Zheng H. Training needs analysis: a VR training tool to improve weight‐related communication across healthcare settings [Poster]. Obes Rev 2020 Aug;21(S1):EP-179. [doi: 10.1111/obr.13118]
32. Quigley F. Design, development, usability testing and feasibility study of VITAL-COMS: a 3D-virtual simulator to improve weight-related communication skills training for healthcare professionals. : Ulster University; 2023.
33. Kelley JF. Wizard of Oz (WoZ): A yellow brick journey. J Usability Stud 2018;13:119-124.
34. Nielsen Norman Group. Wizard of Oz Prototyping. Nielsen Norman Group. 2024. URL: https://www.nngroup.com/articles/wizard-of-oz [accessed 2025-01-03]
35. Schlögl S, Doherty G, Luz S. Wizard of Oz experimentation for language technology applications: Challenges and tools. Interact Comput 2015 Nov;27(6):592-615. [doi: 10.1093/iwc/iwu016]
36. Chaby L, et al. Embodied virtual patients as a simulation-based framework for training clinician-patient communication skills. Frontiers in Virtual Reality 2022;3:827312. [doi: 10.3389/frvir.2022.827312]
37. Katz A, Tepper R, Shtub A. Simulation training: Evaluating the instructor's contribution to a Wizard of Oz simulator in obstetrics and gynecology ultrasound training. JMIR Med Educ 2017 Apr 21;3(1):e8. [doi: 10.2196/mededu.6312] [Medline: 28432039]

38. Eldridge SM, Chan CL, Campbell MJ, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. Pilot Feasibility Stud 2016;2:64. [doi: 10.1186/s40814-016-0105-8] [Medline: 27965879]
39. NIHR guidance on applying for feasibility studies. : NIHR; 2023. URL: https://www.nihr.ac.uk/guidance-applying-feasibility-studies [accessed 2023-11-03]
40. Venkatesh V, Bala H. Technology Acceptance Model 3 and a Research Agenda on Interventions. Decision Sciences 2008 May;39(2):273-315. [doi: 10.1111/j.1540-5915.2008.00192.x]
41. Krippendorff K. Content Analysis: An Introduction to Its Methodology, 4th edition: Sage Publications; 2018. [doi: 10.4135/9781071878781]
42. Setchell J, Gard M, Jones L, Watson BM. Addressing weight stigma in physiotherapy: Development of a theory-driven approach to (re)thinking weight-related interactions. Physiother Theory Pract 2017 Aug;33(8):597-610. [doi: 10.1080/09593985.2017.1328718] [Medline: 28590789]
43. Ledema R, Carroll K, Collier A, Hor S, Mesman J. Video-reflexive ethnography in health research and healthcare improvement. . 2018. [doi: 10.1080/14461242.2019.1651175]
44. McGuigan RD, Wilkinson JM. Obesity and healthcare avoidance: A systematic review. AIMS Public Health 2015;2(1):56-63. [doi: 10.3934/publichealth.2015.1.56] [Medline: 29546095]
45. Alberga AS, Edache IY, Forhan M, Russell-Mayhew S. Weight bias and health care utilization: a scoping review. Prim Health Care Res Dev 2019 Jul 22;20:e116. [doi: 10.1017/S1463423619000227] [Medline: 32800008]
46. Cheater S. The NHS Long-Term Plan. Int J Health Promot Educ 2019 May 4;57(3):174-175 [FREE Full text]
47. Kushner RF, Horn DB, Butsch WS, et al. Development of obesity competencies for medical education: a report from the Obesity Medicine Education Collaborative. Obesity (Silver Spring) 2019 Jul;27(7):1063-1067. [doi: 10.1002/oby.22471] [Medline: 31231957]
48. Busetto L, Dicker D, Frühbeck G, et al. A new framework for the diagnosis, staging and management of obesity in adults. Nat Med 2024 Sep;30(9):2395-2399. [doi: 10.1038/s41591-024-03095-3] [Medline: 38969880]
49. Nadolsky K, Addison B, Agarwal M, et al. American Association of Clinical Endocrinology Consensus Statement: Addressing stigma and bias in the diagnosis and management of patients with obesity/adiposity-based chronic disease and assessing bias and stigmatization as determinants of disease severity. Endocr Pract 2023 Jun;29(6):417-427. [doi: 10.1016/j.eprac.2023.03.272] [Medline: 37140524]
50. Chang YY, Chao LF, Chang W, et al. Impact of an immersive virtual reality simulator education program on nursing students' intravenous injection administration: A mixed methods study. Nurse Educ Today 2024 Jan;132:106002. [doi: 10.1016/j.nedt.2023.106002] [Medline: 37922767]
51. Guetterman TC, Sakakibara R, Baireddy S, et al. Medical students' experiences and outcomes using a virtual human simulation to improve communication skills: Mixed methods study. J Med Internet Res 2019;21(11):e15459. [doi: 10.2196/15459]
52. Cooke S, Davies L, Forrest C, et al. The Virtual Patient: a training tool for healthcare staff to have helpful conversations about healthy weight using a Making Every Contact Count (MECC) approach. The Lancet 2024 Nov;404(S62):S62. [doi: 10.1016/S0140-6736(24)02045-2]
53. Ryan L, Coleman S, Zimmermann T, et al. A pilot feasibility study exploring the preliminary effectiveness of an AI-driven virtual human intervention for general practitioner obesity education and communication-skills training. Obes Sci Pract 2025 Aug;11(4):e70083. [doi: 10.1002/osp4.70083] [Medline: 40630231]
54. World Obesity Atlas 2024. World Obesity Federation. 2024 Mar. URL: https://www.worldobesity.org/resources/resource-library/world-obesity-atlas-2024 [accessed 2025-03-03]
55. Collaboration NRF. Worldwide trends in underweight and obesity from 1990 to 2022: a pooled analysis of 3663 population-representative studies with 222 million children, adolescents, and adults. Lancet 2024 Mar 16;403(10431):1027-1050. [doi: 10.1016/S0140-6736(23)02750-2] [Medline: 38432237]
56. Logue J, O'Donnell J, Brooksbank K, et al. An educational intervention to increase referrals of patients with type 2 diabetes from primary care to weight management (small talk big difference): results of a randomised controlled trial. In: Obesity Facts: 26th European Congress on Obesity: Glasgow United Kingdom; 2019, Vol. 12:1-364.
57. Welzel FD, Bär J, Stein J, et al. Using a brief web-based 5A intervention to improve weight management in primary care: results of a cluster-randomized controlled trial. BMC Fam Pract 2021 Apr 2;22(1):61. [doi: 10.1186/s12875-021-01404-0] [Medline: 33794781]
58. Royal College of General Practitioners RCGP Obesity Hub. 2022. URL: https://elearning.rcgp.org.uk/course/view.php?id=534 [accessed 2025-08-11]
59. SCOPE e-learning. World Obesity Federation. 2023. URL: https://www.scope-elearning.org/Saba/Web_spf/EU2PRD0110/guest/guestlearningcatalog [accessed 2025-08-11]

Abbreviations

AI: artificial intelligence
COM-B: capability, opportunity, and motivation for behavior change

CONSORT: Consolidated Standards of Reporting Trials
HCP: health care professional
VITAL-COMS: virtual training and assessment for communication skills
WoZ: Wizard of Oz

Edited by B Lesselroth; submitted 30.08.24; peer-reviewed by S Ganesh, YY Kristian; revised version received 07.04.25; accepted 10.06.25; published 15.08.25.

Please cite as:
Quigley F, Ryan L, Bond R, McAloon T, Zheng H, Moorhead A
A Virtual Simulator to Improve Weight-Related Communication Skills for Health Care Professionals: Mixed Methods Pre-Post Pilot Feasibility Study
JMIR Med Educ 2025;11:e65949
URL: https://mededu.jmir.org/2025/1/e65949
doi:10.2196/65949

Media-Induced and Psychological Factors That Foster Empathy Through Virtual Reality in Nursing Education: 2×2 Between-Subjects Experimental Study

Kuo-Ting Huang[1], PhD; Zexin Ma[2], PhD; Lan Yao[3], PhD

[1]Department of Information Culture and Data Stewardship, University of Pittsburgh, 135 N Bellefield Ave, Room 616, Pittsburgh, PA, United States

[2]Department of Communication, University of Connecticut, Storrs, CT, United States

[3]School of Nursing, Oakland University, Rochester, MI, United States

Corresponding Author:
Kuo-Ting Huang, PhD
Department of Information Culture and Data Stewardship, University of Pittsburgh, 135 N Bellefield Ave, Room 616, Pittsburgh, PA, United States

Abstract

Background: Virtual reality (VR) has emerged as a promising tool in medical education, particularly for fostering critical skills such as empathy. However, how VR, combined with perspective-taking, influences affective empathy in nursing education remains underexplored.

Objective: This study investigates the influence of VR and perspective-taking on affective empathy in nursing education, focusing on 4 psychological factors: perceived self-location, narrative transportation, emotional engagement, and affective empathy.

Methods: A 2×2 between-subjects design was used, involving 69 nursing undergraduates from two Midwest universities. The participants engaged with a narrative-focused video game, That Dragon, Cancer, in either VR or non-VR conditions and from the perspective of either parents or clinicians.

Results: VR significantly enhanced perceived self-location (*P*=.01), while adopting a clinician's perspective amplified emotional engagement (*P*=.03). However, VR did not significantly influence narrative transportation (*P*=.35). An interaction effect was found between the platform and player's perspective on narrative transportation (*P*=.04). Several indirect effects of media elements on affective empathy were observed via other psychological factors, though the direct effect of VR on affective empathy was not significant (*P*=.84).

Conclusions: These findings underscore the potential of VR in medical education, suggesting that perspective-taking should be carefully considered when designing immersive learning experiences. The study advocates for broader integration of VR technologies into medical curricula to enhance instruction quality and patient-centered care.

(JMIR Med Educ 2025;11:e59083) doi:10.2196/59083

KEYWORDS

nursing education; narrative transportation; presence; virtual reality; game-based learning; affective empathy

Introduction

Background

The domain of health care education is currently undergoing a monumental shift, facilitated by advancements in immersive technologies such as virtual reality (VR). Immersive technologies have demonstrated significant potential in transforming medical and health education, including enhanced training in surgical procedures [1,2], improved understanding of complex biomedical processes through immersive visualization [3,4], and more empathetic patient care through simulated patient interactions [5,6]. Despite its promise, VR technology faces challenges such as system fidelity and presence, which can impact user experience and learning outcomes [7]. Addressing these challenges is essential for the effective design and implementation of VR training modules.

Moreover, VR provides a safe and controlled environment for students to practice and make mistakes without direct consequences on actual patients, increasing their confidence and proficiency before real-world clinical scenarios [8,9]. Concurrent with this paradigm shift, immersive technologies like VR have emerged as an effective tool in the instruction of empathy among nursing students [10-12]. Empathy, being a fundamental aspect of the nursing profession, has been shown to improve patient adherence to treatment plans [13], satisfaction levels [14], and overall health outcomes [15]. Specifically, being

immersed in virtual narratives allows students to navigate and process their own emotions, as well as respond appropriately within simulated scenarios [16]. This innovative approach offers a safe space for students to handle the emotional complexities associated with patient care, thus better preparing them for future clinical encounters involving nuanced emotional interactions.

VR simulation has also been studied in the context of nursing education [17-21]. In our recent study [22], we found that VR-based role-playing games enhanced cognitive empathy among nursing students. However, affective empathy remains underexplored in VR-based nursing education. Cognitive empathy involves understanding another's perspective, whereas affective empathy refers to sharing and responding to another's emotional states [23]. A recent meta-analysis revealed that VR has a significant effect on perspective-taking outcomes (cognitive empathy) but lacks impact on affective empathy [23]. Therefore, the objective of the current research is to investigate the potential of VR to influence affective empathy among nursing students by exploring narrative-related psychological factors.

Research on immersive media indicates that perceived self-location (ie, being there in a virtual environment) is a key mechanism that explains the impact of VR on empathy [24,25]. VR-based empathy training programs often contain a story with plots and characters to help users experience a situation first-hand [26]. They comprise both medium-based (ie, of the virtual environment) and message-based (ie, of the narrative) elements [27].To understand how VR might enhance affective empathy, we explore specific psychological mechanisms: perceived self-location and narrative transportation.

Perceived self-location refers to the sensation of "being there" in a virtual environment, enhancing users' immersion and empathy [24,25], which we hypothesize will lead to greater empathy by allowing users to more deeply understand and share the feelings of virtual characters. Narrative transportation is the cognitive and emotional absorption into a story, where individuals lose awareness of their physical surroundings and form intense connections with the narrative and characters [28,29]. This deep absorption can lead to significant shifts in attitudes and empathy toward others [30]. Research has found that VR narratives lead to higher levels of transportation compared with traditional media, which in turn enhance empathetic responses [31]. Therefore, we hypothesize that VR will enhance narrative transportation, significantly impacting affective empathy.

A noteworthy characteristic of VR-based empathy training programs is their ability to feature multiple characters, thus allowing users to experience a narrative from varying character perspectives [32,33] The perspective-taking aspect significantly impacts users' emotional engagement with the character [34]. Research has found that users are more likely to experience emotional engagement toward characters that are portrayed positively [33] and are similar to them [35]. Therefore, we hypothesize that character perspective will impact emotional engagement, with higher engagement when viewing from the clinician's perspective compared with the parents' perspective due to character-user similarity, subsequently influencing affective empathy.

Moreover, we propose that perceived self-location, narrative transportation, and emotional engagement will form a sequential mediation model to account for the effect of VR-based training. Previous research on video games has found that the feeling of "being there" in the game is a predictor of flow, an experience similar to narrative transportation [36,37]. A recent study obtained similar findings: spatial presence predicted narrative transportation in a VR storytelling experience [38]. Furthermore, existing work suggests that narrative transportation is associated with an increase in emotional responses [39,40]. Hence, we predict that perceived self-location, enhanced by VR, will foster narrative transportation, which will, in turn, promote emotional engagement. Emotional engagement will then lead to affective empathy.

Hypotheses and Research Questions

Based on the literature review, the following hypotheses were proposed to explore the interconnected roles of perceived self-location, narrative transportation, and emotional engagement in enhancing affective empathy through VR interventions:

1. Hypothesis 1: VR conditions will enhance (1) perceived self-location and (2) narrative transportation compared with non-VR conditions.
2. Hypothesis 2: the participants will experience higher emotional engagement when viewing the narrative from a clinician's perspective compared with a parent's perspective.
3. Hypothesis 3: perceived self-location will positively predict narrative transportation in VR-based training programs.
4. Hypothesis 4: narrative transportation will positively predict emotional engagement within VR-based experiences.
5. Hypothesis 5: emotional engagement will positively predict affective empathy.

In addition to these hypotheses, we proposed the following research questions to explore potential interaction and indirect effects:

1. Research question 1: Are there interaction effects between the media platform (VR vs non-VR) and character perspective (clinician vs parents) on psychological factors such as perceived self-location, narrative transportation, and emotional engagement?
2. Research question 2: Do perceived self-location, narrative transportation, and emotional engagement mediate the relationship between the media platform and affective empathy?

Methods

Design

The proposed conceptual framework of this study is illustrated in Figure 1. This study used a 2×2 between-subjects experimental design to investigate the effects of the platform (VR vs non-VR) and perspective (parents vs clinicians) on nursing undergraduates' empathy levels. Participants were randomly assigned to 1 of 4 conditions: VR parents, VR

clinicians, non-VR parents, or non-VR clinicians. Figure 2 shows the 4 conditions of the study.

Figure 1. The proposed conceptual framework.

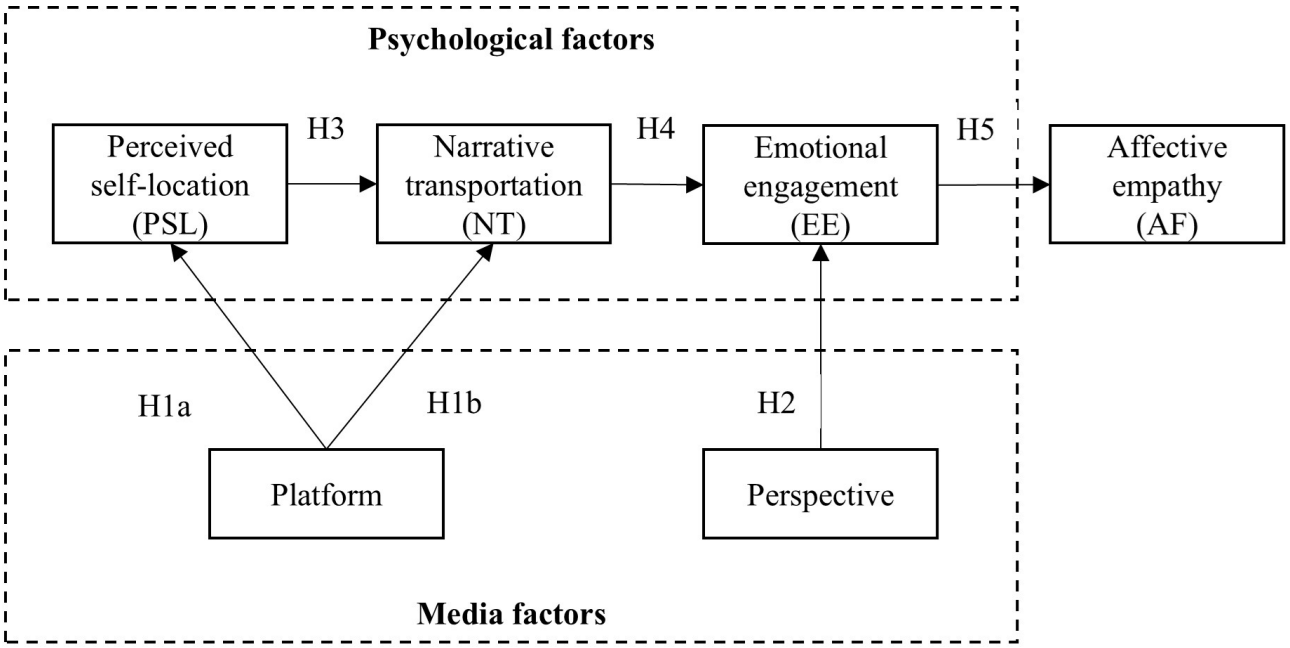
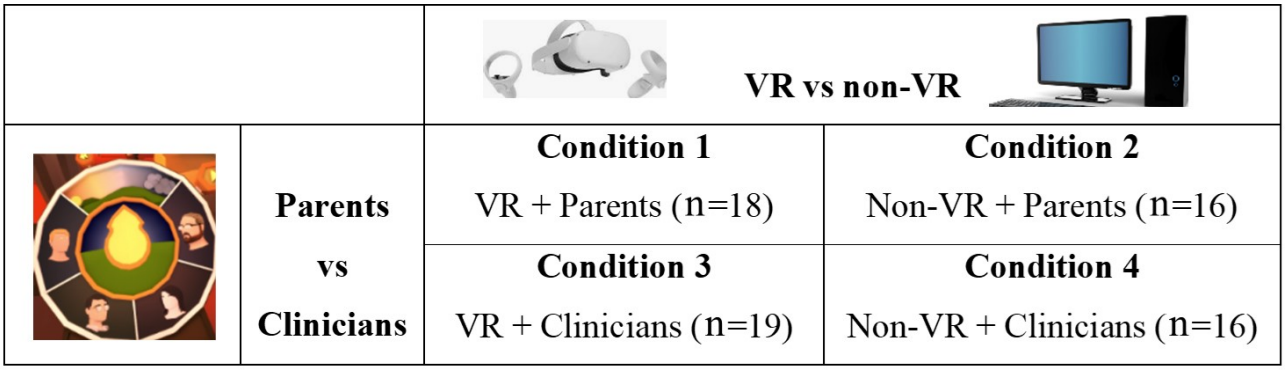


Figure 2. Four conditions of the study.



Participants

A total of 69 nursing undergraduates from two Midwest universities participated in the study, predominantly female (57/69, 82.6%) and White (57/69, 82.6%), mostly juniors (39/69, 56.5%; sophomores: 24/69, 34.8%; seniors: 6/69, 8.7%). The average age was 22.13 (SD 3.70) years. The participants were recruited via university mailing lists and classroom announcements, with inclusion criteria requiring them to be nursing undergraduates aged 18 years or older. Most participants (54/69, 77.6%) reported they were not familiar with VR.

Before participation, all were screened for visual impairments and susceptibility to motion sickness; those with significant issues were excluded to avoid adverse effects during the VR experience.

Randomization was done through a drawing process, assigning participants to 1 of 4 conditions, ensuring balanced allocation between VR and non-VR platforms and the perspectives of parents or clinicians. Given the sample size and complexity of the 2×2 design, the study may be underpowered to detect small effects. No a priori power calculations were conducted; however, this exploratory research aims to investigate initial effects and generate hypotheses for future studies with larger samples.

Experimental Procedures and Stimulus

Upon arrival at the research lab, participants provided informed consent and completed a short pretest questionnaire. They were then assigned to 1 of the 4 gameplay conditions as per the randomization process described above. The participants were informed that they could stop or report at any time if they experienced motion sickness or visual discomfort.

The participants engaged with the seventh chapter of the narrative-focused video game That Dragon, Cancer, titled “I’m Sorry Guys, It’s Not Good.” That Dragon, Cancer is an interactive narrative game developed by Numinous Games (Mainframe Studios) [41] that tells the real-life story of a family’s experience with their son’s terminal cancer diagnosis. The game is designed to evoke emotional responses and foster empathy through immersive storytelling [42]. This chapter was selected based on prior research demonstrating its efficacy in increasing empathy among medical students [42]. The gameplay allowed participants to experience a pivotal moment when Joel’s parents were informed of his terminal cancer diagnosis,

navigating the scene from 4 unique perspectives: dad, mom, doctor, and nurse. The game is designed as a point-and-click adventure, which allows participants to trigger conversations and access a selected character's inner thoughts.

The participants in the VR conditions used Oculus Go headsets, seated in a quiet room to minimize distractions. The headsets provided a 360-degree immersive experience with built-in headphones for audio. For the non-VR conditions, the participants used Dell laptops or iPads (Apple, Inc) with over-ear headphones, seated at a desk in the same room to ensure consistent environmental conditions. The game was presented on a standard screen, and the participants interacted using a mouse or touchscreen, replicating typical non-VR gameplay settings. The gameplay lasted approximately 10 minutes. The selection of this exposure time was based on previous studies indicating that brief VR experiences can effectively elicit emotional and empathetic responses [43]. Immediately after the gameplay, participants completed a posttest questionnaire assessing their empathy and gaming experience. This immediate administration was intended to capture their reactions and reduce potential recall bias.

Instrument

Several validated scales were used to measure the constructs of interest. These measures are specifically applicable to our study's context in nursing education and VR-based empathy training. First, the Spatial Presence Experience Scale, developed by Hartmann et al [44], was used to evaluate self-location and assess nursing students' immersion in the simulated clinical scenarios. This validated scale, widely used in diverse media environments, measures 2 facets of the spatial presence experience: perceived self-location and potential actions, while also considering key factors that influence spatial presence. The study's transportation scale was an adaptation from Green and Brock [28], which measures students' absorption of patient stories that may foster empathy. In addition, the Emotional Engagement scale used in the study was sourced from Knol and Van Linge [45], which captures students' emotional connection with virtual characters, vital for developing affective empathy. Affective empathy was assessed with 3 items from the validated empathy scale by Batson et al [46]. The participants responded to items on a 7-point Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree). The results of these items were averaged to produce a composite score for data analysis. Covariates such as participants' gender, age, race, and university affiliation were included in the following statistical analyses. Descriptive statistics, the items used, and reliability estimates for all scales are presented in Table 1.

Table . List of items used in the study and the descriptive statistics (N=69).

Variable	Statistics	Items
Perceived self-location [44]	• Mean 5.19 (SD 1.12) • Cronbach (α=0.88)	• I felt like I was actually there in the environment of the presentation. • It was as though my true location had shifted into the environment in the presentation. • I felt as though I was physically present in the environment of the presentation. • It seemed as though I actually took part in the action of the presentation
Transportation [28]	• Mean 5.77 (SD 0.93) • Cronbach α=0.83	• I could picture myself in the scene of the events described in the story. • I was mentally involved in the story while watching it. • I wanted to learn how the story ended.
Emotional engagement [45]	• Mean 5.83 (SD 0.98) • Cronbach α=0.83	• The story affected me emotionally. • During the media experience, when the characters suffered in some way, I felt sad. • I felt sorry for some of the characters in the story.
Affective empathy [46]	• Mean 5.65 (SD 1.24) • Cronbach α=0.93	• Did watching/playing this video make you feel __ • softhearted • sympathetic • compassionate

Statistical Analysis

Data analysis was performed using SPSS 29 statistical software. Hypotheses H1, H2, and RQ1 were assessed through a series of analyses of covariance (ANCOVA), controlling for covariates such as participants' gender, age, race, and university affiliation. Covariates such as participants' gender, age, race, and university affiliation were included in the ANCOVA because these demographic factors have been shown to influence empathy levels and responses to VR experiences [23]. Including these covariates helps control for potential confounding variables, ensuring that the effects observed are attributable to the

experimental manipulations rather than demographic differences. Hypotheses H3-H5 and RQ2 were tested with mediation analyses using the PROCESS macro, following a bootstrap estimation approach with 5000 samples, based on Hayes' Process Model 6 [47]. Control variables were also included in these analyses to control for potential confounding influences.

Ethical Considerations

This study was approved by the Institutional Review Board at Ball State University (approval number 1386023 - 1). All participants provided informed consent before data collection. To acknowledge their participation, they received extra course credits as compensation. Confidentiality and privacy were maintained, and participants had the right to withdraw at any time without consequences.

Results

Descriptive Statistics and Correlation Analysis

The descriptive statistics for each condition, including mean and SD, are presented in Table 2. In addition, skewness and kurtosis values were assessed to check the normality of the data distribution, and the results were within the acceptable range confirming the appropriateness of the data for further analysis. A correlation analysis was conducted to identify any potential multicollinearity issues among the variables. The results indicated that while variables were correlated, they did not exceed the threshold that would suggest multicollinearity, thus ensuring the independence of predictors. The means for all variables were above the midpoint of the scale, indicating generally high levels of perceived self-location, narrative transportation, emotional engagement, and affective empathy among participants. The SD values ranged from 0.83 to 1.24, suggesting moderate variability in responses.

Table . Means and SEs by experimental conditions (N=69). Participants' race, gender, school year, and university affiliation were controlled.

Platform and perspective	Perceived self-location	Narrative transportation	Emotional engagement	Affective empathy
VR, mean (SE)				
Parents (n=18)	5.58 (1.07)	6 (0.87)	5.85 (0.72)	5.76 (1.43)
Clinicians (n=19)	5.66 (0.98)	5.81 (1.07)	6.02 (1.04)	5.84 (1.15)
Non-VR, mean (SE)				
Parents (n=16)	4.50 (1.34)	5.27 (1.16)	5.31 (1.26)	5.23 (1.41)
Clinicians (n=16)	4.89 (0.65)	5.96 (0.59)	6.13 (0.7)	5.75 (0.9)

Analysis of Covariance

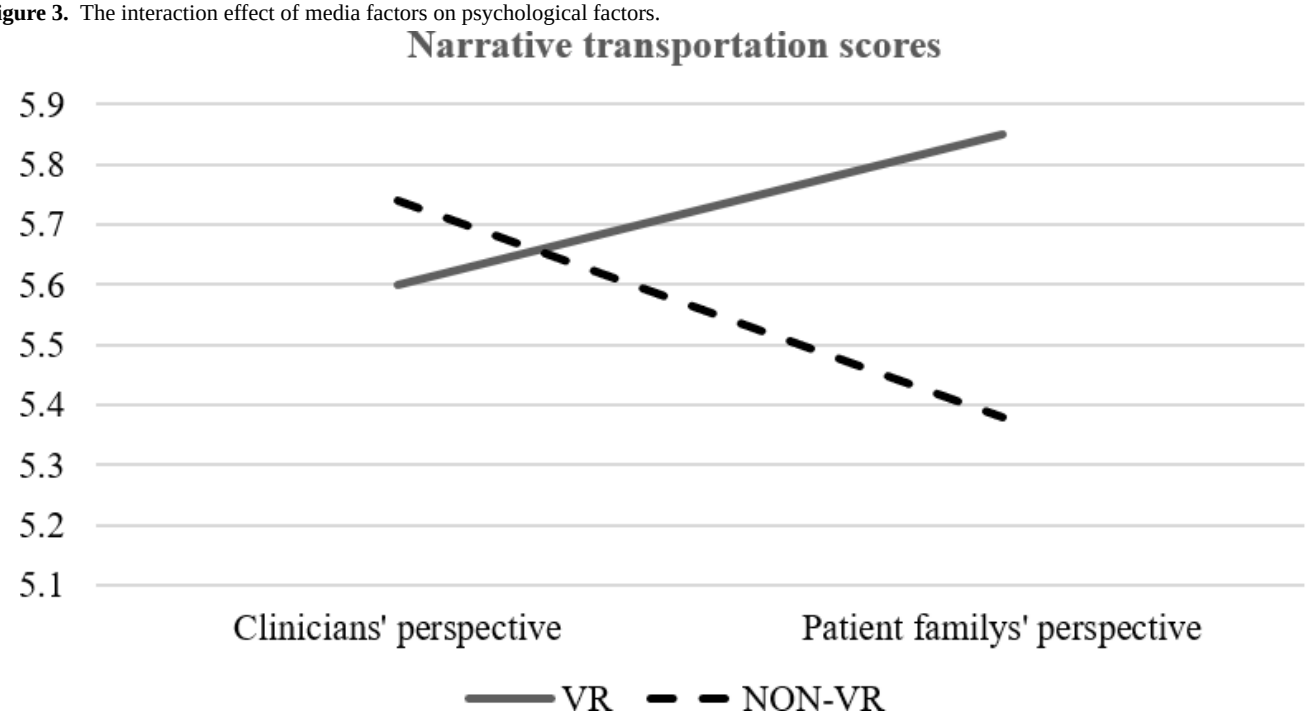
The analysis of covariance (ANCOVA) included checks for assumptions such as homogeneity of variances, assessed using the Levene test. The results of these tests confirmed that the assumptions of ANCOVA were met across all variables of interest. Specifically, Levene test results for homogeneity of variances were $F_{3,65}$=2.146, P=.10 for self-location; $F_{3,65}$=1.566, P=.20 for narrative transportation; $F_{3,65}$=0.904, P=.44 for emotional engagement; and $F_{3,65}$=0.620, P=.60 for affective empathy. These results suggest that the variances of the residuals were not significantly different from each other across groups for each variable, thus fulfilling one of the key assumptions for conducting ANCOVA and lending validity to the subsequent analyses.

Based on our findings (Table S1 in Multimedia Appendix 1), the platform of the game had a significant influence on perceived self-location ($F_{1,61}$=6.60, P=.01, partial η^2=0.098), indicating that VR has a stronger influence on perceived self-location compared with non-VR environments. This substantial difference in perceived self-location depending on the platform used provided support for hypothesis H1a, suggesting VR's unique capacity to enhance users' sense of presence within the virtual environment. We also examined the effect of the adopted perspective on participants' emotional engagement. The results revealed a significant effect ($F_{1,61}$=4.76, P=.03, partial η^2=0.072), indicating that participants who assumed the clinician's perspective exhibited greater emotional engagement compared with those adopting a patient's perspective. This finding supported hypothesis H2, highlighting the importance of perspective in influencing emotional responses in VR settings. However, the effect of VR on narrative transportation did not yield significant results ($F_{1,61}$=0.90, P=.35, partial η^2=0.014), thereby not supporting hypothesis H1b.

To answer the first research question, we also test whether there is an interaction effect of media factors on psychological factors. The results revealed that the platform and perspective had an interaction effect on narrative transportation ($F_{1,61}$=4.68, P=.04, partial η^2=0.070). The post hoc analysis indicated that participants experiencing the game from the perspective of patients' families in a non-VR platform exhibited the lowest level of narrative transportation. This nuanced finding sheds light on how different combinations of platform and perspective can uniquely affect the immersive experience of users. These interaction effects are further elucidated in Figure 3, providing

a visual representation of these dynamics. Specifically, the narrative transportation scores for clinicians and patient families under VR (solid line) and non-VR (dashed line) conditions showed diverging trends between the two perspectives.

Figure 3. The interaction effect of media factors on psychological factors.

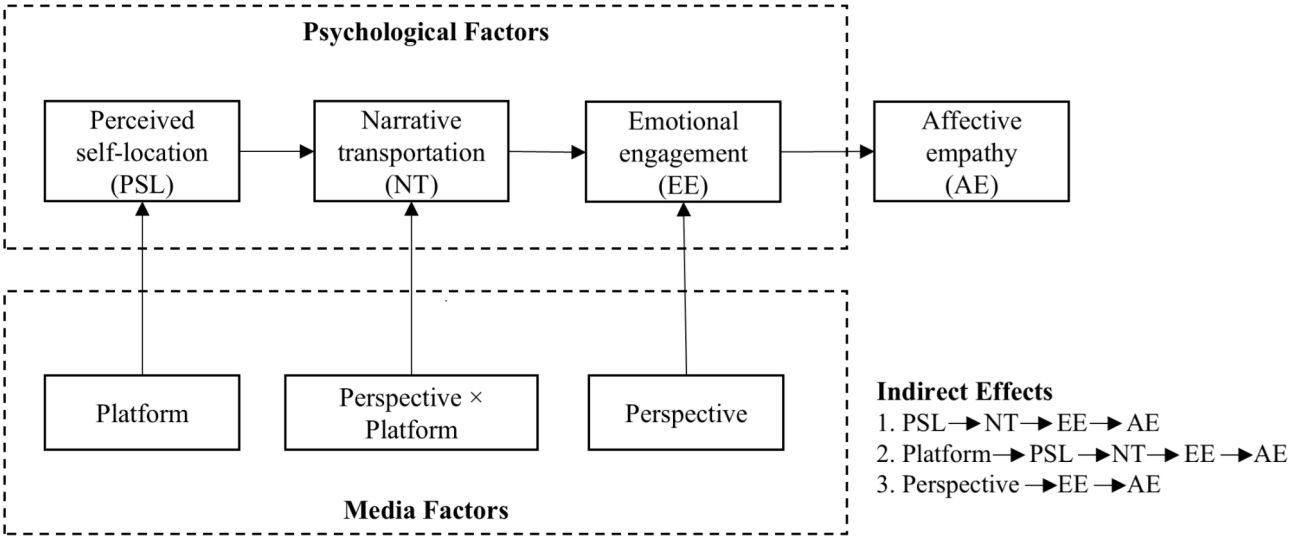


Mediation Analysis

The mediation analyses using Hayes' PROCESS model 6 revealed several direct and indirect effects. Regarding direct effects, self-location was found to have a significant positive effect on transportation (b=0.54, P<.001). Transportation had a positive effect on emotional engagement (b=0.74, P<.001). Emotional engagement was found to have a strong positive effect on affective empathy (b=0.84, P<.001). These findings, supporting H3-H5, illustrate the psychological process from self-location in VR, through narrative transportation and emotional engagement, to the ultimate development of affective empathy. We used 5000 bootstrap samples to generate bias-corrected confidence intervals for the indirect effects. The significance of the mediation pathways was determined by examining whether the confidence intervals excluded zero. All variables were included in the model simultaneously, and control variables were accounted for in the analysis. The model fit was assessed, and all pathways were found to be significant, confirming the validity of the mediation pathways.

The results also uncovered three sets of indirect effects, which answered our second research question. First, perceived self-location exerted an indirect effect on affective empathy through narrative transportation and emotional engagement (indirect effect=0.25, 95% CI 0.0610-0.5048). Second, the platform had an indirect effect on affective empathy through perceived self-location, narrative transportation, and emotional engagement (indirect effect=0.25, 95% CI 0.0610-0.5048). Finally, the perspective adopted by the participants was found to have an indirect effect on affective empathy through emotional engagement (indirect effect=0.27, 95% CI 0.0059-0.6153). These indirect effects, along with the direct effects, are represented in the revised conceptual framework presented in Figure 4.

Figure 4. The revised conceptual framework.



Discussion

Summary and Interpretations of the Findings

The study investigates the impact of media platforms and players' perspectives on perceived self-location, narrative transportation, emotional engagement, and affective empathy within medical education, which capture key psychological processes essential for developing empathy in clinical practice. The results indicate that all psychological factors were influenced by media elements, albeit through different mechanisms.

Our hypothesis H1a was supported, demonstrating that playing the game in VR (vs non-VR) significantly increased perceived self-location. This finding is consistent with prior studies [24,25]. However, contrary to our hypothesis H1b, VR (vs non-VR) did not differ in narrative transportation, which echoes findings from a recent meta-analysis [27]. This nonsignificant finding suggests that while VR enhances the sense of presence or self-location, it may not necessarily increase narrative transportation compared with non-VR platforms. One possible explanation is that narrative transportation is more strongly influenced by the quality of the narrative itself rather than the medium through which it is delivered [28]. It is possible that the narrative content was equally engaging in both VR and non-VR formats, resulting in similar levels of transportation. Future research could explore how different narrative structures or content types interact with VR to affect narrative transportation.

Regarding our hypothesis H2 adopting a clinician's perspective during the VR experience significantly influenced emotional engagement, the result is consistent with previous research [32,33]. This insight emphasizes the value of learning from other clinicians in strengthening emotional ties with patients, thereby fostering affective empathy. The findings showed the potential of VR in medical education by enhancing perceived self-location, which can improve empathy toward patients. The significant influence of perspective adopted by players reiterates its role in enhancing emotional engagement, crucial for fostering empathy in health care professionals.

The proposed direct relationships between psychological factors and affective empathy (H3-H5) were all supported. The findings demonstrate that a sense of perceived self-location in VR can enhance narrative transportation, leading to increased emotional engagement. In turn, this fosters affective empathy, a critical skill for health care professionals to understand and respond to patients' emotional experiences effectively, a critical skill for health care professionals to understand and respond to patients' emotional experiences effectively. This sequential process confirms the potential of VR in facilitating immersive, emotionally engaging learning experiences in medical training, promoting the development of affective empathy.

The exploration of the research questions (RQ1 and RQ2) yielded compelling insights. The first research question investigates potential interaction effects. Our findings indicate a significant interaction effect among these conditions, specifically revealing that scores on narrative transportation were significantly lower in the condition using non-VR with a patient's perspective compared with the other 3 conditions. This interaction effect suggests that the combination of platform and perspective plays a crucial role in influencing narrative transportation. One possible explanation is that adopting the patient's perspective in a non-VR environment may not provide sufficient immersion or sensory cues to facilitate deep engagement with the narrative. In contrast, VR may compensate for the less immersive perspective by enhancing sensory immersion, while adopting a clinician's perspective may align more closely with the students' professional identity, facilitating engagement even in non-VR settings. This finding indicates that the effectiveness of narrative transportation may depend on the congruence between the medium, the perspective adopted, and the user's own identity and experiences. Future studies could explore how personal relevance and role identification influence narrative engagement across different platforms.

For the second research question, the study uncovered 3 indirect effects: perceived self-location impacted affective empathy via narrative transportation and emotional engagement, the platform influenced affective empathy through self-location, narrative transportation, and emotional engagement, and the perspective affected affective empathy through emotional engagement.

Theoretical Contributions and Practical Implications

This study makes significant theoretical contributions to the fields of empathy research and narrative communication. It demonstrates how VR can influence affective empathy through mechanisms such as perceived self-location, narrative transportation, and emotional engagement, thereby deepening our understanding of the integral role immersive technologies play in fostering critical emotional competencies. By identifying the sequential mediation of these psychological factors, our findings extend existing theories on empathy development and immersive media, providing empirical evidence within the context of nursing education.

These findings add valuable empirical evidence to empathy research and highlight the importance of immersive, technology-enabled experiences in shaping affective responses. Specifically, our study fills a gap in the literature by focusing on affective empathy rather than cognitive empathy, which has been less examined in VR research. Furthermore, by exploring the interaction between character perspective and affective empathy within VR environments, the study offers a novel perspective on empathy development and enriches narrative communication research. This contributes to a more nuanced understanding of how perspective-taking in VR can differentially impact emotional engagement and empathy outcomes, which has practical implications for designing effective educational interventions.

From a practical standpoint, the findings offer actionable insights for integrating VR into nursing education. We propose that nursing programs should incorporate VR experiences that emphasize perspective-taking from a clinician's viewpoint to enhance emotional engagement and affective empathy among students. To address potential barriers such as cost, accessibility, and technological limitations, nursing programs could start by incorporating affordable VR solutions like stand-alone VR headsets or 360-degree video experiences, which are more feasible than high-end VR systems or simulation stations. For accessibility, it is important to ensure that VR experiences are also designed for students in the classroom settings. Technological limitations, such as a lack of technical expertise among faculty, can be mitigated through training workshops and technical support services. In addition, curricula should be designed to seamlessly integrate VR experiences into existing courses, perhaps starting with pilot programs to evaluate effectiveness before broader implementation. By proactively addressing these barriers, educators can more effectively leverage VR technology to enhance empathy training in nursing education. By implementing these recommendations, educators and institutions can leverage VR technology to significantly enhance the quality of medical education and training, especially in the domain of empathy development.

Limitations and Future Research Directions

This study has several limitations. First, while the sample size was sufficient to yield statistical power, it was relatively small. The participants were primarily female, white nursing undergraduates from two Midwestern universities. The small sample size may have also limited our ability to detect smaller effect sizes. The homogeneity in gender and race may have influenced the results, as previous research suggests that empathy levels and responses to VR experiences can vary across different demographic groups. For instance, gender differences have been observed in emotional processing and empathetic responses, which could affect how participants engage with VR-based empathy training. Therefore, our sample may limit the generalizability of the results.

Future studies should prioritize recruitment strategies that enhance demographic diversity to ensure the broader applicability of findings. One approach is expanding outreach to institutions with more diverse student populations, such as historically Black colleges and universities, Hispanic-serving institutions, and community colleges. Establishing partnerships with nursing programs in urban and rural areas can also help reach a wider range of participants with different socioeconomic and educational backgrounds. In addition, leveraging professional nursing associations, student organizations, and social media platforms can improve the recruitment of participants from underrepresented groups. Providing flexible participation options, such as internet-based study components or varied scheduling, may further increase accessibility and encourage participation from nontraditional students, working professionals, or those with caregiving responsibilities. Implementing these strategies can enhance inclusivity and variability, ultimately strengthening the generalizability of future research.

Second, the research design offered only a brief, single-session exposure to VR and non-VR platforms and varying character perspectives. This short exposure duration may not have allowed sufficient time for participants to fully immerse themselves in the VR experience or for the effects on empathy to fully manifest. This limited interaction may not fully capture the potential effects of extended VR-based training on empathy. Longer or repeated exposures could provide a more accurate assessment of VR's impact on empathy development. Therefore, longitudinal studies are recommended to investigate the long-term effects of VR on empathy, providing insights into the sustainability and potential long-term integration of VR into medical education.

Third, while we used self-reported scales that have been validated and widely used, these measures may introduce a certain level of response bias. Self-reported data can be influenced by social desirability or participants' subjective interpretations of the questions, which may affect the accuracy of the results. Future investigations could stand to gain significantly from incorporating more objective evaluative methods, such as physiological and behavioral observations, that would serve to substantiate the self-reported data. For example, using biometric measures like heart rate variability or skin conductance could provide objective insights into emotional engagement and empathy responses. In addition, behavioral assessments during simulated interactions could offer tangible evidence of empathy development. Implementing mixed-method approaches would mitigate reliance on self-reports and enhance the validity of future research findings.

Finally, although the usability of the VR interface is important, our study did not assess usability using standard questionnaires

such as the System Usability Scale [48] or the Usefulness, Satisfaction, and Ease of use [49] questionnaire. Future research could incorporate these usability measures to provide a more comprehensive evaluation of VR interfaces in educational settings. Usability testing plays a crucial role in ensuring that VR interventions in nursing education are intuitive, user-friendly, and meet the needs of learners. A system that is difficult to navigate or provides a poor user experience can disrupt engagement and limit the effectiveness of the intervention. Future research should integrate standardized usability assessments, such as the System Usability Scale or Usefulness, Satisfaction, and Ease of use questionnaire, to systematically evaluate the user experience. These measures will help identify areas for improvement, offering insights into how the VR interface aligns with learning goals and how usability affects student engagement and comprehension. By incorporating such evaluations, design improvements can be made—whether refining the interface, enhancing interaction features, or adjusting the VR experience to better suit diverse learning styles. Ultimately, addressing usability issues can improve the practical application of VR in nursing education, ensuring that immersive learning experiences are both accessible and impactful for real-world clinical practice.

Conclusion

In this study, we explored the interactive roles of media platforms and perspective-taking in shaping key psychological factors, including perceived self-location, narrative transportation, emotional engagement, and affective empathy, in nursing education settings. The initial findings provide empirical evidence for the potential of immersive technologies as applicable pedagogical tools, particularly in teaching and training future health care providers. The capacity of virtual reality to facilitate the feeling of presence, build emotional engagement, and foster empathy among nursing students shows the potential to foster a greater degree of patient-centered care. Therefore, this study advocates for a wider consideration of integrating technologies into health care education curriculum design and development. The potential benefits and financial viability of VR technologies could enrich pedagogical experiences and pave the way for the emergence of competent health care professionals, well-equipped to navigate different scenarios in health care delivery.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Analysis of covariance (ANCOVA) results for multiple outcome measures.
[DOCX File, 17 KB - mededu_v11i1e59083_app1.docx]

References

1. Gupta A, Cecil J, Pirela-Cruz M, Ramanathan P. A virtual reality enhanced cyber-human framework for orthopedic surgical training. IEEE Systems Journal 2019;13(3):3501-3512. [doi: 10.1109/JSYST.2019.2896061]
2. Jiang H, Vimalesvaran S, Wang JK, Lim KB, Mogali SR, Car LT. Virtual reality in medical students' education: scoping review. JMIR Med Educ 2022 Feb 2;8(1):e34860. [doi: 10.2196/34860] [Medline: 35107421]
3. González Izard S, Juanes Méndez JA, Ruisoto Palomera P, García-Peñalvo FJ. Applications of virtual and augmented reality in biomedical imaging. J Med Syst 2019 Mar 14;43(4):1-5. [doi: 10.1007/s10916-019-1239-z] [Medline: 30874965]
4. Manuel JK, Purcell N, Abadjian L, et al. Virtual worlds technology to enhance training for primary care providers in assessment and management of posttraumatic stress disorder using motivational interviewing: pilot randomized controlled trial. JMIR Med Educ 2023 Aug 28;9:e42862. [doi: 10.2196/42862] [Medline: 37639299]
5. Dyer E, Swartzlander BJ, Gugliucci MR. Using virtual reality in medical education to teach empathy. J Med Libr Assoc 2018 Oct;106(4):498-500. [doi: 10.5195/jmla.2018.518] [Medline: 30271295]
6. Helle N, Vikman MD, Dahl-Michelsen T, Lie SS. Health care and social work students' experiences with a virtual reality simulation learning activity: qualitative study. JMIR Med Educ 2023 Sep 20;9:e49372. [doi: 10.2196/49372] [Medline: 37728988]
7. Cummings JJ, Cahill TJ, Wertz E, Zhong Q. Psychological predictors of consumer-level virtual reality technology adoption and usage. Virtual Real 2023;27(2):1357-1379. [doi: 10.1007/s10055-022-00736-1] [Medline: 36597421]
8. Brown L, Ilhan E, Pacey V, Hau W, Van Der Kooi V, Dale M. The effect of high-fidelity simulation–based learning in acute cardiorespiratory physical therapy—a mixed-methods systematic review. JOPTE 2021;35(2):146-158. [doi: 10.1097/JTE.0000000000000183]
9. Liu JYW, Yin YH, Kor PPK, et al. The effects of immersive virtual reality applications on enhancing the learning outcomes of undergraduate health care students: systematic review with meta-synthesis. J Med Internet Res 2023 Mar 6;25:e39989. [doi: 10.2196/39989] [Medline: 36877550]
10. Saab MM, Hegarty J, Murphy D, Landers M. Incorporating virtual reality in nurse education: a qualitative study of nursing students' perspectives. Nurse Educ Today 2021 Oct;105:105045. [doi: 10.1016/j.nedt.2021.105045] [Medline: 34245956]
11. Plotzky C, Lindwedel U, Sorber M, et al. Virtual reality simulations in nurse education: a systematic mapping review. Nurse Educ Today 2021 Jun;101:104868. [doi: 10.1016/j.nedt.2021.104868] [Medline: 33798987]

12. Dean S, Halpern J, McAllister M, Lazenby M. Nursing education, virtual reality and empathy? Nurs Open 2020 Nov;7(6):2056-2059. [doi: 10.1002/nop2.551] [Medline: 33072391]
13. Wu D, Lowry PB, Zhang D, Tao Y. Patient trust in physicians matters-understanding the role of a mobile patient education system and patient-physician communication in improving patient adherence behavior: field study. J Med Internet Res 2022 Dec 20;24(12):e42941. [doi: 10.2196/42941] [Medline: 36538351]
14. Wu DC, Zhao X, Wu J. Online physician-patient interaction and patient satisfaction: empirical study of the internet hospital service. J Med Internet Res 2023 Aug 24;25:e39089. [doi: 10.2196/39089] [Medline: 37616031]
15. Yu J, Parsons GS, Lancastle D, Tonkin ET, Ganesh S. "Walking in Their Shoes": the effects of an immersive digital story intervention on empathy in nursing students. Nurs Open 2021 Sep;8(5):2813-2823. [doi: 10.1002/nop2.860] [Medline: 33743185]
16. Calvert J, Abadia R. Impact of immersing university and high school students in educational linear narratives using virtual reality technology. Comput Educ 2020 Dec;159:104005. [doi: 10.1016/j.compedu.2020.104005]
17. Yeo JY, Nam H, Park JI, Han SY. Multidisciplinary design-based multimodal virtual reality simulation in nursing education: mixed methods study. JMIR Med Educ 2024 Jul 26;10:e53106. [doi: 10.2196/53106] [Medline: 39058550]
18. Chan K, Kor PPK, Liu JYW, Cheung K, Lai T, Kwan RYC. The use of immersive virtual reality training for developing nontechnical skills among nursing students: multimethods study. Asian Pac Isl Nurs J 2024 Jul 10;8:e58818. [doi: 10.2196/58818] [Medline: 38986130]
19. Son H, Ross A, Mendoza-Tirado E, Lee LJ. Virtual reality in clinical practice and research: viewpoint on novel applications for nursing. JMIR Nurs 2022 Mar 16;5(1):e34036. [doi: 10.2196/34036] [Medline: 35293870]
20. Buchanan C, Howitt ML, Wilson R, Booth RG, Risling T, Bamford M. Predicted influences of artificial intelligence on nursing education: scoping review. JMIR Nurs 2021;4(1):e23933. [doi: 10.2196/23933] [Medline: 34345794]
21. Lange AK, Koch J, Beck A, et al. Learning with virtual reality in nursing education: qualitative interview study among nursing students using the unified theory of acceptance and use of technology model. JMIR Nurs 2020;3(1):e20249. [doi: 10.2196/20249] [Medline: 34345791]
22. Ma Z, Huang KT, Yao L. Feasibility of a computer role-playing game to promote empathy in nursing students: the role of immersiveness and perspective. Cyberpsychol Behav Soc Netw 2021 Nov;24(11):750-755. [doi: 10.1089/cyber.2020.0371] [Medline: 33989057]
23. Ventura S, Badenes-Ribera L, Herrero R, Cebolla A, Galiana L, Baños R. Virtual reality as a medium to elicit empathy: a meta-analysis. Cyberpsychol Behav Soc Netw 2020 Oct;23(10):667-676. [doi: 10.1089/cyber.2019.0681] [Medline: 32757952]
24. Cummings JJ, Tsay-Vogel M, Cahill TJ, Zhang L. Effects of immersive storytelling on affective, cognitive, and associative empathy: The mediating role of presence. New Media Soc 2022 Sep;24(9):2003-2026. [doi: 10.1177/1461444820986816]
25. Sundar SS, Kang J, Oprean D. Being there in the midst of the story: how immersive journalism affects our perceptions and cognitions. Cyberpsychol Behav Soc Netw 2017 Nov;20(11):672-682. [doi: 10.1089/cyber.2017.0271] [Medline: 29125787]
26. Herrera F, Bailenson J, Weisz E, Ogle E, Zaki J. Building long-term empathy: A large-scale comparison of traditional and virtual reality perspective-taking. PLoS One 2018;13(10):e0204494. [doi: 10.1371/journal.pone.0204494] [Medline: 30332407]
27. Ma Z, Ma R, Chen M, Walter N. Present, empathetic, and persuaded: a meta-analytic comparison of storytelling in high versus low immersive mediated environments. Hum Commun Res 2023 Dec 26;50(1):27-38. [doi: 10.1093/hcr/hqad030]
28. Green MC, Brock TC. The role of transportation in the persuasiveness of public narratives. J Pers Soc Psychol 2000 Nov;79(5):701-721. [doi: 10.1037//0022-3514.79.5.701] [Medline: 11079236]
29. Ma Z, Yang G. Show me a photo of the character: exploring the interaction between text and visuals in narrative persuasion. J Health Commun 2022 Feb 1;27(2):125-133. [doi: 10.1080/10810730.2022.2065387] [Medline: 35422202]
30. Green MC, Tesser A, Wood JV, Stapel DA. Transportation into narrative worlds: implications for the self. In: On Building, Defending and Regulating the Self: A Psychological Perspective: Psychology Press; 2005:53-75. [doi: 10.4324/9780203998052]
31. Ahn SJ(, Bailenson JN, Park D. Short- and long-term effects of embodied experiences in immersive virtual environments on environmental locus of control and behavior. Comput Human Behav 2014 Oct;39:235-245. [doi: 10.1016/j.chb.2014.07.025]
32. Kors MJ, van der Spek ED, Ferri G, Schouten BA. You; the observer, partaker or victim. delineating three perspectives to empathic engagement in persuasive games using immersive technologies. Presented at: CHI PLAY '18 Extended Abstracts: Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in PLAY Companion Extended Abstracts; Oct 28-31, 2018; Melbourne, VIC, Australia. [doi: 10.1145/3270316.3271547]
33. Ma Z, Zytko D. Designing immersive stories for health: choosing character perspective based on the viewer's modality. International Journal of Human–Computer Interaction 2021 Sep 14;37(15):1423-1435. [doi: 10.1080/10447318.2021.1886486]
34. Beverly E, Rigot B, Love C, Love M. Perspectives of 360-degree cinematic virtual reality: interview study among health care professionals. JMIR Med Educ 2022 Apr 29;8(2):e32657. [doi: 10.2196/32657] [Medline: 35486427]
35. Hoeken H, Kolthoff M, Sanders J. Story perspective and character similarity as drivers of identification and narrative persuasion. Hum Commun Res 2016 Apr;42(2):292-311 [FREE Full text] [doi: 10.1111/hcre.12076]

36. Jin SAA. “I feel present. therefore, I experience flow:” a structural equation modeling approach to flow and presence in video games. J Broadcast Electron Media 2011 Feb 25;55(1):114-136. [doi: 10.1080/08838151.2011.546248]
37. Weibel D, Wissmath B. Immersion in computer games: the role of spatial presence and flow. International Journal of Computer Games Technology 2011;2011:1-14. [doi: 10.1155/2011/282345]
38. Ma Z. Effects of immersive stories on prosocial attitudes and willingness to help: testing psychological mechanisms. Media Psychol 2020 Nov 1;23(6):865-890. [doi: 10.1080/15213269.2019.1651655]
39. Liu S, Yang JZ. Incorporating message framing into narrative persuasion to curb e-cigarette use among college students. Risk Anal 2020 Aug;40(8):1677-1690. [doi: 10.1111/risa.13502] [Medline: 32390210]
40. Murphy ST, Frank LB, Moran MB, Patnoe-Woodley P. Involved, transported, or emotional? exploring the determinants of change in knowledge, attitudes, and behavior in entertainment-education. J Commun 2011 Jun;61(3):407-431. [doi: 10.1111/j.1460-2466.2011.01554.x]
41. That dragon, cancer. Numinous Games. 2016. URL: https://www.thatdragoncancer.com/ [accessed 2025-03-18]
42. Chen A, Hanna JJ, Manohar A, Tobia A. Teaching empathy: the implementation of a video game into a psychiatry clerkship curriculum. Acad Psychiatry 2018 Jun;42(3):362-365. [doi: 10.1007/s40596-017-0862-6] [Medline: 29204755]
43. Martingano AJ, Hererra F, Konrath S. Virtual reality improves emotional but not cognitive empathy: A meta-analysis. Technol Mind Behav 2021;2(1). [doi: 10.1037/tmb0000034]
44. Hartmann T, Wirth W, Schramm H, et al. The spatial presence experience scale (SPES). J Media Psychol 2016 Jan;28(1):1-15. [doi: 10.1027/1864-1105/a000137]
45. Knol J, van Linge R. Innovative behaviour: the effect of structural and psychological empowerment on nurses. J Adv Nurs 2009 Feb;65(2):359-370. [doi: 10.1111/j.1365-2648.2008.04876.x] [Medline: 19191936]
46. Batson CD, Fultz J, Schoenrade PA. Distress and empathy: two qualitatively distinct vicarious emotions with different motivational consequences. J Pers 1987 Mar;55(1):19-39. [doi: 10.1111/j.1467-6494.1987.tb00426.x] [Medline: 3572705]
47. Hayes AF. Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach: Guilford Publications; 2017.
48. Vlachogianni P, Tselios N. Perceived usability evaluation of educational technology using the System Usability Scale (SUS): A systematic review. J Res Technol Educ 2022 May 27;54(3):392-409. [doi: 10.1080/15391523.2020.1867938]
49. Gao M, Kortum P, Oswald F. Psychometric evaluation of the USE (usefulness, satisfaction, and ease of USE) questionnaire for reliability and validity. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 2018 Sep;62(1):1414-1418. [doi: 10.1177/1541931218621322]

Abbreviations

ANCOVA: analysis of covariance
VR: virtual reality

Edited by A Hasan Sapci; submitted 02.04.24; peer-reviewed by B Garrett, CV Gomez, S Benbelkacem; revised version received 07.11.24; accepted 02.01.25; published 31.03.25.

Please cite as:
Huang KT, Ma Z, Yao L
Media-Induced and Psychological Factors That Foster Empathy Through Virtual Reality in Nursing Education: 2×2 Between-Subjects Experimental Study
JMIR Med Educ 2025;11:e59083
URL: https://mededu.jmir.org/2025/1/e59083
doi:10.2196/59083

Original Paper

Case-Based Virtual Reality Simulation for Severe Pelvic Trauma Clinical Skill Training in Medical Students: Design and Pilot Study

Peng Teng[1], MM; Youran Xu[2], BS; Kaoliang Qian[3], BS; Ming Lu[4*], MD; Jun Hu[3*], MD

[1]Department of Teaching Resources Management, Teaching Management Office of Nanjing Medical University, Nanjing, China

[2]School of Stomatology, Nanjing Medical University, Nanjing, China

[3]Department of Orthopedics, First Affiliated Hospital of Nanjing Medical University, Nanjing, China

[4]Department of Pharmacology & Jiangsu Key Laboratory of Neurodegeneration, Nanjing Medical University, Nanjing, China

[*]these authors contributed equally

Corresponding Author:
Jun Hu, MD
Department of Orthopedics
First Affiliated Hospital of Nanjing Medical University
Guang Zhou Road 300
Nanjing,
China
Phone: 86 02568303196
Email: junhu89@vip.sina.com

Abstract

Background: Teaching severe pelvic trauma poses a significant challenge in orthopedic surgery education due to the necessity of both clinical reasoning and procedural operational skills for mastery. Traditional methods of instruction, including theoretical teaching and mannequin practice, face limitations due to the complexity, the unpredictability of treatment scenarios, the scarcity of typical cases, and the abstract nature of traditional teaching, all of which impede students' knowledge acquisition.

Objective: This study aims to introduce a novel experimental teaching methodology for severe pelvic trauma, integrating virtual reality (VR) technology as a potent adjunct to existing teaching practices. It evaluates the acceptability, perceived ease of use, and perceived usefulness among users and investigates its impact on knowledge, skills, and confidence in managing severe pelvic trauma before and after engaging with the software.

Methods: A self-designed questionnaire was distributed to 40 students, and qualitative interviews were conducted with 10 teachers to assess the applicability and acceptability. A 1-group pretest-posttest design was used to evaluate learning outcomes across various domains, including diagnosis and treatment, preliminary diagnosis, disease treatment sequencing, emergency management of hemorrhagic shock, and external fixation of pelvic fractures.

Results: A total of 40 students underwent training, with 95% (n=38) affirming that the software effectively simulated real-patient scenarios. All participants (n=40, 100%) reported that completing the simulation necessitated making the same decisions as doctors in real life and found the VR simulation interesting and useful. Teacher interviews revealed that 90% (9/10) recognized the VR simulation's ability to replicate complex clinical cases, resulting in enhanced training effectiveness. Notably, there was a significant improvement in the overall scores for managing hemorrhagic shock (t_{39}=37.6; 95% CI 43.6-48.6; P<.001) and performing external fixation of pelvic fractures (t_{39}=24.1; 95% CI 53.4-63.3; P<.001) from pre- to postsimulation.

Conclusions: The introduced case-based VR simulation of skill-training methodology positively influences medical students' clinical reasoning, operative skills, and self-confidence. It offers an efficient strategy for conserving resources while providing quality education for both educators and learners.

(JMIR Med Educ 2025;11:e59850) doi:10.2196/59850

KEYWORDS

case-based learning; virtual reality; pelvic fracture; severe pelvic trauma; hemodynamic instability; clinical skill training; VR; pelvic trauma; medical student; pilot study; orthopedic surgery; theoretical teaching; acceptability

Introduction

Severe pelvic trauma is characterized by unstable pelvic fractures resulting from high-energy impacts, typically accompanied by complications such as fatal massive bleeding and organ injuries. The mortality rate in China is as high as 10%-30% [1-3]. Managing pelvic fractures with hemodynamic instability poses a significant challenge within the orthopedic surgery discipline [4,5]. The current gap that exists in the field of the effectiveness of severe pelvic trauma clinical skill training is multifaceted. It includes the inaccessibility of clinical teaching in hospitals and the constraints of traditional classroom instruction. In addition, owing to the complexity, unpredictability of treatment locations, the rarity of typical cases, reluctance to cooperate, and ethical concerns surrounding clinical teaching, traditional methods of clinical skill training in diagnosing and treating severe pelvic trauma are limited to theoretical instruction and mannequin-based simulation. These methods face limitations, including disproportionate teacher-student ratios, high model consumption, insufficient training spaces, the absence of comprehensive and immediate feedback, and inadequate training overall. Specifically, on-site mentoring can be challenging to achieve effectively and efficiently. Therefore, a teaching model for severe pelvic trauma needs to integrate new teaching strategies and computer technology to address these issues. This study addresses gaps in the current training methods and tools for managing severe pelvic trauma by designing an innovative virtual reality (VR)–based simulation platform.

In recent years, VR technology has gained widespread acceptance in orthopedic surgery education because of its multisensory immersive experience, the convenience of real-time interaction, and a psychologically secure experimental environment [6,7]. These technologies have mitigated the limitations of time, space, and teaching resources in orthopedic surgery education and addressed issues such as simulated patients and the difficulty of repeated practice. To some extent, this has enhanced the efficiency and quality of clinical practice. Digital patient simulators have become valuable tools in medical education, offering a standardized method of patient simulation [8,9]. However, virtual patients typically only exhibit clinical symptoms and signs, with minimal explanation of the underlying fundamental medical knowledge of primary symptoms and positive or negative signs. Furthermore, there is a scarcity of research on virtual simulation teaching models that integrate basic and clinical medicine for severe pelvic trauma. To mimic the real teaching scenario of the disease and capture the sudden and variable condition of patients in clinical settings, we developed an electronic standardized patient (ESP) for severe pelvic trauma, capable of replicating both macroscopic changes, such as monitor display data, and microscopic changes, including alterations in blood circulation, organs, tissues, and cells. The ESP model facilitates the integration and application of clinical and basic knowledge. The original design and technology of the ESP were conceived by Professor Xing Ya Gao's team at Nanjing Medical University, Department of Physiology [10]. An ESP is grounded in contemporary theories of human systems physiology and incorporates relevant clinical literature and data through analog circuits, physics, and other methods to formulate a mathematical model. Artificial intelligence and data analytics are used to refine and adjust the simulation data. In summary, the ESP represents a web-based intelligent standardized patient, enabling students to interact with it in a virtual hospital setting from a physician's perspective. To date, the ESP has not been fully used in the clinical skills education for severe pelvic trauma.

Case-based learning (CBL) is recognized as an effective teaching strategy in clinical skills training [11,12]. CBL is particularly valued in orthopedic surgery education for its ability to enhance learner participation, foster active learning, and develop critical thinking and problem-solving skills [13]. CBL is an educational method designed to analyze medical records, recreate real clinical scenarios, and engage learners in addressing actual clinical challenges, thereby stimulating their curiosity and promoting active learning [14].

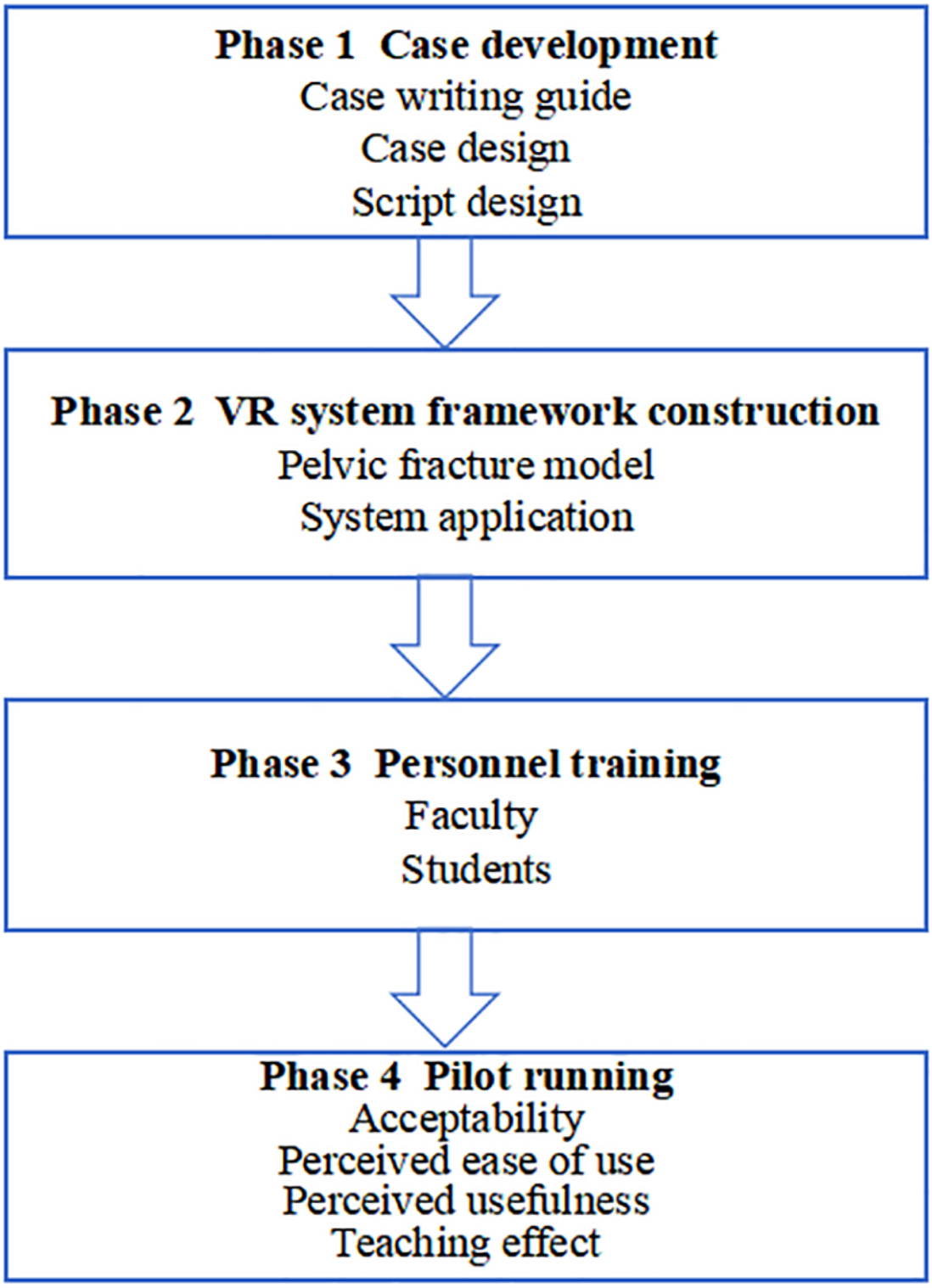
CBL, combined with VR simulation technology, has been successfully applied in midwifery laboratory courses, with its effectiveness widely acknowledged by students [15]. However, the application of these combined methodologies in teaching clinical skills for severe pelvic trauma has yet to be explored. Therefore, in this study, we introduced a case-based digital skill training program for severe pelvic trauma and conducted mixed methods to evaluate its acceptance among users. Additionally, we implemented a pretest-posttest design to investigate the potential impact of this clinical skills training on undergraduate and graduate students. We hypothesize that the case-based VR simulation teaching method will effectively complement current training practices for severe pelvic trauma, enhancing knowledge, procedural skills, and confidence, while also improving instructional efficiency and effectiveness for educators.

Methods

Study Design

We conducted a 4-phase study (Figure 1). First, we created the simulated teaching case base and adapted typical clinical cases based on the case writing guide [16,17]; subsequently, the case script was developed. Next, we established the framework of the VR system, comprising 3 ports and 3 system modules. To facilitate the effective implementation of the teaching system, we formulated a comprehensive training plan for department administrators, course instructors, and medical students. A pilot test of the system was conducted on a limited scale, followed by a 1-group pretest-posttest design to assess its acceptability, potential application, and existing limitations.

Figure 1. A flow diagram illustrating the steps involved in developing a case-based VR simulation for severe pelvic trauma clinical skill training. VR: virtual reality.



Phase 1: Case Development

Case Design

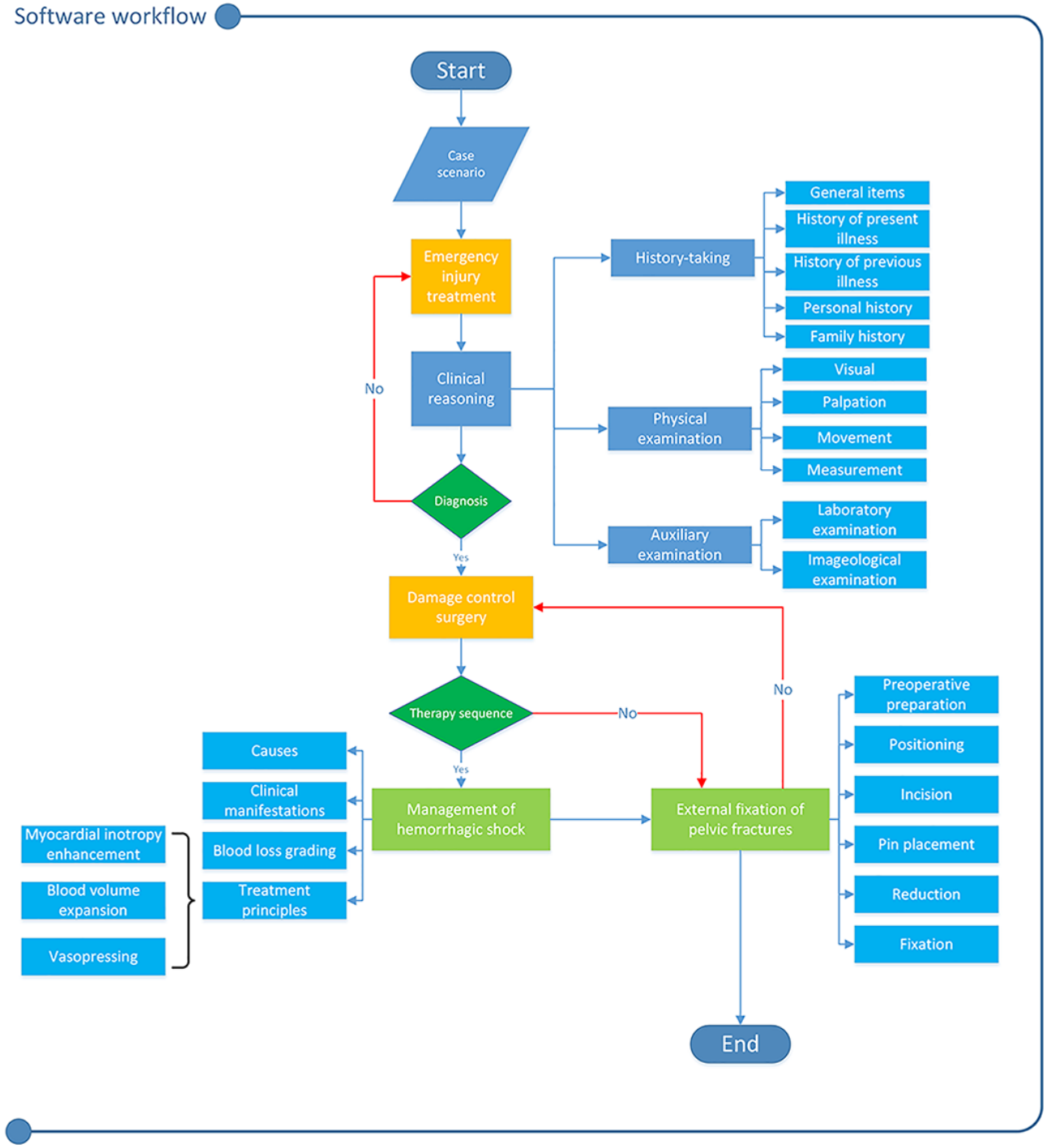
Specialists in basic and clinical medicine, drawing on a literature review, real patient cases, and the course syllabus, designed the cases. A representative case involved a worker who fell from a high platform and was sequentially evaluated in the emergency department, admitted to the intensive care unit, and taken to the operating room as his condition worsened. The diagnostic and treatment processes for this condition were standardized. Learners were guided through different scenarios to acquire both declarative and procedural knowledge for managing patients with severe pelvic trauma. We developed 3 initial cases, each representing a different stage of trauma and treatment approach (Table 1). Learning paths for diagnosis and treatment were outlined in a flowchart (Figure 2), based on clinical

treatment guidelines [18,19] to serve as a framework for evaluating student performance. We collected clinical manifestations, relevant imaging, and laboratory examination results of real patients.

Table 1. Three cases with patients diagnosed with pelvic trauma.

	Case 1	Case 2	Case 3
Diagnosis	Pelvic fracture	Pelvic fracture with mild hemorrhagic shock	Pelvic fracture with severe hemorrhagic shock
Injury evaluation	Mild	Moderate	Severe
Diagnostics	Standard	Plus computed tomography of the pelvis	Plus computed tomography of the pelvis and abdomen
Therapy	Surgery	Antihemorrhagic shock therapy, surgery	Antihemorrhagic shock therapy, surgery

Figure 2. An optimal template for the diagnosis and treatment process of severe pelvic trauma (blue arrows for the correct path, red arrows for the wrong path).



The 3 cases varied in the severity of pelvic trauma, necessitating different diagnostic and treatment strategies.

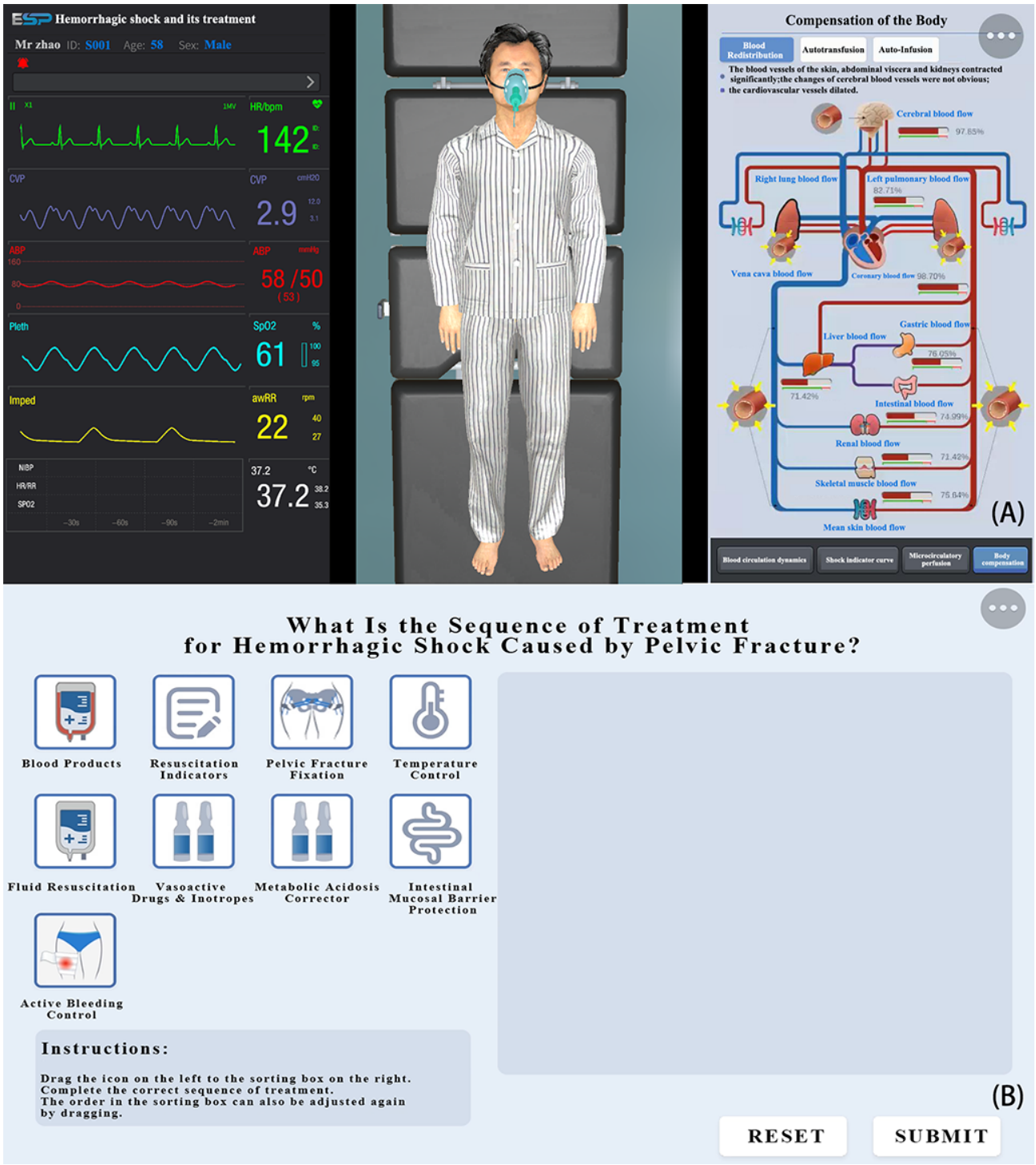
Script Design

Scripts focused on scenarios in emergency rooms, intensive care units, and operating rooms. In a virtual emergency room, users could interact with the ESP, using various diagnostic and treatment methods, such as history taking, physical examination, and auxiliary examination, to formulate a preliminary diagnosis based on evidence gathered during the process. Learners engaged with the ESP through text input or voice chat during the history-taking session. They then entered pertinent information into the system's modules for general items, history of present illness, personal history, marital history, and family history. The system evaluated 4 aspects: completeness of the inquiry framework, logical order of inquiries, communication skills, and awareness of humanistic care, assigning scores based on the accuracy of the information collected. The system also recorded the total duration of inquiries and the time spent on each module for later analysis. Physical examination, a fundamental skill for diagnosing diseases, involves comparing and identifying positive signs. The system assessed the patient examination position, completeness and sequence of operational steps, standardization of procedures, comprehensiveness of examination content, accurate reporting of results, and basic professional quality. Auxiliary examination, crucial for diagnosis and treatment planning, involves evaluating laboratory tests and computed tomography images. Upon completing these steps, students made a preliminary diagnosis and a prioritized list of differential diagnoses. Correct diagnoses led to immediate treatment initiation; otherwise, students continued trial and error in this module (Figure 3).

Figure 3. Screenshot of the emergency room. (A) The user acts as a doctor in the emergency room while interacting with the ESP by selecting different diagnoses and treatment icons. (B) Trainees can choose the present history module in history taking to communicate with the ESP. (C) During the physical examination session, students can use a tape measure to measure the distance between the bilateral anterior superior iliac spine and the xiphoid process to determine whether the pelvis was displaced. (D) The students order a computed tomography scan to evaluate the location and severity of the ESP's injury. ESP: electronic standardized patient. Please note that a higher resolution version of this image can be found in Multimedia Appendix 1.

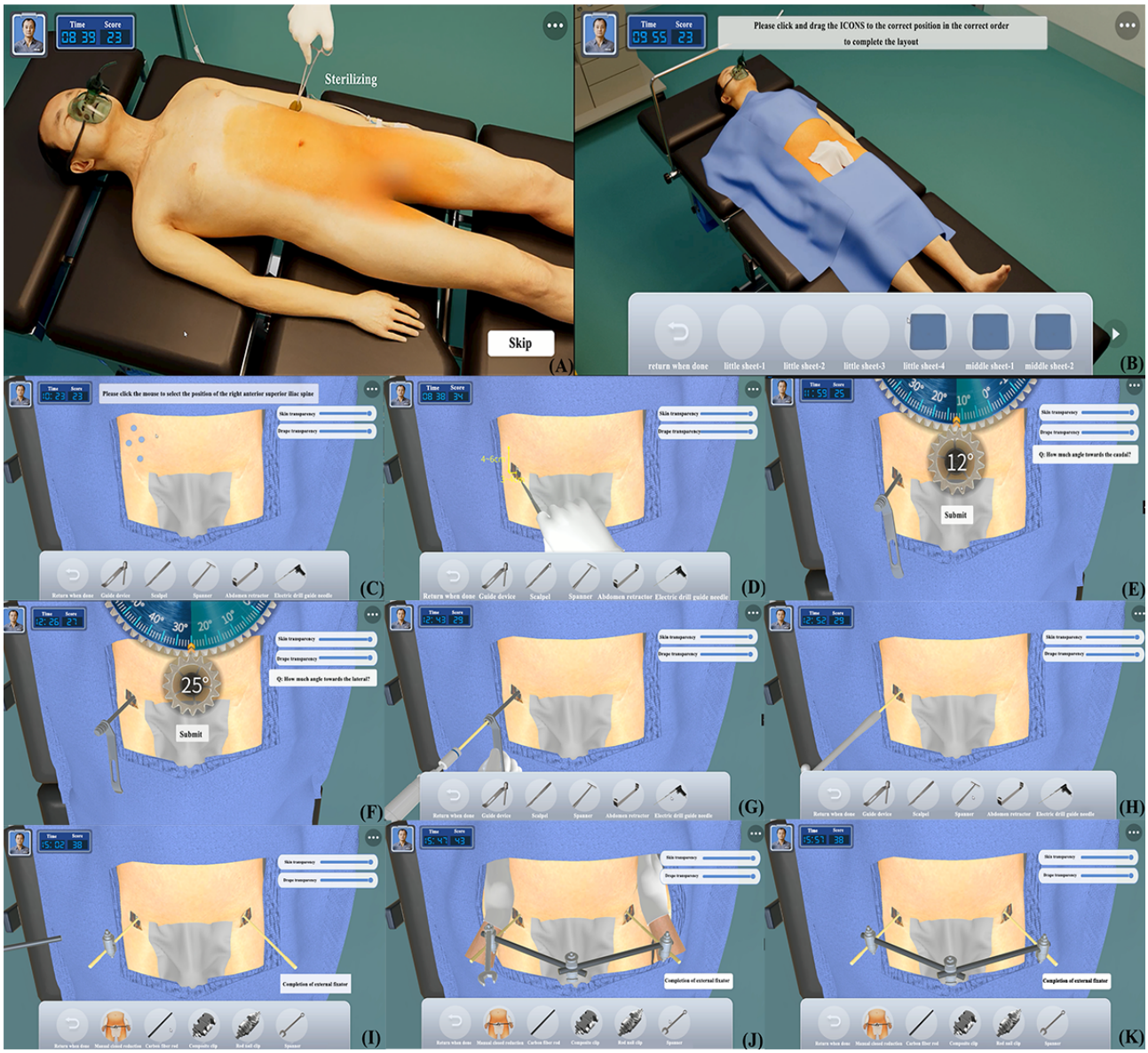
Upon initiating appropriate immediate treatment measures, the ESP was transferred to the intensive care unit. Vital sign monitors reflected real-time changes based on the condition and treatment progression. In this module, users learned about the causes, clinical manifestations, severity classifications, and management of hemorrhagic shock (Figure 4).

Figure 4. Screenshot of hemorrhagic shock and its treatment. (A) Monitors in the intensive care unit show decreased blood pressure, increased heart rate, and subnormal central venous pressure and arterial oxygen saturation in the ESP. By users' observation, the ESP at this time showed pallor, decreased urine output, and cold sweats. Users can also learn about the compensatory mechanisms of ESP in terms of blood redistribution, autotransfusion, and auto-infusion at the system, organ, and tissue levels after hemorrhagic shock. Screenshot of treatment principles for severe pelvic trauma. (B) In this module, the user should drag the corresponding icons from all the treatment measures given on the left to the right blank according to the treatment principle. When the dragged content matches the built-in answer of the system, the next section will be entered. ESP: electronic standardized patient. Please note that a higher resolution version of this image can be found in Multimedia Appendix 1.



During the surgical procedure for external fixation of a pelvic fracture, users first completed preparatory tasks such as handwashing and donning surgical attire. Then, the ESP underwent anesthesia, sterilization, positioning, drilling, and bracket installation according to surgical standards (Figure 5).

Figure 5. Screenshot of the surgical procedure of external fixation of pelvic fractures. (A) According to the sequence of operation, the users need to complete disinfection, (B) laying, (C) positioning, (D) incising, (E,F) positioning and protection of the guide, (G) electric insertion of screw, (H) manual insertion of screw, (I) use of rod-nail clamp and composite clamps, and (J,K) adjustment of the external fixator after manual reduction. Please note that a higher resolution version of this image can be found in Multimedia Appendix 1.



Phase 2: Framework Construction

Construction of Pelvic Fracture Model

To provide learners with a comprehensive understanding of pelvic anatomy and fracture morphology, we collaborated with software engineers to develop the relevant content. The development process involved several steps. Initially, 3D modeling data of pelvic fractures were extracted from real clinical cases, and a preliminary geometry was constructed using the Maya (Autodesk) development tool to complete the foundational model of pelvic fractures. Subsequently, the ZBrush (Pixologic) tool was used to sculpt and refine the fracture site shape, fracture line position, and bone surface structure of the basic pelvic fracture model, minimizing the detailed morphology of the pelvic fracture. Next, we used the support model subdivision level adjustment feature to enhance the pelvic fracture model's subdivision level, showcasing the complex structure of the fracture site. The addition of materials and textures in Maya, such as bone texture and skin color and texture, further improved the visual realism. The established skeletal system was then integrated into the pelvic fracture model, and animations were created based on real pelvic fracture case data, including the degree of fracture displacement and the relative position of fracture blocks. Finally, the pelvic fracture model was exported from Maya to Unity (Unity Technologies) for digital and real-time rendering, allowing learners to interact with a pelvic fracture model in a virtual environment to simulate actual surgical procedures. The development and construction process of the pelvic fracture model is elaborated in Multimedia Appendix 2.

Construction of System Application

The digital simulation system was segmented into 3 ports: the department administrator, course leader, and user ports, constructed on a website platform. The department's primary role is to create the appropriate courses in the system, either

compulsory or optional. Once the courses are established, they should be linked to the digital simulation skills training system, and relevant instructors assigned to teach all enrolled students for the semester. Course instructors are responsible for preparing preview materials, learning videos, self-assessment questions, and postclass surveys. They also review experimental outcomes and evaluate their performance within the system. Users can engage with the digital simulation system for severe pelvic trauma by enrolling in or attending a scheduled course. Upon system login, the interface offers 2 modes: training and assessment. Due to the digital simulation experiment teaching integration with the virtual simulation experiment teaching sharing platform of Nanjing Medical University, campus users are not required to register or authenticate before use, thanks to unified identity verification and experimental data integration.

Phase 3: Personnel Training

Faculty Training

Faculty training targeted course directors and instructors from various disciplines within basic and clinical medicine, aiming to familiarize them with the digital simulation experimental teaching system. They received training on case materials, experimental teaching objectives, principles, teaching processes and methods, steps, outcomes, and conclusions. Additionally, they learned to address technical system issues, respond to student inquiries, and interact with students on the platform during experiments.

Student Training

Before software utilization, students were informed about the operating system and hardware configuration requirements. The application runs on a Windows 7 64-bit or higher PC, equipped with a 3.60 GHz Intel i5 processor, 8 GB RAM, NVidia GTX 2060 graphic card, and a 1920×1080 display resolution. The application supports various browsers on different operating systems, such as Google Chrome, the 360 browser, and Firefox. After accessing a specific URL, users must install MengooLauncher, requiring less than 100 MB of plug-in capacity, as indicated. Before proceeding to the autonomous training or assessment interface, users familiarize themselves with the experimental teaching objectives and principles through introductory and instructional videos.

Phase 4: Pilot Running Evaluation of Digital Simulation Software

Design

A self-controlled teaching comparison study was conducted at Nanjing Medical University, China, from October 2023 to January 2024, to examine the impact on knowledge, skills, and confidence before and after using virtual simulation experimental teaching software. All participants underwent a knowledge assessment of equal difficulty before and after system engagement.

In the initial design phase, the teaching and research teams sought input from the software development team and feedback from various users through internal reviews. A case-based VR simulation of severe pelvic trauma was tested by students majoring in clinical medicine, clinical teachers, and basic medicine instructors. We distributed the ESP digital simulation teaching web link via WeChat (Tencent) to pertinent users, soliciting face-to-face or written feedback on case and script design and software development, including aspects related to clinical and basic medicine education. Additionally, the teaching research group reviewed classical cases of severe pelvic trauma and questionnaire responses.

Sampling and Recruitment

The recruitment criteria were as follows: (1) students must have completed courses in diagnostics, internal medicine, and surgery; (2) they needed to have a laptop for the study; (3) they should not have participated in any form of digital simulation software training for clinical skills prior; and (4) they agreed to participate in the pilot study and signed an informed consent form. We invited a purposive sample of 20 fourth-year undergraduates and 20 first-year graduate students majoring in clinical medicine from the First Clinical Medical College of Nanjing Medical University to test the case-based VR simulation software. Based on sample size requirements previously reported in the literature for evaluating data collection materials, a minimum of 10 samples is necessary to ensure the adequacy and validity of the assessment instrument [20]. The evaluation sought feedback from a diverse group of users, including undergraduate and graduate students, as well as teachers with various professional titles. The qualitative study used a representative population most familiar with the study topic, comprising 5 orthopedic teaching teachers and 5 basic medicine teachers.

Data Collection

Participants were required to complete the training and assessment using the digital simulation software for severe pelvic trauma treatment. Our data collection involved a repeated measurement approach to assess knowledge test scores before and after the simulation. Feedback on the simulation teaching tools was collected through a single questionnaire. Participation in the survey was entirely voluntary, and students were informed that their decision to participate would not impact their academic standing. The survey distribution was conducted independently, with no direct or known ties between the distributors and the students, ensuring an unbiased and pressure-free environment for participants. Participants were recruited and invited to complete the survey through the WeChat platform using the Wenjuanxing applet. The survey was administered anonymously to encourage honest feedback on their experience with the VR simulation software.

Outcome Assessment

Although evaluation questionnaires are commonly used to compare learning tools, there is a dearth of validated tools for assessing the ESP digital simulation software as a learning instrument. Consequently, the questionnaire was adapted from a validated assessment tool in educational literature, offering a resource for future research on the perception in clinical medical professional education. The questionnaire comprised 15 Likert-scale statements (1=strongly disagree to 5=strongly agree), assessing accessibility and usability. The teaching team and subject experts reviewed the questionnaire. Cronbach α for the questionnaire was 0.85 (n=15). The questionnaire was

adapted from a validated assessment tool widely used in educational literature for clinical medical professional education [21]. The instrument used for assessing system acceptability was based on a modified version of the Technology Acceptance Model questionnaire, with validation provided by Balki et al [22]. The Cronbach α coefficient, which reflects internal consistency, was calculated jointly for both the usability and acceptability surveys, yielding a consistent value for both aspects.

Data Analysis

Descriptive analysis was applied to the quantitative data obtained from the Likert scale. For qualitative data, which included responses to 7 open-ended questions, we used a validated content analysis method as described by Elo and Kyngäs [23]. All participant comments were transcribed and imported into Excel (Microsoft Corp) for coding. The content analysis process consists of several steps: (1) familiarizing oneself with the data and the hermeneutic spiral, (2) dividing up the text into meaning units and subsequently condensing those meaning units, (3) formulating codes, and (4) developing categories and themes [24]. Initially, the primary investigator analyzed the content, and the research team subsequently reviewed and discussed the codes to achieve consensus. In cases of disagreement, group discussions were held, and, if necessary, a third-party opinion was sought to ensure triangulation and enhance reliability. The 7 open-ended survey questions are provided in Multimedia Appendix 3 for reference.

For user acceptance analysis of the ESP platform among undergraduates, graduates, and tutors, descriptive statistics were used. To compare the mean rating scales of each survey item between groups, an independent 2-tailed *t* test was performed with a significance threshold of $P<.05$. Prior to conducting parametric tests, we verified the assumption of normality using the Shapiro-Wilk test, confirming that the data met the requirements for a parametric approach. This method was chosen over nonparametric tests due to the normal distribution of the data, making it suitable for our sample size and study design.

Ethical Considerations

The Nanjing Medical University ethics committee approved this study (2023418). During the informed consent process, participants were made aware that no incentives were provided for participation in the survey. All methods were implemented in accordance with the Helsinki declaration. All participants were voluntary in the study.

Results

Demographic Results

Of the 56 students enrolled in the optional course on the integrated case of severe pelvic trauma in October 2023, 40 students consented to participate in the pilot study. Among these participants, 50% (n=20) of the students were senior-year undergraduates, 50% (n=20) of the students were first-year graduates (n=20), 45% (n=18) of students were men, and 55% (n=22) of the students were women. The mean age was 22.9 (SD 1.3) years. Among the undergraduate participants (n=20), there were 10 male and 10 female students, with a mean age of 21.9 (SD 0.9) years. For the graduate participants (n=20), there were 8 male and 12 female students, with a mean age of 24.0 (SD 0.8) years. Ten faculty members with at least 5 years of teaching experience in orthopedic surgery or basic medicine were also invited to participate. Neither the students nor the faculty had prior experience with this type of digital simulation platform. All participants were required to complete the questionnaire shortly after finishing the training tasks.

Questionnaire Results

A 5-point Likert scale assessed perceptions of the acceptability, effectiveness, and applicability, summarized in Table 2.

Table 2. Average rating scores of survey questions (Q1-Q15) by all students.

Survey questions	Average rating scores, mean (SD)	
	Undergraduate students	Graduate students
(1) The digital software provides a simulation of a real patient	4.40 (0.66)[a]	4.80 (0.40)
(2) During the simulation, I felt like a doctor taking care of this patient	3.90 (0.99)	3.90 (1.14)
(3) When I finished the simulation, I felt I had to make the same decisions as doctors in real life	4.55 (0.50)	4.80 (0.40)
(4) The VR[b] simulation is interesting and useful	4.85 (0.36)	4.70 (0.46)
(5) The difficulty of the VR simulation is appropriate to my own level of knowledge and skills	4.10 (0.94)[a]	4.60 (0.49)
(6) The feedback from the system adequately reflected my actual performance	4.80 (0.40)	4.60 (0.66)
(7) The goals of scenario simulation are clear and easy to understand	4.55 (0.59)	4.80 (0.40)
(8) I can access the system anytime and anywhere for simulation training	4.75 (0.54)	4.90 (0.30)
(9) The VR simulation can help me to use basic medical knowledge to explain clinical manifestations of clinical reasoning skills	4.80 (0.40)[c]	4.20 (0.87)
(10) The ESP[d] simulator can help me develop clinical operation skills	4.30 (0.78)	4.60 (0.49)
(11) I feel more confident about working with hospital colleagues	4.40 (0.80)	4.70 (0.46)
(12) The VR simulation increased my confidence as a practicing physician	4.40 (0.66)	4.60 (0.49)
(13) The VR simulation can support courses and exams	4.60 (0.49)	4.75 (0.43)
(14) Compared with traditional teaching practice training methods, VR simulation can reduce my training cost and risk	4.90 (0.30)[c]	4.50 (0.50)
(15) In general, this VR simulation training should enhance my learning	4.80 (0.40)	4.80 (0.40)

[a]$P<.05$ compared to the graduate student group.

[b]VR: virtual reality.

[c]$P<.01$ compared with the graduate student group.

[d]ESP: electronic standardized patient.

The respondents showed strong agreement; 95% (n=38) agreed or strongly agreed that the interactive software simulated a real patient scenario (Q1 in Figures 6 and 7). However, 68% (n=27) agreed or strongly agreed that “During the simulation, I felt like a doctor caring for this patient” (Q2), with 18% (n=7) neutral and 15% (n=6) disagreeing. All students (n=40, 100%) felt they had to make real-life doctor decisions by the end of the simulation (Q3) and found the VR simulation interesting and useful (Q4).

Figure 6. Acceptability, effectiveness, and applicability of the case-based VR software by undergraduate students. ESP: electronic standardized patient; VR: virtual reality. Please note that a higher resolution version of this image can be found in Multimedia Appendix 1.

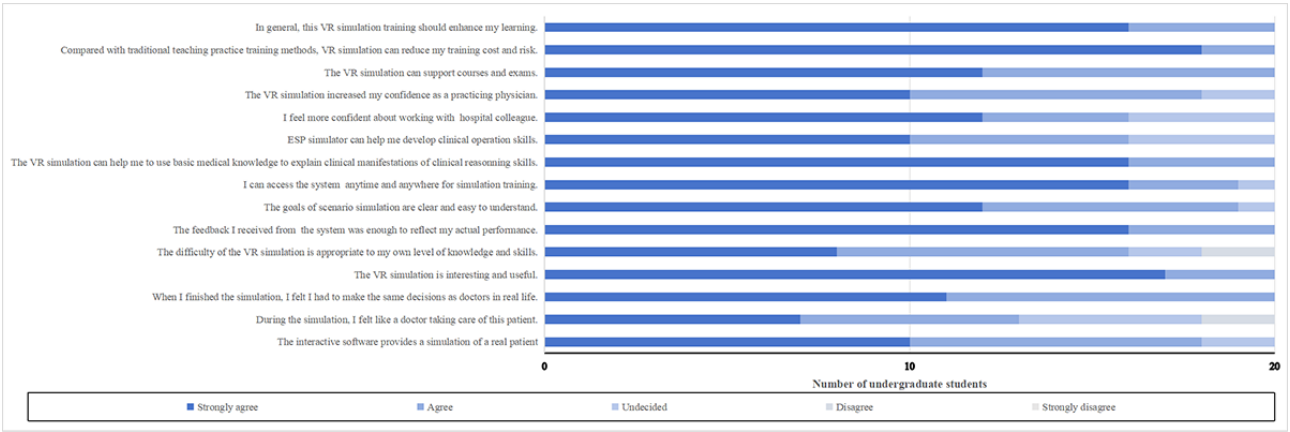
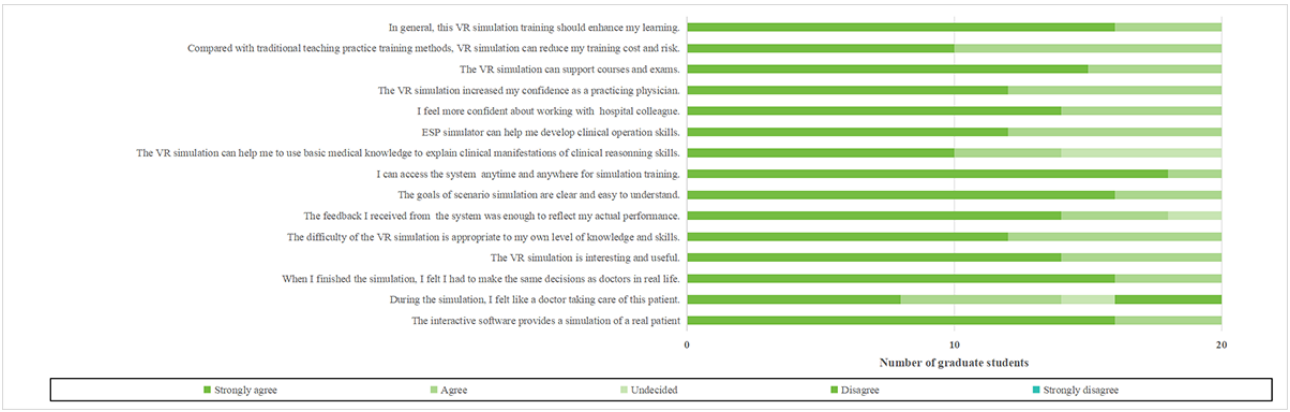


Figure 7. Acceptability, effectiveness, and applicability of the case-based VR software by graduate students. ESP: electronic standardized patient; VR: virtual reality. Please note that a higher resolution version of this image can be found in Multimedia Appendix 1.



Additionally, 90% (n=36) believed the VR simulation's difficulty was appropriate for their knowledge and skills (Q5), while 5% (n=2) disagreed. Moreover, 95% (n=38) reported that the feedback from the system sufficiently reflected their performance (Q6). Most students (95%, n=38) understood the goals of the scenario simulation clearly (Q7). Nearly all students (98%, n=39) could access the system anytime for training (Q8), and 98% (n=39) agreed that "The VR simulation can help apply basic medical knowledge to clinical reasoning skills" (Q9).

When inquired if the ESP simulator aided in developing clinical operational skills (Q10), 85% (n=34) agreed. Regarding confidence in collaborating with hospital colleagues (Q11) and functioning as practicing physicians (Q12), 90% (n=36) agreed or strongly agreed. All students (n=40) concurred that the VR simulation supports courses and exams (Q13), is cost-effective compared to traditional training (Q14), and enhances learning overall (Q15).

Impact of Training Level on Questionnaire Answers

Finally, a 2-tailed *t* test was used to compare the average rating scales between undergraduate and graduate students. This analysis aimed to determine if the responses varied according to their academic level. Specifically, for Questions 1 and 5, undergraduate students exhibited significantly stronger disagreement than their graduate counterparts, as indicated by the *P* values (Q1, *P*=.03; Q5, *P*=.047). Moreover, when compared to graduate students, a larger proportion of undergraduates believed that VR simulation could enhance their clinical reasoning abilities (Q9, *P*=.009) and decrease their training costs and associated risks (Q14, *P*=.004). Additionally, all participants indicated a low level of agreement with the statement "During the simulation, I felt like a doctor taking care of this patient" (Q2, *P*=.99). No significant differences were observed in responses to the remaining questions when analyzed based on academic level.

Qualitative Analysis

Overview

The members of the research team conducted a 1-to-1 structured interview with the participating teachers around the interview outline of 7 open questions formulated in advance, as shown in Table 3.

Table 3. Themes of teacher groups' interview on the application and research of digital simulation teaching curriculum system.

Theme	Teacher
Perceived benefits of systematic teaching	"As one of the most complex and urgent diseases in orthopedics, severe pelvic trauma often fails to receive on-site teaching from teachers in the tense emergency treatment site. In addition, due to the long treatment period of this disease, it takes a long time to fully learn the diagnosis and treatment process of this disease. However, in real life, learners only spend limited time rotating with one department and cannot follow through the entire disease process and treatment course. By creating typical cases of severe pelvic trauma and constructing a virtual clinical diagnosis and treatment environment based on ESP, this system shortens the learning cycle of students, enables learners to experience different treatment settings, allows a large scale concurrent online participation breaking through the limitations of traditional teaching in time and space and improving the efficiency of teaching organizations."
The appropriateness of case application subjects	"It helps me conduct classified teaching according to the basic knowledge level of undergraduates and postgraduates. The basic medical and clinical medical knowledge involved in the disease set in the system is suitable for students at different undergraduate and postgraduate levels to learn. At the same time, the extensibility of the system enriches the flexibility and innovation of students' training and assessment."
The extendibility of course application	"The teaching design of this case is very suitable for the objective structured clinical examination scenario, which is closer to reality than traditional simulation training scenario. In addition, it introduces ESP, without the need for on-site re-placement of exam environments and standardized patient training."
The limitations of systematic research	"In the early stage of communication and interaction with the ESP speech inquiry, it was found that the ESP lacked a large sample of language training model library, so it could not recognize the semantics of the trainer. In addition, the ESP has not achieved the language style characteristics of different types of characters at this stage."
Recommendations for enhancing systematic research	"To enhance the virtual ESP simulation system, five improvements were suggested: enlarging the ESP case database, incorporating a feature for automatic and manual responses to technical queries, broadening the range of disease diagnosis and treatment simulations, expanding the ESP history collection database, and advancing the ESP's artificial intelligence for inquiry processing."

Theme 1: Perceived Benefits of Systematic Teaching

Participants noted that the digital simulation experimental teaching system for severe pelvic trauma significantly improved teaching efficiency and effectiveness, overcoming the traditional teaching constraints related to time and space.

Theme 2: The Appropriateness of Case Application Subjects

Most participants pointed out the variable difficulty of teaching cases within the system for different learning groups, highlighting the advanced design. The freedom for students to interact with the ESP in the system underscores its innovative development. The immersive simulation for diagnosis, treatment training, and assessment allowed students to thoroughly apply their theoretical knowledge and skills, presenting a notable challenge.

Theme 3: The Extendibility of Course Application

The majority of participants regarded case-based digital simulation systems as potent educational tools for both undergraduate and graduate training. A substantial number of participants viewed the system as suitable for integration into an objective structured clinical examination.

Theme 4: The Limitations of Systematic Research

Some limitations of the digital simulation software were reported by participants, particularly issues with the ESP not always accurately recognizing the semantics and tone of the inquiries.

Theme 5: Recommendations for Enhancing Systematic Research

A large case base, different training paths, and smarter ESP interaction can enhance the freshness, challenge, and realism of the ESP experience for the trainers.

Theoretical Knowledge Level of Severe Pelvic Trauma

A comparison of theoretical examination scores before and after participants used the digital simulation software for severe pelvic trauma showed significant improvements in their overall scores for diagnosing and treating the condition, making preliminary diagnoses, the sequence of disease treatment, emergency management of hemorrhagic shock, and performing external fixation of pelvic fractures (Table 4). The IQR box plots for the theoretical knowledge levels of severe pelvic trauma, both pretest and posttest, are provided in Multimedia Appendix 4.

Table 4. Mean scores at presimulation and postsimulation for the 5 uncoached assignments.

Severe pelvic trauma clinical skill training	Presimulation score, % (SEM)	Postsimulation score, % (SEM)	Mean difference[a] (95% CI)	*t* test (df=39)	Cohen *d*	*P* value[b]
Order of diagnosis and treatment	49.9 (2.0)	85.5 (1.4)	35.5 (32.7-38.3)	25.9	3.4	.001
Make a preliminary diagnosis	45.1 (1.4)	89.4 (1.0)	44.4 (42.2-46.5)	41.8	6.0	.001
Order of disease treatment	69.4 (1.8)	95.2 (0.5)	25.8 (22.2-29.5)	14.3	3.2	.001
Emergency treatment of hemorrhagic shock	39.7 (0.9)	85.8 (0.8)	46.1 (43.6-48.6)	37.6	8.5	.001
External fixation operation of pelvic fracture	32.7 (2.3)	91.1 (0.7)	58.4 (53.4-63.3)	24.1	5.3	.001

[a]The analysis included only paired data. The mean difference is the difference in mean presimulation score and mean postsimulation score.

[b]*P* value obtained from a paired 2-tailed *t* test.

Discussion

Principal Findings

This study yielded 3 primary findings. First, we developed a case-based digital simulation teaching system for severe pelvic trauma, incorporating principles of basic and clinical medicine. In contrast to traditional training methods, the VR system allows students to engage in repeated practice at their own pace, providing immediate and standardized feedback after each interaction. This feature overcomes challenges like high teacher-student ratios and insufficient feedback, which are common in traditional training environments. Furthermore, the use of a computer model to demonstrate physiological hemodynamic changes has been shown to aid in understanding the connection between clinical phenomena and underlying knowledge. Second, the software's acceptability, perceived ease of use, and perceived usefulness were highly regarded by users. Finally, the application of this digital simulation teaching system resulted in a significant improvement in all participating knowledge and skill scores. These findings contribute to the innovation in severe pelvic trauma skills training and may offer guidance for the development of enhanced training strategies and the revision of orthopedic surgery training standards.

Comparison to Prior Work

Although severe pelvic trauma is relatively rare in China, our hospital, being an orthopedic center of excellence, sees a higher incidence, treating over 100 patients annually and performing more than 20 external pelvic fixation procedures. The design of our case-based VR simulation curriculum for severe pelvic trauma draws from real cases, expert consensus, and literature reviews. Although previous research has demonstrated the efficacy of integrated learning [25], simulation training [26,27], traditional CBL [13], online learning [28], and digital patient simulator-assisted learning [29] in orthopedic clinical skills training, few studies have combined these methodologies. To our knowledge, this research is the inaugural study to amalgamate these proven effective training methods to enhance severe pelvic trauma clinical skill training, using a hybrid approach to assess the digital simulation efficacy.

Participants' acceptance of this clinical skills training was evident in several areas. Most participants felt the simulation training provided a compelling immersion experience, was accessible at any time and location, and had clear and understandable case scenario goals. Previous studies indicate that digital simulation software can significantly impact learning success [30]. Moreover, the degree of immersion is crucial in VR software, as identification with a digital character directly influences learning motivation and effectiveness [31]. Concerning the utility, all students concurred that the self-directed exploration learning method facilitated a deeper understanding of the knowledge and skills necessary for treating severe pelvic trauma and bolstered their confidence in handling similar conditions in real-life scenarios.

However, acceptance of clinical skill training was lower in certain aspects. A minority of trainees felt that the digital simulation technology's construction of the clinical environment and ESP allowed for experimentation within a safe psychological space. Studies suggest that digital simulators are effective for training doctor-patient communication skills [32]. Nevertheless, the discrepancy between virtual scenarios and real-life situations led to challenges in caring for actual patients. In our pilot study, most learners reported difficulty in direct communication with the ESP in the virtual environment, including eye and body language, and in discerning the nuances of the real language environment (such as tone and intonation); hence, they did not fully practice effective doctor-patient communication skills. Future educational efforts in all hospital departments should prioritize teaching doctor-patient communication skills to students.

Differences in the efficacy of case-based VR in pelvic trauma clinical skills training were observed between undergraduate and graduate students. Undergraduate respondents felt that after undergoing training with the digital simulation system, they solidified their basic medical knowledge and mastered the diagnostic and treatment processes for severe pelvic trauma; however, they expressed a lack of confidence in performing pelvic fracture external fixation. Graduate respondents believed that systematic training deepened their understanding of the diagnosis, treatment, and operational procedures for antihemorrhagic shock therapy and pelvic fracture external fixation. These variances are attributable to their respective stages of learning: undergraduates possess a stronger foundation in basic medical knowledge, while graduates have more opportunities to apply clinical knowledge in practice. Furthermore, undergraduate students outperformed graduate students in retaining disease-related knowledge due to their firmer grasp of basic medical principles. Conversely, their skills

were slightly inferior to those of graduate students, a disparity linked to the latter's greater internship experience and the number of surgical procedures conducted in the hospital. Thus, the training focus should be tailored to each need during severe pelvic trauma clinical skill training.

Quantitative outcome analyses revealed an overall improvement in pass rates at crucial assessment points posttraining with the digital simulation system, with external fixation of pelvic fractures displaying the most significant enhancement. Participants identified the realism and interactivity of the pelvic fracture model within the virtual environment as pivotal in elevating their learning experiences and assessment scores. Despite this, scores for external fixation of pelvic fractures were not the highest due to the inherent complexity and the necessity for interns to rotate through the orthopedic department to gain familiarity with patient care and the surgical technique [33]. Although the scores for diagnosing and treating the disease were the lowest presimulation, participants' scores in these 2 categories were the highest.

Personal interviews confirmed that teaching software facilitates large-scale online student learning in terms of ease and effectiveness. Tailoring cases to different learner groups introduced high levels of order, innovation, and challenge. The experiment also addressed challenges such as prolonged real-world teaching durations, access to actual patients, and teaching environment constraints. However, 1 instructor cautioned that this innovative skill training should complement, rather than replace, traditional teaching methods, a sentiment echoed by other educational research [21,34].

Implications

The digital simulation software for severe pelvic trauma provides undergraduates with an immersive learning experience that bridges theoretical knowledge and practical skills. For graduate students, it offers targeted preclinical training, preparing them for real-world trauma care. This approach enhances skill acquisition and promotes standardized training, potentially improving patient outcomes in severe trauma cases.

Limitations

Limitations include the inability to directly compare the digital simulation teaching system for severe pelvic trauma with traditional teaching models. Moreover, being self-controlled, participants' preexisting knowledge about the digital simulation system may have biased the outcomes. Although the participant count was sufficient for statistical analysis in this pilot study, the sample size remains limited.

Conclusions

Case-based VR simulation of skill training is an effective educational approach for medical students learning about severe pelvic trauma. It presents a potentially resource-efficient approach to delivering high-quality education for both educators and learners.

Acknowledgments

We express our gratitude to the 2021 first-class undergraduate curriculum project at Nanjing Medical University for its financial support. Additionally, we thank the undergraduate and graduate students of the First Clinical Medical College of Nanjing Medical University for their involvement, as well as the teaching development teams of the First Clinical Medical College and the Basic Medical College for their role in instructional design, software development, and production. The corresponding author may be contacted via email for further details or to request permission for educational tool use. Our appreciation also extends to Editage for their assistance with English language editing.

Data Availability

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

Authors' Contributions

All the authors (PT, YX, KQ, ML, and JH) were involved in the study design and methods. PT conducted the investigation and formal analysis of the data and was responsible for writing the first draft of the manuscript. ML and JH supervised the study. All the authors contributed to and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Higher Resolution versions of Figures 3-7.
[DOCX File , 32564 KB - mededu_v11i1e59850_app1.docx]

Multimedia Appendix 2
Software development and construction process of pelvic fracture model.
[DOC File , 2820 KB - mededu_v11i1e59850_app2.doc]

Multimedia Appendix 3
An interview outline of 7 open-ended questions.
[DOC File , 32 KB - mededu_v11i1e59850_app3.doc]

Multimedia Appendix 4
Box plot with the median scores and IQR for pre- and posttest.
[DOC File , 148 KB - mededu_v11i1e59850_app4.doc]

References

1. Zong ZW, Chen SX, Qin H, Liang HP, Yang L, Zhao YF, Representing the Youth Committee on Traumatology branch of the Chinese Medical Association, PLA Professional Committee and Youth Committee on Disaster Medicine, Traumatology branch of the China Medical Rescue Association, Disaster Medicine branch of the Chongqing Association of Integrative Medicine. Chinese expert consensus on echelons treatment of pelvic fractures in modern war. Mil Med Res 2018 Jun 30;5(1):21 [FREE Full text] [doi: 10.1186/s40779-018-0168-3] [Medline: 29970166]
2. Heetveld MJ, Harris I, Schlaphoff G, Balogh Z, D'Amours SK, Sugrue M. Hemodynamically unstable pelvic fractures: recent care and new guidelines. World J Surg 2004 Sep;28(9):904-909. [doi: 10.1007/s00268-004-7357-9] [Medline: 15593465]
3. Coccolini F, Stahel PF, Montori G, Biffl W, Horer TM, Catena F, et al. Pelvic trauma: WSES classification and guidelines. World J Emerg Surg 2017;12:5 [FREE Full text] [doi: 10.1186/s13017-017-0117-6] [Medline: 28115984]
4. Marmor M, El Naga AN, Barker J, Matz J, Stergiadou S, Miclau T. Management of pelvic ring injury patients with hemodynamic instability. Front Surg 2020;7:588845 [FREE Full text] [doi: 10.3389/fsurg.2020.588845] [Medline: 33282907]
5. Bagaria D. The past, present, and future management of hemodynamic instability in patients with unstable pelvic ring injuries. Injury 2022 Mar;53(3):1294. [doi: 10.1016/j.injury.2021.11.061] [Medline: 34865818]
6. Cevallos N, Zukotynski B, Greig D, Silva M, Thompson RM. The utility of virtual reality in orthopedic surgical training. J Surg Educ 2022;79(6):1516-1525 [FREE Full text] [doi: 10.1016/j.jsurg.2022.06.007] [Medline: 35821110]
7. Keith K, Hansen D, Johannessen M. Perceived value of a skills laboratory with virtual reality simulator training in arthroscopy: a survey of orthopedic surgery residents. J Am Osteopath Assoc 2018 Oct 01;118(10):667-672 [FREE Full text] [doi: 10.7556/jaoa.2018.146] [Medline: 30264141]
8. Salem J, Fukuta J, Coombs A, Morgan J. Virtual patient journey: a novel learning resource. Clin Teach 2020 Jun;17(3):315-319. [doi: 10.1111/tct.13101] [Medline: 31680422]
9. Kononowicz AA, Woodham LA, Edelbring S, Stathakarou N, Davies D, Saxena N, et al. Virtual patient simulations in health professions education: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019 Jul 02;21(7):e14676 [FREE Full text] [doi: 10.2196/14676] [Medline: 31267981]
10. Yuan YB, Wang JJ, Lin MH, Gao XY. Building virtual simulation teaching platform based on electronic standardized patient. Sheng Li Xue Bao 2020 Dec 25;72(6):730-736 [FREE Full text] [Medline: 33349830]
11. Peñuela-Epalza M, De la Hoz K. Incorporation and evaluation of serial concept maps for vertical integration and clinical reasoning in case-based learning tutorials: perspectives of students beginning clinical medicine. Med Teach 2019 Apr;41(4):433-440. [doi: 10.1080/0142159X.2018.1487046] [Medline: 30091645]
12. Wei F, Sun Q, Qin Z, Zhuang H, Jiang G, Wu X. Application and practice of a step-by-step method combined with case-based learning in Chinese otoendoscopy education. BMC Med Educ 2021 Feb 04;21(1):89 [FREE Full text] [doi: 10.1186/s12909-021-02513-1] [Medline: 33541330]
13. Demetri L, Donnelley CA, MacKechnie MC, Toogood P. Comparison of case-based learning and traditional lectures in an orthopedic residency anatomy course. J Surg Educ 2021;78(2):679-685. [doi: 10.1016/j.jsurg.2020.08.026] [Medline: 32888846]
14. Gartmeier M, Pfurtscheller T, Hapfelmeier A, Grünewald M, Häusler J, Seidel T, et al. Teacher questions and student responses in case-based learning: outcomes of a video study in medical education. BMC Med Educ 2019 Dec 05;19(1):455 [FREE Full text] [doi: 10.1186/s12909-019-1895-1] [Medline: 31805913]
15. Zhao L, Dai X, Chen S. Effect of the case-based learning method combined with virtual reality simulation technology on midwifery laboratory courses: a quasi-experimental study. Int J Nurs Sci 2024 Jan;11(1):76-82 [FREE Full text] [doi: 10.1016/j.ijnss.2023.12.009] [Medline: 38352279]
16. Wang D, Kim L, Gronberg M, Stambaugh C, AAPM Medical Physics Leadership Academy (MPLA) Cases Subcommittee. A brief guide to writing a medical physics leadership case. J Appl Clin Med Phys 2021 Mar;22(3):285-286 [FREE Full text] [doi: 10.1002/acm2.13186] [Medline: 33739581]
17. ten Cate O, Custers EJFM, Durning SJ. Principles and Practice of Case-based Clinical Reasoning Education: A Method for Preclinical Students. Cham: Springer; 2018:95-108.
18. Encinas-Ullán CA, Martínez-Diez JM, Rodríguez-Merchán EC. The use of external fixation in the emergency department: applications, common errors, complications and their treatment. EFORT Open Rev 2020 Apr;5(4):204-214 [FREE Full text] [doi: 10.1302/2058-5241.5.190029] [Medline: 32377388]

19. Mi M, Kanakaris NK, Wu X, Giannoudis PV. Management and outcomes of open pelvic fractures: an update. Injury 2021 Oct;52(10):2738-2745. [doi: 10.1016/j.injury.2020.02.096] [Medline: 32139131]
20. Hertzog MA. Considerations in determining sample size for pilot studies. Res Nurs Health 2008 Apr;31(2):180-191. [doi: 10.1002/nur.20247] [Medline: 18183564]
21. Mahling M, Wunderlich R, Steiner D, Gorgati E, Festl-Wietek T, Herrmann-Werner A. Virtual reality for emergency medicine training in medical school: prospective, large-cohort implementation study. J Med Internet Res 2023 Mar 03;25:e43649 [FREE Full text] [doi: 10.2196/43649] [Medline: 36867440]
22. Balki E, Holland C, Hayes N. Use and acceptance of digital communication technology by older adults for social connectedness during the COVID-19 pandemic: mixed methods study. J Med Internet Res 2023 Aug 02;25:e41535 [FREE Full text] [doi: 10.2196/41535] [Medline: 37531187]
23. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008 Apr;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
24. Erlingsson C, Brysiewicz P. A hands-on guide to doing content analysis. Afr J Emerg Med 2017 Sep;7(3):93-99 [FREE Full text] [doi: 10.1016/j.afjem.2017.08.001] [Medline: 30456117]
25. Boller E, Courtman N, Chiavaroli N, Beck C. Design and delivery of the clinical integrative puzzle as a collaborative learning tool. J Vet Med Educ 2021 Apr;48(2):150-157. [doi: 10.3138/jvme.2019-0036] [Medline: 33861187]
26. Marchand LS, Sciadini MF. Simulation training in fracture surgery. J Am Acad Orthop Surg 2020 Nov 01;28(21):e939-e947. [doi: 10.5435/JAAOS-D-20-00076] [Medline: 32796368]
27. Tillander B, Ledin T, Nordqvist P, Skarman E, Wahlström O. A virtual reality trauma simulator. Med Teach 2004 Mar;26(2):189-191. [doi: 10.1080/0142159042000192037] [Medline: 15203531]
28. Brennan JN, Hall AJ, Baird EJ. Surgeon 2023 Oct;21(5):e263-e270. [doi: 10.1016/j.surge.2023.02.005] [Medline: 36914519]
29. Huber M, Katzky U, Müller K, Blätzinger M, Goetz W, Grechenig P, et al. Evaluation of a new virtual reality concept teaching k-wire drilling with force feedback simulated haptic in orthopedic skills training. J Hand Surg Am 2022 Dec;47(12):1225.e1-1225.e7. [doi: 10.1016/j.jhsa.2021.09.008] [Medline: 34857404]
30. Howard T, Iyengar KP, Vaishya R, Ahluwalia R. High-fidelity virtual reality simulation training in enhancing competency assessment in orthopaedic training. Br J Hosp Med (Lond) 2023 Sep 02;84(9):1-8 [FREE Full text] [doi: 10.12968/hmed.2022.0360] [Medline: 37769263]
31. Thompson J, White S, Chapman S. Interactive clinical avatar use in pharmacist preregistration training: design and review. J Med Internet Res 2020 Nov 06;22(11):e17146 [FREE Full text] [doi: 10.2196/17146] [Medline: 33155983]
32. Guetterman TC, Sakakibara R, Baireddy S, Kron FW, Scerbo MW, Cleary JF, et al. Medical students' experiences and outcomes using a virtual human simulation to improve communication skills: mixed methods study. J Med Internet Res 2019 Nov 27;21(11):e15459 [FREE Full text] [doi: 10.2196/15459] [Medline: 31774400]
33. Ferede B, Ayenew A, Belay W. Pelvic fractures and associated injuries in patients admitted to and treated at emergency department of Tibebe Ghion Specialized Hospital, Bahir Dar University, Ethiopia. Orthop Res Rev 2021;13:73-80 [FREE Full text] [doi: 10.2147/ORR.S311441] [Medline: 34140815]
34. Kleinert R, Heiermann N, Plum PS, Wahba R, Chang DH, Maus M, et al. Web-based immersive virtual patient simulators: positive effect on clinical reasoning in medical education. J Med Internet Res 2015 Nov 17;17(11):e263 [FREE Full text] [doi: 10.2196/jmir.5035] [Medline: 26577020]

Abbreviations

CBL: case-based learning
ESP: electronic standardized patient
VR: virtual reality

Edited by SR Mogali; submitted 24.04.24; peer-reviewed by Q Yan, N Jiang; comments to author 28.10.24; revised version received 11.11.24; accepted 15.12.24; published 17.01.25.

Please cite as:
Teng P, Xu Y, Qian K, Lu M, Hu J
Case-Based Virtual Reality Simulation for Severe Pelvic Trauma Clinical Skill Training in Medical Students: Design and Pilot Study
JMIR Med Educ 2025;11:e59850
URL: https://mededu.jmir.org/2025/1/e59850
doi:10.2196/59850
PMID:

Original Paper

Extended Reality–Enhanced Mental Health Consultation Training: Quantitative Evaluation Study

Katherine Hiley[1,2], BSc; Zanib Bi-Mohammad[3,4], PhD; Luke Taylor[1], BSc; Rebecca Burgess-Dawson[5], MSc; Dominic Patterson[5], MSc, MBChB; Devon Puttick-Whiteman[5], BA; Christopher Gay[5], MSc; Janette Hiscoe[5]; Chris Munsch[5], MB, ChM, FRCS; Sally Richardson[5]; Mark Knowles-Lee[6]; Celia Beecham[6], BA; Neil Ralph[5], DClinPsych; Arunangsu Chatterjee[7], PhD; Ryan Mathew[1,8], FRCS, PhD; Faisal Mushtaq[1,2], PhD

[1]Centre for Immersive Technologies, HELIX, University of Leeds, Leeds, United Kingdom
[2]School of Psychology, Faculty of Medicine & Health, University of Leeds, Leeds, United Kingdom
[3]School of Science, Technology and Health, York St John University, York, United Kingdom
[4]School of Healthcare, Faculty of Medicine & Health, University of Leeds, Leeds, United Kingdom
[5]NHS England, England, United Kingdom
[6]Fracture Reality, Brighton, United Kingdom
[7]School of Medicine, Faculty of Medicine & Health, University of Leeds, Leeds, United Kingdom
[8]Department of Neurosurgery, Leeds Teaching Hospitals NHS Trust, Leeds, United Kingdom

Corresponding Author:
Faisal Mushtaq, PhD
School of Psychology
Faculty of Medicine & Health
University of Leeds
Woodhouse
Leeds, LS2 9JT
United Kingdom
Phone: 44 07525418924
Email: f.mushtaq@leeds.ac.uk

Abstract

Background: The use of extended reality (XR) technologies in health care can potentially address some of the significant resource and time constraints related to delivering training for health care professionals. While substantial progress in realizing this potential has been made across several domains, including surgery, anatomy, and rehabilitation, the implementation of XR in mental health training, where nuanced humanistic interactions are central, has lagged.

Objective: Given the growing societal and health care service need for trained mental health and care workers, coupled with the heterogeneity of exposure during training and the shortage of placement opportunities, we explored the feasibility and utility of a novel XR tool for mental health consultation training. Specifically, we set out to evaluate a training simulation created through collaboration among software developers, clinicians, and learning technologists, in which users interact with a virtual patient, "Stacey," through a virtual reality or augmented reality head-mounted display. The tool was designed to provide trainee health care professionals with an immersive experience of a consultation with a patient presenting with perinatal mental health symptoms. Users verbally interacted with the patient, and a human instructor selected responses from a repository of prerecorded voice-acted clips.

Methods: In a pilot experiment, we confirmed the face validity and usability of this platform for perinatal and primary care training with subject-matter experts. In our follow-up experiment, we delivered personalized 1-hour training sessions to 123 participants, comprising mental health nursing trainees, general practitioner doctors in training, and students in psychology and medicine. This phase involved a comprehensive evaluation focusing on usability, validity, and both cognitive and affective learning outcomes.

Results: We found significant enhancements in learning metrics across all participant groups. Notably, there was a marked increase in understanding (P<.001) and motivation (P<.001), coupled with decreased anxiety related to mental health consultations (P<.001). There were also significant improvements to considerations toward careers in perinatal mental health (P<.001).

Conclusions: Our findings show, for the first time, that XR can be used to provide an effective, standardized, and reproducible tool for trainees to develop their mental health consultation skills. We suggest that XR could provide a solution to overcoming the current resource challenges associated with equipping current and future health care professionals, which are likely to be exacerbated by workforce expansion plans.

(JMIR Med Educ 2025;11:e64619) doi:10.2196/64619

KEYWORDS

mental health; training; consultation; extended reality; virtual reality; augmented reality

Introduction

As the demand for mental health services in health care systems continues to rise, the need for skilled professionals capable of providing effective mental health consultation and support also increases [1,2]. In the face of changing workforce training requirements (coupled with significant health care workforce expansion plans), there is a growing recognition that the effective implementation of emerging technologies could help overcome some of the logistical and resource-related barriers involved in education and training.

Mental health nursing, in particular, faces distinct challenges that necessitate specialized training solutions. Mental health nurses encounter unique stressors, including high levels of emotional exhaustion, moral distress, and exposure to patient-initiated violence, all of which contribute to job dissatisfaction and high turnover rates [3], which in turn negatively impact workforce stability, patient outcomes, and overall health care service quality [4]. Additionally, mental health nurses often report insufficient opportunities for continuing professional development and limited support from leadership, further compounding retention challenges. Addressing these issues through targeted and innovative training approaches is essential for fostering resilience, enhancing job satisfaction, and improving workforce retention.

Beyond specialist mental health settings, primary care physicians or general practitioners (GPs) also play key roles in managing mental health conditions, with more than a third of general practice consultations involving mental health issues [5]. Effective communication and therapeutic relationships have been shown to significantly influence outcomes, emphasizing the need for better training in interpersonal and empathetic skills for managing mental health conditions in primary care. However, variability in the ability of GPs to detect and manage mental health issues highlights gaps in current training models [6]. As communication forms a central part of mental health treatment, poorly trained clinicians may inadvertently block disclosure of emotional distress, potentially delaying critical interventions [5]. Therefore, innovative training approaches are crucial not only for mental health nurses but also for GPs and other health care professionals involved in mental health consultations.

Traditional training for health professionals in managing mental health problems typically relies on a combination of in-person placements, which employ observation-based learning, and actor-based simulations. While in-person placements provide valuable real-world experience, they often present challenges, such as unpredictable exposure to a diverse range of patient demographics, risks to both students and vulnerable service users, and limited opportunities for structured feedback. Actor-based simulations, on the other hand, are difficult to scale and standardize due to variability in actors' interpretations of scripts and inconsistencies in their familiarity with specific case studies. These limitations make it challenging to provide health care professionals with the comprehensive training necessary to handle the complexities of mental health consultations. Effective and compassionate mental health consultations require more than procedural knowledge. They demand the ability to empathize, engage in therapeutic communication, and establish a strong patient-provider relationship. To address these needs, training must focus on promoting empathy and compassion while preparing health care professionals to navigate the diverse backgrounds and emotional experiences of patients. However, traditional training methods often struggle to meet these goals due to ethical concerns around exposing students to sensitive cases and the inherent difficulty in replicating the unpredictable dynamics of real-life mental health scenarios.

Advances in a suite of new immersive technologies that go under the banner of extended reality (XR) and include virtual reality (VR) and augmented reality (AR) could be particularly well-suited to address these challenges by providing interactive, standardized, repeatable learning experiences that bridge the gap between theory and practice. VR presents users with a computer-generated environment that immerses them in a fully digitally simulated environment, while AR overlays virtually generated elements onto the real world. The value of XR for health care training has already been demonstrated across various domains, such as surgery [7], physical rehabilitation [8], anatomy [9], and the training of practical skills in nurses [10]. However, the implementation of XR in the training of mental health professionals has lagged.

Given the importance and complexity of training for mental health consultations, coupled with the increasing workload pressure on GPs and mental health nurses to meet the population's mental health support needs [7], we set out to test whether XR technology could be used to create a training environment to support the development of mental health consultation skills. We reasoned that the ability to deliver standardized repeatable experiences of varied patient encounters (including more rare presentations) in a safe and controlled environment could provide a learning experience that nurtures confidence and competence in consultation skills that augment traditional training.

To assess the potential efficacy of XR in mental health consultation training, we focused on perinatal mental health

training, a subspecialty supporting women navigating mental health challenges during pregnancy or the initial postpartum year. This is an area of the mental health service with an urgent training need. The recent report of the Royal College of Psychiatrists [8] highlighted a critical need for comprehensive perinatal training programs across both specialized and general health care services. There is also a notable lack of confidence among perinatal mental health nurses in their capacity to deliver care to women with perinatal mental health challenges, with only a quarter feeling well-equipped to support these women [9]. Trainees also have relatively limited opportunities to train, with a shortage of placement opportunities. A recent review of perinatal mental health education across 32 UK medical schools [10] found that perinatal mental health was not considered a core curriculum topic. Instead, it was typically incorporated as a subtopic within broader topic areas, such as lectures on depression. Given the shortage of staff and limited placements in perinatal mental health, a new training tool that could support the development of the next generation of health care staff could have an immediate impact.

Here, we report on the validation and evaluation of a novel XR training tool developed through a collaboration between software developers and health care staff, including nurses specializing in perinatal mental health and GPs. The simulation presents an interactive virtual patient ("Stacey") with severe perinatal mental health problems. Stacey is a mother of 2 children, with her youngest child only 4 weeks old, and has a record of mild postnatal depression following her first birth. Low mood, suicidal ideation, and episodes of psychosis add complex layers to her clinical presentation. Users interact with Stacey verbally, and her responses are selected by a human instructor from a range of prerecorded voice-acted clips in an audio repository. We explore the utility of this tool for supporting social and emotional interactions with the simulation, investigate the ease of use for trainers and trainees, and evaluate the impact on cognitive and affective learning.

Methods

Overall Approach

We undertook a 2-stage evaluation process that included a pilot study exploring feasibility and a subsequent evaluation of the perinatal mental health XR training experience in terms of learning outcomes and perceptions. In this section, we introduce the simulation platform and training experience and subsequently detail the methods and procedures common to and distinct for each phase of the experiment. It should be noted that the authors involved in developing the content played no role in the evaluation. The analysis was carried out independently by the authors KH, LT, and FM.

This study was not designed as a head-to-head comparison with traditional training approaches. Some participants, particularly those in mental health nursing, had previously received standard training methods (eg, classroom-based teaching, in-person clinical placements, or actor-based role plays), but these forms of training were not systematically assessed here. Instead, the primary aim was to evaluate the feasibility and potential impact of XR as a supplementary training tool. We have thus included the details of traditional training experiences for context but did not incorporate a direct comparative arm in this work.

XR Simulation

The simulation was built on a platform ("JoinXR") created by the software developer Fracture Reality. The JoinXR platform was designed to enable multi-user simulation training environments over a range of head-mounted VR or AR displays. In the evaluation, we used the Meta Quest 2 headset (Meta Platforms, Inc) for the VR version of the platform and Microsoft HoloLens 2 (Microsoft Corp) for the AR version.

The human-computer interface within the JoinXR platform was central to facilitating immersive lifelike interactions with the virtual patient. The interface allowed learners to engage through natural voice-based dialogue, processed in real time using instructor-guided responses. The integration of audio feedback, gesture recognition (HoloLens 2), and hand controllers (Meta Quest 2) enabled users to navigate the virtual space intuitively. Learners interacted directly with Stacey, the patient avatar, whose responses, including eye contact, subtle emotional cues, and body movements, were programmed to mimic real-world patient behavior, creating a realistic and contextually relevant learning experience. Users engaged with the simulation through natural voice-based dialogue and through VR controllers (Meta Quest 2) or hand-tracking gestures (HoloLens 2). Nonverbal communication, such as the avatar's facial expression, body language, and spatial audio, enhanced the simulation's realism.

The JoinXR platform was designed to be a conversational engine enabling "human to digital avatar" interactions in a multi-user, real-time environment. In this way, it could facilitate remote participation by learners, instructors, and observers, supporting the practice and refinement of nonroutine clinical skills. The learner-instructor dynamic was a crucial component of the simulation, incorporating both real-time guidance and postsimulation feedback. Instructors played an active role during the interaction by interpreting learner inputs and controlling Stacey's responses using a soundboard system (Figure 1B). This allowed for dynamic adaptations, where learners could engage organically with the avatar and explore different conversational pathways. After the simulation, instructors conducted debrief sessions using performance analytics that tracked response accuracy, emotional sensitivity, and decision-making, providing learners with targeted feedback to refine their clinical competencies.

Figure 1. Development of the learning platform. (A) Wireframe of the patient avatar, Stacey; (B) Soundboard for instructors to control Stacey's responses; (C) Setting for the consultation, showing the learner (blue) and the patient avatar.

During the simulation, learners interacted with Stacey by asking questions that were either processed by conversational artificial intelligence (AI) or directly controlled by the instructor for tailored responses. Stacey's reactions were designed to simulate real-world patient behaviors, including nuanced emotional expressions and gestures. Figure 2B illustrates an example consultation scenario in which the learner uses voice input to ask about Stacey's symptoms, prompting verbal and nonverbal responses (eg, maintaining eye contact and gesturing to emphasize a point).

Learners using VR devices (Meta Quest 2) could navigate the virtual consultation room using handheld controllers to manipulate objects, such as a clipboard or a stethoscope, or to adjust their position relative to Stacey. In contrast, AR users (HoloLens 2) experienced a blended environment where Stacey's avatar appeared within a real-world room.

The clinical simulations were developed through collaboration between Fracture Reality and a panel of subject-matter experts from the National Health Service (NHS), including mental health clinicians, GPs, and psychologists specializing in perinatal mental health. These experts supported the design of all aspects of the simulations, from character development and storyline construction to ensuring the accurate portrayal of medical conditions. Prior to the present evaluation, the development process included an iterative feedback process involving clinicians with primary care and perinatal mental health experience, software developers, and intended end users.

The specific focus of our evaluation is the first clinical simulation scenario developed using this new platform (Figure 1). The simulation is centered around a female patient avatar ("Stacey"). The aforementioned clinical experts contributed to the development of her patient history and personal attributes. Digital reference photos were then gathered to build a montage of the patient. A base model was built by taking a full body scan of a human model and was modified using a combination of 3D modeling software. The 3D models were created using a combination of Maya (Autodesk) and Blender (Blender Foundation). Clothing was designed, and then, the digital model was dressed. Bespoke custom lighting and skin rendering pipelines were developed to deliver realistic digital human features that could be rendered on headsets using low-powered graphics processing units.

Multiple script iterations were recorded, and the dialogue was reviewed and refined for clinical authenticity by the Fracture Reality team in consultation with the aforementioned subject-matter experts. Auditions were held to select actors. Studio sessions and spatial audio engineering rebalanced vocals to realistically imitate the patient avatar. Animations combined motion capture, hand animation, and lip-syncing for seamless responses. A custom Unity system facilitated quick and accurate lip-syncing to facial expressions and body poses (Figure 2). Reference photos from NHS facilities were used, and lighting was tailored for realistic environments, focusing on meaningful prop placement.

Two scenarios were designed, each tailored to address the needs of 2 primary but distinct target groups: mental health nursing students and primary care trainees (postgraduate doctor in GP training). While both scenarios feature a patient named Stacey presenting with a similar mental health condition, contextual variations were introduced to align more closely with the necessary professional capabilities of the respective trainee groups.

In the mental health nursing scenario, Stacey Morris is introduced as an emergency referral from her GP for a comprehensive assessment. Stacey, a 32-year-old mother of 2 children, with a 4-week-old newborn, has a history of postpartum depression following the birth of her first child. The primary objective for the student in this scenario is to conduct an initial mental health examination of Stacey.

In the primary care scenario, following a telephone conversation with her husband, Josh, who expressed concerns about her behavior, the postgraduate doctor in GP training agrees to meet Stacey in her home. Stacey in this scenario has a similar profile as in the mental health nursing scenario. She is a 32-year-old mother of 2 children, with her youngest child being 4 weeks old. In this context, the role of the postgraduate doctor in GP training centers on conducting a comprehensive mental health assessment with Stacey.

For each context, specific learning outcomes were defined by subject-matter experts. For the mental health nursing scenario, learners were expected to (1) understand and reflect on the lived experience of assessing the mental health of a patient with perinatal mental health problems; (2) identify signs and symptoms of perinatal mental ill health in acute assessment presentation; (3) apply the skills, knowledge, and abilities relevant to one's own profession in the assessment of mental health; and (4) have an appropriate reflected and evaluated performance of the task in a supported reflection. For the primary care setting, learners were expected to be able to (1) take history from a patient presenting with an acute psychotic illness; (2) ascertain and evaluate information relating to safeguarding; and (3) assess suicide and homicide risk.

Figure 2. Development of the verbal storyboard and voice integration for patient avatar interactions. (A) Demonstration of the process of integrating voice actors' performances into the patient avatar through a custom Unity-based lip-syncing system. Multiple iterations of dialogue scripts were recorded, and voice actors were selected via auditions, with audio engineering applied to simulate realistic patient speech patterns. Motion capture, hand animation, and spatial audio balancing enhanced the avatar's authenticity. (B) Demonstration of the verbal storyboard for the virtual patient, displaying categorized responses covering key clinical themes such as psychosis, self-harm, and family concerns. The storyboard guided the avatar's realistic conversational flow, informed by clinical expertise and reference material from National Health Service facilities to create immersive training scenarios.

General Methods

Following study advertisement, interested participants were screened for physical conditions that would exclude them from participation, including physical and auditory impairments and epilepsy. Included participants met with the instructor for a one-to-one session in a quiet room located on the university campus or at a local NHS hospital. Multimedia Appendix 1 outlines the study procedure. At the beginning of the session, participants had the opportunity to read the information sheet

and ask questions related to the study. Participants provided their consent to the study once they had been informed of their right to withdraw.

After consenting, participants were asked to complete a baseline questionnaire, capturing demographic and attitudinal data. Participants were then randomly allocated to 1 of 2 immersive environments: VR (Meta Quest 2) or mixed reality (HoloLens 2). Participants were subsequently exposed to either "primary care" Stacey (targeted at medical students and postgraduate doctors in GP training) or "mental health" Stacey (for mental health nursing or psychology students), contingent upon their current training program. These simulations share identical features, with the sole distinction lying in the introductory context of the consultation process. Both simulations were configured to align with the familiar protocols of health care trainees, specifically in terms of patient reception. Importantly, the responses of Stacey and the trajectory of the consultation remained consistent across the 2 scenarios.

Trainees verbally interacted with Stacey, who was in turn controlled by an instructor through the navigation of a soundboard, which triggered prerecorded audio clips from Stacey (see Figure 1B). The conversation journey would typically begin with general introductions; discussion of Stacey's relationships with her daughter, son, and husband; discussion of her birthday; discussion of experiences of psychosis and self-harm; and finally, formulation of a care plan. If the instructor felt the student was unable to lead the conversation or the student expressed having difficulty conversing with the avatar, prompts could be provided within the simulation. Multimedia Appendix 2 shows the prompts available for the early, mid, and late stages of the conversation that could be made visible to the students by the instructor.

Following the experience, the instructor carried out a postexperience debrief session with the trainee, including a critical discussion of the experience and the participant's performance. Following this, participants completed a postexperience survey measuring attitudinal domains and career considerations alongside measures of usability, presence, discomfort, and preference. The total session lasted approximately 1 hour (Multimedia Appendix 1).

Pilot Study

Participants

In the pilot study, we recruited 9 subject-matter experts from primary care and mental health disciplines. This included a consultant perinatal psychiatrist, a GP, 4 mental health nurses, a specialist perinatal mental health nurse, a psychiatry trainee (ST4), and a mental health nursing lecturer. All had more than 5 years of experience in their respective fields, with 8 having 10 or more years of experience. The purpose of this study was to formally test face and content validity and usability, and to support the latter, we included 5 undergraduate university students (mean age 22.4 years, SD 0.8 years).

Participants followed the procedure outlined in Multimedia Appendix 1. We evaluated face and content validity, usability, and utility as reported by a group of nonnursing or medical students and subject matter experts in the postexperience questionnaire. Face validity was assessed using a scale applied previously to expert evaluations of VR health care training [11]. This original 13-item scale was adapted to this study, and 11 items analyzed the ease of use, effectiveness, and immersion of the XR simulation on a 4-point Likert scale (strongly agree to strongly disagree).

In addition, the Lawshe method [12] also known as the content validity ratio (CVR) method was used. This is a method used to assess the content validity of a measurement instrument or a test, especially in the context of psychological, educational, or health care research, using expert opinion. Here, experts rate each item on a 3-point scale: (1) Essential: if the item is crucial and necessary for measuring the construct; (2) Useful but not essential: if the item is relevant but not critical for measuring the construct; and (3) Not necessary: if the item is irrelevant or not needed for measuring the construct. The CVR is calculated using the equation:

where *Ne* represents the count of experts who have deemed the item as "essential," and *N* denotes the total number of experts who have participated in the rating process. The CVR is a numerical value that quantifies the consensus among experts regarding the essential nature of the items under consideration. The critical value is a benchmark used to assess the appropriateness of items included in a content validity assessment. If the number of experts who agree on the relevance of an item meets or exceeds the critical value, the item is deemed valid; otherwise, it may be considered for revision or removal from the assessment. According to the values calculated previously [13] with a panel of 8 subject matter experts, this study's critical value was 0.75. Thus, constructs must surpass a CVR of 0.75 to be deemed essential to the procedure.

We also captured usability through the 10-item System Usability Scale (SUS) [14] as it has widely been used to evaluate XR as a tool for health care training [15-17]. Scores of more than 80 indicate excellence, between 70 and 80 are considered good, and less than 50 are not acceptable [18].

We assessed user discomfort through the Virtual Reality Sickness Questionnaire (VRSQ) [19]. As a more context appropriate adaptation of the validated Simulator Sickness Questionnaire (SSQ) [20], the VRSQ was designed to minimize burden on participants. The VRSQ sums the scores of oculomotor and disorientation discomfort items to generate an overall total. While there are no widely agreed bounds of acceptability for the VRSQ, we set out to compare the scores of the Meta Quest 2 and HoloLens 2 to assess relative differences in physical discomfort between the 2 devices.

Experiment

Following the demonstration of the feasibility of the use of XR in consultation training, we undertook a larger-scale evaluation. Here, we continued to collect measures of usability and supplemented them with surveys exploring cognitive and affective learning, training preference, presence, and career considerations.

Participants

The experiment involved 123 participants (mean age 24.3 years, SD 7.86 years; 97 female participants, 22 male participants, 3 nonbinary or third gender participants, and 1 who did not disclose gender). No participants had known health conditions, such as epilepsy, or visual, auditory, or cognitive disorders that would prevent participation in XR-based activities. They were drawn from a range of health care disciplines, including postgraduate doctors in GP training (n=18; mean age 38.2 years, SD 6.38 years), mental health nursing students (n=30; mean age 25.9 years, SD 7.6 years) from the Universities of Leeds and Huddersfield, and undergraduate medical students (n=28; mean age 19.8 years, SD 3.12 years) and psychology students (n=47; mean age 19.8 years, SD 3.2 years) recruited from the University of Leeds.

Participants were approached via their institutions, through the distribution of emails including information sheets. Participants were offered a monetary voucher incentive where appropriate (ie, for registered students), while clinical staff were asked to undertake the study voluntarily with no remuneration. Participants were randomly allocated to the AR (n=63, 51.2%) and VR training groups (n=60, 48.8%).

Measures

In phase 2, the VRSQ and usability continued to be evaluated as described in the pilot study. The experiment extended the evaluation to capture attitudes, cognitive and affective learning, career aspirations, presence, and conversation fluency. These self-reported measures were implemented to provide insights into the user's social and emotional interactions with the simulation, as well as any reported enhancements in knowledge, understanding, motivation, learning satisfaction, and learning confidence, further assessing the effectiveness of XR mental health consultations.

Attitudes

Cognitive Learning

Success and confidence in practical situations are often predicted by possessing knowledge, familiarity, and understanding of the themes and techniques embedded in a course curriculum [21]. Conversely, a deficiency in such familiarity may hinder the ability to apply theoretical knowledge in practice [22].

To capture this, a 14-item Perinatal Mental Health Familiarity and Awareness Scale (PMHAFS) (Multimedia Appendix 3) was developed by the study team with subject-matter experts. Participants were asked to evaluate their knowledge with, awareness of, and understanding of the perinatal mental health assessment conditions and care on a 5-point Likert scale (strongly disagree to strongly agree).

Affective Learning

Intrinsic and extrinsic motivation for learning was assessed through a 6-item scale [23], developed based on the Motivated Strategies for Learning Questionnaire Manual [24]. Evaluating intrinsic and extrinsic constructs provides a holistic examination of the influences of learner engagement from within the learner and from the learning environment [25]. Higher scores on each item suggest a greater motivation for learning.

To assess self-confidence and learning satisfaction, a 12-item variant [26] of the original Student Satisfaction and Self-Confidence in Learning Scale was used [27]. This instrument has been shown to be highly reliable, with a Cronbach α of .92 for the presence of features and .96 for their importance. Each item on the Likert scale was coded from 1 to 5 (strongly disagree to strongly agree), with 5 items reverse coded to prevent acquiescent responding. Higher scores on the scale indicate greater satisfaction and self-confidence with learning [28].

Career Attitudes

To assess students' considerations of health care specialization, we assessed 9 items across 3 affective domains of motivation, preparedness, and sense of support toward perinatal mental health specialization. Higher scores on each 5-point Likert scale indicate a greater desire to consider perinatal mental health upon graduation.

Presence

The construct of presence is regularly evaluated in studies involving virtual environments. Defined as the subjective experience of being in one place or environment, even when physically in another [29], there has been an active debate on its contribution to learning [30-32]. High levels of presence are speculated to be associated with deeper cognitive engagement, a cornerstone for effective learning [29], increasing intrinsic motivation and creating an environment where learners are more likely to integrate and retain new information [32]. A high degree of presence may help to minimize the impact of real-world distractions, allowing learners to fully immerse themselves in the task at hand [33]. Presence has also been proposed to be instrumental for the transfer of skills from the virtual to the real world [31]. We sought to measure presence through the previously validated iGroup Presence Questionnaire (IPQ) [34,35], a 14-item scale capturing spatial presence, realness, and involvement.

Ethical Considerations

Approval for the study was granted by the School of Psychology Ethics Committee (approval number: PSYC-615; date of approval: November 13, 2022). Consent was obtained from participants at the beginning of the session.

Statistical Analysis

ANOVA tests were performed to examine the effect of the XR training tool on the ratings of improvement in cognitive learning of conditions, assessment, and care. This same technique was applied to attitude changes in career motivation, support and preparedness, learning confidence, and learning satisfaction. Where appropriate, a between-subjects variable was introduced in the ANOVA when comparing population groups: GP postgraduate doctor in training, mental health nursing student, psychology student, or medical student.

For presence, specific data items related to presence were filtered and selected to include measures, such as "general," "spatial," "involvement," and "realism," as defined in the IPQ. The presence scores were reported across different devices and groups, examining how users experienced each of these presence

measures. An ANOVA assessed differences in presence scores between devices and measures (device [AR vs VR] × iGroup construct [general vs spatial vs involvement vs realism]). Post-hoc tests were applied to decompose interaction effects for VR and AR where appropriate.

For each family of tests (per construct), *P* values were corrected for multiple comparisons using the Bonferroni method. Corrected *P* values below an α threshold of .05 were considered statistically significant. All data analyses were performed in R (version 4.2.2) using RStudio (version 2022.12.0.353; Posit).

Results

Pilot Study

All experts, across both VR and AR systems (n=9, 100%), felt actively involved and in charge of the situation. The simulation software responded adequately and did not lag according to 8 of the experts, while all 9 experts reported that it was easy to learn how to interact with the software. Notably, all were interested in the progress of events throughout the simulation, suggesting high engagement. Additionally, all stated that it was easy to move around in the virtual environment, and the same amount of people reported that the controller buttons responded adequately.

Using the Lawshe method, we calculated the CVR for each step of the simulation process. These steps were: briefing instructions, medical notes, in-simulation prompts, instructor prompts, and postsimulation debrief. Briefing instructions provided the user with the necessary context for the forthcoming consultation. Medical notes, collaboratively developed with subject-matter experts, provided a comprehensive medical history for the virtual character, Stacey, to enhance the contextual richness of the consultation. In-simulation text prompts, illustrated in Multimedia Appendix 2, could be administered within the XR environment by the instructor, without verbal disruption to the ongoing consultation. In contrast, instructor prompts denoted verbal interventions made by the instructor at any time during the simulation. The postsimulation debrief is an opportunity for the user to reflect and for both the user and instructor to critically evaluate the consultation. The critical value in our study for the content validity of a construct and the component part of the simulation was 0.75. The obtained CVR scores for simulation outcomes, briefing instructions, and postexperience debrief were all 1, indicating that these processes were all rated as essential by all experts.

Some parts of the procedure, including previewing medical notes and using prompts during sessions, were considered optional by design. Our evaluation revealed that all content experts rated it as either essential or useful. In the case of in-simulation and instructor prompts, the majority found them essential or useful, but some considered them “not necessary,” as indicated by a score of 0.75.

The SUS scores were 78.75 for VR and 73.75 for AR, indicating good usability for both systems. The VRSQ scores were 0 for VR and 4.17 for AR, suggesting a negligible amount of discomfort for participants.

Pilot Study Summary

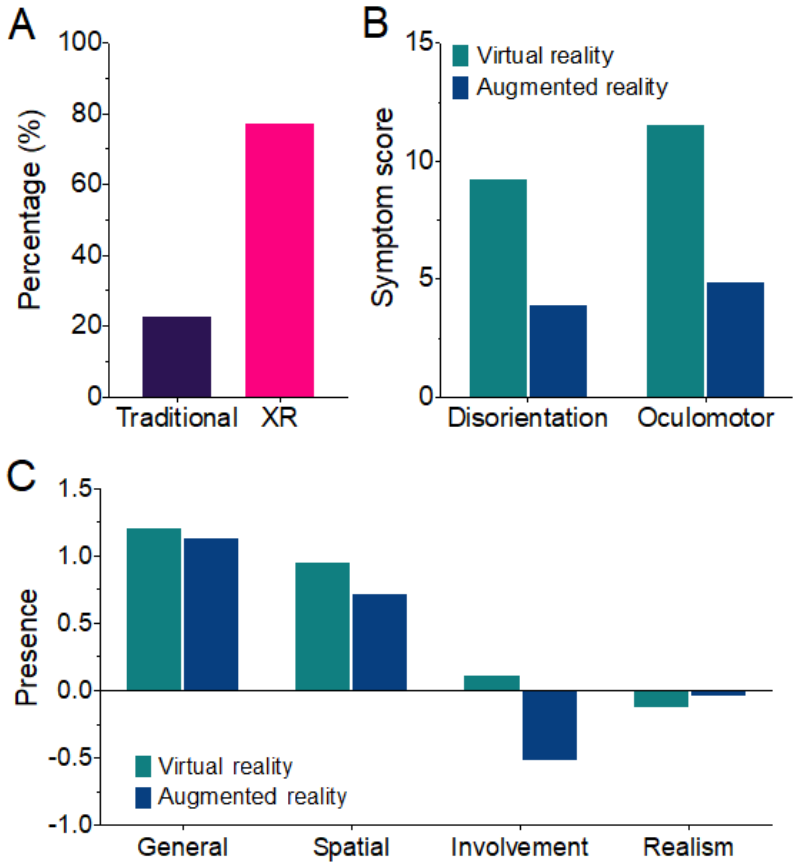
Participants provided positive feedback, reporting high usability levels for both VR and AR systems and minimal discomfort. Subject-matter experts rated the XR simulation highly in terms of engagement, involvement, and simulation quality, particularly within the context of perinatal training. They found the content and procedures valid, aligning with their expectations for an effective training session. These results suggest that the XR simulation has the potential to serve as a learner-centered training tool and provide the basis for conducting a larger-scale evaluation with health care trainees.

Experiment

Preference

Participants were asked whether they preferred the XR simulation or traditional approach to training that they had been exposed to. Overall, 77.2% (95/123) of participants preferred XR over traditional training methods (28/123, 22.8%) (Figure 3A).

Figure 3. Usability and preference. (A) User preference toward traditional learning for perinatal mental health or the integration of extended reality (XR) to augment perinatal mental health learning. (B) Symptom scores for the disorientation and oculomotor domains of the Virtual Reality Sickness Questionnaire for virtual reality (VR) and augmented reality (AR). (C) Self-reported experience of presence, illustrating that participants felt less involved in AR relative to VR. Error bars represent ±1 SEM.



Feasibility

The overall SUS score was 81.6 (SD 11.1), with no difference (t_{73}=0.75; *P*=.45) between the scores for AR (mean 82.3, SD 10.9) and VR (mean 80.3, SD 12.8), which translates to an excellent usability rating for both systems.

Simulator Sickness

In an analysis designed to understand the impact of different devices on simulator sickness, a 2-way ANOVA revealed a significant interaction between device and symptom ($F_{3,312}$=6.41; *P*<.001). There were greater sickness scores in the disorientation domain in VR (mean 9.22, SD 1.06) than in AR (mean 3.92, SD 0.81) (t_{208}=3.47; *P*<.001) and greater scores in the oculomotor domain in VR (mean 11.53, SD 1.33) than in AR (mean 4.89, SD 1.01) (t_{208}=4.34; *P*<.001). The analysis suggests that VR is more likely to cause symptoms of disorientation and oculomotor discomfort than AR.

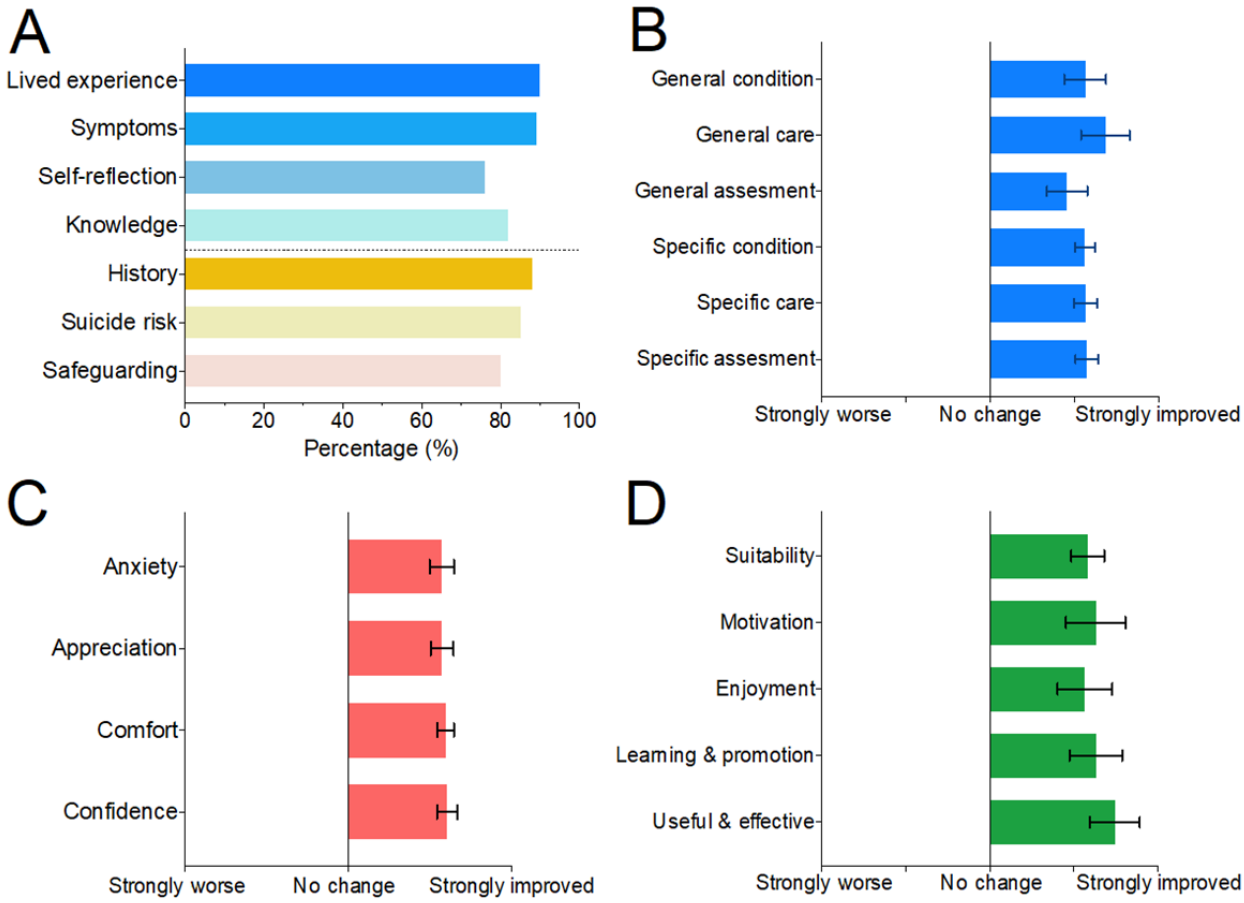
Presence

IPQ scores were compared between VR and AR. There was a statistically significant interaction between presence measure and device ($F_{3,327}$=5.78; *P*=.02; η^2_G=0.025). Post-hoc analysis revealed a statistically improved sense of involvement for those using VR relative to AR (t_{484}=3.18; *P*=.002). There were no significant differences between the systems in general (t_{484}=0.13; *P*=.90) and in the spatial (t_{484}=–1.11; *P*=.27) and experienced realism (t_{484}=1.17; *P*=.24) domains of presence.

Learning Outcomes

For the 2 simulations evaluated in this study, specific learning outcomes were defined by subject matter experts from perinatal mental health and primary care. We report these separately for each group. Figure 4A shows the percentage of sessions in which the learning outcome was achieved across all groups, as reported by the instructor.

Figure 4. Learning outcomes, cognitive and affective changes, and learning satisfaction. (A) Percentage of achievement of learning objectives for perinatal mental health and primary care simulations within sessions across all participants. (B) Improvements in understanding across perinatal conditions, assessment, and care domains following the simulation for general practitioner (GP) trainees, and improvements in understanding across perinatal conditions, assessment, and care domains following the simulation for mental health nursing students. (C) Improvements in the affective domains of confidence, comfort, appreciation for the challenges in providing support to perinatal cases, and reduction in anxiety among perinatal cases for GP trainees. (D) Improvements in the utility and feasibility domains for mental health consultation training following XR simulations compared with the current training approach. Error bars represent the SEM for domain change responses.



In the primary care simulation, instructors rated that they were able to achieve learning objective 1 (able to take history from a patient presenting with an acute psychotic illness) in 100% of sessions, learning objective 2 (able to ascertain and evaluate information relating to safeguarding) in 80% of sessions, and learning objective 3 (able to assess suicide and homicide risk) in 80% of sessions.

In the perinatal mental health simulation, instructors rated that they were able to achieve learning objective 1 (understand and reflect on the lived experience of assessing the mental health of a patient with perinatal mental health problems) in 100% of sessions, learning objective 2 (identify the signs and symptoms of perinatal mental ill health in acute assessment presentation) in 90% of sessions, learning objective 3 (apply the skills, knowledge, and abilities relevant to one's own profession in the assessment of mental health) in 89% of sessions, and learning objective 4 (have appropriate reflected and evaluated performance of the task in a supported reflection) in 80% of sessions.

Changes in Cognitive and Affective Attitudes

Primary Care

At baseline, 22% of participants stated that they had "no experience" with perinatal mental health cases, 61% expressed "little experience," and only 17% expressed "some experience." Understanding of complex mental health (general) and perinatal mental health (specific) was measured at baseline, revealing that 59% of GP trainees expressed an understanding of complex mental health at a general level and 58% expressed an understanding of perinatal mental health specifically.

Regarding affective constructs, 44% of trainees expressed anxiety around complex mental health cases and 50% expressed anxiety around perinatal mental health cases. Following the simulation, participants reported a statistically significant improvement in cognitive attitudes (mean 0.91, SD 0.86; t_{17}=4.47; P=.003; d=1.05).

Participants further reported a statistically significant improvement in affective attitudes following the simulation (mean 0.92, SD 0.74; t_{17}=5.27; P<.001; d=1.17). Across the affective domain, participants reported an improvement in confidence (mean 0.83, SD 1.04; t_{17}=3.39; P=.004; d=0.79),

comfort (mean 0.89, SD 0.76; t_{17}=4.97; P<.001), appreciation for the challenges of providing perinatal mental health support (mean 0.94, SD 1.00; t_{17}=4.01; P=.001; d=0.95), and reduced anxiety toward perinatal mental health cases (mean 1.00, SD 1.09; t_{17}=3.91; P=.001; d=0.92).

Medical Students

Following the simulation, medical students reported an improvement in cognitive attitudes (mean 1.38, SD 0.40; t_{27}=18.14; P<.001; d=3.42). This group also reported a statistically significant improvement in affective attitudes (mean 1.35, SD 0.46; t_{27}=10.01; P<.001; d=1.89). Across the affective domain, students reported an improvement in confidence (mean 1.64, SD 0.58; t_{27}=13.45; P<.001; d=2.54), comfort (mean 1.39, SD 0.57; t_{27}=13.00; P<.001; d=2.46), appreciation (mean 1.29, SD 0.90; t_{27}=7.59; P<.001; d=1.43), and reduced anxiety toward perinatal mental health cases (mean 1.25, SD 0.97; t_{27}=6.84; P<.001; d=1.29).

Mental Health Students and Psychology Students

Mental health and psychology students reported a significantly improved understanding of perinatal mental health conditions (mean 1.01, SD 0.62; t_{76}=14.26; P<.001; d=1.63), assessment (mean 1.21, SD 0.74; t_{76}=14.60; P<.001; d=1.62), and care (mean 1.09, SD 0.65; t_{26}=14.82; P<.001; d=1.69) following the simulation.

Within the mental health student group, improvements were seen following the simulation across the domains of perinatal mental health conditions (mean 0.86, SD 0.65; t_{29}=7.26; P<.001; d=1.32), assessment (mean 0.91, SD 0.67; t_{29}=7.41; P<.001; d=1.35), and care (mean 0.76, SD 0.60; t_{29}=6.88; P<.001; d=1.26).

Improvements were also seen in the psychology group across the domains of conditions (mean 1.11, SD 0.59; t_{46}=12.85; P<.001; d=1.87), assessment (mean 1.39, SD 0.73; t_{46}=13.10; P<.001; d=1.91), and care (mean 1.31, SD 0.59; t_{26}=15.32; P<.001; d=2.23).

Across all mental health and psychology students, we found a significant increase in learning confidence (mean 1.14, SD 0.49; t_{76}=20.32; P<.001; d =2.32). Students further reported a significant increase in learning satisfaction (mean 1.33, SD 0.69; t_{76}=16.51; P<.001; d=1.88). There was a similar finding within groups, as mental health students reported a significant increase in learning confidence following the simulation (mean 1.13, SD 0.57; t_{29}=10.83; P<.001; d=1.98). Psychology students also reported a significant increase in learning confidence following the simulation (mean 1.13, SD 0.43; t_{46}=15.32; P<.001; d=2.61).

For learning satisfaction, mental health students reported a significant increase following the simulation (mean 1.25, SD 0.82; t_{29}=8.33; P<.001; d=1.52), and psychology students also reported a significant increase following the simulation (mean 1.37, SD 0.62; t_{46}=15.12; P<.001; d=2.20).

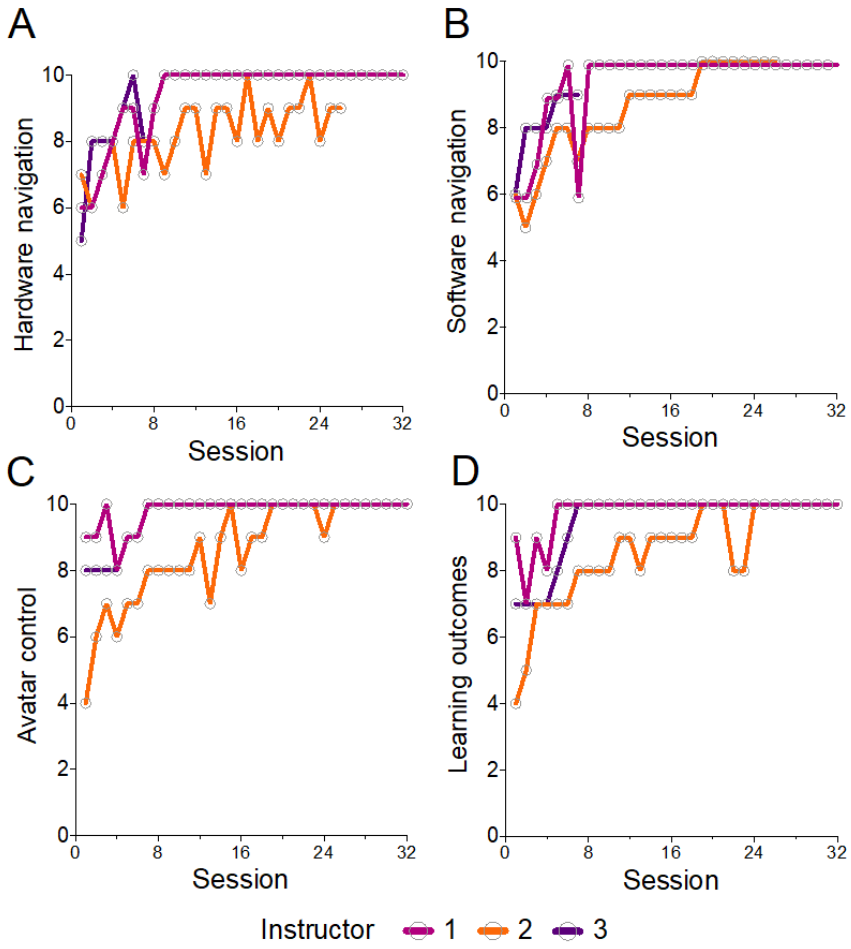
Career Considerations

At baseline, 49% of mental health nursing students stated that they were motivated to pursue a career in perinatal mental health, while 30% agreed that they felt prepared to pursue a career in perinatal mental health and 24% felt supported to pursue a career in perinatal mental health. Only 25% of psychology students were considering a career in perinatal mental health. Following the simulation, mental health nursing students felt significantly more motivated (mean 0.73, SD 0.65; t_{29}=6.15; P<.001; d=1.12), prepared (mean 1.10, SD 0.52; t_{29}=11.61; P<.001; d=2.12), and supported (mean 0.74, SD 0.74; t_{29}=5.79; P<.001; d=1.06) to pursue a career in perinatal mental health. Similarly, psychology students also reported a significantly greater likelihood of considering a career in perinatal mental health following the simulation (t_{46}=7.04; P<.001; d=1.03).

Instructor Training

In addition to assessing the benefits for participants, to better understand how much time it would take to train staff without previous XR experience to become comfortable with navigating through this training platform, we asked our instructors to document their degree of confidence on a scale from 0 to 10 regarding four key dimensions: (1) hardware navigation, (2) software navigation, (3) avatar control, and (4) delivery of session learning outcomes. Following each session, instructors assigned ratings to these constructs, thereby creating a subjective trajectory of their session delivery proficiency (Figure 5). Notably, these ratings rose rapidly and plateaued after approximately 6 to 8 sessions across all key constructs, suggesting that it will take multiple training sessions before instructors feel that they can deliver reliably consistent training sessions. We also observed some variation from session to session, which may be accounted for by a combination of measurement errors and technical and logistical factors. While not amenable to formal statistical analysis, instructors reported lower scores when they experienced Wi-Fi dropouts or software crashes.

Figure 5. Instructor development over session delivery. Following each session, instructors self-reported the following on a scale of 1 to 10: (A) ease of hardware handling; (B) ease of software navigation; (C) confidence in controlling the avatar; and (D) comfort in achieving learning outcomes.



Discussion

Principal Findings

We explored the idea that XR technologies could support the delivery of mental health training through a simulated mental health consultation, in which a trainee interacts with a human-controlled virtual avatar. An initial feasibility pilot with subject-matter experts and students demonstrated potential efficacy worthy of further investigation. We subsequently followed this up with a comprehensive evaluation of its impact on trainees from across mental health nursing, medical doctors training to be GPs, and undergraduate psychology and medicine students. Our findings demonstrate the significant potential of XR as a pedagogical tool in supporting the development of mental consultation delivery skills.

We observed notable enhancements in cognitive and affective learning across all health care trainee groups. Instructors reported high rates of successful delivery of learning objectives, while participant groups reported increased knowledge in diverse perinatal domains, including the recognition of conditions, such as depression and anxiety, during pregnancy and in the postpartum period. Trainees demonstrated proficiency in systematic evaluation using diagnostic tools to assess severity. We also observed improvements in knowledge and confidence, specific to perinatal mental issues and broader issues of working with complex mental health challenges.

The integration of XR into mental health training represents a significant advancement, offering immersive, interactive, and repeatable learning environments that traditional methods often fail to replicate effectively. Conventional training typically relies on static case studies, peer-based role play, or interactions with real patients, each of which presents limitations. XR, however, combines high-quality instructional content with advanced technological features, including real-time feedback, iterative “fail and retry” opportunities, and high-fidelity simulations. This integration is not merely a shift in delivery medium but a holistic synthesis of content and technology, fostering experiential and contextually relevant learning.

The educational content needed to address complex and sensitive scenarios, such as perinatal mental health, is notably limited within mainstream mental health nursing curricula and GP training. Traditional training often emphasizes general psychiatric principles or common conditions, leaving significant gaps in specialized instructions for nuanced cases like acute postpartum psychosis and perinatal depression. This lack of exposure to high-risk sensitive clinical contexts underscores the need for innovative training solutions that can bridge this gap. XR-based simulations offer a tailored and immersive approach, allowing learners to engage with realistic perinatal

mental health cases and gain practical experience-driven insights beyond what conventional programs typically provide.

Although this study did not conduct a direct comparison with conventional approaches, XR's ability to standardize and replicate complex scenarios addresses many logistical and ethical challenges associated with actor- or patient-based training. By facilitating autonomous practice, enhancing critical competencies, and building learner confidence in a psychologically safe environment, XR provides a valuable environment for high-stake contexts such as mental health consultations. Effective training in this domain is essential, as errors can negatively impact both therapeutic relationships and patient outcomes [36]. XR's capacity to support the development of these therapeutic relationships is key to achieving improved health outcomes for individuals with mental illness [37]. This work also suggests that immersive educational technologies might be able to influence career planning and specialization. Our study found an increase in the reported interest among trainees considering a career in perinatal mental health. This positive shift in attitude toward perinatal mental health careers is particularly significant given the documented shortage in this specialty [9]. Such tools may extend beyond traditional educational outcomes to influence career aspirations and potentially bridge the gap between abstract career concepts and tangible professional identity formation.

Immersive educational technologies, exemplified by XR simulations, possess the potential to not only shape career preferences but also address significant concerns regarding the cultivation of empathetic connections and the practical application of theoretical knowledge during training. In mental health training, a crucial aspect involves nurturing the user's ability to establish therapeutic relationships. This necessitates engaging in specific scenarios and subsequent reflection to ensure nurses can comprehensively apply theoretical knowledge effectively [38]. We were concerned that the interaction with a virtual avatar may be a poor substitute for the development of this relationship and that it may be difficult to empathize with. However, our investigation into users' social and emotional interactions within the simulation revealed positive indicators, including general and spatial presence and improvements across cognitive and affective domains. These promising outcomes suggest that immersive technologies may not act as barriers but instead as facilitators in establishing effective therapeutic relationships.

Further grounds for our concerns about the feasibility of this tool in this context came from the "Uncanny Valley" [39] phenomenon, which describes the sense of unease or discomfort experienced when an artificial representation closely resembles a human but is not quite convincingly lifelike. Stacey had indeed been designed to be as realistic as possible (working within the graphical constraints of today's technology). Our outcomes indicate that the design quality and the method for interacting with the avatar were sufficient to circumvent this effect, allowing users to transcend potential unease and engage meaningfully with the simulation. Nevertheless, somewhat paradoxically, as the graphical capabilities of XR technology increase, this area will become increasingly more important to monitor in the design and implementation of patient avatars until they become indistinguishable from real humans. This necessitates a careful iterative approach in the design and implementation of patient avatars, one that is cognizant of these psychological effects. Future iterations of XR simulations must be not only technically advanced but also underpinned by a deep understanding of user psychology to ensure that they support rather than detract from the learning objectives [40].

In looking to the future, the rapid advances being made in generative AI provide an avenue for such training tools to become increasingly autonomous, which could significantly alleviate the workload of instructors while simultaneously enhancing the dynamic interactivity of training sessions through the development of bespoke patient avatars tailored to the needs of learners. AI analysis of utterance-response pairs could predict context-specific reactions, enabling intelligent and adaptive XR training tools. XR training tools could leverage this "generative" AI to create dynamic and realistic scenarios for training health care professionals in mental health consultations, thereby enhancing their ability to understand and respond to a wide range of patient interactions. The use of generative AI could also democratize access to high-quality training resources, making them available across different geographies and socioeconomic contexts, thereby potentially reducing disparities in mental health training quality globally. Instructors could personalize scenarios, offer real-time feedback, and adapt to unique learner needs. Such potential advances do, however, raise ethical concerns [41], including the risk of bias that would need to be tackled for effective, efficient, and inclusive training.

Limitations

It is important to note that this study does not suggest that XR learning can replace traditional placements or direct learning opportunities and experiences or that simulation avatars can fully replicate real patients. What it does show is that XR could be a valuable tool for providing standardized training experiences to mental health trainees across different institutions and professional domains. The simulation employed in this study serves as a potential solution for exposing trainees to complex and nonroutine patient presentations. Going a step further, we suggest that the tool could also offer an opportunity to explore underrepresented scenarios, including those involving minoritized populations, and could be a useful vehicle for promoting cultural competence and enhancing the overall diversity of training scenarios. We propose that by using XR technology, mental health training programs may be able to bridge gaps in exposure to various clinical scenarios and populations, contributing to a more comprehensive and inclusive approach to mental health training.

While this study demonstrated significant improvements in various aspects of trainee confidence and perceived competence, it is important to clarify that the study's primary aim was not to evaluate current educational provisions or compare XR training directly to traditional methods. Instead, the focus was on assessing the feasibility and potential benefits of an XR-based tool as a supplementary learning aid within existing training frameworks. The intention was to explore how XR could augment current educational experiences rather than to position it as a replacement for established training methods. Future

research should consider comparative studies that directly assess the effectiveness of XR against traditional pedagogical approaches to determine the conditions under which XR-based learning is most beneficial. Incorporating controlled trials and longitudinal assessments would further strengthen the understanding of XR's role in skill retention and clinical application. It is also important to note that our evaluation only involved a single session and an examination of changes immediately after the session. This has shown substantial promise and must be followed up with an examination of any longer-term changes, capturing skill retention and whether this knowledge and confidence can be translated to clinical practice. Equally, the implementation of this technology into the curriculum should not be a "one-shot" standalone affair. Instead, we propose that it should be integrated systematically across multiple sessions to reinforce and build upon the acquired knowledge and skills. Long-term evaluations, including follow-up assessments at intervals beyond the immediate postsession period, are imperative to gauge the durability and sustainability of the observed impacts. Additionally, future research endeavors should explore the application of XR technology in diverse clinical scenarios to assess its versatility and effectiveness across various health care contexts. The iterative and continuous integration of XR simulations into the curriculum, coupled with ongoing assessments, will contribute to a more comprehensive understanding of its benefits and practical applicability in real-world health care settings.

While we focused on evaluating one-to-one sessions, the platform also affords the delivery of one-to-many training sessions and the opportunity for group-led discussion. One-to-one sessions in XR offer personalized interactions where trainees can practice engaging with virtual patients in a safe controlled environment, receiving tailored feedback from instructors. This approach allows for intensive skill development, particularly in handling complex or sensitive mental health scenarios. On the other hand, one-to-many sessions leverage XR's multi-user capabilities to enable group training, where multiple participants can observe and interact within the same virtual environment. By leveraging its multi-user capabilities, XR training tools could be used to create an environment conducive to collaboration, group discussion, and the promotion of intra- and interprofessional discussions.

Furthermore, in a world where hybrid (or blended) learning has started to become a norm, XR provides a practical solution for overcoming resource and time constraints faced by training programs. The ability to access training sessions and share the same learning space from anywhere in the world could provide a practical solution to the resource and time constraints faced by training programs, promoting both inclusivity and efficiency in health care education. This flexibility ensures that trainees across diverse locations and professional domains can participate in standardized training experiences, contributing to equitable and scalable mental health education.

While immersive technologies present transformative opportunities as learning tools, accessibility for individuals with visual and auditory impairments remains a critical concern. XR environments heavily rely on visual and auditory inputs, which can exclude users with disabilities if not adequately addressed. For visual impairments, accessibility may involve features, such as screen reader compatibility, audio descriptions, and haptic feedback, to convey spatial or contextual information. For auditory impairments, captions, subtitles, and integration with assistive hearing devices, such as cochlear implants, are essential. To address these challenges, the software in this study integrates specific accessibility features. For users with hearing impairments, the JoinXR platform includes automatic captioning, enabling subtitles to appear beneath a user's avatar during interactions. For users with visual impairments, the design process emphasized hardware compatibility, recommending the HTC Vive Focus 3 for VR due to its adjustable lenses and focal settings and the Microsoft HoloLens 2 for AR, which allows users to keep their glasses on. These measures reflect a commitment to inclusivity, though further advancements are needed to fully overcome accessibility barriers in XR technologies.

Finally, important considerations for the implementation of XR training tools are the economic cost and the return on investment. There are significant start-up expenditures, including the procurement of XR hardware and software licenses. In addition, adopting XR technology requires appropriate technical infrastructure, such as accessible, reliable, and reasonably fast internet connectivity, along with a long-term strategy for sustainable implementation. The rapid pace of technological advancement poses the risk of hardware and software becoming quickly obsolete, compelling organizations to contemplate strategies for regular updates and maintenance to keep pace with technological innovations. On the other hand, XR technology affords numerous opportunities to enhance educational experiences, reducing training time and improving learning efficacy [42]. Additionally, the potential of XR to facilitate remote learning could reduce the necessity for travel and accommodation expenses, which are traditionally associated with centralized training programs [43]. A critical next step for advancing this field is the development of a return-on-investment framework. This framework should account for the wide spectrum of benefits as well as the initial and ongoing expenses. In this way, organizations will have clear insights into the viability and value of adopting tools, such as the one introduced here, as they address the escalating demands of health care workforce training.

Conclusions

The use of an XR-based simulated mental health consultation scenario, where trainees interacted with a human-controlled virtual avatar, showed promise in an initial feasibility pilot and was further substantiated by a comprehensive evaluation across various health care trainee groups. Our findings indicate significant enhancements in cognitive and affective learning, with high rates of successful delivery of learning objectives. These findings show, for the first time, that XR can be used to provide an effective, standardized, and reproducible tool for trainees to develop their mental health consultation skills. We suggest that XR could provide a solution to overcome the current resource challenges associated with equipping current and future health care professionals, which are likely to be exacerbated by workforce expansion plans.

Acknowledgments

This project was funded by Health Education England, which is now part of National Health Service (NHS) England. Authors RM and FM are supported in part by the National Institute for Health and Care Research (NIHR) Leeds Biomedical Research Centre (NIHR203331). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care. FM is further supported by the European Union's Horizon Research and Innovation Programme under grant agreement number 101070155 and UK Research and Innovation (UKRI) through the Horizon Europe Guarantee (#10039307). The authors would like to thank Dr Amy Micklethwaite, Dr James Bullock, Dr Mayur Vibhuti, Dr William Edney, Dr Matthew Bull, Dr Richard Elliott, Dr Giles Berrisford, and Dr Jelena Jankovic for their contributions as subject matter experts in the development of the simulation evaluated in this paper.

Authors' Contributions

Conceptualization: RBD, DP, DPW, CG, JH, MKL, CB
Data curation: KH, ZBM, LT, FM
Formal analysis: KH, ZBM, FM
Funding acquisition: RM, FM
Investigation: KH, ZBM, LT, FM
Methodology: KH, ZBM, LT, RBD, DP, DPW, RM, FM
Project administration: KH, ZBM, JH, SR, RM, FM
Resources: RBD, DP, MKL, CB
Software: MKL, CB
Supervision: ZBM, FM
Validation: FM
Visualization: KH, FM
Writing – original draft: KH, RBD, FM
Writing – review & editing: KH, ZBM, LT, RBD, DP, CG, CM, SR, MKL, CB, NR, AC, RM, FM

Conflicts of Interest

Authors from the University of Leeds (KH, ZBM, LT, RM, AC, and FM) declare no conflicts of interest relating to this study and undertook data collection and analysis independent of the rest of the authorship team. Authors MKL (Founder, Fracture Reality) and CB (Product Manager, Fracture Reality) led the development of the application. They were not involved in data collection or analysis and did not contribute to the discussion section of this manuscript. Co-authors RBD and DP, Health Education England (now NHS England) contributed to the development of the simulation scenarios created by Fracture Reality and to the development of the research project and were not involved in data collection or analysis. DPW, CG, JH, and SR were involved at the project supervisory level from Health Education England (now NHS England) and were not involved in data collection or analysis.

Multimedia Appendix 1
Process flow chart of the hour-long evaluation session.
[DOCX File , 37 KB - mededu_v11i1e64619_app1.docx]

Multimedia Appendix 2
In-simulation prompts available to the instructors to support users.
[DOCX File , 48 KB - mededu_v11i1e64619_app2.docx]

Multimedia Appendix 3
Perinatal Mental Health Familiarity and Awareness Scale (PMHAFS).
[DOCX File , 15 KB - mededu_v11i1e64619_app3.docx]

References

1. Committee of Public Accounts. Progress in improving NHS mental health services. UK Parliament. URL: https://publications.parliament.uk/pa/cm5803/cmselect/cmpubacc/1000/report.html [accessed 2025-02-24]
2. Mental health services: addressing the care deficit. NHS Providers. URL: https://nhsproviders.org/mental-health-services-addressing-the-care-deficit [accessed 2025-02-24]

3. Adams R, Ryan T, Wood E. Understanding the factors that affect retention within the mental health nursing workforce: a systematic review and thematic synthesis. Int J Ment Health Nurs 2021 Dec 28;30(6):1476-1497 [FREE Full text] [doi: 10.1111/inm.12904] [Medline: 34184394]
4. Oates J, Jones J, Drey N. Subjective well-being of mental health nurses in the United Kingdom: Results of an online survey. Int J Ment Health Nurs 2017 Aug 23;26(4):391-401. [doi: 10.1111/inm.12263] [Medline: 27878917]
5. Imrie R. Communication skills for consultations about mental health problems. InnovAiT: Education and Inspiration for General Practice 2015 Mar 03;8(4):246-251. [doi: 10.1177/1755738015570980]
6. Mitchell AJ, Rao S, Vaze A. International comparison of clinicians' ability to identify depression in primary care: meta-analysis and meta-regression of predictors. Br J Gen Pract 2011 Feb 01;61(583):e72-e80. [doi: 10.3399/bjgp11x556227]
7. Magnée T, de Beurs DP, de Bakker DH, Verhaak PF. Consultations in general practices with and without mental health nurses: an observational study from 2010 to 2014. BMJ Open 2016 Jul 18;6(7):e011579 [FREE Full text] [doi: 10.1136/bmjopen-2016-011579] [Medline: 27431902]
8. Perinatal mental health services: Recommendations for the provision of services for childbearing women. Royal College of Psychiatrists. URL: https://www.rcpsych.ac.uk/docs/default-source/improving-care/better-mh-policy/college-reports/college-report-cr232---perinatal-mental-heath-services.pdf [accessed 2025-02-24]
9. Noonan M, Galvin R, Jomeen J, Doody O. Public health nurses' perinatal mental health training needs: A cross sectional survey. J Adv Nurs 2019 Nov 27;75(11):2535-2547. [doi: 10.1111/jan.14013] [Medline: 30937923]
10. King JD, Crowley G, El-Maraghy M, Davis W, Jauhari A, Wilson-Jones C. Perinatal mental health in medical school curricula: a national scoping survey of British universities and student psychiatry societies. BJPsych Bull 2024 Feb 12;48(1):51-56 [FREE Full text] [doi: 10.1192/bjb.2022.91] [Medline: 36632805]
11. Sadeghi AH, Peek JJ, Max SA, Smit LL, Martina BG, Rosalia RA, et al. Virtual reality simulation training for cardiopulmonary resuscitation after cardiac surgery: face and content validity study. JMIR Serious Games 2022 Mar 02;10(1):e30456 [FREE Full text] [doi: 10.2196/30456] [Medline: 35234652]
12. Lawshe C. A quantitative approach to content validity. Personnel Psychology 2006 Dec 07;28(4):563-575. [doi: 10.1111/j.1744-6570.1975.tb01393.x]
13. Ayre C, Scally AJ. Critical values for Lawshe's Content Validity Ratio. Measurement and Evaluation in Counseling and Development 2017 Mar 08;47(1):79-86. [doi: 10.1177/0748175613513808]
14. Brooke J. SUS: A 'Quick and Dirty' Usability Scale. In: Usability Evaluation In Industry. Boca Raton, FL: CRC Press; 1996.
15. Taylor L, Dyer T, Al-Azzawi M, Smith C, Nzeako O, Shah Z. Extended reality anatomy undergraduate teaching: A literature review on an alternative method of learning. Ann Anat 2022 Jan;239:151817. [doi: 10.1016/j.aanat.2021.151817] [Medline: 34391910]
16. Arpaia P, De Benedetto E, De Paolis L, D'Errico G, Donato N, Duraccio L. Performance and usability evaluation of an extended reality platform to monitor patient's health during surgical procedures. Sensors (Basel) 2022 May 21;22(10):3908 [FREE Full text] [doi: 10.3390/s22103908] [Medline: 35632317]
17. Donekal Chandrashekar N, Manuel M, Park J, Greene A, Safford S, Gračanin D. An Extended Reality Simulator for Advanced Trauma Life Support Training. In: Chen JYC, Fragomeni G, editors. Virtual, Augmented and Mixed Reality: Applications in Education, Aviation and Industry. HCII 2022. Lecture Notes in Computer Science, vol 13318. Cham: Springer; 2022:31-44.
18. Bangor A, Kortum P, Miller J. Determining what individual SUS scores mean: adding an adjective rating scale. Journal of Usability Studies 2009;4(3):114-123 [FREE Full text] [doi: 10.5555/2835587.2835589]
19. Kim SK, Lee Y, Yoon H, Choi J. Adaptation of extended reality smart glasses for core nursing skill training among undergraduate nursing students: usability and feasibility study. J Med Internet Res 2021 Mar 02;23(3):e24313 [FREE Full text] [doi: 10.2196/24313] [Medline: 33650975]
20. Kennedy RS, Lane NE, Berbaum KS, Lilienthal MG. Simulator sickness questionnaire: an enhanced method for quantifying simulator sickness. The International Journal of Aviation Psychology 1993 Jul;3(3):203-220. [doi: 10.1207/s15327108ijap0303_3]
21. Tasdemir C, Gazo R. Integrating sustainability into higher education curriculum through a transdisciplinary perspective. Journal of Cleaner Production 2020 Aug;265:121759. [doi: 10.1016/j.jclepro.2020.121759]
22. Hansen J, Ramachandran R, Vockley J. Survey of health care provider understanding of gene therapy research for inherited metabolic disorders. Clin Ther 2022 Aug;44(8):1045-1056 [FREE Full text] [doi: 10.1016/j.clinthera.2022.07.002] [Medline: 35927093]
23. Wang L, Chen M. The effects of game strategy and preference‐matching on flow experience and programming performance in game‐based learning. Innovations in Education and Teaching International 2010 Feb;47(1):39-52. [doi: 10.1080/14703290903525838]
24. Pintrich PR, Smith D, Garcia T, McKeachie W. A Manual for the Use of the Motivated Strategies for Learning Questionnaire (MSLQ). ERIC. URL: https://eric.ed.gov/?id=ED338122 [accessed 2025-02-24]

25. Ryan RM, Deci EL. Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. Contemporary Educational Psychology 2020 Apr;61:101860. [doi: 10.1016/j.cedpsych.2020.101860]
26. Unver V, Basak T, Watts P, Gaioso V, Moss J, Tastan S, et al. The reliability and validity of three questionnaires: The Student Satisfaction and Self-Confidence in Learning Scale, Simulation Design Scale, and Educational Practices Questionnaire. Contemp Nurse 2017 Feb 10;53(1):60-74. [doi: 10.1080/10376178.2017.1282319] [Medline: 28084900]
27. Jeffries PR, Rizzolo MA. Designing and Implementing Models for the Innovative Use of Simulation to Teach Nursing Care of Ill Adults and Children: A National, Multi-Site, Multi-Method Study. National League for Nursing. URL: https://www.nln.org/docs/default-source/uploadedfiles/professional-development-programs/read-the-nln-laerdal-project-summary-report-pdf.pdf [accessed 2025-02-24]
28. Franklin AE, Burns P, Lee CS. Psychometric testing on the NLN Student Satisfaction and Self-Confidence in Learning, Simulation Design Scale, and Educational Practices Questionnaire using a sample of pre-licensure novice nurses. Nurse Educ Today 2014 Oct;34(10):1298-1304. [doi: 10.1016/j.nedt.2014.06.011] [Medline: 25066650]
29. Witmer BG, Singer MJ. Measuring presence in virtual environments: a presence questionnaire. Presence 1998 Jun;7(3):225-240. [doi: 10.1162/105474698565686]
30. Berkman MI, Akan E. Presence and Immersion in Virtual Reality. In: Lee N, editor. Encyclopedia of Computer Graphics and Games. Cham: Springer; 2019:1-10.
31. Mania K, Chalmers A. The effects of levels of immersion on memory and presence in virtual environments: a reality centered approach. Cyberpsychol Behav 2001 Apr;4(2):247-264. [doi: 10.1089/109493101300117938] [Medline: 11710251]
32. Slater M, Neyret S, Johnston T, Iruretagoyena G, Crespo M, Alabèrnia-Segura M, et al. An experimental study of a virtual reality counselling paradigm using embodied self-dialogue. Sci Rep 2019 Jul 29;9(1):10903 [FREE Full text] [doi: 10.1038/s41598-019-46877-3] [Medline: 31358846]
33. Makransky G, Terkildsen TS, Mayer RE. Adding immersive virtual reality to a science lab simulation causes more presence but less learning. Learning and Instruction 2019 Apr;60:225-236. [doi: 10.1016/j.learninstruc.2017.12.007]
34. Schubert TW. The sense of presence in virtual environments:. Zeitschrift für Medienpsychologie 2003 Apr;15(2):69-71. [doi: 10.1026//1617-6383.15.2.69]
35. Schubert T, Friedmann F, Regenbrecht H. The experience of presence: factor analytic insights. Presence: Teleoperators & Virtual Environments 2001 Jun;10(3):266-281. [doi: 10.1162/105474601300343603]
36. Nyttingnes O, Ruud T, Rugkåsa J. 'It's unbelievably humiliating'-Patients' expressions of negative effects of coercion in mental health care. Int J Law Psychiatry 2016 Nov;49(Pt A):147-153. [doi: 10.1016/j.ijlp.2016.08.009] [Medline: 27726890]
37. Wolsing SK, Hjorth P, Løkke A, Hilberg O, Frølund J. Experiences of receiving a medical consultation - an interview study among hospitalized psychiatric patients. Nord J Psychiatry 2024 Oct 22;78(7):583-590 [FREE Full text] [doi: 10.1080/08039488.2024.2373251] [Medline: 39037071]
38. Skår R. Knowledge use in nursing practice: the importance of practical understanding and personal involvement. Nurse Educ Today 2010 Feb;30(2):132-136. [doi: 10.1016/j.nedt.2009.06.012] [Medline: 19631424]
39. Mori M. The Uncanny Valley: The Original Essay by Masahiro Mori. IEEE Spectrum. URL: http://umnikizdes.ru/aways/web.ics.purdue.edu/~drkelly/MoriTheUncannyValley1970.pdf [accessed 2025-02-24]
40. MacDorman KF, Chattopadhyay D. Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. Cognition 2016 Jan;146:190-205 [FREE Full text] [doi: 10.1016/j.cognition.2015.09.019] [Medline: 26435049]
41. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022 Jan 20;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
42. Huang CL, Luo YF, Yang SC, Lu CM, Chen A. Influence of students' learning style, sense of presence, and cognitive load on learning outcomes in an immersive virtual reality learning environment. Journal of Educational Computing Research 2019 Aug 05;58(3):596-615. [doi: 10.1177/0735633119867422]
43. Cheng K, Tsai C. A case study of immersive virtual field trips in an elementary classroom: Students' learning experience and teacher-student interaction behaviors. Computers & Education 2019 Oct;140:103600. [doi: 10.1016/j.compedu.2019.103600]

Abbreviations

AI: artificial intelligence
AR: augmented reality
CVR: content validity ratio
GP: general practitioner
IPQ: iGroup Presence Questionnaire
NHS: National Health Service
SUS: System Usability Scale
VR: virtual reality

VRSQ: Virtual Reality Sickness Questionnaire
XR: extended reality

Edited by J Moen; submitted 22.07.24; peer-reviewed by S Smith, M Pritchard; comments to author 18.12.24; revised version received 07.02.25; accepted 11.02.25; published 02.04.25.

Please cite as:
Hiley K, Bi-Mohammad Z, Taylor L, Burgess-Dawson R, Patterson D, Puttick-Whiteman D, Gay C, Hiscoe J, Munsch C, Richardson S, Knowles-Lee M, Beecham C, Ralph N, Chatterjee A, Mathew R, Mushtaq F
Extended Reality–Enhanced Mental Health Consultation Training: Quantitative Evaluation Study
JMIR Med Educ 2025;11:e64619
URL: https://mededu.jmir.org/2025/1/e64619
doi:10.2196/64619
PMID:

Original Paper

Exploring the Impact of the COVID-19 Pandemic on Learning Experience, Mental Health, Adaptability, and Resilience Among Health Informatics Master's Students: Focus Group Study

Nadia Davoody[1], MSc, PhD; Natalia Stathakarou[1], MSc; Cara Swain[1,2], MBChB; Stefano Bonacina[1], MSc, PhD

[1]Health Informatics Centre, Department of Learning, Informatics, Management and Ethics, Karolinska Institutet, Stockholm, Sweden

[2]Academic Department of Military Surgery & Trauma, Royal Centre for Defence Medicine, Birmingham, United Kingdom

Corresponding Author:
Nadia Davoody, MSc, PhD
Health Informatics Centre, Department of Learning, Informatics, Management and Ethics
Karolinska Institutet
Tomtebodavägen 18 A
Stockholm, S-17177 Stockholm
Sweden
Phone: 46 08 524 864 86
Email: nadia.davoody@ki.se

Abstract

Background: The shift to online education due to the COVID-19 pandemic posed significant challenges and opportunities for students, affecting their academic performance, mental well-being, and engagement.

Objective: This study aimed to explore the overall learning experience among health informatics master's students at Karolinska Institutet, Sweden, and the strategies they used to overcome learning challenges posed by the COVID-19 pandemic.

Methods: Through 3 structured focus groups, this study explored health informatics master's students' experiences of shifting learning environments for classes that started in 2019, 2020, and 2021. All focus group sessions were recorded and transcribed verbatim. Inductive content analysis was used to analyze the data.

Results: The results highlight the benefits of increased autonomy and flexibility and identify challenges such as technical difficulties, diminished social interactions, and psychological impacts. This study underscores the importance of effective online educational strategies, technological preparedness, and support systems to enhance student learning experiences during emergencies. The findings of this study highlight implications for educators, students, and higher education institutions to embrace adaptation and foster innovation. Implications for educators, students, and higher education institutions include the need for educators to stay current with the latest educational technologies and design teaching strategies and pedagogical approaches suited to both online and in-person settings to effectively foster student engagement. Students must be informed about the technological requirements for online learning and adequately prepared to meet them. Institutions play a critical role in ensuring equitable access to technology, guiding and supporting educators in adopting innovative tools and methods, and offering mental health resources to assist students in overcoming the challenges of evolving educational environments.

Conclusions: This research contributes to understanding the complexities of transitioning to online learning in urgent circumstances and offers insights for better preparing educational institutions for future pandemics.

(JMIR Med Educ 2025;11:e63708) doi:10.2196/63708

KEYWORDS

COVID-19 pandemic; eHealth; blended learning; health informatics; higher education adaptation

Introduction

Background

The COVID-19 pandemic disrupted education worldwide, and many universities were required to shift to online education despite most being unprepared for such a shift [1]. The education response during the early phase of the COVID-19 pandemic focused on implementing remote learning modalities as an emergency response and mostly on the online delivery of educational material. The result of these efforts was a substantive rise in e-learning, whereby teaching and learning

activities took place remotely via digital platforms. After the pandemic, the use of e-learning was expected to grow [1].

e-Learning is the use of internet technologies to enhance knowledge and performance. e-Learning technologies offer learners control over the content, learning sequence, pace of learning, time, and often media, allowing them to tailor their experiences to meet their learning objectives [2]. Within the context of the COVID-19 pandemic, e-learning and online learning refer to remote teaching strategies and methods that universities used to urgently respond to the requirements of health protocols and restrictions in mobility [3].

While e-learning holds significant potential, the rapid deployment of strategies during the pandemic showed mixed results. Most of the swift adoption focused on reaching all students and enhancing accessibility, but educators lacked the time to develop and implement pedagogical strategies that could enhance the learning experience [3]. Strategies used for short-term online education may not have been suitable for the prolonged disruption caused by the pandemic.

Student experiences with e-learning during the pandemic were mixed. Some studies reported a "new normal" that students viewed positively, citing benefits such as education continuity, increased accessibility, stronger learner-lecturer interactions, and greater confidence in expressing themselves in an online learning environment [1,4-7].

Within the field of medical education, one study reported that students responded positively to the transition to online learning methods, with a notable improvement in student satisfaction related to course structure [8]. Another study highlighted medical students' generally positive attitudes toward e-learning during the pandemic. However, while e-learning was seen as a necessary and beneficial alternative, challenges were recognized, such as increased stress and anxiety, limited internet access, technical difficulties, and reduced hands-on clinical training [9]. Health profession education programs have reported issues in transitioning to online learning and maintaining continuity in education [10]. There have been difficulties related to adapting traditional teaching methods to online formats and the stress this placed on both students and educators [11]. Stress, unfamiliarity with online classrooms, uncertainty about academic futures, and the rapid shift to e-learning contributed to negative experiences for many students [12,13]. In addition, insufficient training, inadequate internet infrastructure in some countries, and lack of preparation led to poor student experiences, undermining the sustainability of e-learning [14].

The contradictory evidence on diverse experiences of e-learning during the pandemic underscores the need for further investigation within this area. To better prepare for future crises, higher education institutions need to understand the experiences of both educators and students and develop educational strategies for online learning during emergencies. Although many studies have investigated students' learning experiences during the pandemic, there are limited insights into the specific challenges and strategies used to overcome them. Contextual factors such as internet access, cultural differences, and subject of study likely influenced the experiences and perceptions of both students and educators. Therefore, there is a need to further explore e-learning experiences and strategies across different contexts and subjects of study. Health informatics education uniquely integrates theoretical knowledge with applied technical and health care–related skills, which may have been significantly affected by the shift to online learning.

Study Aim and Research Question

This study aimed to explore the overall learning experience of health informatics master's students at Karolinska Institutet (KI), Sweden, and the strategies they used to overcome the learning challenges posed by the COVID-19 pandemic. This study addressed the following research question: How did health informatics master's students at KI experience learning during the COVID-19 pandemic, and what strategies did they use to overcome related challenges?

Methods

Study Design

In this study, we conducted 3 semistructured focus group interviews comprising a total of 16 registered students and alumni of the Master's Programme in Health Informatics. The COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines [15] were used for reporting the results. The process was piloted, and questions for the participants were determined in advance by authors ND, SB, and NS adapted from previous studies on the impact of the COVID-19 pandemic on medical education [9] and on teaching and learning in health professional education [11].

In this explorative study, focus group interviews were chosen as a data collection technique as they allow for the generation of rich qualitative data through group interactions and dynamics, which is beneficial when addressing an exploratory question. Participants can build on each other's ideas, leading to more nuanced insights [16,17]. During these focus groups, the students could share their experiences and perspectives as part of an open-ended discussion, leading to a deeper understanding of their learning experiences and the strategies used to overcome challenges during the COVID-19 pandemic.

Given the exploratory study design, focus groups are ideal as a data collection method as they facilitate open-ended discussion and are effective in collecting rich data that might not have been possible to collect through more structured data collection methods. The focus groups were held remotely using the Zoom platform (Zoom Video Communications) with CS in the role of interviewer.

Study Setting and Participants

The context of the study was the Master's Programme in Health Informatics provided by the Department of Learning, Informatics, Management, and Ethics at KI. It is a 2-year global program run jointly with Stockholm University. The program is designed for students with an interest in IT and how it can be applied to the fields of medicine and health care. As such, the students may have either a technical or health care background.

Convenience sampling was used to recruit participants. Individuals who had participated in the master's program during the COVID-19 pandemic were invited to contribute. Therefore,

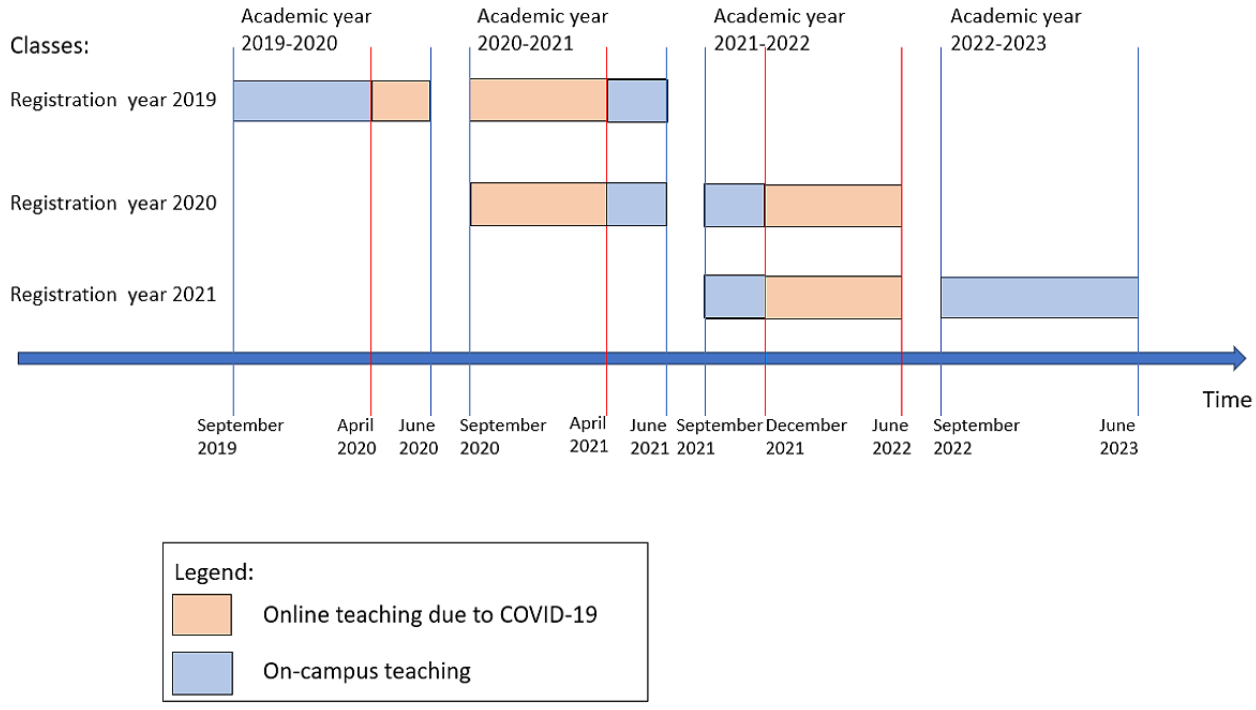
all participants were alumni or active registered students split into 3 cohorts: students who had registered in 2019 and graduated in 2021 (cohort 1), students who started in 2020 and graduated in 2022 (cohort 2), and students who started in 2021 and graduated in 2023 (cohort 3). Table 1 provides some characteristics of the participants. Most participants (13/16, 81%) were living in Sweden; however, as some of the students had relocated to their native countries or were working overseas, this informed the decision to conduct online focus groups.

Figure 1 shows the timeline of how the Master's Programme in Health Informatics from the academic year 2019-2020 to 2022-2023 was conducted. On-campus teaching and online teaching due to the COVID-19 pandemic are indicated using distinct colors. Students that started the program in September 2020 mostly received online teaching.

Table 1. Participant characteristics.

	Background	Experience with online learning
2019_Participant 1	Health care	No
2019_Participant 2	Health care	Yes
2019_Participant 3	Health care	No
2020_Participant 1	IT or technical	No
2020_Participant 2	Health care	Yes
2020_Participant 3	Health care	Yes
2020_Participant 4	Health care	Yes
2020_Participant 5	Health care	Yes
2020_Participant 6	Health care	No
2020_Participant 7	IT or technical	No
2020_Participant 8	Health care	Yes
2021_Participant 1	IT or technical	Yes
2021_Participant 2	IT or technical	Yes
2021_Participant 3	Health care	Yes
2021_Participant 4	Health care	Yes
2021_Participant 5	Health care	Yes

Figure 1. On-campus and online teaching during the COVID-19 pandemic.



Data Collection and Analysis

A total of 3 online focus group interviews were conducted during 2023 using the Zoom platform; the interviews were recorded and transcribed verbatim. To reduce familiarity bias in the study, CS, who had not been involved in teaching within the health informatics program, conducted the focus groups. The first focus group included 3 students from cohort 1, the second focus group included 8 students from cohort 2, and the third and final focus group included 5 students from cohort 3. The first focus group had a duration of 45 minutes, with the other 2 focus groups lasting approximately 1 hour and 30 minutes. The piloting process did not generate any data and was not included in the analysis.

Generated data were analyzed using inductive content analysis [18]. The inductive content analysis allowed for themes and patterns to be constructed directly from the data that were grounded in the students' actual experiences rather than imposing preconceived categories. A combination of coding methods was used: descriptive coding summarizes the main topics of the text, and pattern coding was used to condense meaning units into broader patterns to group the initial codes into broader themes. Pattern coding helps identify and understand the broader patterns and relationships within the data [19]. ND, NS, and SB conducted the initial coding. Each coder was instructed to familiarize themselves with the data, read through the entire dataset to gain an overall understanding before starting the coding process, identify meaning units, condense them, and assign codes that captured the essence of each unit. They independently reviewed all transcribed interviews, dividing the responsibilities for identifying relevant meaning units and conducting the initial coding. CS joined the analytical process once initial coding and subcategories had been generated. Upon identifying subcategories, a comparative analysis was conducted to reveal similarities and differences among student groups. Comparative analysis was conducted through peer debriefing sessions in which ND, NS, and SB compared their assigned codes and discussed differences in interpretations. All authors reviewed discrepancies collaboratively until a consensus was reached. Discrepancies were resolved by re-examining the raw data for the meaning unit in question, discussing interpretations considering the research aim, and refining the coding if needed. At this point, it was decided to collaboratively proceed with the categorization and identification of subthemes for all 3 student groups. This not only minimized the recurrence of redundant findings resulting from similarities among groups but also empowered us to emphasize the subcategories in which different student groups expressed distinct experiences or opinions.

Ethical Considerations

This research was carried out in Sweden. According to the Swedish Ethical Review Act, the research presented in our submitted manuscript did not require ethics approval as it did not handle sensitive personal information (as understood by the European General Data Protection Regulation). However, ethical requirements still apply, and written informed consent to take part in the study was obtained from all participants. The consent form outlined the study's purpose, potential risks and discomforts, the voluntary nature of participation, and the right to withdraw at any time. It also stated that no compensation would be provided for participation. Participants were assured that their confidentiality and privacy would be preserved.

Results

Overview of Data Collection and Data Analysis

Our analysis of the generated focus group data identified 1 main theme—*adapting to hybrid learning*—and three subthemes: (1) students' considerations of learning during the pandemic, (2) moving between learning environments, and (3) students' well-being and engagement in learning. Each subtheme included several categories relevant to all student groups, summarized in Table 2.

The overarching theme of *adapting to hybrid learning* highlights the challenges and experiences faced by participants navigating the transition from traditional on-campus education to online learning and vice versa. It includes the autonomy that students seek in shaping their learning experiences, the integration of learning into their daily lives, and the varying perceptions of online education as an obligation rather than a choice during the COVID-19 pandemic. It further explores the changed aspects of social interaction, the struggle with engagement in online learning, the technical challenges faced, and the diverse technological readiness levels for different learning environments. In addition, this theme addresses the psychological impact of remote learning and the need for adequate support, the varied levels of student motivation, the influence of family and pet support, and the observed lack of networking opportunities and social interaction in online educational settings.

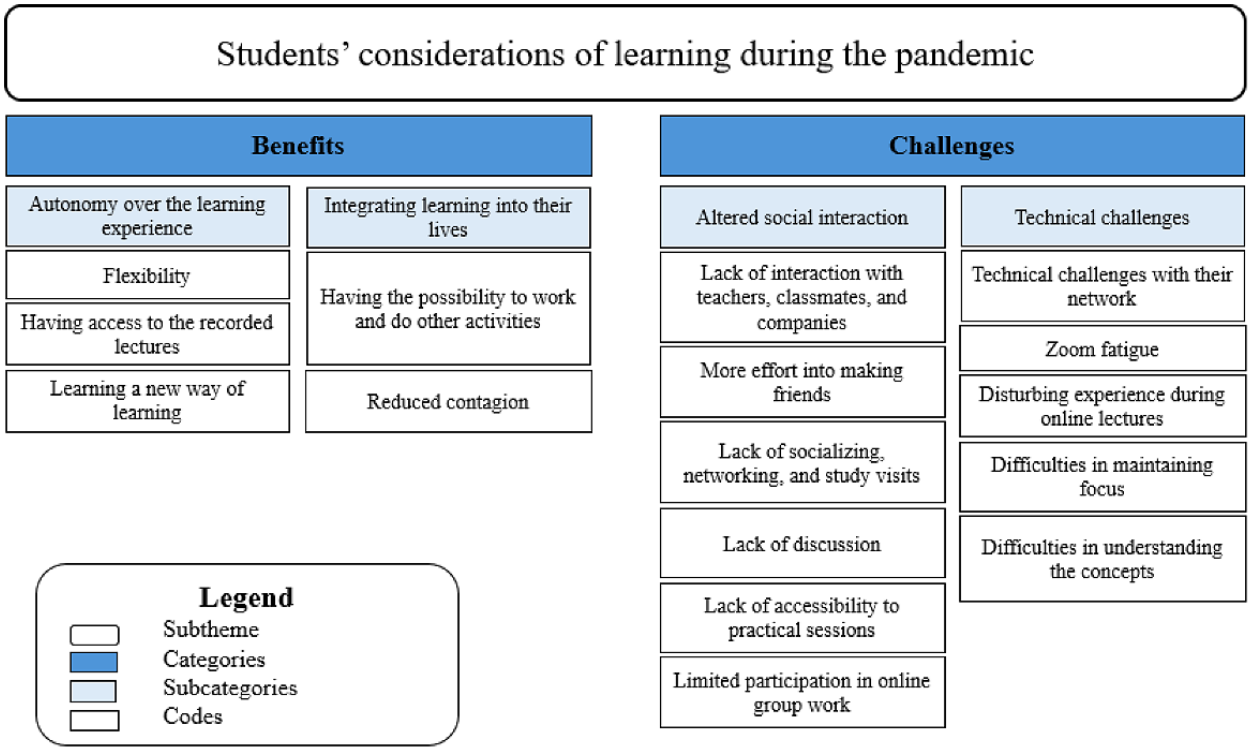
Table 2. An overview of the subcategories, categories, subthemes, and themes.

Theme, subtheme, and category	Subcategories
Adapting to hybrid learning	
Students' considerations of learning during the pandemic	
• Benefits	• Autonomy over the learning experience • Integrating learning into their lives
• Challenges	• Altered social interaction • Technical challenges
Moving between learning environments	
• Transition—campus to online education	• Online education as an obligation, not an option • Technological readiness for different learning environments • Psychological impact and support
• Transition—online to campus education	• Technological readiness for different learning environments • Lack of engagement in online learning • Different experiences in interaction with classmates and teachers between online and campus education
Students' well-being and engagement in learning	
• Motivation level • Strategies to stay motivated	• Varied levels of motivation • Group work as a connecting element • Family and pet support
• The impact of online learning on academic performance	• Autonomy • Lack of networking (social interaction)
• Overall experience and recommendations	• Improving online sessions • Improving educational support and delivery

Students' Considerations of Learning During the Pandemic

The students recognized that the changes and requirements necessitated by the COVID-19 pandemic brought both benefits and challenges. Figure 2 illustrates the relationship among the codes, subcategories, and categories for the subtheme *students' consideration of learning during the pandemic*.

Figure 2. A hierarchical structure displaying the subtheme and codes for theme 1—students' considerations related to learning during the pandemic.



Benefits

Benefits were characterized by increased *autonomy over their learning experience* and an ability to *integrate learning more readily into their lives*. Most of the participants mentioned that *flexibility* was one of the biggest benefits during the COVID-19 pandemic. Some mentioned that they could work from home or, in the case of many international students attending the Master's Programme in Health Informatics at KI, even work in their home country:

> *[The] main benefit is that it was less time-consuming. I was living at that time in [area_1] and I would have to travel to [area_2], that time was saved for me.* [2019_Participant 3]

> *I can start with one of the big benefits I had, especially in the beginning was that I was able to start the master's program from Germany where I am originally coming from. So, I moved I think 2 weeks in the program to Sweden. So, that was a big benefit to be more flexible in terms of the location for sure.* [2020_Participant 1]

Having access to the recorded lectures was appreciated as it made it possible for the students to go through the material whenever it was suitable for them. In addition, some of the students believed that they had *learned a new way of learning* and could now learn anything online. Online learning was believed to be more *efficient* as they could save time without the requirement to commute to campus:

> *...For me, it helped me to have the independence to learn as [participant] said. And yeah, and like now I can I have the feeling that I can learn anything online, I can just sign up for a course and in some videos and do it. And yeah, and in my current job, I did this a lot during the last six months. New technologies, new programming languages, everything...* [2020_Participant 7]

> *...I do watch tutorials on YouTube right now, but you get used to it. To manage to solve problems alone actually, which is really interesting because it's something that I do in my current job. If I don't know how to do something, I watch YouTube. I'm not going to ask anybody, and I think that's something with the pandemic as well...I managed to solve the problems myself.* [2020_Participant 4]

Having the possibility to work and do other activities in parallel with their studies was also appreciated by some of the participants:

> *In terms of benefits, it was very beneficial to save the commute time to the university by at least two hours a day. I was able to work in parallel with my studies. That wasn't a big deal. The third benefit, I would say [was] the flexibility to schedule meetings with colleagues in Group work. It's much easier than scheduling a physical meeting...* [2020_Participant 7]

As the risk of being infected and getting sick was high during the COVID-19 pandemic, the *reduced contagion* was also perceived as a benefit by many students as they lived with their families and wished to protect or shield them or, at least, lower the transmission risk:

> *Yes, of course. But before that, I would like to add one thing that I think it hasn't been mentioned by my fellow friend is that the advantages of online class or*

> *online learning during the pandemic is that we were able to refrain from the infection for COVID cases, of course, and I do believe that in Sweden it's well managed about the cases or the infection rate is remain low, but you know the preventive measures that we stay in our home and limiting interaction that's is also one.* [2021_Participant 5]

Challenges

The challenges that students experienced were divided into 2 distinct categories: *altered social interaction* and *technical challenges*. Most of the participants perceived the *lack of interaction with teachers, classmates, and companies* as a significant challenge of online learning. They experienced that the *group feeling* and the feeling of belonging to a bigger group was missing. Some students mentioned that they needed to put *more effort into making friends* and developing collegial relationships with their classmates. Several participants mentioned the challenges with the lack of *socializing, networking, and study visits* through online learning. This was noted through the *lack of discussion* due to people being shy and not turning on their cameras and by students leaving the online lectures directly after they finished:

> *...In this case, it's an international program. And I was not living there, so I moved to this country to learn. But also like to meet new people, make friends, to expand my network. The interaction with the teachers. And as yeah, [participant] said to others you just attend the meeting, and then you close it, and you don't have this interaction discussion afterward or during the class. So that was also a point, and I also like making friends. Of course, I made a lot of friends in this program. But you like the effort was bigger, you know? So, you have to be proactive to let's meet. But in an in-person or personal program, you have more facilities...* [2020_Participant 5]

> *...another aspect that at least made me feel a bit I shouldn't say depressed, but not as happy as I was not meeting all the people in the class. Partly built on that, you should form a team over two years, and that was for me very, very obvious that I almost missed that team spirit or the team.* [2021_Participant 1]

Lack of accessibility to practical sessions on campus and *limited participation in online group work* that might have resulted in the low quality of group work were other challenges mentioned by the students:

> *...I don't have technical problems with attending online lectures...the only thing, as I mentioned before, is we had no opportunity to have someone...to have an instructor while we were doing these programming things...having online lectures is fine but having a technical lab doesn't always work online.* [2019_Participant 2]

> *...regarding the challenge of collaboration...we had like several group assignments where we needed to have a lot of discussions...that was fairly difficult because we tend to have passive collaboration, I mean like normally one or two people lead while the others just agree it's totally different from what we have in the class where everyone can jump in and share their talks during the group work.* [2021_Participant 1]

Technical challenges associated with online learning were discussed by students of all cohorts. Some students mentioned that they were worried about the reliability of their network along with some *technical challenges with their network*. For some students, it was difficult to work with Zoom in the beginning. There was also an adjustment period required to use online platforms for longer periods, such as experiencing *Zoom fatigue* from having several hours of online classes. There were additional technical problems with hybrid sessions, and some of the participants mentioned the *disturbing experience* that they had during online learning as some people spoke at the same time, which disrupted the flow of the conversation. During hybrid learning sessions, there was a lack of interaction between those online and those present in the room on campus:

> *So, in the beginning, I mean [during] the lectures sometimes there were some network issues, and then when you are the first time sitting on Zoom, and everybody has their cameras turned off.* [2019_Participant 2]

> *...Zoom fatigue because I think during that period of time, we had two or three classes in one day. Even with a normal session like from 9:00 to 3:00, we felt really exhausted and it was worsened when we had the online classes, and I was thinking that perhaps some sessions could be minimized. I mean like normally we will like 90 minutes for a single theoretical class. But if online sessions can be reduced to about 30 or 40 minutes and the rest of the time, we could do our self-learning. I think that would have been more beneficial.* [2021_Participant 1]

> *...One negative aspect I noticed was that I didn't feel comfortable interrupting to ask a question. I know I would get distracted because it's an online setting, and sometimes when two people speak at the same time, it can be very disruptive...also there are always some technical difficulties.* [2021_Participant 2]

While some students found online learning beneficial as it allowed them to concentrate better during lectures by multitasking and engaging in other activities simultaneously, others reported *difficulties in maintaining focus*. These students struggled with reduced attention spans, often due to the distractions of being at home with their families or lack of camera use during online lectures, which contributed to a sense of disengagement:

> *...Then the other thing is that I cannot focus properly because my kids are small. When I was at home, it was distracting many, many times and difficult...* [2019_Participant 3]

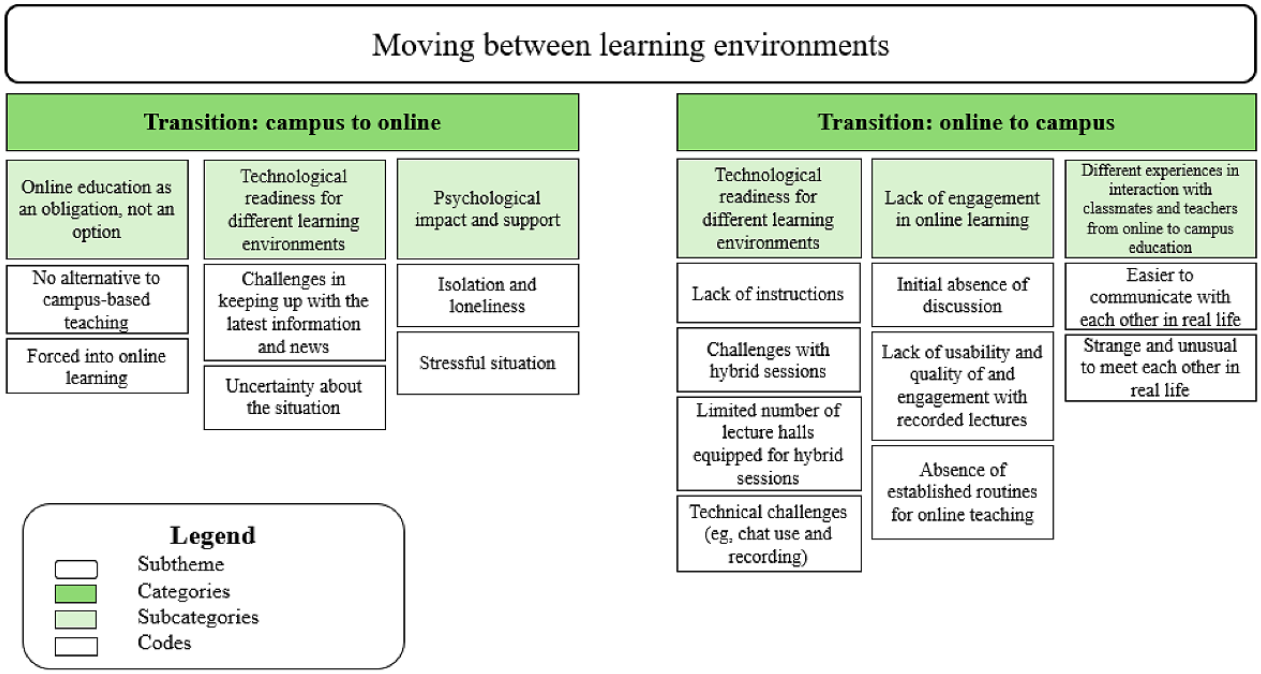
As a result, some students experienced *difficulties in understanding the concepts* and needed some assistance:

> *...Well, I think the plus point was that my husband works in technology, so I could take his help to understand the concepts and I think that was the main part that kept me going because if I don't understand things, I just...I can't focus and I can't move forward...* [2019_Participant 2]

Moving Between Learning Environments

Figure 3 illustrates the relationship among the codes, subcategories, and categories for the subtheme *moving between learning environments*.

Figure 3. A hierarchical structure displaying the subtheme and codes for theme 2—moving between learning environments.



Transition: Campus to Online Education

The COVID-19 restrictions prompted universities to shift from on-campus to online teaching. The disease spread rapidly, rendering on-campus lectures unfeasible. Numerous international students faced *challenges in keeping up with the latest information and news*. Despite the continuous updates available on the university's website, many international students experienced *uncertainty about the situation* and when everything would go back to normal:

> *...and I think that the transition between these different ways of working, I think that was hard for all of us because, well, especially for me, since I was studying before. So, I got used to online learning and then I went back to traditional learning and was very excited about that. But then it went back to online learning again and I was like, huh, OK...* [2021_Participant 2]

Working *online was an obligation rather than an option* for students, which may have made it easier to accept and manage, at least in the early stages of the pandemic. Having *no alternative to campus-based teaching* and *being forced into it* made online learning an obligation rather than a choice:

> *Well, I mean, I think there was no other choice. Everything was shut down, and there was no question of going back to campus. Nobody knew when things would reopen again, so in the beginning, we were forced into it, left with no other option. But then, I think the transition was OK.* [2019_Participant 1]

However, the psychological impact and requirement of support were important when considering the transition from on-campus to online environments. *Isolation and loneliness* and *a stressful situation* were common themes among students who did not have significant social support, such as family members living in Sweden. Most of the international students leaving their home countries to study the program in Sweden experienced that they were isolated and felt lonely during the COVID-19 pandemic:

> *...Then we changed our routines, but always we had some fear because of the unknown and what would happen next. It was always stressful, and we didn't meet our friends and teachers because from my side I don't have any relatives here, just only people from the university then I felt isolation and loneliness, and something, something alwayswas in the back of my mind and a little difficult to focus on my studies...* [2019_Participant 3]

Transition: Online to Campus Education

Technological readiness for different learning environments was a subcategory of transitioning from on-campus to an online environment and on the return to on-campus learning. At the onset of the pandemic, some students encountered challenges with Zoom due to a *lack of instructions*, particularly affecting those without a technical background. Numerous students encountered issues with audio quality during lectures and faced

challenges with hybrid sessions. The university was not fully prepared for online or hybrid formats due to the *limited number of lecture halls equipped for hybrid sessions*:

> *...I was a nurse, so I wasn't used to using it [Zoom]...No, it was difficult in the beginning to get used to using Zoom and the online campus. Nobody was explaining to you how to use it in person, so you have to do tutorials online for everything you do...* [2020_Participant 4]

> *We would have that many problems in the classroom technology-wise, like not being able to record the meetings, for instance. That was a big loss. Not being able to join functionally from home because it was not great for you. You didn't actually know what was happening in the classroom, so it was. I just wasn't expecting that. The classrooms weren't equipped to handle hybrid learning. If that was presented as an option, so.* [2020_Participant 8]

Some participants noted the *initial absence of discussion* on enhancing online learning at the onset of the COVID-19 pandemic. Nevertheless, they observed an increase in discussions on how online learning worked and how it could be improved as the pandemic progressed. In addition, at the beginning of the pandemic, technical challenges (eg, Zoom difficulties, uncertainties about chat use, and initial problems with recordings) were encountered. As the pandemic continued, these issues were addressed, leading to a gradual improvement in the online learning experience over time:

> *So, I think by the time we started [the online lectures] everything got improved especially when they started to record the lectures. Because I remember at the beginning the lecture was not recorded. But later lectures were recorded, and this was good.* [2019_Participant 2]

> *I mean for the teachers also, in the beginning, it was difficult...now if you see Zoom meetings are usually facilitated by one person who keeps an eye on the chat while the teacher is teaching. And then in between they ask questions. I mean it took a few months...and during the third semester, it became better towards the end of it. So, I think yes, I mean everybody adapted to it because that was the only way left.* [2019_Participant 1]

Other *technical challenges* included the requirement for extra tools such as headsets and screens, issues with hybrid teaching, and inadequate online content delivery, marked by a *lack of usability and quality of and engagement with recorded lectures* and an *absence of established routines for online teaching*:

> *...but that isn't very equitable for those who might not have had headphones or something like that, or maybe lived in a very small apartment, even though I live in a small apartment. Well, then maybe have a family and other people at home, maybe they didn't have any headphones, so maybe that's an aspect of the required equipment for the most optimal way of learning remotely...* [2021_Participant 2]

> *...if you maybe go and look at a lecture or recorded video on YouTube. They are very different, and they are a lot more engaged. Even though they're not live, so I do think that there should be some sort of, yeah, look over, like how you're supposed to have an online lecture to maximize learning. Because, yeah, I've, I found it quite strange that the prerecorded ones were better than the live ones.* [2021_Participant 4]

Some students encountered challenges with active participation in online and hybrid sessions, which they experienced as a lack of engagement in online learning from learners and, at times, educators. They emphasized the need for implementing strategies to enhance engagement during online sessions:

> *It lies on the lecturers. I mean, they're supposed to engage the students, and I understand from their perspective that it's so hard to engage people on a computer, because the students usually don't have their cameras on, usually, maybe even sit in their beds, listening to the lecture. And there are some lecturers who really, really try to be like: please turn on your cameras...I think that it requires maybe some standard routine for how we're supposed to do remote learning, and I mean, just like any technical product that goes into implementation, we need to have change management afterward. We need to maybe have some implementation consultants. So, both students and teachers or professors need to learn how they are supposed to teach, or for students how we are supposed to learn.* [2021_Participant 2]

Students had *different experiences in interaction with classmates and teachers between online and campus education*. Several of them noted that *it was easier to communicate with each other* during the transition from online to on-campus learning given their previous digital interaction through lectures and the WhatsApp group, although some students found it *strange and unusual to meet their classmates* in person for the first time. However, the students experienced limited interaction with teachers within the program. They found it more challenging to engage with teachers compared to their classmates:

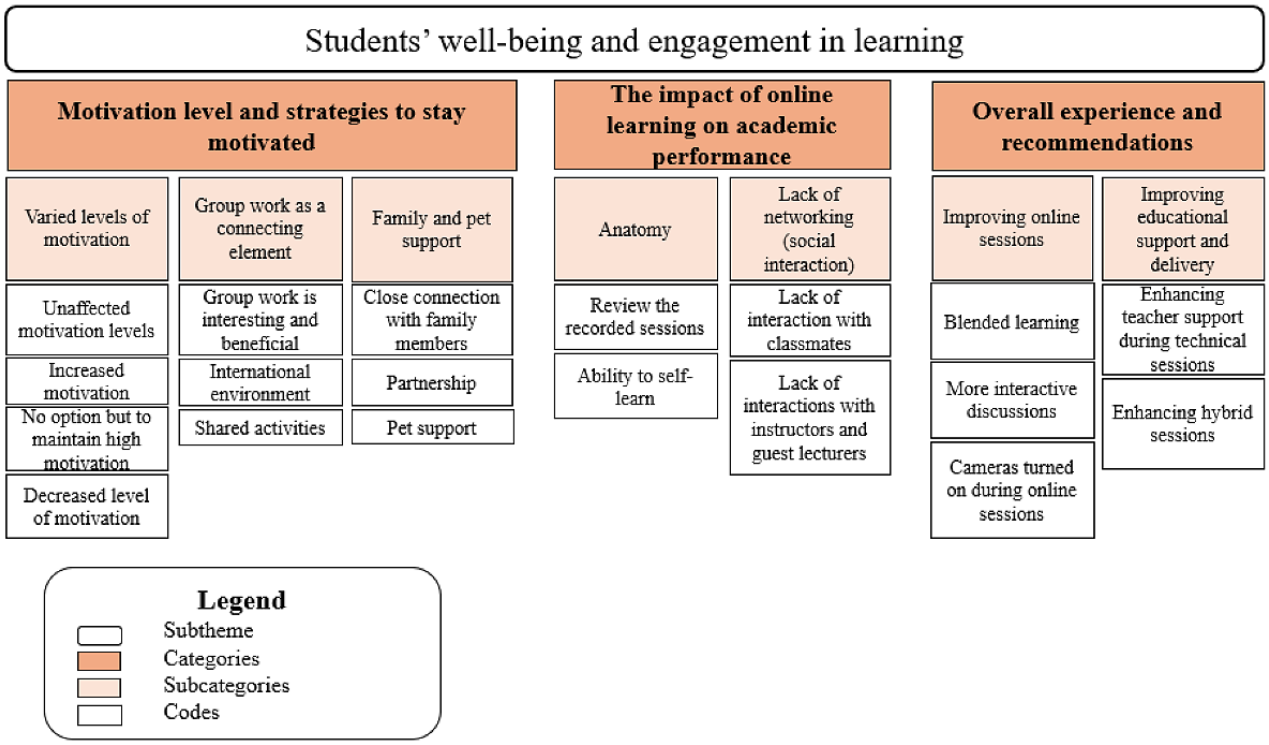
> *I think for me I can kind of remember that. Since we first started digitally and then we changed to the campus, we knew each other's faces and how we interacted. So, in a way, it was kind of easier to change...After you've trained a little bit in the digital parts, then it gives you more confidence to talk in person. On the other side maybe, I was more shy to talk to the teachers...it was harder. I felt more the distance in a way, so it's kind of the different sides, but maybe I felt nearer to my colleagues...* [2020_Participant 2]

> *I would say I mean. The first day when we met. It was, it was very weird to meet our colleagues. I mean, I'm not used to seeing them in real life. Some were shorter, some were taller, and you know it was a weird experience for the first day, I would say...* [2020_Participant 7]

Students' Well-Being and Engagement in Learning

Figure 4 illustrates the relationship among the codes, subcategories, and categories for the subtheme *students' well-being and engagement in learning.*

Figure 4. A hierarchical structure displaying the subtheme and codes for theme 3—students' well-being and engagement in learning.



Motivation Level and Strategies to Stay Motivated

The *motivationlevels* of students varied depending on their individual circumstances and also throughout the pandemic. Some students noted that their *motivation levels remained unaffected*, maintaining the same commitment to learning despite the pandemic. However, they expressed an *increased motivation* to meet people in person. Several students found that their motivation improved during the pandemic as they had more free time to plan upcoming courses, learn new skills, and engage in activities beyond their studies. Other students expressed that they had *no option but to maintain high motivation* due to visa requirements, family responsibilities, and future career plans. They highlighted that their relocation to Sweden was driven by the pursuit of a better future:

> *I would definitely say that when we moved to the campus, I felt more motivated not because of the learning, maybe because the motivation might be exactly the same, but also more, more motivated to meet people, to actually have an opportunity to meet different people because some of you have lived in Sweden and it's hard most of the time because of the difference in the cultural aspects. So, I think that's something that I remember being nice that I was happy to go to the campus and see these people and talk and learn with them.* [2020_Participant 2]

> *I think we as international students are here because you have your student visa, you have to study, and you have to pass to have your visa renewed. So, we don't have this luxury...you know, I'm bored...I don't want to study; I will skip it. No. We have to do this...We came here to study this program, and you have your plans, and you need to work to achieve your plans.* [2019_Participant 2]

> *...It's the same motivation because I brought my kids [to Sweden] and I have to make the future. That is the motivation.* [2019_Participant 3]

Nevertheless, some students reported a *decreased level of motivation*. Some related this loss of motivation to their backgrounds, particularly feeling uncomfortable with certain courses such as programming and machine learning. Others mentioned that the darkness negatively affected their mood, although it did not impact their motivation to study:

> *If I talk about the motivation, yes it was reduced because I come from a medical background, and the technical things, especially the courses like programming and machine learning...There were so many new things that sometimes I felt I was lost...Within different things, all alone, and I don't know if I'm the only one feeling like this or [if] there's somebody else, or I mean, what else do I need to make it work better? But then I didn't want to stop doing this...Because my main motivation was to learn technology and I wanted to complete it so that I know...* [2019_Participant 1]

> *I think the majority of the time that we had online lectures was during the darkest period in Sweden. And I think that affected the mood also because you then sat in front of your computer and it was dark, and then you sat inside the whole day pretty much.*

> *And then it was dark again. And so obviously that has an impact too, that it just happened to be during the most like the darkest period.* [2021_Participant 4]

Some students preferred online lectures due to the time and energy saved from morning routines such as waking up early, having breakfast, and preparing lunch. However, other students believed that lectures and interactions in person gave them more energy:

> *I really like remote learning and additionally, I'd like to add as I was saying in the beginning, I was able to be in another country because at that time my parents were living in another country and then me and [participant] actually could collaborate because we were at the same time zone.* [2021_Participant 2]

> *I will actually agree here with [participant] that also the lectures in person and interactions in person actually gave me more energy than they took, due to the fact that I was way more active. For example, I cycled to university. Then uh each morning and after class. Which always gives me an energy boost when I move more than also like just walking together on the campus, going for lunch, or getting a coffee or whatsoever. That also gave me I would stay way more energy than it took from me compared to when I would just be in my apartment. Pick up a coffee and then sit back again at the table...* [2020_Participant 1]

Group work was identified as a key factor for gaining and maintaining motivation during the pandemic. Many students believed that *engaging in group work was interesting and beneficial* throughout the program. The *international environment* played a crucial role in keeping students motivated. In addition, maintaining positive relationships with classmates and enjoying *shared activities* helped students stay motivated during the COVID-19 pandemic:

> *I would say that the group work made me more motivated. If I felt down. Honestly, we had very nice groups that we worked with. I enjoyed working with them and I learned a lot from them. But yeah, when you feel down and then we get a challenge to work on some project or some research or yeah, it was studying having some fun, and some jokes here and there. So, it wasn't 100% serious mode. So yeah, group work affected...* [2020_Participant 7]

> *If it was just entirely remote, and I never had the opportunity to meet my classmates one-on-one, I don't think I would be this motivated. Meeting them for the first time, seeing their various background, a lot of experience, and being able to tap into their own personal experiences and professional experiences gave me at least a lot of motivation, and a lot of interest in how better I can be...Or if it was just purely digital, or that involved me not traveling, I don't think I would have been this motivated.* [2021_Participant 3]

Several students noted that having close connections with their *family, a partner, or a pet* played a significant role in keeping them motivated during the pandemic:

> *Yeah, I agree. Actually, that happened to me too with my husband. He was more of a body shadowing. Like, if someone else is working on your side like you, I don't know. You get more folks. Somehow with cats, it's different though. It may not be a routine or like force it to go out. To walk the cat...So, it would actually force me to get up, play a little bit with the cat, and then come back and it was really good.* [2020_Participant 8]

The Impact of Online Learning on Academic Performance

While some students reported no change in their academic performance, others observed an improvement in their overall achievement, among other reasons due to having the opportunity to *review the recorded sessions*. The students expressed a sense of adaptability to online learning, emphasizing their *ability to self-learn*. However, the students found it challenging to grasp technical aspects solely through online learning:

> *For me, I think that the pandemic situation increased my performance. Because lectures were recorded and then I used them again and again until I understood them. And then especially for the technical part...because we are not familiar with technology, then I used again and again for all those recorded things to understand...that was very effective for me.* [2019_Participant 3]

Students experienced a *lack of interaction with classmates, instructors, and guest lecturers*, hindering the exchange of ideas and discussion of questions. This lack of networking could potentially lead to delayed job opportunities:

> *...But yes, I agree with my colleagues that some from the social side there was a lack. Having a bigger network of friends or socializing with your friends and visiting companies in person, those who were giving us lectures like different healthcare companies were giving us guest lectures, but we couldn't go in person to the companies to visit them...Because it was a pandemic, there were no, not many opportunities for summer jobs and internships, but not much open during that time. So, it was kind of that gap, which was present during [the pandemic]...* [2020_Participant 6]

Overall Experience and Recommendations

The students noted that the program altered their approach to problem-solving. However, they expressed a preference for *blended learning* over online learning alone as the latter had a negative impact on some, leading to feelings of isolation. In addition, the students provided suggestions for improvement in both online and campus-based sessions. They desired *more interactive discussions*, especially for technical sessions, to strengthen their knowledge after online lectures. Students also

believed that having *cameras turned on during online sessions* would enhance the learning experience:

> *...So overall, it was a positive experience, in my opinion. I would say I learned a new way of learning I wasn't used to, but sort of adaptability I would say. In terms of challenges, I would say I would agree with my colleagues about the social part. We were not able to socialize as compared to the normal or the offline study. It was very important. As well as building relationships with the instructors, teachers, and companies. And we were not able to do that. One other disadvantage was the interactive discussion with colleagues...* [2020_Participant 7]

> *...I can mention that not having the cameras on for example is another aspect that didn't help in concentration, also being able to socialize with people because this is a very international program, and would be amazing just to be able to have been in the classroom from the beginning to the end...* [2020_Participant 2]

> *We got to see our peers in person and kind of create a connection and sit and eat lunch with each other and talk about things to get to know each other more, which actually can not only motivate us but also affect us in our ways of learning because we can get another perspective on a certain topic or something that we wouldn't have gotten without talking to them during an informal event such as lunch, for instance...Again, I think a mix, a mix of online and on-site learning has made me at least reach my performance goals.* [2021_Participant 2]

The importance of having *teacher support* during the practical sessions in technical courses in the master's program during the pandemic highlights a key factor in improving educational support in online learning:

> *Also, we have some practical sessions for our program, so we must have someone with us in the room to ask because there are technical things in programming. I don't know where the error is in what I am doing, so I need a next eye to see what I'm doing.* [2019_Participant 2]

Enhancing hybrid sessions was mentioned as an area for improvement as students encountered numerous challenges during these sessions throughout the pandemic:

> *Yeah, I missed something. I just remembered that when we switched to in-person, there were some meetings hybrids. So people joined online while we were in the class and this was not really well organized or I don't know if it was our problem or if it could be better, but I think it's something that can be improved or just not having hybrid meetings, but I mean, I think we're not prepared and we just had like 1 computer with the meeting, so only the teacher could hear the student who was at home.* [2020_Participant 5]

Discussion

Principal Findings

This paper reports the experience of master's program students at a higher education institution due to the COVID-19 pandemic requiring adaptation to hybrid learning. Students experienced both benefits and challenges in relation to this. The transition between on-campus and online learning resulted in a hybrid learning era, which is likely here to stay. The students in this study appreciated the flexibility and autonomy provided by online learning, enabling them to integrate their studies into their daily lives. However, the rapid shift to online learning caused significant challenges, such as changes in social interactions, technical difficulties, and feelings of obligation rather than choice in adopting e-learning. The transition from on-campus to online learning caused a psychological impact on students and highlighted the need for better support and technological readiness to adapt to different learning environments. Returning to campus presented mixed experiences, with some students struggling to re-engage in face-to-face learning, whereas others faced challenges in adjusting to renewed forms of social and academic interactions. Students' motivation varied during the pandemic. Group work, family support, and having pets played crucial roles in maintaining students' moods. Despite the challenges, students were positive about hybrid learning to enhance future experiences, emphasizing the need for better networking opportunities and innovative strategies.

The findings of this study highlight implications for educators, students, and higher education institutions to embrace adaptation and foster innovation. These implications include the need for educators to stay current with the latest educational technologies and design teaching strategies and pedagogical approaches suited to both online and in-person settings to effectively foster student engagement. Students must be informed about the technological requirements for online learning and adequately prepared to meet them. Institutions play a critical role in ensuring equitable access to technology, guiding and supporting educators in adopting innovative tools and methods, and offering mental health resources to assist students in overcoming the challenges of evolving educational environments.

Comparison to the Literature

Studies have previously reported both positive and negative consequences of the shift in learning environments. Naciri et al [10] have suggested that students generally responded positively to the rapid shift to online health science education during this crisis, expressing views on aspects such as acceptance, motivation, and engagement. Despite varying socioeconomic conditions across countries, certain key factors such as access to technology, basic computer literacy, well-designed online course pedagogy, and flexibility in learning consistently supported online education. However, students encountered challenges such as inconsistent internet access, difficulties with educational platforms, and hurdles in acquiring clinical skills online. These insights are crucial for enhancing the integration of these technologies into educational frameworks [10]. This is congruent with our results, where

overall the students responded well to the rapid shift to online learning but reported both positive and negative outcomes, with challenges especially in computer laboratory sessions. Our findings also confirm the results of other studies [9,11] on the difficulties regarding limited internet access, technical problems, challenges related to adjustments from traditional to online formats, and impact on students' well-being.

Students in our study were not entirely satisfied with the practical sessions in technical courses, highlighting the need for access to teacher assistance during these sessions. This contrasts with a study that reported satisfaction with clinical teaching and practical sessions remaining adequately high during the shift to e-learning. They also reported notable improvement in student satisfaction concerning course structure, instructor expertise, learning materials, and overall contentment with the courses, as well as a tendency for student grades to improve in the online format [8].

A review study [20] revealed that motivation and self-regulated learning were significant challenges, impacting students' ability to engage critically with the material. It also showed varied attitudes toward online learning, with decreased satisfaction and emotional well-being in many students due to feelings of isolation and increased stress. These findings align with our results as students in this study also experienced loneliness and isolation, which led to a lack of focus on the studies.

Findings regarding the use of technology show that competence, perceived usefulness, ease of use, and facilitated implementation are predictors of learners' attitudes toward and intentions to use technology [18,21,22]. Technologically capable students are likely to associate poor digital implementation by educators with lower satisfaction and self-efficacy. Integration of technology with a student-centered focus is likely to influence the development of autonomy and self-regulated learning [23,24]. For students enrolled in a health informatics program, where technology is a significant aspect of the learning curriculum, it could be extrapolated that they would likely be comfortable and be able to adapt to the use of technologies for learning. Therefore, the reported themes associated with technology use may be more significant for other groups of learners in different educational fields.

In addition to these findings, there has been a significant association reported between instructors' use of effective teaching practices and student motivation. Alongside the quality of teaching and challenges associated with technology, prepandemic studies have indicated that motivation in online learning can be affected by demographic characteristics such as age, gender, employment status, income, and family obligations [22]. However, most studies were conducted at higher education institutions where students had a choice to enroll in online learning; learning in a hybrid environment may be different.

Online laboratory sessions in technical courses have shown some drawbacks in managing requests for help from student groups. During in-person sessions, teaching assistants can manage the requests of a group but, at the same time, give feedback to others (spontaneous interactions or questions that require a short answer). Using online learning tools with students divided into rooms, requests are managed only sequentially without allowing for spontaneous interactions. In addition, managing the request queue was more challenging for teaching assistants due to the unavailability of a specific function in the online learning tools. This seems to be in accordance with previous research on the topic [25-27].

Implications for the Educators, Students, and Higher Education Institutions

The transition to online learning during the COVID-19 pandemic had significant implications for educators, students, and universities. Both educators and students had to make substantial adjustments in their teaching methods and learning styles, respectively. This study highlighted a greater emphasis on active learning strategies, self-discipline and time management, active participation in discussions and group work, and the development of flexibility and resilience.

Although our study focused on the experiences of students, the importance of educators in both the design and delivery of hybrid learning must be recognized. An online replication of a physical classroom can result in bored students and exhausted teachers, with both experiencing *Zoom fatigue*. A change in pedagogy is required to teach in the hybrid learning environment. Hybrid pedagogy is a development from blended learning using elements of both online and face-to-face learning, resulting in no separation between learners in the physical or digital space [13].

A significant issue related to hybrid learning is the required use of technologies. The health informatics students recognized both individual and organizational gaps in infrastructure and a lack of technological preparedness, which was experienced by many institutions. After the pandemic, it is critical that higher education institutions ensure that technology is leveraged to provide students with a good online, physical, or hybrid experience. The pandemic has highlighted the importance of preparedness and resilience in higher education for future crises. To support online and hybrid education, institutions need to invest in robust technological infrastructure and prioritize continuous professional development and training for educators. As highlighted in our study and similar research [9,10,28,29], there is a significant need to equip educators with the skills to adapt to the evolving educational landscape, particularly in the use of digital technologies. Educators need to develop skills to effectively use online platforms and tools to create engaging and interactive learning experiences. Common themes when designing quality online instruction that engages and motivates students are those of interaction, collaboration, communication, and discussion [21,30]. Online and hybrid learning require a greater focus on interactivity; this helps break the monotony and allows for student socialization. Therefore, engaging in ongoing professional development to stay updated with the latest educational technologies is a necessity for educators. Those students who had previously had the opportunity to meet in person on campus before transitioning to online learning may have had a closer sense of community than those whose educational experience transitioned in the opposite direction. Therefore, it is vital for instructors to build and foster a sense of community to keep students motivated. Overall, the onus is

on educators to design and deliver quality teaching appropriate to the environment (web based or in person) to promote student engagement and encourage motivation to learn. To support these efforts, it is equally important that students are informed in advance of the requirements to attend classes online and be advised on the necessary technology (such as headphones, microphone, and camera). Furthermore, institutions should ensure access to technology by providing it on a short-term loan basis or guiding students in applying for grant funding, thereby ensuring equitable access to learning resources. Institutions should also guide and support educators to learn about and incorporate new technologies into their teaching practices [31-33]. In addition, providing extra resources and support to help students navigate the challenges of online learning is crucial. Given the impact that the shift had on students' mental health, higher education institutions should provide mental health resources and support systems to assist students in managing these challenges. Students, educators, and institutions play a crucial role in informing policy makers regarding challenges and implications of future crises by sharing best practices for managing education emergencies. Providing feedback, sharing research findings, and offering concrete examples are essential for ensuring that policy makers are well informed and able to respond to the evolving needs of the educational community.

Strengths and Limitations of This Study

This study is limited by its context-specific nature, focusing on a particular group of students within a specific educational environment (health informatics master's students at KI), which may not fully capture the diverse experiences of learners in other regions, disciplines, or institutional settings. This may limit the generalizability of the results. However, the recommendations in the results can apply to other programs and institutions with similar educational systems and resources. The sample size in this study could be considered a limitation. However, saturation was achieved across the entire sample as no additional new themes emerged during the data analysis process. Although the number of participants from the 2019 cohort was low, the insights gathered from this group were consistent with those gathered from students from the other cohorts.

Conclusions

The shift to hybrid learning in response to the COVID-19 pandemic presented both benefits and challenges for postgraduate students in higher education institutions. Technological preparedness, equitable access to technology, and educator training are crucial factors that institutions must address to support students' learning experiences. In addition, fostering interaction, collaboration, communication, and discussion in online and hybrid learning environments is essential for engaging and motivating students. Ultimately, educators play a key role in designing and delivering quality teaching that promotes student engagement and encourages motivation to learn regardless of the learning environment.

Future studies should aim to explore the impact of the COVID-19 pandemic on students' learning experiences across varied contexts, incorporating cross-institutional comparisons to develop a more comprehensive understanding of how different factors influence learning experiences and outcomes in hybrid and online education. In addition, it is of great importance that future studies examine new teaching methods and pedagogical approaches that can enhance students' engagement and increase communication and interaction between students in hybrid education. Finally, it is crucial to develop strategies and policies that prepare higher education institutions for crises, ensuring that they can effectively transition between educational environments—whether shifting from in-person to online or hybrid modes—while maintaining educational continuity and quality. Future research could use case studies to investigate institutional responses and apply principles of action research to collaboratively develop and refine teaching strategies and crisis management policies.

Acknowledgments

The authors would like to thank all master's students and alumni who participated in this study for their time and valuable insights. The authors received no specific funding for this work. Artificial intelligence technology such as Grammarly (Grammarly Inc) and ChatGPT-3.5 (OpenAI) was used in the preparation of this manuscript to harmonize the text. After using these tools, the authors reviewed and edited the content as needed, and they take full responsibility for the content of the publication.

Authors' Contributions

ND, SB, and NS were involved in the design of the study. These authors prepared the interview guide and piloted it. The focus groups were held online with CS in the role of interviewer. ND, NS, and SB conducted the initial coding. CS joined the analytical process once the initial coding and subcategories had been generated. After identifying subcategories, a comparative analysis was conducted to highlight similarities and differences among the student groups. We then collaboratively categorized and identified subthemes for all 3 groups. All authors contributed to the subsequent writing and review of the manuscript.

Conflicts of Interest

None declared.

References

1. Li D. The shift to online classes during the COVID-19 pandemic: benefits, challenges, and required improvements from the students' perspective. Electron J E Learn 2022 Jan 21;20(1):pp1-p18. [doi: 10.34190/ejel.20.1.2106]
2. Ruiz JG, Mintzer MJ, Leipzig RM. The impact of E-learning in medical education. Acad Med 2006 Mar;81(3):207-212. [doi: 10.1097/00001888-200603000-00002] [Medline: 16501260]
3. Barrot JS, Llenares II, Del Rosario LS. Students' online learning challenges during the pandemic and how they cope with them: the case of the Philippines. Educ Inf Technol (Dordr) 2021 May 28;26(6):7321-7338 [FREE Full text] [doi: 10.1007/s10639-021-10589-x] [Medline: 34075300]
4. Lederman D. Most teaching is going remote. Will that help or hurt online learning? Inside Higher Ed. URL: https://www.insidehighered.com/digital-learning/article/2020/03/18/most-teaching-going-remote-will-help-or-hurt-online-learning [accessed 2024-04-29]
5. Lin Y, Nguyen H. International students' perspectives on e-learning during COVID-19 in higher education in Australia: a study of an Asian student. Electron J E Learn 2021 Aug 11;19(4):241-251. [doi: 10.34190/ejel.19.4.2349]
6. Wu Z. How a top Chinese university is responding to coronavirus. World Economic Forum. URL: https://www.weforum.org/agenda/2020/03/coronavirus-china-the-challenges-of-online-learning-for-universities/ [accessed 2024-04-29]
7. Worth D. Coronavirus: how to maximise distance learning. Tes Global Ltd. URL: https://www.tes.com/magazine/article/coronavirus-how-maximise-distance-learning [accessed 2024-04-29]
8. Alenezi S, Bahathig A, Soliman M, Alhassoun H, Alkadi N, Albarrak M, et al. Performance and satisfaction during the E-learning transition in the COVID-19 pandemic among psychiatry course medical students. Heliyon 2023 Jun;9(6):e16844 [FREE Full text] [doi: 10.1016/j.heliyon.2023.e16844] [Medline: 37303529]
9. Alsoufi A, Alsuyihili A, Msherghi A, Elhadi A, Atiyah H, Ashini A, et al. Impact of the COVID-19 pandemic on medical education: medical students' knowledge, attitudes, and practices regarding electronic learning. PLoS One 2020;15(11):e0242905 [FREE Full text] [doi: 10.1371/journal.pone.0242905] [Medline: 33237962]
10. Naciri A, Radid M, Kharbach A, Chemsi G. E-learning in health professions education during the COVID-19 pandemic: a systematic review. J Educ Eval Health Prof 2021 Oct 29;18:27 [FREE Full text] [doi: 10.3352/jeehp.2021.18.27] [Medline: 34710319]
11. Kumar A, Sarkar M, Davis E, Morphet J, Maloney S, Ilic D, et al. Impact of the COVID-19 pandemic on teaching and learning in health professional education: a mixed methods study protocol. BMC Med Educ 2021 Aug 19;21(1):439 [FREE Full text] [doi: 10.1186/s12909-021-02871-w] [Medline: 34412603]
12. Nuryana Z, Xu W, Kurniawan L, Sutanti N, Makruf SA, Nurcahyati I. Student stress and mental health during online learning: potential for post-COVID-19 school curriculum development. Compr Psychoneuroendocrinol 2023 May;14:100184 [FREE Full text] [doi: 10.1016/j.cpnec.2023.100184] [Medline: 37038597]
13. Haikalis M, Doucette H, Meisel MK, Birch K, Barnett NP. Changes in college student anxiety and depression from pre- to during-COVID-19: perceived stress, academic challenges, loneliness, and positive perceptions. Emerg Adulthood 2022 Apr 27;10(2):534-545 [FREE Full text] [doi: 10.1177/21676968211058516] [Medline: 35382515]
14. Lederman D. The shift to remote learning: the human element. Inside Higher Ed. URL: https://www.insidehighered.com/digital-learning/article/2020/03/25/how-shift-remote-learning-might-affect-students-instructors-and [accessed 2024-04-29]
15. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
16. Collins III JW, O'Brien NP. The Greenwood Dictionary of Education. Santa Barbara, CA: Greenwood Press; 2003.
17. Gundumogula M. Importance of focus groups in qualitative research. Int J Humanit Soc Sci 2020 Nov 30;8(11):299-302. [doi: 10.24940/theijhss/2020/v8/i11/hs2011-082]
18. Graneheim UH, Lundman B. Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. Nurse Educ Today 2004 Feb;24(2):105-112. [doi: 10.1016/j.nedt.2003.10.001] [Medline: 14769454]
19. Saladana J. The Coding Manual for Qualitative Researchers. 2nd edition. Thousand Oaks, CA: Sage Publications; 2013.
20. Salas-Pilco SZ, Yang Y, Zhang Z. Student engagement in online learning in Latin American higher education during the COVID-19 pandemic: a systematic review. Br J Educ Technol 2022 May 15;53(3):593-619 [FREE Full text] [doi: 10.1111/bjet.13190] [Medline: 35600418]
21. Munday D. Hybrid pedagogy and learning design influences in a higher education context. Stud Technol Enhanc Learn 2022;2(2):25 [FREE Full text] [doi: 10.21428/8c225f6e.b5af8bae]
22. Hanham J, Lee CB, Teo T. The influence of technology acceptance, academic self-efficacy, and gender on academic achievement through online tutoring. Comput Educ 2021 Oct;172:104252. [doi: 10.1016/j.compedu.2021.104252]
23. Franchi T. The impact of the COVID-19 pandemic on current anatomy education and future careers: a student's perspective. Anat Sci Educ 2020 May 05;13(3):312-315 [FREE Full text] [doi: 10.1002/ase.1966] [Medline: 32301588]
24. Mukhtar K, Javed K, Arooj M, Sethi A. Advantages, limitations and recommendations for online learning during COVID-19 pandemic era. Pak J Med Sci 2020 May 18;36(COVID19-S4):S27-S31. [doi: 10.12669/pjms.36.covid19-s4.2785]
25. McKnight K, O'Malley K, Ruzic R, Horsley MK, Franey JJ, Bassett K. Teaching in a digital age: how educators use technology to improve student learning. J Res Technol Educ 2016 May 21;48(3):194-211. [doi: 10.1080/15391523.2016.1175856]

26. Austen L, Parkin HJ, Jones-Devitt S, McDonald K, Irwin B. Digital capability and teaching excellence: an integrative review exploring what infrastructure and strategies are necessary to support effective use of technology enabled learning (TEL). Sheffield Hallam University Research. URL: https://shura.shu.ac.uk/13750/1/Digital-capability-and-teaching-excellence-2016.pdf [accessed 2024-04-29]
27. May D, Morkos B, Jackson A, Hunsu NJ, Ingalls A, Beyette F. Rapid transition of traditionally hands-on labs to online instruction in engineering courses. Eur J Eng Educ 2022 Mar 05;48(5):842-860. [doi: 10.1080/03043797.2022.2046707]
28. Rashid S, Yadav SS. Impact of COVID-19 pandemic on higher education and research. Indian J Hum Dev 2020 Aug 23;14(2):340-343. [doi: 10.1177/0973703020946700]
29. Jakoet-Salie A, Ramalobe K. The digitalization of learning and teaching practices in higher education institutions during the Covid-19 pandemic. Teach Public Adm 2022 Apr 21;41(1):59-71. [doi: 10.1177/01447394221092275]
30. Teodorescu D, Aivaz KA, Amalfi A. Factors affecting motivation in online courses during the COVID-19 pandemic: the experiences of students at a Romanian public university. Eur J High Educ 2021 Aug 30;12(3):332-349. [doi: 10.1080/21568235.2021.1972024]
31. Turnbull D, Chugh R, Luck J. Transitioning to e-learning during the COVID-19 pandemic: how have higher education institutions responded to the challenge? Educ Inf Technol (Dordr) 2021 Jun 23;26(5):6401-6419 [FREE Full text] [doi: 10.1007/s10639-021-10633-w] [Medline: 34177349]
32. El Said GR. How did the COVID-19 pandemic affect higher education learning experience? An empirical investigation of learners' academic performance at a university in a developing country. Adv Hum Comput Interact 2021 Feb 8;2021:1-10. [doi: 10.1155/2021/6649524]
33. Oliveira G, Grenha Teixeira J, Torres A, Morais C. An exploratory study on the emergency remote education experience of higher education students and teachers during the COVID-19 pandemic. Br J Educ Technol 2021 Jul 18;52(4):1357-1376 [FREE Full text] [doi: 10.1111/bjet.13112] [Medline: 34219758]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
KI: Karolinska Institutet

Edited by B Lesselroth; submitted 27.06.24; peer-reviewed by H Abuhassna, H Gray; comments to author 10.11.24; revised version received 13.12.24; accepted 02.01.25; published 10.02.25.

Please cite as:
Davoody N, Stathakarou N, Swain C, Bonacina S
Exploring the Impact of the COVID-19 Pandemic on Learning Experience, Mental Health, Adaptability, and Resilience Among Health Informatics Master's Students: Focus Group Study
JMIR Med Educ 2025;11:e63708
URL: https://mededu.jmir.org/2025/1/e63708
doi:10.2196/63708
PMID:

Digital Dentists: A Curriculum for the 21st Century

Michelle Mun[1,2], DDS; Samantha Byrne[1], PhD; Louise Shaw[2], PhD; Kayley Lyons[2], PhD

[1]Melbourne Dental School, Faculty of Medicine, Dentistry and Health Sciences, The University of Melbourne, 720 Swanston Street, Melbourne, Australia

[2]Centre for Digital Transformation of Health, Faculty of Medicine, Dentistry and Health Sciences, The University of Melbourne, Melbourne, Australia

Corresponding Author:
Michelle Mun, DDS
Melbourne Dental School, Faculty of Medicine, Dentistry and Health Sciences, The University of Melbourne, 720 Swanston Street, Melbourne, Australia

Abstract

Future health professionals, including dentists, must critically engage with digital health technologies to enhance patient care. While digital health is increasingly being integrated into the curricula of health professions, its interpretation varies widely depending on the discipline, health care setting, and local factors. This viewpoint proposes a structured set of domains to guide the designing of a digital health curriculum tailored to the unique needs of dentistry in Australia. The paper aims to share a premise for curriculum development that aligns with the current evidence and the national digital health strategy, serving as a foundation for further discussion and implementation in dental programs.

(JMIR Med Educ 2025;11:e54153) doi:10.2196/54153

KEYWORDS

digital health; digital transformation; informatics; ehealth; dentistry; dental informatics; curriculum; competence; capability; dental education

Introduction

As the world continues to be digitally transformed, there are increasing expectations for health care providers to use technology and handle health information safely and ethically [1]. It is likely that future health professionals will also need to think critically about how digital health technologies can be used to transform models of care [2].

Digital health and informatics remain relatively new curriculum topics for many health professions, including dentistry. Defining the relevant curricular objectives in entry-to-practice degrees can be particularly challenging for several reasons. First, there are several definitions and conceptualizations of the term “digital health” [1,3-6]. Second, the implementation of digital health education in health profession degrees has largely been ad hoc, with different schools adopting varied approaches [7]. This has resulted in inconsistent learning outcomes and a fragmented understanding of digital health competencies for health profession graduates. Third, although a multitude of digital health competency frameworks exist [8-11], there is a notable absence of shared curriculum models specific to dentistry. Therefore, dentistry educators are not aware of or struggle to adopt best practices in the teaching of digital health.

In this viewpoint, we argue that there is a need to integrate digital health education into dentistry curricula to prepare future practitioners for the increasingly digitized health care environment. Specifically, we propose a distinct point of view for defining "digital health" in dental education, a structured set of domains to guide the design of digital health curriculum, and a framework for curriculum development that aligns with current evidence.

Beginning With the End in Mind

In Australia, a country that ranks consistently high for digital health maturity [12], clear digital health objectives are set out through national strategies such as the Australian Digital Health Agency (ADHA) Capability Action Plan (2023) and the National Healthcare Interoperability Plan (2023 - 2028) [1,13]. The government body for digital health (ie, the ADHA), peak bodies such as the Australasian Institute of Digital Health (AIDH), and digital health innovation centers all identify building workforce capability as a critical part of achieving digital transformation of health care [1,13-15]. These organizations envision a future where health professionals will work in integrated and multidisciplinary environments. Digital health education in entry-to-practice degrees is thus a core element of advancing workforce capability; however, the specific content, including priority areas of knowledge and skills, must be tailored to the unique demands of each discipline and local context. For dentistry in Australia, this means aligning the curriculum with the national digital health strategy while addressing the current maturity level and future needs of the dental profession.

Reframing "Digital Health" for the Next Generation of Australian Dental Practitioners

A lack of standardized digital health education in entry-to-practice degrees in Australia has been recognized for decades [16-18] and has been demonstrated by gaps in workforce competency [19]. Global interest in the digital transformation of health care is accelerating, catalyzed by the COVID-19 pandemic and advances in artificial intelligence (AI). However, not all health professions have advanced equally considering their digital transformation. In Australia, as in many countries, the dental sector remains traditionally siloed from the rest of the health care system and faces fragmentation in its information systems [20]. In the dental sector, the most progress in digital health has been observed in restorative and surgical procedures, where technology is directly integrated into clinical workflows [21]. For example, conventional manual techniques and laboratory workflows for the design and fabrication of dental restorations have evolved into in-house, fully digitized workflows with the application of intraoral scanners and chairside milling machines [22]. In contrast, dentistry has a relatively nascent data culture [23], with less focus on the broader scope of digital health, which we define in this viewpoint to encompass virtual care, remote monitoring, mobile health (mHealth), wearables, big data analytics, platforms, and "the exchange of data and sharing of relevant information across the health ecosystem creating a continuum of care" [4].

Developing digital health–capable dentists thus involves more than simply teaching the technical aspects of digital tools used in service delivery; it requires a shift towards understanding how digital data can inform clinical decisions, enhance patient care, and contribute to system-wide improvements. This conceptual change is crucial for moving from a service-focused practice to one that leverages digital health as an integral part of modern dental care. The Learning Health Systems (LHS) framework [24] is one example of how to help dental professionals characterize digital health. LHS are health care environments where science, informatics, incentives, and culture align to promote continuous improvement and innovation. In these systems, best practices are embedded in care, patients actively participate, and new knowledge is generated from every care experience [25]. Building on this vision, dental education should emphasize models where digital health is central to both practice and continuous improvement. This approach will foster digital health capability by cultivating a deeper understanding of how and why digital health technologies enable the delivery of high-quality, safe, and sustainable care.

Textbox 1 outlines the questions that can guide the development of a digital health curriculum for entry-to-practice dental education. These questions are intended to help educators and curriculum developers define clear goals aligned with the specific needs of the discipline and the local health care context.

Textbox 1. Defining the goals of digital health curriculum for entry-to-practice degrees.

- What is outlined in national and local digital health strategies for the next 5-10 years? What does the political and funding environment look like?
- What are the digital health-related accreditation standards of the profession?
- What does the current and future digital health maturity of the primary work environments of your graduates look like?
  - Consider the difference in goals for:
    - A rural school where graduates may work in areas with limited digital maturity
    - A health discipline or specialty where graduates will typically work in tertiary care rather than primary care

Considerations for Curriculum Development

The Australian Dental Council (ADC) recently revised its competencies for newly qualified dental practitioners; they updated the requirement to include "using digital technologies and informatics to manage health information and inform person-centred care" [26]. This prompted the authors to develop a digital health curriculum to be implemented in a higher education institution that has graduating dental professionals in Australia. As per the best practice in curriculum development [27], we considered the existing digital health competency and capability frameworks as part of our curricular needs assessment. An environmental literature scan found that only a few frameworks had been created specifically for dentistry or involved dental experts in their consultations, reflecting a lag in dentistry's digital health participation (Multimedia Appendix 1). As a result, not all topics in these existing frameworks were relevant or current to the reality of training dental professionals in Australia, who will predominantly work in small clinics in primary care, in practices with varying digital health maturity [28,29]. An exception was the digital dentistry curriculum proposed by the American College of Prosthodontists [30], which is well-researched but focused solely on digital skills for prosthodontics. This highlighted a gap in resources to support the broader skill set of graduating dentists in Australia, as outlined by the ADC.

The process of designing higher education courses aims to align industry standards with a scaffolded approach for developing effective learning outcomes that produce work-ready graduates. While the ADC's revised competency served as a catalyst for curriculum development, our efforts extended beyond the ADC's scope to meet standards such as those overseen by the Tertiary Education Quality and Standards Agency (TEQSA), which

performs the quality assurance checks for all participants, delivered as part of higher education in Australia. TEQSA's emphasis on authenticity in curricula design, as well as contemporary leading practice [31,32], influenced our approach towards designing a curriculum that not only meets regulatory competencies but also prepares students for practical, professional challenges in the evolving digital health landscape.

Finally, a critical component of our approach was to tailor the curriculum to the local context. While internationally recognized informatics competencies [33] often underpin digital health capability frameworks, they do not alone fully capture the breadth and nuances of digital health proficiency. Digital health encompasses a range of skills, including digitally enabled clinical processes, care pathways, and behavior change management, all of which are shaped by local variations in digital health maturity and sociocultural contexts. Furthermore, curriculum development often occurs under significant time and resource constraints, requiring an approach that is rigorous but targeted. For example, rural schools may not yet prioritize AI competencies if electronic health records are not yet in use locally.

Key Domains

Two frameworks were selected to inform the development of the dental digital health curriculum, both of which are government-sponsored, peer-reviewed, and directly relevant to the Australian setting [Textbox 2]. The domains in Table 1 are an abridged synthesis created by the authors, drawing on elements from the two selected frameworks. This reimagined structure is intended to facilitate the development of a digital health curriculum for dentistry, aligning learning objectives, instruction, and assessment with the national strategy in Australia.

Textbox 2. Frameworks selected to inform development of the dental digital health curriculum.

1. Framework 1 (2018): eHealth Capabilities Framework for Graduates and Health Professionals [34]. This framework was developed by the University of Sydney and eHealth New South Wales, consisting of a tri-phase literature review, focus groups with faculty and government representatives (n=23), and a Delphi method refinement with 4 iterations. The framework is structured in 4 domains and describes recommended knowledge and skills for health professions graduates in digital health.
2. Framework 2 (2021): Digital Health Capability Framework for Allied Health Professionals [35]. This framework was developed by the Department of Health, Victoria, and consisted of a 3-part development program including a competency framework review, expert discussion panel interviews (n=28), and an online survey of Victorian allied health professionals (n=164). This document draws from Framework 1 and is similarly structured into 4 domains of 3-6 subdomains, with the addition of levels of digital health proficiency ranging from Foundation, Consolidation, Expert, and Leadership.

Table . Domains and goals for digital health curriculum for an entry-to-practice dental degree.

Domain	Learning goal	Suggested learning topics
1. Digital transformation of health	Newly graduated dental practitioners will actively lead the digital transformation of dentistry by using technology to deliver patient-centred care and by recognizing the role of data and analytics in improving it.	Electronic health records, digital dentistry (radiography, intraoral scanning, CAD/CAM[a], and other digital workflows) data, interoperability and learning health systems, artificial intelligence
1. Legislation, policy, and governance	Newly graduated dental practitioners will drive improvements in the privacy and security of patient data, and model the safe, ethical, and responsible use of digital health technologies in the dental practice.	Data privacy and cybersecurity
1. Digital health for patients	Newly graduated dental practitioners will promote patient engagement in health care, prescribe appropriate digital resources, and support digital health literacy.	Digital health literacy, patient engagement in health care, and digital health equity
1. Digital professionalism	Newly graduated dental practitioners will model a professional and appropriate digital identity.	Social media and digital professionalism

[a]CAD/CAM: computer-aided design/computer-aided manufacturing.

The first domain recognizes that along with technical proficiency in digital clinical workflows, dental practitioners must be able to think in multidisciplinary terms of the flow of data and information across health care [13]. Dental practitioners must understand the importance of informatics, interoperability, and a quality improvement mindset to be the building blocks for creating LHS [24,25].

The second domain recognizes the role of the dental practitioner in safe and ethical governance of patient data across digital workflows, noting that health care is the consistently top-reporting sector for data breaches in Australia [36].

The third domain recognizes the shift from the paternalistic model of health care towards a person-centered one where the person receiving care plays an active role in shared health care decision-making. The OpenNotes mandate in the US is a good

example of this [37]. This domain is also particularly relevant to the rapid pace of AI development and the accessibility of generative AI models that patients may use to access health (mis)information. Dental practitioners must understand digital health literacy; how patients may engage with digital health technologies and services; and the uses, ethics, benefits, and risks of AI in health care.

The fourth domain recognizes that dental practitioners must develop a professional identity, which is multidimensional across social media and the internet. The obligation for a dental practitioner to uphold their professional code of conduct is binding for both their in-person and digital profiles [38].

This holistic overview of digital health in dentistry is a step towards addressing the observation that digital health education tends to be focused on medical degrees—mostly in electives or single-unit areas such as telehealth—and in utilizing diverse approaches for delivery, development, and assessment [39]. A similar observation was found during our curricular needs assessment, revealing a strong focus in single content areas such as telehealth and digital dentistry, but confirming opportunities to facilitate a more coordinated and comprehensive learning pathway to support full digital health competency.

Final Thoughts

Viewing dentistry through a "digital health" lens may seem like a small matter. However, the change in perspective for dental educators is important. Dentistry has traditionally focused on individual patient care and procedural intervention, but contemporary health care is increasingly shaped by system-level forces. AI, interoperability, value-based care, and increasing consumer participation are now current realities [40-43]. The potential for digital health to drive meaningful systemic improvements in oral health and health care cannot be truly realized without first building the necessary capability at the graduate level. Consequently, these topics can and should be taught in a structured manner in entry-to-practice dental education.

Critically, although newer generations are often seen as digitally adept, they do not automatically master the necessary digital skills simply from being exposed to technology [44]. This gap in digital competency underscores the importance of intentional curriculum design. Universities are increasingly using the approach of constructive alignment to enhance outcome-based education [45], and this approach should be used to design a longitudinal digital health curriculum that can align with the intended graduate attributes.

This viewpoint has outlined the premise for designing a digital health curriculum in dentistry, using a structured set of domains based on current evidence and adapted to the Australian context. The proposed domains provide a foundation for educators to build a curriculum that aligns with the unique needs of dental professionals and the national strategy for digital health. This approach is intended for integration into the University of Melbourne's dentistry program and aims to encourage the further development and discussion of digital health education within dental programs, both nationally and globally.

Conclusion

It can be difficult for educators to define digital health curriculum that is both evidence-based and relevant to their discipline and local context; to design it is to predict the future. However, keeping pace involves changing our view of digital health in dentistry. A common understanding about the language of digital health is important for developing health professionals who will be able to navigate the environment of the modern health care system. We found that existing digital health capability frameworks were useful to define a view of digital health across an entry-to-practice dental degree, and high level roadmaps and frameworks are valuable to envision a future-ready dental graduate who can embrace the next wave of digital transformation. This perspective will be useful for developing the curriculum aligned with the national vision of building workforce capability and realizing the aim of safe, connected care.

Acknowledgments

Our sincere thanks to Alastair Sloan, Melbourne Dental School's Head of School, and the Centre for Digital Transformation of Health for supporting this educational initiative.

Authors' Contributions

Conceptualization: MM
Supervision: SB, KL
Writing – original draft: MM
Writing – review & editing: MM, SB, LS, KL

Conflicts of Interest

None declared.

Multimedia Appendix 1
Digital health capability and competency frameworks considered for curriculum development in dentistry.
[DOCX File, 22 KB - mededu_v11i1e54153_app1.docx]

References

1. Australian Digital Health Agency. National digital health capability action plan. 2022. URL: https://www.digitalhealth.gov.au/sites/default/files/documents/national-digital-health-capability-action-plan.pdf [accessed 2023-08-31]
2. Kraus S, Schiavone F, Pluzhnikova A, Invernizzi AC. Digital transformation in healthcare: Analyzing the current state-of-research. J Bus Res 2021 Feb;123:557-567. [doi: 10.1016/j.jbusres.2020.10.030]
3. Fatehi F, Samadbeik M, Kazemi A. What is Digital health? Review of definitions. Stud Health Technol Inform 2020 Nov 23;275:67-71. [doi: 10.3233/SHTI200696] [Medline: 33227742]
4. World Health Organization. Global strategy on digital health 2020–2025. 2021. URL: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2023-08-31]
5. European Commission. EHealth: digital health and care: overview. 2017. URL: https://health.ec.europa.eu/ehealth-digital-health-and-care/overview_en [accessed 2024-09-19]
6. U.S. Food & Drug Administration. What is digital health?. 2020. URL: https://www.fda.gov/medical-devices/digital-health-center-excellence/what-digital-health [accessed 2024-09-19]
7. Gray K, Dattakumar A, Maeder A, Butler-Henderson K, Chenery H. Advancing eHealth education for the clinical health professions. 2014. URL: http://clinicalinformaticseducation.pbworks.com/w/file/fetch/74500403/PP10_1806_Gray_report_2014.pdf [accessed 2024-07-26]
8. Brice S, Almond H. Health professional digital capabilities frameworks: a scoping review. J Multidiscip Healthc 2020;13:1375-1390. [doi: 10.2147/JMDH.S269412] [Medline: 33173300]
9. Longhini J, Rossettini G, Palese A. Digital health competencies among health care professionals: systematic review. J Med Internet Res 2022 Aug 18;24(8):e36414. [doi: 10.2196/36414] [Medline: 35980735]
10. Jimenez G, Spinazze P, Matchar D, et al. Digital health competencies for primary healthcare professionals: a scoping review. Int J Med Inform 2020 Nov;143:104260. [doi: 10.1016/j.ijmedinf.2020.104260] [Medline: 32919345]
11. Nazeha N, Pavagadhi D, Kyaw BM, Car J, Jimenez G, Tudor Car L. A digitally competent health workforce: scoping review of educational frameworks. J Med Internet Res 2020 Nov 5;22(11):e22706. [doi: 10.2196/22706] [Medline: 33151152]
12. Global Digital Health Monitor. State of global health report 2023. URL: https://digitalhealthmonitor.org/stateofdigitalhealth23 [accessed 2024-07-26]
13. Australian Digital Health Agency. Connecting Australian healthcare – national healthcare interoperability plan 2023-2028. 2023. URL: https://www.digitalhealth.gov.au/sites/default/files/documents/national-healthcare-interoperability-plan-2023-2028.pdf [accessed 2023-10-16]
14. Australasian institute of digital health. About workforce advancement. URL: https://digitalhealth.org.au/ [accessed 2024-07-26]
15. Centre for Digital Transformation of Health. Connecting digital innovation to health. URL: https://mdhs.unimelb.edu.au/digitalhealth [accessed 2024-07-26]
16. Garde S, Harrison D, Huque M, Hovenga EJS. Building health informatics skills for health professionals: results from the Australian Health Informatics Skill Needs Survey. Aust Health Rev 2006 Feb;30(1):34-45. [Medline: 16448376]
17. Gray K, Sim J. Factors in the development of clinical informatics competence in early career health sciences professionals in Australia: a qualitative study. Adv Health Sci Educ Theory Pract 2011 Mar;16(1):31-46. [doi: 10.1007/s10459-010-9238-3] [Medline: 20544387]
18. Dattakumar A, Gray K, Henderson KB, Maeder A, Chenery H. We are not educating the future clinical health professional workforce adequately for e-health competence: findings of an Australian study. Stud Health Technol Inform 2012;178:33-38. [Medline: 22797016]
19. Butler-Henderson K, Dalton L, Probst Y, Maunder K, Merolli M. A meta-synthesis of competency standards suggest allied health are not preparing for a digital health future. Int J Med Inform 2020 Dec;144:104296. [doi: 10.1016/j.ijmedinf.2020.104296] [Medline: 33091830]
20. Kalenderian E, Zouaidi K, Yeager J, et al. Learning from data in dentistry: Summary of the third annual openwide conference. Learn Health Syst 2024 Apr;8(2):e10398. [doi: 10.1002/lrh2.10398] [Medline: 38633022]
21. de Ahumada Servant P, Martín-Martín D, Romero I. Digital transformation of oral health care: measuring the digitalization of dental clinics. Soc Indic Res 2024. [doi: 10.1007/s11205-024-03366-z]
22. Mourouzis P, Tolidis K. CAD/CAM systems. In: Delantoni A, Orhan K, editors. Digital Dentistry: An Overview and Future Prospects, 1st edition: Springer; 2024:47-66. [doi: 10.1007/978-3-031-52826-2_5]
23. Schwendicke F, Krois J. Data dentistry: how data are changing clinical care and research. J Dent Res 2022 Jan;101(1):21-29. [doi: 10.1177/00220345211020265] [Medline: 34238040]
24. Institute of Medicine (US). Institute of medicine roundtable on value and science-driven health care. In: Grossmann C, Powers B, McGinnis JM, editors. Digital Infrastructure for the Learning Health System: The Foundation for Continuous Improvement in Health and Health Care: Workshop Series Summary: National Academies Press; 2011. URL: https://www.ncbi.nlm.nih.gov/books/NBK83568/ [accessed 2024-10-30]
25. Smith M, Saunders R, Stuckhardt L, McGinnis JM. Best Care at Lower Cost: The Path to Continuously Learning Health Care in America: National Academies Press (US); 2013. [doi: 10.17226/13444] [Medline: 24901184]

26. Australian Dental Council. Professional competencies of the newly qualified dental practitioner. 2022. URL: https://www.adc.org.au/files/accreditation/competencies/ADC_Professional_Competencies_of_the_Newly_Qualified_Practitioner.pdf [accessed 2023-08-31]
27. Schneiderhan J, Guetterman TC, Dobson ML. Curriculum development: a how to primer. Fam Med Community Health 2019;7(2):e000046. [doi: 10.1136/fmch-2018-000046] [Medline: 32148703]
28. Richardson A. Dental services in Australia: industry report Q8531. : IBISWorld; 2022. URL: https://www.ibisworld.com/ [accessed 2023-08-31]
29. Nanayakkara S, Zhou X, Spallek H. Impact of big data on oral health outcomes. Oral Dis 2019 Jul;25(5):1245-1252. [doi: 10.1111/odi.13007] [Medline: 30474902]
30. American College of Prosthodontists. Digital dentistry curriculum for predoctoral and advanced education in prosthodontics. 2018. URL: https://www.prosthodontics.org/assets/1/7/ACP_Digital_Dentistry_Curriculum.pdf [accessed 2023-08-31]
31. Orrel J. Expert advice on designing authentic assessments for online delivery. : Tertiary Education Quality and Standards Agency; 2022. URL: https://www.teqsa.gov.au/sites/default/files/2022-10/assessment-for-online-delivery-orrell-4may2020.pdf [accessed 2024-10-30]
32. Contact North. How assessment is changing in the digital age – five guiding principles. 2020. URL: https://teachonline.ca/tools-trends/how-assessment-changing-digital-age-five-guiding-principles [accessed 2024-10-30]
33. Valenta AL, Berner ES, Boren SA, et al. AMIA Board White Paper: AMIA 2017 core competencies for applied health informatics education at the master's degree level. J Am Med Inform Assoc 2018 Dec 1;25(12):1657-1668. [doi: 10.1093/jamia/ocy132] [Medline: 30371862]
34. Brunner M, McGregor D, Keep M, et al. An eHealth capabilities framework for graduates and health professionals: mixed-methods study. J Med Internet Res 2018 May 15;20(5):e10229. [doi: 10.2196/10229] [Medline: 29764794]
35. Littlewood N, Downie S, Sawyer A, Feely K, Govil D, Gordon B. Development of a digital health capability framework for allied health practitioners: an Australian first. IJAHSP 2021;20(3):22. [doi: 10.46743/1540-580X/2022.2234]
36. Office of the Australian Information Commissioner. Notifiable data breaches report july to december 2023. 2023. URL: https://www.oaic.gov.au/privacy/notifiable-data-breaches/notifiable-data-breaches-publications/notifiable-data-breaches-report-july-to-december-2023 [accessed 2024-07-26]
37. OpenNotes. Putting patient safety first with OpenNotes. URL: https://www.opennotes.org/ [accessed 2024-07-26]
38. Australian Health Practitioner Regulation Agency. Social media: how to meet your obligations under the national law. URL: https://www.ahpra.gov.au/Resources/Social-media-guidance.aspx [accessed 2024-07-26]
39. Tudor Car L, Kyaw BM, Nannan Panday RS, et al. Digital health training programs for medical students: scoping review. JMIR Med Educ 2021 Jul 21;7(3):e28275. [doi: 10.2196/28275] [Medline: 34287206]
40. Shan T, Tay FR, Gu L. Application of artificial intelligence in dentistry. J Dent Res 2021 Mar;100(3):232-244. [doi: 10.1177/0022034520969115] [Medline: 33118431]
41. Vujicic M, David G. Value-based care in dentistry: Is the future here? J Am Dent Assoc 2023 Jun;154(6):449-452. [doi: 10.1016/j.adaj.2023.04.001] [Medline: 37097278]
42. Childers C, Marron J, Meyer EC, Abel GA. Clinical ethics consultation documentation in the era of open notes. BMC Med Ethics 2023 May 3;24(1):27. [doi: 10.1186/s12910-023-00904-1] [Medline: 37138339]
43. Schwendicke F, Chaurasia A, Wiegand T, et al. IADR e-oral health network and the ITU/WHO focus group AI for health. Artif Intell Oral Dent Healthc: Core Educ Curr J Dent 2023;128:104363. [doi: 10.1016/j.jdent.2022.104363] [Medline: 36410581]
44. Cham K, Edwards ML, Kruesi L, Celeste T, Hennessey T. Digital preferences and perceptions of students in health professional courses at A leading Australian university: a baseline for improving digital skills and competencies in health graduates. AJET 2021;38(1):69-86. [doi: 10.14742/ajet.6622]
45. Biggs JB, Tang CS. Teaching for Quality Learning at University: What the Student Does, 4th edition: Society for Research into Higher Education& Open University Press; 2011.

Abbreviations

ADC: Australian Dental Council
ADHA: Australian Digital Health Agency
AI: artificial intelligence
AIDH: Australasian Institute of Digital Health
CAD/CAM: computer-aided design/computer-aided manufacturing
LHS: learning health system
mHealth: mobile health
TEQSA: Tertiary Education Quality and Standards Agency

Edited by P Kanzow; submitted 31.10.23; peer-reviewed by M Pang, S Brice; revised version received 30.10.24; accepted 31.10.24; published 08.01.25.

Please cite as:
Mun M, Byrne S, Shaw L, Lyons K
Digital Dentists: A Curriculum for the 21st Century
JMIR Med Educ 2025;11:e54153
URL: https://mededu.jmir.org/2025/1/e54153
doi:10.2196/54153

Feedback From Dental Students Using Two Alternate Coaching Methods: Qualitative Focus Group Study

Lulwah Alreshaid[1,2,3], BDS, PhD; Rana Alkattan[1,2,3], BDS, MSD, PhD

[1]Department of Restorative and Prosthetic Dental Sciences, College of Dentistry, King Saud bin Abdulaziz University for Health Sciences, P.O. Box 22490, Riyadh, Saudi Arabia

[2]Ministry of National Guard Health Affairs, King Abdullah International Medical Research Centre, Riyadh, Saudi Arabia

[3]Dental Services, King Abdulaziz Medical City, Ministry of the National Guard Health Affairs, Riyadh, Saudi Arabia

Corresponding Author:
Rana Alkattan, BDS, MSD, PhD
Department of Restorative and Prosthetic Dental Sciences, College of Dentistry, King Saud bin Abdulaziz University for Health Sciences, P.O. Box 22490, Riyadh, Saudi Arabia

Abstract

Background: Student feedback is crucial for evaluating the effectiveness of institutions. However, implementing feedback can be challenging due to practical difficulties. While student feedback on courses can improve teaching, there is a debate about its effectiveness if not well-written to provide helpful information to the receiver.

Objective: This study aimed to evaluate the impact of coaching on proper feedback given by dental students in Saudi Arabia.

Methods: A total of 47 first-year dental students from a public dental school in Riyadh, Saudi Arabia, completed 3 surveys throughout the academic year. The surveys assessed their feedback on a Dental Anatomy and Operative Dentistry course, including their feedback on the lectures, practical sessions, examinations, and overall experience. The surveys focused on assessing student feedback on the knowledge, understanding, and practical skills achieved during the course, as aligned with the defined course learning outcomes. The surveys were distributed without coaching, after handout coaching and after workshop coaching on how to provide feedback, designated as survey #1, survey #2, and survey #3, respectively. The same group of students received all 3 surveys consecutively (repeated measures design). The responses were then rated as neutral, positive, negative, or constructive by 2 raters. The feedback was analyzed using McNemar test to compare the effectiveness of the different coaching approaches.

Results: While no significant changes were found between the first 2 surveys, a significant increase in constructive feedback was observed in survey #3 after workshop coaching compared with both other surveys (P<.001). The results also showed a higher proportion of desired changes in feedback, defined as any change from positive, negative, or neutral to constructive, after survey #3 (P<.001). Overall, 20.2% reported desired changes at survey #2% and 41.5% at survey #3 compared with survey #1.

Conclusions: This study suggests that workshops on feedback coaching can effectively improve the quality of feedback provided by dental students. Incorporating feedback coaching into dental school curricula could help students communicate their concerns more effectively, ultimately enhancing the learning experience.

(JMIR Med Educ 2025;11:e68309) doi:10.2196/68309

KEYWORDS

student feedback; coaching; dental education; student evaluation; teaching methods; educational intervention

Introduction

Feedback, a cornerstone of effective performance improvement, plays a crucial role in various domains, including education. Understanding how feedback is delivered and received is essential to maximize its impact. Several models provide frameworks for analyzing feedback processes, such as Hattie and Timperley's [1] model, which categorizes feedback based on its focus (task, process, and self-regulation), and Kluger and DeNisi's [2] Feedback Intervention Theory, which explores the instructional, motivational, emotional, and learning effects of feedback. These models highlight the complexities of feedback delivery and the importance of considering the recipient's needs and the specific context. These models go beyond simple evaluation and rather focus on providing actionable information that supports student learning and development.

Feedback is a critical method of measuring the effectiveness of performance and outcome of any institution. More importantly, if these institutions play an important role in education, health, or essential services, it is crucial to use student feedback to ensure the successful performance of these institutions. Feedback is often challenging to execute due to interaction issues or practical applicability [3,4]. Challenges arise from a complex interaction between the providers and recipients'

performance [4]. An example of these challenges could be the fear of recognizing unsatisfactory performance, discouragements, and liability. However, feedback's primary purpose is to improve the outcome. Delivering productive feedback to assess teaching procedures and students' experience is critical for effective learning and developing a solid connection between feedback providers and recipients [5-7]. In addition, it serves to evaluate teaching strategies. By aligning with the principles of key feedback models, the overall learning experience can be enhanced for both students and faculty.

Giving feedback to recipients can be complex; however, various techniques have been reported in the literature; 1 of the popular techniques is the "compliment sandwich," in which the recipient receives 1 criticism between 2 positive comments [8]. In contrast, another effective technique is to eliminate the negative connotation of feedback, in which the feedback provider mentions the mistakes and provides some solutions [9]. In any case, it is important to note that effective feedback comprises structure, content, and time [10]. When this feedback is expected from students to instructors, another level of challenge can be anticipated [11]; certain boundaries between the students and their instructors may restrict students' ability to express themselves freely. Students may also perceive end-of-course feedback as a mere administrative requirement fulfilling curricular mandates, potentially diminishing their perceived value and engagement in the process. Hence, to give constructive feedback, it is essential to guide students to the fact that the goal is not to deliver the feedback by criticizing but to enhance the feedback process to be more effective and constructive [12,13].

Many educational institutions imply student and professor feedback concerning courses in which they are both involved [14]. The feedback from the students usually involves a set of surveys to rate a course and the instructor giving that course. This process could assist the instructors in better recognizing areas of strengths and weaknesses, ultimately improving the educational experience [15-17]. Debate emerges that questions the effectiveness of such feedback [15,18-20]. A recent study found that implementing feedback could be beneficial if incorporated into the curriculum while also providing instructors with how to receive such feedback and how to adapt to these comments [17,21,22]. Furthermore, another author highlighted the importance of student evaluation and excelling in education, which could provide the instructor with minor adjustments to reform the course [21-23]. In contrast, some instructors note that this feedback will not encourage them to modify their courses [23]. Furthermore, some instructors might find it difficult to solely base altering decisions on input provided by students, arguing that some aspects will affect the student's ability to provide trustworthy information based on factors such as the ability to construct critical feedback or complex circumstances, including age, gender, or educational background [18,24].

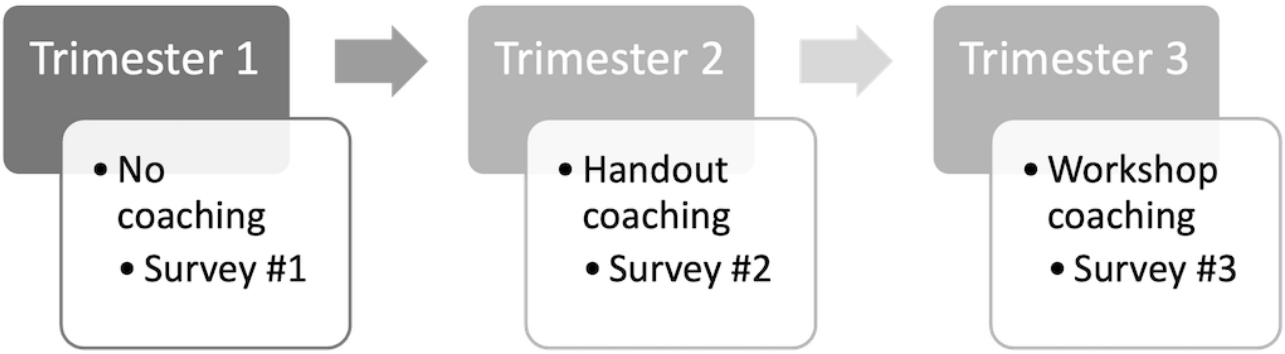
Although previous studies assessed the effect of feedback given by students on teaching quality and the improvement of feedback over a certain period [14,18], the relations between coaching to give and receive feedback and the feedback received from students after coaching have not been investigated among dental students in Saudi Arabia. Teaching students how to provide reflective, constructive feedback to elicit better outcomes for course, curriculum, and general educational development would be significant. Thus, the primary objective of this study was to evaluate the feedback given by students in the College of Dentistry, King Saud bin Abdulaziz University for Health Sciences (KSAU-HS), after using 2 different coaching approaches on how to provide feedback. The secondary objective was to improve the effectiveness of the feedback given by dental students after exposing them to 2 different coaching approaches on how to provide feedback. The null hypothesis of the current study is twofold as there is no difference in the nature of the feedback given by the students using 1 coaching methods, and both the coaching methods increase the proportion of constructive student feedback equally.

Methods

Overview

In total, 50 students were invited to participate in the study. Students were asked to complete 3 surveys in open-ended question format at the end of each trimester (repeated measures design, Figure 1). These surveys asked the same questions but were specific to each trimester. Invited students were asked to provide feedback on the RSTO 311 course. This study focused on first-year dental students enrolled in their introductory dental course, which serves as their initial exposure to fundamental dental concepts, including tooth anatomical landmarks, cavity preparation techniques, and restorative procedures in both theoretical and simulated clinical environments. Each survey consisted of 5 questions, adopted by Hajhamid and Somogyi - Ganss [16]. The first question was to indicate the 2-digit number assigned to each student by a research assistant who never interacted with the students to ensure anonymity. The second question was about the lectures given during the trimester. The third question was about the practical sessions taken during the trimester. The fourth question was about the quizzes, written and practical exams taken during the trimester. The fifth and last question was about the overall course (Multimedia Appendix 1). Based on the course learning outcomes, the students are expected to meet minimum criteria of knowledge and understanding as well as practical skills; thus, the survey focused on these aspects of the course.

Figure 1. Study design showing coaching methods for feedback given in surveys #1, #2, and #3.



At the beginning of the course, an invitation was sent to all students taking the RSTO 311 course. The students were offered a bonus (2 grades) if they participated in the study. If they wished not to participate, they could write an essay about a topic related to their course and get the same bonus grades. The survey was designed using Google Forms and was sent by email to all participating students; the survey link was sent by the same research assistant who assigned the 2-digit numbers to participating students. Consent was obtained from all participating students at the beginning of each survey. Before completing the first survey at the end of the first trimester, no coaching or instructions were given to the students on how to receive and provide feedback. Before completing the second survey at the end of the second trimester, students were coached by reading a 2-page handout on how to receive and provide feedback, which can be covered in approximately 10 minutes (Multimedia Appendix 2). The handout explains the different types of feedback, as well as steps and examples for giving constructive feedback. Before completing the third and last survey at the end of the third trimester, students were coached by attending a 1-hour workshop on how to receive and provide constructive feedback. The workshop was given by a faculty member who was not involved in the course or the research study. The workshop similarly explained different theories on giving and receiving feedback and demonstrated to the students how to improve feedback with examples. Both the handout and the workshop were based on a previously published paper [16].

All 3 surveys' answers were evaluated independently by 2 raters, the course director and the co-course director of the course. The answers were rated as either neutral, positive, negative, or constructive feedback. Any disagreement between the 2 evaluators' ratings of the survey answers was discussed and agreed upon before the analysis. Answers were considered neutral feedback if there were no positive, negative, or constructive comments. Answers were deemed positive if general praise was included. Answers were considered negative if the provided nonconstructive criticism. Finally, answers were considered constructive if there were any suggestions to improve the course in any aspect, even if they contained any positive or negative comments. When rater disagreement was noted in cases where responses were not clearly considered as positive or negative feedback and did not include any constructive comment or intent for improvement, responses were rated as neutral. Data were collected and analyzed based on the ratings given by the 2 evaluators and then compared between surveys.

Student feedback was collected in textual form and subsequently coded into 4 categories: positive, negative, neutral, and constructive. The reliability of this coding process was ensured through independent assessments by 2 raters. Kappa statistics was used to assess inter-rater reliability between the 2 raters. The ratings followed a nominal scale (1=neutral, 2=positive, 3=negative, and 4=constructive); hence, frequency and proportions were reported for the ratings as descriptive statistics. Inferential statistical analysis was used to test rating changes over time (McNemar test). The level of significance of .05 was used for inferential analysis with P values <.05 reported as statistically significant. Analysis was performed combined for 4 questions as well as separately for each question. IBM SPSS Statistics software (version 29) was used for descriptive and inferential analysis.

Ethical Considerations

An institutional review board ethical approval was obtained from King Abdullah International Medical Research Center for this cross-sectional study (IRB/3004/23). This study was conducted during the academic year 2023 - 2024 among first-year dental students who took the Dental Anatomy and Operative Dentistry course (RSTO 311) at the College of Dentistry, KSAU-HS, a public dental school in Riyadh, Saudi Arabia. The RSTO 311 is a yearly course divided over 3 trimesters. The course has theoretical and practical components: 30 lectures and 40 practical sessions. The students were assessed based on weekly continuous assessment, 3 quizzes, 3 written exams, and 3 practical exams.

Results

Of the 50 students in the class, 47 participants (25 male and 22 female participants) were included who completed all 3 surveys at 3 different time points, giving a participation rate of 94%. Out of 50, 1 student dropped the course, 1 refused to participate, and 1 failed to complete the third survey.

The 2 raters provided a total of 564 ratings each. Overall, 541 out of 564 ratings matched, suggesting a 95.9% level of agreement. The κ value was 0.941, which, being above 0.9, indicates an almost perfect level of agreement between the raters, demonstrating a high degree of reliability in the classification of responses. Discrepancy in data was discussed, re-evaluated, and a final agreement was reached and recorded. The following are randomly selected examples presented from students' feedback:

Neutral feedback: "No complaints about it"

Positive feedback: "The course provided a solid foundation in the subject matter, it was a valuable learning opportunity"

Negative feedback: "The work was hard and tiring and the time was not enough"

Constructive feedback: "In some anatomy lectures, clearer explanations were needed. Providing a short video would offer better visualization for students"

Within-subject analysis was conducted separately for each of the 4 questions in the 3 surveys. No significant changes were observed between survey #1 and #2 in any of the 4 questions, separately or combined. However, there were statistically significant changes between survey #1 and #3 with regards to increase in proportion of constructive ratings for questions 2 - 4 as well as for the 4 questions combined. Significant change in ratings was also found in survey #3 relative to survey #2 for questions 1 - 3 as well as for the 4 questions combined (Table 1).

Table . Ratings for each of the 4 questions at each of the 3 surveys.

Question and rating		Survey #1, n (%)	Survey #2, n (%)	Survey #3, n (%)
#1[a]				
	Neutral	10 (21.3)	9 (19.1)	13 (27.7)
	Positive	4 (8.5)	2 (4.3)	0 (0)
	Negative	15 (31.9)	25 (53.2)	9 (19.1)
	Constructive	18 (38.3)	11 (23.4)	25 (53.2)
#2[b]				
	Neutral	13 (27.7)	11 (23.4)	11 (23.4)
	Positive	3 (6.4)	8 (17)	1 (2.1)
	Negative	15 (31.9)	14 (29.8)	1 (2.1)
	Constructive	16 (34)	14 (29.8)	34 (72.3)
#3[c]				
	Neutral	27 (57.4)	24 (51.1)	15 (31.9)
	Positive	6 (12.8)	2 (4.3)	1 (2.1)
	Negative	4 (8.5)	10 (21.3)	6 (12.8)
	Constructive	10 (21.3)	11 (23.4)	25 (53.2)
#4[d]				
	Neutral	17 (36.2)	16 (34)	12 (25.5)
	Positive	1 (2.1)	1 (2.1)	2 (4.3)
	Negative	18 (38.3)	12 (25.5)	8 (17)
	Constructive	11 (23.4)	18 (38.3)	25 (53.2)
All 4 combined[e]				
	Neutral	67 (35.6)	60 (31.9)	51 (27.1)
	Positive	14 (7.4)	13 (6.9)	4 (2.1)
	Negative	52 (27.7)	61 (32.4)	24 (12.8)
	Constructive	55 (29.3)	54 (28.7)	109 (58)

[a] McNemar test for survey 2 vs survey 1: χ^2_5=5.86, *P*=.32; McNemar test for survey 3 vs survey 1: χ^2_5=6.64, *P*=.25; McNemar test for survey 3 vs survey 2: χ^2_3=13.07, *P*=.004

[b] McNemar test for survey 2 vs survey 1: χ^2_6=4.63, *P*=.59; McNemar test for survey 3 vs survey 1: χ^2_5=18.26, *P*=.003; McNemar test for survey 3 vs survey 2: χ^2_5=23.30, *P*<.001

[c] McNemar test for survey 2 vs survey 1: χ^2_6=6.10, *P*=.41; McNemar test for survey 3 vs survey 1: χ^2_5=15.87, *P*=.01; McNemar test for survey 3 vs survey 2: χ^2_5=14.53, *P*=.006

[d] McNemar test for survey 2 vs survey 1: χ^2_5=5.31, *P*=.38; McNemar test for survey 3 vs survey 1: χ^2_4=10.81, *P*=.03; McNemar test for survey 3 vs survey 2: χ^2_5=5.62, *P*=.35

[e] McNemar test for survey 2 vs survey 1: $\chi^2(6)$=5.28, *P*=.51; McNemar test for survey 3 vs survey 1: $\chi^2(5)$=33.43, *P*<.001; McNemar test for survey 3 vs survey 2: $\chi^2(5)$=45.28, *PP*<.001

Table 2 shows the proportion of constructive versus nonconstructive (positive, negative, or neutral) ratings for each question and for all 4 questions combined. A significant increase in the proportion of constructive ratings was found between survey #1 and survey #3 for questions 2 - 4 as well as for the 4 questions combined. A significant increase in the proportion of constructive ratings was also found between survey #2 and survey #3 for questions 1 - 3 as well as for the 4 questions combined.

Table . Proportion of constructive ratings.

Question and rating		Survey #1	Survey #2	Survey #3
#1				
	Nonconstructive[a]	29 (61.7)	36 (76.6)	22 (46.8)
	Constructive	18 (38.3)	11 (23.4)	25 (53.2)
			MN_1(b) *P*=.19	MN_1[b](b) *P*=.23
				MN_2[c](b) *P*=.003
#2				
	Nonconstructive	31 (66)	33 (70.2)	13 (27.7)
	Constructive	16 (34)	14 (29.8)	34 (72.3)
			MN_1(b) *P*=.83	MN_1(b) *P*<.001
				MN_2(b) *P*<.001
#3				
	Nonconstructive	37 (78.7)	36 (76.6)	22 (46.8)
	Constructive	10 (21.3)	11 (23.4)	25 (53.2)
			MN_1(b) *P*>.99	MN_1(b) *P*=.003
				MN_2(b) *P*<.001
#4				
	Nonconstructive	36 (76.6)	29 (61.7)	22 (46.8)
	Constructive	11 (23.4)	18 (38.3)	25 (53.2)
			MN_1(b) *P*=.14	MN_1(b) *P*=.01
				MN_2(b) *P*=.14
All 4 combined				
	Nonconstructive	133 (70.7)	134 (71.7)	79 (42)
	Constructive	55 (29.3)	54 (28.7)	109 (58)
			MN_1(b) *P*>.99	MN_1(b) *P*<.001
				MN_2(b) *P*<.001

[a] nonconstructive ratings include positive, negative and neutral.

[b] MN_1(b)=McNemartest using binomial distribution to examine change from survey #1.

[c] MN_2(b)=McNemartest using binomial distribution to examine change from survey #2.

For each question, the change from survey #1 was coded as desired versus not desired. Desired change was defined as any change from positive, negative, or neutral to constructive. All other changes were coded as not desired. The proportion of desired changes is summarized in Table 3. Survey #3 showed a higher proportion of desired changes compared with survey #2. For the 4 questions combined, 20.2% had desired changes at survey #2% and 41.5% at survey #3 compared with survey #1. In survey #3, the most frequent changes reported overall for the 4 questions combined were: neutral to constructive (17.6%), negative to constructive (16.5%) and constructive to constructive (16.5%).

Table . The proportion of desired changes in surveys #2 and #3 compared with survey #1.

Proportion of desired changes	Survey (#2 versus #1), n (%)	Survey (#3 versus #1), n (%)
Question 1	7 (14.9%)	16 (34%)
Question 2	10 (21.3%)	23 (48.9%)
Question 3	9 (19.1%)	19 (40.4%)
Question 4	12 (25.5%)	20 (42.6%)
Four questions combined	38 (20.2%)	78 (41.5%)

Discussion

Principal Findings

This study compared student responses without coaching, coaching using a feedback handout, or coaching using a feedback workshop before completing the surveys. Results demonstrate that handout coaching showed no significant difference compared with no coaching with respect to the number of neutral, positive, negative, or constructive ratings. However, workshop coaching significantly increased the number of constructive ratings compared with both no coaching and handout coaching ($P<.001$, Table 1). Therefore, the null hypothesis was rejected. The reason for these results could be due to the fact that handouts were distributed to the students, and they were asked to read the 2-page document independently. This method does not involve student and instructor interaction and is hence, less engaging. There was also no measure of whether the students in fact read the handout and grasped the information. Thus, no significant changes were noted between survey #1 and survey #2. Workshop coaching, on the other hand, was done in a classroom setting with 1 faculty member present, ensuring a 100% attendance rate of all participating students. Furthermore, the students were able to ask questions regarding the information presented in the workshop and were asked to fill out survey #3 immediately after the workshop, before leaving the classroom.

The proportion of constructive feedback, compared to nonconstructive feedback, significantly increased after workshop coaching (Table 2). The workshop-based format provided multiple examples in a story format from past student feedback, whereas the handout only stated the description of proper feedback writing without detailed examples compared with the examples presented in the workshop. The educational value of workshop coaching has been previously established, wherein the students are “active learners” and can engage in asking questions during the learning process [25,26]. Information presented in video format can also enhance information retention, owing to reduced student cognitive loading and optimized use of visual learner memory [27]. Furthermore, the key learning points are emphasized during the workshop, and audio-visual learning is more likely to keep the students more attentive and engaged in the content being delivered [28]. This is also demonstrated in Table 3, where the most frequently reported changes in feedback from survey #1 (no coaching) to survey #3 (workshop coaching) were from neutral and negative to constructive, reported in this study as “desired changes”.

The effectiveness of workshop coaching can also be understood through several educational and psychological frameworks. For example, the constructivist learning theory emphasizes the importance of social interaction and guided learning in developing cognitive skills [29]. The workshop format, which encourages active participation and immediate feedback from the instructor, aligns with this theory by fostering an environment where students engage with and construct their knowledge of feedback writing through scaffolding, wherein support is provided by a more knowledgeable person. This approach helps students internalize new feedback techniques through direct interaction and reflection on real examples. Furthermore, the importance of emotional intelligence in feedback delivery cannot be overlooked. According to Goleman [30], empathy and self-regulation are key components of emotional intelligence that influence how feedback is communicated. In the workshop setting, students are not only taught the mechanics of constructive feedback but also how to consider the emotional impact of their words, enhancing their ability to offer feedback that is both critical and supportive. This connection to emotional intelligence helps explain why the workshop coaching produced a higher proportion of constructive feedback compared with the handout coaching.

Comparison With Previous Work

In any educational environment, student satisfaction is an essential criterion for quality assessment [31]. Student evaluations of teaching are surveys typically used to collect, analyze, and interpret teaching quality [32]. Hence, every year, students are asked to evaluate the course material and provide feedback. In this study, the survey questions provided to the students concerned the lectures, practical sessions, and examinations at KSAU-HS. They were distributed immediately after the end of each trimester to ensure the feedback was relevant and firsthand. The purpose of these distributed surveys was to gather information on the course teaching, practical sessions, and facilities so that an action plan may be set to ensure improvement. However, most student feedback tends to be general or rely on their personal experience rather than providing helpful information related to the learning experience [33]. As this study is based on open-ended questions, analyzing responses can be quite intricate unless the process is made more structured. Hence, this study evaluated student responses after a handout and workshop coaching.

Written comments add value to both students and educators when compared with scale-type questions [34]. The students are given the possibility to explain their perspective beyond Likert-type scales and raise further topics that may not have

been covered in closed-ended questions [35]. Written comments are more informative for educators, and suggestions are beneficial when compared with receiving a statistical summary of quantitative results [36]. “Student evaluations of teaching” instruments can be a source of valuable thoughts from students and can help educators gain insight into how students perceive their learning experience and how different students learn best in a given setting [37]. However, these benefits can only be reliable after bringing a little order to the chaos of written responses.

The main purpose of the study was to improve the quality of feedback provided by the students. To the best of our knowledge, this is the first study introducing interactive workshop coaching for proper feedback among teaching institutes in Saudi Arabia. The workshop was able to improve the constructive criticism given by the students compared with self-learning using the handout. It is likely that the lower performance with handout coaching reflected less motivation, responsibility, or independence of the students [38]. These results are contrary to a previous similar study, in which both the handout and workshop coaching similarly improved student feedback [16]. The difference in results could be attributed to the nature of the dental school between both studies. This study was performed in a governmental dental school where students are not obliged to pay tuition fees. On the contrary, since their education is financed largely by loans, students from the Canadian private dental school may be more encouraged to commit to assigned tasks [39,40]. It is also worth noting that dental students at our institution are more familiar with lecture- and workshop-based learning as opposed to self-directed learning; as most dental schools in Saudi Arabia have not completely shifted from teacher-centered learning to a more interactive or evidence-based style [41]. Furthermore, culturally, expressing opinions, especially those with negative connotations or suggestive tones, may not necessarily be favored [42]. However, the results of this study clearly show the benefits of workshop coaching in directing students to provide their perception towards the course. This emphasizes the importance of including such a coaching approach for first-year students as part of the academic curriculum at the beginning of their studies.

Limitations

One of the limitations of this study was the inclusion of only first year students, as students in older years may have responded differently to the handout coaching, likely being more familiar with independent self-learning. Students in older years may also be more exposed to course-based surveys compared with first-year students. This also reduced the sample size of the participants. Furthermore, the difference between the topics covered over the 3 trimesters of the course may have influenced the feedback given by the students. In addition, when the students were given the third survey, they had already been exposed to both handout and workshop coaching on proper feedback, and this emphasis on appropriate feedback writing may have led to the higher number of constructive comments in survey #3. Furthermore, self-reported student feedback is subject to various biases, such as recall bias, acquiescence bias, social desirability bias, and cultural influences, which could impact the accuracy of the responses. Finally, the incentive of the bonus grades may have introduced self-selection bias; however, as the incentive was offered to all students equally, whether they participated in the survey or chose to submit an essay assignment, this may have mitigated the bias.

Conclusions

This study compared the effectiveness of 3 approaches, no coaching, handout coaching, and workshop coaching, on improving the quality of feedback provided by dental students. The results show that workshop coaching significantly increased the number of constructive feedback ratings, compared with both no coaching and handout coaching. This study encourages a more expressive feedback culture that facilitates student or instructor interaction in a constructive manner, wherein instructors can receive and implement feedback to improve the educational process. This suggests that interactive, instructor-led workshops foster a more engaged learning environment, encouraging students to provide higher-quality feedback. Given these findings, educators can implement interactive workshops focused on teaching students how to provide constructive feedback. These workshops should encourage active engagement through real-life examples and peer discussions. Given that the study shows significant benefits in first-year students, feedback coaching can be introduced early in the academic program. Building on the concept of scaffolding, educators could start with guided feedback exercises during the workshop, gradually increasing the level of independence as students become more proficient. Educators can also integrate emotional intelligence training into feedback workshops by helping students understand how to express feedback empathetically and how to regulate their emotions while providing feedback. Further studies evaluating different coaching methods to enhance student feedback are needed, with consideration to assign different methods to each study group. Future research should also investigate the impact of standardized coaching protocols on the quality of student feedback and use the data to improve assessment and learning outcomes.

Acknowledgments

We would like to thank the research assistant, Dr Jabir Al Humaid, for his valuable contribution.

Conflicts of Interest

None declared.

Multimedia Appendix 1

RSTO 311 course evaluation survey
[PDF File, 78 KB - mededu_v11i1e68309_app1.pdf]

Multimedia Appendix 2
How to provide constructive feedback
[PDF File, 91 KB - mededu_v11i1e68309_app2.pdf]

References

1. Hattie J, Timperley H. The power of feedback. Rev Educ Res 2007 Mar;77(1):81-112. [doi: 10.3102/003465430298487]
2. Kluger AN, DeNisi A. The effects of feedback interventions on performance: A historical review, A meta-analysis, and A preliminary feedback intervention theory. Psychol Bull 1996 Mar;119(2):254-284. [doi: 10.1037/0033-2909.119.2.254]
3. Jonsson A. Facilitating productive use of feedback in higher education. Active Learning in Higher Education 2013 Mar;14(1):63-76. [doi: 10.1177/1469787412467125]
4. Kowalski K. Giving and receiving feedback: part II. J Contin Educ Nurs 2017 Oct 1;48(10):445-446. [doi: 10.3928/00220124-20170918-04] [Medline: 28954179]
5. Biggs J. Teaching For Quality Learning at University: What the Student Does, 2nd edition: Open University Press; 2003.
6. Harris LR, Brown GTL, Harnett JA. Understanding classroom feedback practices: a study of New Zealand student experiences, perceptions, and emotional responses. Educ Asse Eval Acc 2014 May;26(2):107-133. [doi: 10.1007/s11092-013-9187-5]
7. McCarthy J. Evaluating written, audio and video feedback in higher education summative assessment tasks. Iss Educa Res 2015;25(2):153-169.
8. Cannon MD, Witherspoon R. Actionable feedback: Unlocking the power of learning and performance improvement. AMP 2005 May;19(2):120-134. [doi: 10.5465/ame.2005.16965107]
9. Emory CL. Pearls: giving and receiving feedback. Clin Orthop Relat Res 2019 Jan;477(1):35-36. [doi: 10.1097/CORR.0000000000000538] [Medline: 30586065]
10. Kruidering-Hall M, O'Sullivan PS, Chou CL. Teaching feedback to first-year medical students: long-term skill retention and accuracy of student self-assessment. J Gen Intern Med 2009 Jun;24(6):721-726. [doi: 10.1007/s11606-009-0983-z] [Medline: 19384559]
11. Ben-Porath S, Webster D. Free Speech and Education, 1st edition: Routledge; 2022.
12. van de Ridder JMM, Stokking KM, McGaghie WC, ten Cate OTJ. What is feedback in clinical education? Med Educ 2008 Feb;42(2):189-197. [doi: 10.1111/j.1365-2923.2007.02973.x] [Medline: 18230092]
13. Bienstock JL, Katz NT, Cox SM, et al. To the point: medical education reviews--providing feedback. Am J Obstet Gynecol 2007 Jun;196(6):508-513. [doi: 10.1016/j.ajog.2006.08.021] [Medline: 17547874]
14. Kember D, Leung DYP, Kwan KP. Does the use of student feedback questionnaires improve the overall quality of teaching? Assessment & Evaluation in Higher Education 2002 Sep;27(5):411-425. [doi: 10.1080/0260293022000009294]
15. Boerboom TBB, Jaarsma D, Dolmans DHJM, Scherpbier AJJA, Mastenbroek NJJM, Van Beukelen P. Peer group reflection helps clinical teachers to critically reflect on their teaching. Med Teach 2011;33(11):e615-e623. [doi: 10.3109/0142159X.2011.610840] [Medline: 22022915]
16. Hajhamid B, Somogyi-Ganss E. Improving effectiveness of dental students' feedback and course evaluation. J Dent Educ 2021 Jun;85(6):794-801. [doi: 10.1002/jdd.12548] [Medline: 33502807]
17. Arreola RA. Developing a Comprehensive Faculty Evaluation System: Magna Publications; 2004.
18. Gormally C, Evans M, Brickman P. Feedback about teaching in higher ed: neglected opportunities to promote change. CBE Life Sci Educ 2014;13(2):187-199. [doi: 10.1187/cbe.13-12-0235] [Medline: 26086652]
19. Smither JW, London M, Reilly RR. Does performance improve following multisource feedback? A Theoretical model, meta‐analysis, and review of empirical findings. Pers Psychol 2005 Mar;58(1):33-66. [doi: 10.1111/j.1744-6570.2005.514_1.x]
20. Overeem K, Wollersheim H, Driessen E, et al. Doctors' perceptions of why 360-degree feedback does (not) work: a qualitative study. Med Educ 2009 Sep;43(9):874-882. [doi: 10.1111/j.1365-2923.2009.03439.x] [Medline: 19709012]
21. Ward JR, McCotter SS. Reflection as a visible outcome for preservice teachers. Teac Teacher Educ 2004 Apr;20(3):243-257. [doi: 10.1016/j.tate.2004.02.004]
22. Watts M, Lawson M. Using a meta-analysis activity to make critical reflection explicit in teacher education. Teach Teacher Educ 2009 Jul;25(5):609-616. [doi: 10.1016/j.tate.2008.11.019]
23. Schneider G. Student evaluations, grade inflation and pluralistic teaching: moving from customer satisfaction to student learning and critical thinking. In: Forum for Social Economics: Taylor & Francis; 2013.
24. McColskey W, Leary MR. Differential effects of norm-referenced and self-referenced feedback on performance expectancies, attributions, and motivation. Contemp Educ Psychol 1985 Jul;10(3):275-284. [doi: 10.1016/0361-476X(85)90024-4]

25. Mahler SA, Wolcott CJ, Swoboda TK, Wang H, Arnold TC. Techniques for teaching electrocardiogram interpretation: self-directed learning is less effective than a workshop or lecture. Med Educ 2011 Apr;45(4):347-353. [doi: 10.1111/j.1365-2923.2010.03891.x] [Medline: 21401682]
26. Haidet P, Morgan RO, O'Malley K, Moran BJ, Richards BF. A controlled trial of active versus passive learning strategies in A large group setting. Adv Health Sci Educ Theory Pract 2004;9(1):15-27. [doi: 10.1023/B:AHSE.0000012213.62043.45] [Medline: 14739758]
27. Chen CM, Wu CH. Effects of different video lecture types on sustained attention, emotion, cognitive load, and learning performance. Comput Educ 2015 Jan;80:108-121. [doi: 10.1016/j.compedu.2014.08.015]
28. Shqaidef AJ, Abu-Baker D, Al-Bitar ZB, Badran S, Hamdan AM. Academic performance of dental students: A randomised trial comparing live, audio recorded and video recorded lectures. Eur J Dent Educ 2021 May;25(2):377-384. [doi: 10.1111/eje.12614] [Medline: 33021047]
29. Vygotsky LS. In: Cole M, Jolm-Steiner V, Scribner S, Souberman E, editors. Mind in society: development of higher psychological processes: Harvard University Press; 1978.
30. Goleman D. Emotional intelligence, 10th edition: Bantam Books; 2007.
31. Douglas JA, Douglas A, McClelland RJ, Davies J. Understanding student satisfaction and dissatisfaction: an interpretive study in the UK higher education context. Stu High Educ 2015 Feb 7;40(2):329-349. [doi: 10.1080/03075079.2013.842217]
32. Zabaleta F. The use and misuse of student evaluations of teaching. Teach High Educ 2007 Feb;12(1):55-76. [doi: 10.1080/13562510601102131]
33. Richardson JTE. Instruments for obtaining student feedback: a review of the literature. Assess Evalu Higher Educ 2005 Aug;30(4):387-415. [doi: 10.1080/02602930500099193]
34. Pan D, Tan GSH, Ragupathi K, Booluck K, Roop R, Ip YK. Profiling teacher/teaching using descriptors derived from qualitative feedback: formative and summative applications. Res High Educ 2009 Feb;50(1):73-100. [doi: 10.1007/s11162-008-9109-4]
35. Spooren P, Brockx B, Mortelmans D. On the validity of student evaluation of teaching. Rev Educ Res 2013 Dec;83(4):598-642. [doi: 10.3102/0034654313496870]
36. Svinicki MD. Encouraging your students to give feedback. New Drctns for Teach & Learn 2001 Sep;2001(87):17-24 [FREE Full text] [doi: 10.1002/tl.24]
37. Lewis KG. Making sense of student written comments. New Drctns for Teach & Learn 2001 Sep;2001(87):25-32 [FREE Full text] [doi: 10.1002/tl.25]
38. Beckert L, Wilkinson TJ, Sainsbury R. A needs-based study and examination skills course improves students' performance. Med Educ 2003 May;37(5):424-428. [doi: 10.1046/j.1365-2923.2003.01499.x] [Medline: 12709183]
39. Karibe H, Suzuki A, Sekimoto T, et al. Cross-cultural comparison of the attitudes of dental students in three countries. J Dent Educ 2007 Nov;71(11):1457-1466. [doi: 10.1002/j.0022-0337.2007.71.11.tb04417.x] [Medline: 17971576]
40. Matthew IR, Walton JN, Dumaresq C, Sudmant W. The burden of debt for Canadian dental students: part 3. Student indebtedness, sources of funding and the influence of socioeconomic status on debt. J Can Dent Assoc 2006 Nov;72(9):819. [Medline: 17109801]
41. Ahmad MS, Bhayat A, Fadel HT, Mahrous MS. Comparing dental students' perceptions of their educational environment in Northwestern Saudi Arabia. Saudi Med J 2015 Apr;36(4):477-483. [doi: 10.15537/smj.2015.4.10754] [Medline: 25828286]
42. Ahmad M, Al Shorman H, Mahrous M. Assessment of the educational environment in a newly established dental college. J Educ Ethics Dent 2013;3(1):6. [doi: 10.4103/0974-7761.126935]

Abbreviations

KSAU-HS: King Saud bin Abdulaziz University for Health Sciences
KSAU-HS: King Saud bin Abdul-Aziz University

Edited by S Nedunchezhiyan, T Gladman; submitted 02.11.24; peer-reviewed by D Jocic, MZ Nassani; revised version received 07.02.25; accepted 11.02.25; published 18.03.25.

Please cite as:
Alreshaid L, Alkattan R
Feedback From Dental Students Using Two Alternate Coaching Methods: Qualitative Focus Group Study
JMIR Med Educ 2025;11:e68309
URL: https://mededu.jmir.org/2025/1/e68309
doi:10.2196/68309

Alignment Between Classroom Education and Clinical Practice of Root Canal Treatment Among Dental Practitioners in China: Cross-Sectional Study

XinYue Ma, MA; JingShi Huang, PhD

Humanomics Science Center, International Institute of Creative Design, Shanghai University of Engineering Science, 350 Xianxia Road, Shanghai, China

Corresponding Author:
JingShi Huang, PhD
Humanomics Science Center, International Institute of Creative Design, Shanghai University of Engineering Science, 350 Xianxia Road, Shanghai, China

Abstract

Background: This cross-sectional study assessed the perceived alignment between preclinical education and clinical practice in root canal treatment (RCT) among dental practitioners in China, aiming to identify systemic gaps in dental curricula and their clinical implications.

Objective: Dental professionals in Eastern Coastal China. This study distributed questionnaires through hospital dental specialties and medical forums, covering the Southeastern Region of China.

Methods: A validated, web-based survey was distributed to 90 dental professionals in Eastern Coastal China, focusing on 9 key stages of RCT, preoperative preparation, intraoperative procedures, postoperative care, and clinician-patient communication. Responses were measured using a 7-point Likert scale to evaluate perceived discrepancies between education and clinical practice.

Results: A total of 83 valid questionnaires were recovered, which revealed significant disparities between academic training and clinical demands. The survey showed that the specialized practitioners identified pronounced mismatches in RCT operative techniques and doctor-patient communication (P<.05). Participants aged ≤29 years demonstrated heightened awareness of discrepancies in disinfection protocols and temporary filling procedures (P<.05). Shanghai-trained practitioners reported fewer educational-clinical gaps across multiple procedural stages (P<.05). Notably, 82% of respondents rated comprehensive RCT implementation as more challenging than individual procedural components. Curriculum deficiencies were identified in treatment indication diagnostics (56.6% agreement) and communication training (43.4% agreement). Emerging technologies like virtual reality and augmented reality (VR and AR) showed minimal educational penetration (3.7% exposure rate). In the free-response section, qualitative feedback highlighted equipment accessibility issues (eg, thermal gutta-percha tools) and instructor-dependent learning outcomes.

Conclusions: Structural discrepancies exist in Chinese preclinical RCT education, influenced by factors such as experience level, age, and region. These findings underscore the need for curriculum reforms, emphasizing competency-based training, enhanced simulation technologies, and standardized clinical protocols, particularly in areas like periodontal pathology and communication skills.

(JMIR Med Educ 2025;11:e65534) doi:10.2196/65534

KEYWORDS

root canal treatment; endodontic education; dental education; clinical practice; dental medical workers

Introduction

Periapical lesions and endodontic diseases are highly prevalent among the global adult population [1], with root canal treatment (RCT) being one of the most common interventions provided by general dentists in clinical practice and the preferred treatment method for treating endodontic diseases.

Studies have shown that RCT is a challenging clinical intervention in dentistry, leading to varying degrees of practical difficulty for dental practitioners. Over 70% of dental practitioners express a desire for more clinical training [2]. Unlike most surgical dental procedures, RCT involves multiple steps, and any errors can result in unsatisfactory outcomes. RCT is performed in a concealed space, with limited space for operation and without visual control. It involves a variety of equipment, requires fine competence of the dentist, and involves complex operational procedures. According to dental practitioners, they often perform RCT procedures feeling a lack of control. In an emotional survey report related to RCT, nearly

all dentists expressed feelings such as anxiety, frustration, stress, or exhaustion [3].

Compared with other dental disciplines, endodontic therapy demands higher levels of manual dexterity and independent operational skills. Educational institutions play a crucial role in nurturing individuals to meet societal needs [4]. Optimal dental education should produce competent general practitioners in dentistry to ensure patient safety [5,6]. However, the difference in education among dental schools is considered a major obstacle to standardizing and ensuring the quality of dental education [5]. Early research on endodontic education focused on improving teaching methods, while later studies began to focus on different learning periods and the introduction of new technologies [7]. Authoritative organizations such as the European Society of Endodontology (ESE) have established guidelines to ensure educational standards. Furthermore, most dental schools in China offer a 5-year dental training program leading to a Bachelor of Dental Surgery degree, with the emphasis on clinical practice training in the final year of the dental education plan (following the "Clinical Practice for Chinese Undergraduate Students Majoring in Stomatology" standards established by Chinese Stomatological Association) [8]. Due to differences in economic and cultural backgrounds between Western countries and China, disparities exist in dental education programs, licenses, curricula, and facilities [9].

Regarding the allocation of teaching time in endodontic education, significant differences exist among countries. Previous studies show substantial variations in preclinical endodontic education at dental schools. In Germany, the average time spent on theoretical courses is 13.3 hours, practical courses require an average of 45.4 hours, and the total time for the endodontics course averages 56 hours [10,11]. In Spain, 95% of schools allocate over 20 hours for preclinical training, with 60% of schools dedicating over 50 hours [12]. However, in a study in China, 71.99% of Chinese schools spend less than 4 hours per week on endodontics education, totaling no more than 80 hours per semester, with the lowest training time allocated to periodontics and significant shortages in facilities for dental surgery courses [13]. This indicates that education emphasis and methods vary greatly across regions.

In terms of educational systems, comparing China and the United States illustrates 2 distinct teaching models. In the United States, after completing 4 years of general education, students are required to independently prepare some practice equipment, have relatively ample time to focus on coursework and preclinical training at schools, and enhance their skills through dental practice [9]. In China, dental schools are divided into three categories: 8-year program, 5-year program, and 3-year program. The 8-year program is only held in a few renowned dental schools in China, while the majority of dental students pursue the 5-year Bachelor program [14]. Although the order and proportion of courses may differ across schools in China, the undergraduate curriculum generally includes basic courses and dental courses. The undergraduates take public courses and basic courses of medicine in the first 2 years, focusing on clinical medical courses in the junior and senior years, with the opportunity to do a rotation in several departments such as Otolaryngology and Endocrinology. In the final year, there is no theoretical research, and all the time belongs to clinical practice. In America, clinical practice typically takes 2 years. In addition, China has 93 dental institutions offering a shorter 3-year training program for dental assistants [15]. Upon completion of the 3-year program and assessments, students are awarded the junior college in China, comparable to Bachelor of Science in Dental Hygiene in the United States (70% superposable curriculum). Differences in the dental degree concepts between China and the United States might lead to inequity situations [16]. Dental education in China is mainly government-supported, with schools providing equipment for students. Some dental schools cannot afford to provide advanced materials for all students [9].

Due to the aforementioned reasons, China exhibits certain differences from other countries, both in the content and discipline orientation of endodontic education. Past research on endodontic teaching primarily focused on undergraduate and postgraduate students, leading to the limitations of research not addressing specialized students and lacking feedback from clinical practitioners. This study is designed as a random sampling survey targeting clinical health care professionals to gather evaluation data on clinical RCT and corresponding teaching practices. This aims to identify issues related to the differences between clinical practices and classroom education, the effectiveness of teaching content, difficulties in practical components, and the bias in educational training.

Methods

Questionnaire Design

The design of the questionnaire referred to the fuzzy Delphi method and formed an expert group consisting of experienced oral physician (1), university professor (1), and dental nurses (2). The questionnaire validity was verified through 2 rounds of structured opinion solicitation. Expert discussions focused on the professionalism of the questions (accuracy of instrument references and rationality of operational sequences), representativeness (coverage of essential procedures of standard root canal treatment), and breadth (including preoperative preparation, intraoperative emergencies, and postoperative evaluations). At the same time, the wording of the questions was screened for semantic ambiguity and reading complexity calibration (Flesch-Kincaid level≤8). After 3 iterative modifications, a final consensus rate of 86.4% at the item level was achieved. To prevent respondent misunderstandings, nonconsensual operational terms were annotated in the questionnaire in Multimedia Appendix 1. Participants were required to read the survey instructions before participating in the survey, which outlined the basic process and main content of the questionnaire, as well as clarified the voluntary nature of participation and the noncommercial research purposes of the data.

The questionnaire consists of 30 questions in 4 categories. The basic questions follow a 7-point Likert scale to categorize the sensitivity of the responses. Some questions are designed as multiple-choice and free-response options. The main content focuses on the differences between in-school learning content and clinical practice, the practical difficulty of clinical RCT,

and issues related to educational methods and approaches in RCT. The questionnaire refers to the uniformly used endodontic teaching textbooks in the China region, dividing the entire process of clinical root canal treatment into 9 stages: determining indications, x-ray photography and observation, equipment adjustments, local anesthesia, root canal preparation, root canal disinfection and temporary sealing, root canal filling, posttreatment supplies organization, and clinical doctor-patient communication and doctor-nurse cooperation. To prevent any cognitive misinterpretation, we provided annotations for nonconsensus operations (Multimedia Appendix 1). Before participating in the questionnaire survey, participants are required to read the survey notice, which outlines the basic processes and main content of the survey and clarifies the voluntary nature of participation and the noncommercial research purpose of the data.

Data Source

This study distributed questionnaires through hospital dental specialties and medical forums, covering the Southeastern Region of China.

Statistical Analysis

Data collection was managed using internet-based survey software. Descriptive statistics were applied to each question to obtain basic distribution characteristics, and data were cross-compared for different grouping situations. The statistical analysis in the study was carried out using SPSS 26 (IBM), with 1-way analysis of variance used for certain differential issues and the K-W independent sample test and nonparametric test methods used for questions related to the difficulty of RCT operations. The level of statistical significance was set at $P<.05$.

Ethical Considerations

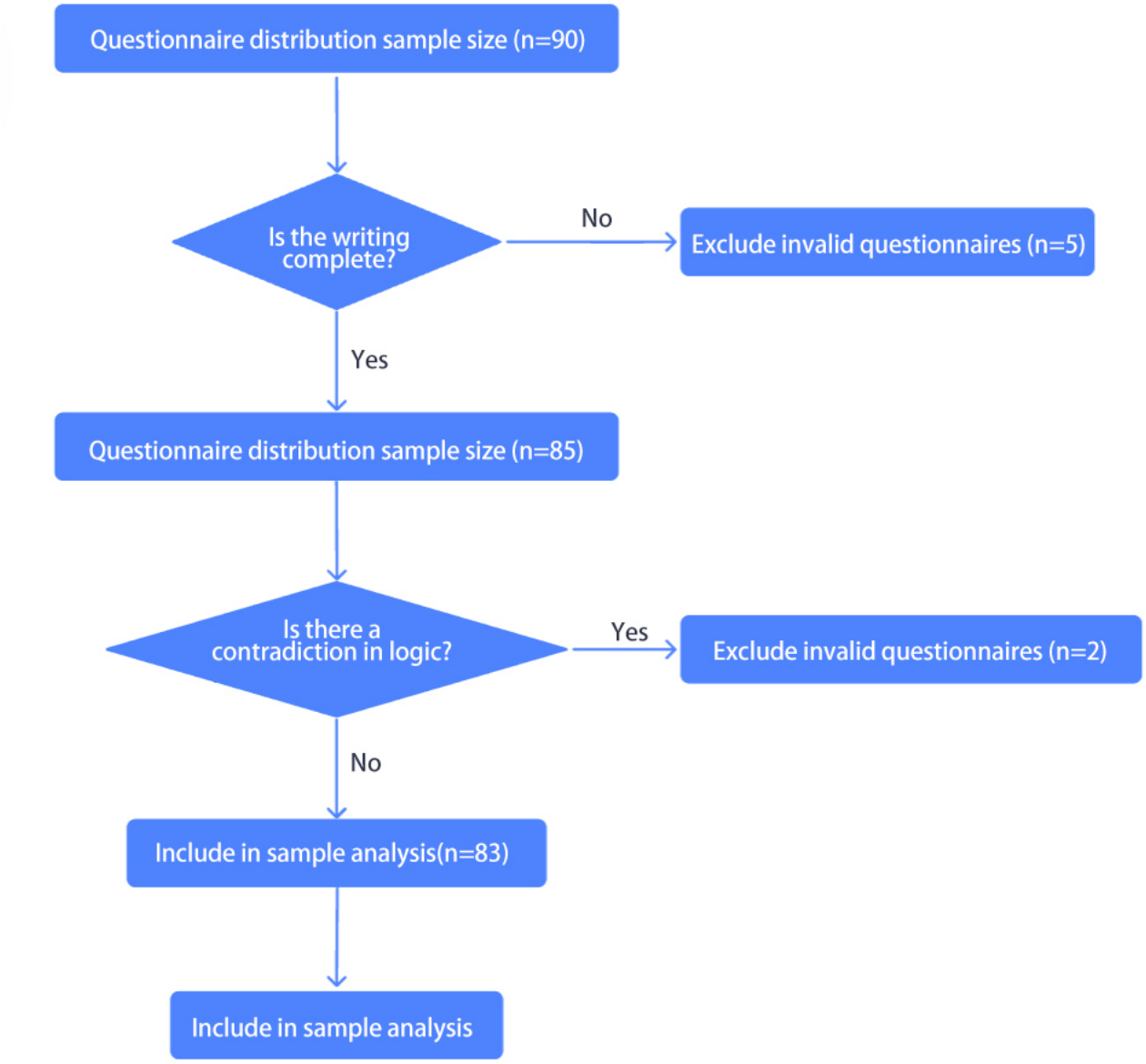
The questionnaire was distributed and collected in May 2024, over a period of 30 days. A total of 90 questionnaires were distributed, yielding 83 valid responses. A total of 7 incomplete or unsuccessfully retrieved questionnaires were excluded from the data analysis, resulting in a response rate of 92.22%. The questionnaire was filled out anonymously, with no personal names or specific hospital names disclosed. This study was conducted in the form of a questionnaire and was approved by the Ethics Committee of Shanghai University of Engineering Science (Approval No. EST-2024‐027). Written informed consent of the participants was obtained prior to enrolment in the study.

Results

Overview

The questionnaires were distributed and collected from May 1 to May 30, 2024, lasting for 30 days. A total of 90 questionnaires were collected, of which 7 were determined to be invalid due to missing key variables (such as education and age) or logical contradictions. Ultimately, 83 complete data sets were included (missing rate of 7.8%), with an effective recovery rate of 92.22%, as shown in Figure 1. The questionnaires were filled out anonymously without involving personal names or specific hospital names.

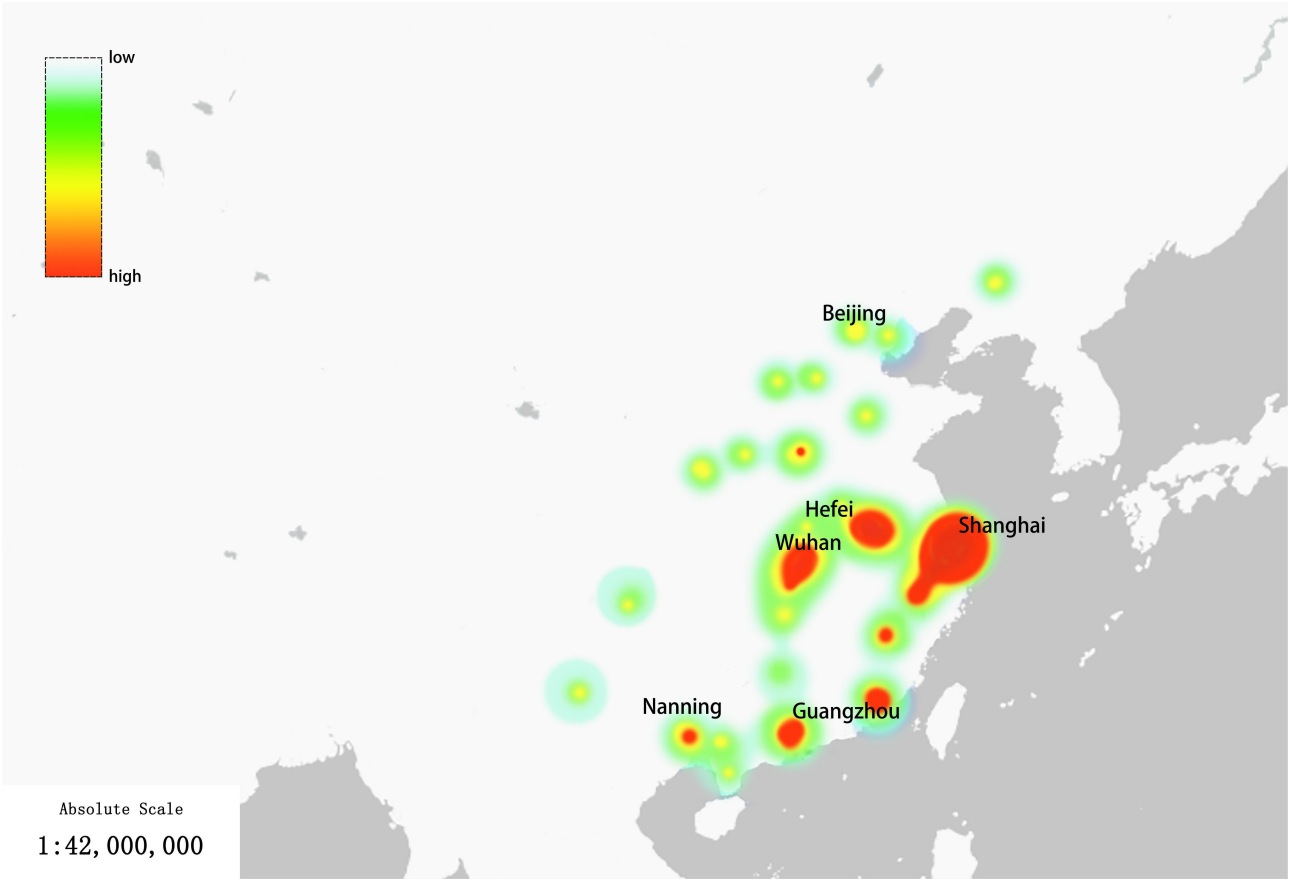
Figure 1. Data screening flowchart.



Through reliability and validity tests, the standardized reliability coefficient of the questionnaire items was 0.845, and the KMO test coefficient result was 0.87, indicating that the reliability and validity of this questionnaire were good. To evaluate the robustness of the results, the following sensitivity analyses were conducted, using parametric tests (ANOVA) and nonparametric tests (Mann-Whitney U) to analyze the intergroup differences. The results were highly consistent (the direction of *P* values was consistent, and the level of significance did not change). After excluding samples with >10 years of work experience (n=12) and reanalyzing, the impact of education and region on teaching satisfaction remained significant ($P<.05$), indicating that the results were not sensitive to outliers. Therefore, the collected data is meaningful for reference.

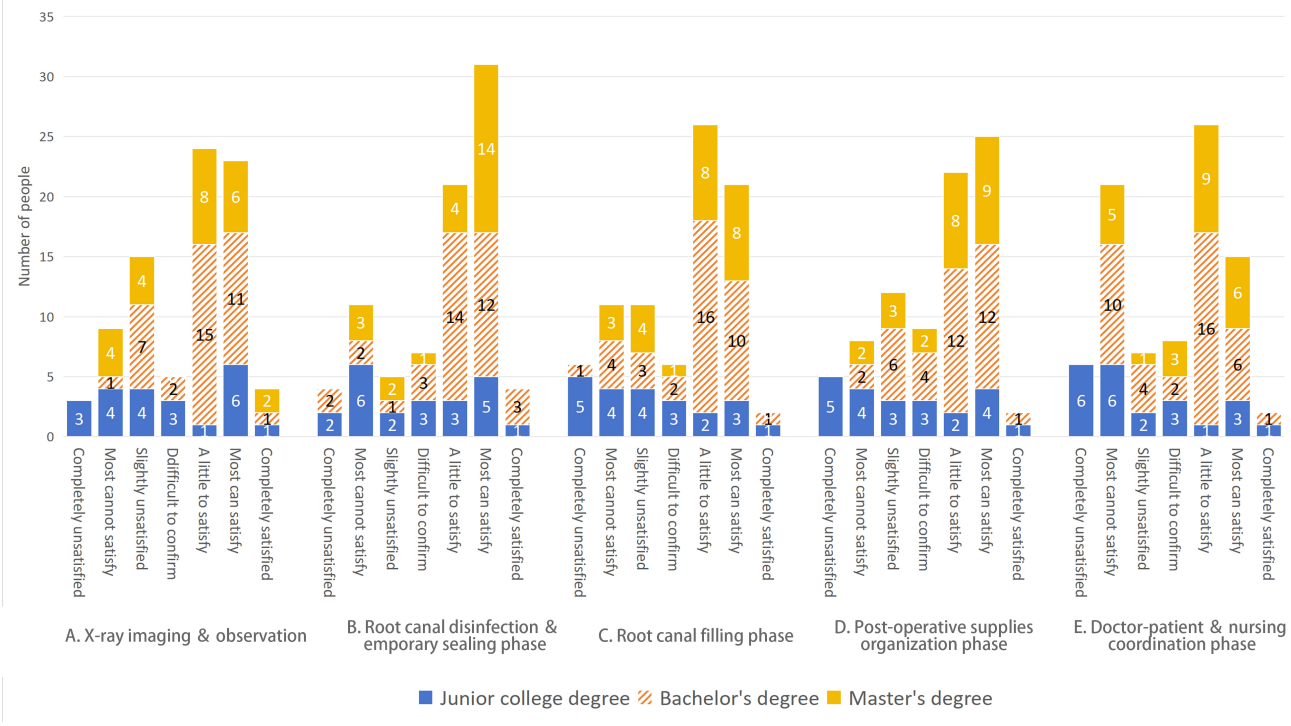
The participants in the study were exclusively practicing dental professionals, with 45.78% hailing from the Shanghai region, as depicted in the density distribution map (Figure 2). These survey participants all have undergone university-level education in the field of dentistry, possess a minimum of 1 year of clinical dental work experience, and have engaged in the performance of RCT during their professional tenure. The age demographic of participants ranged from 21 to 60 years, comprising 36.1% males and 63.9% females. In terms of their educational attainment, 26.5% held junior college degrees, 44.6% possessed bachelor's degrees, and 28.9% had obtained master's degrees. The spectrum of clinical dental work experience extended from a minimum of 1 year to a maximum of 36 years.

Figure 2. Density visualization of the regional distribution of participants: color indicates the number of participants from less (green) to more (red).



Regarding the congruence between specific aspects of school instruction and clinical practices, respondents' answers exhibited notable disparities. This investigation revealed that in the x-ray imaging and observation stage, the congruence between bachelor's and master's degree holders was notably superior to that of junior college graduates (P<.05). Participants with junior college degrees account for 50% (11/22) of the total who expressed negative evaluations of school teaching (selecting "completely unsatisfied," "most cannot satisfy," and "slightly unsatisfied" options). In contrast, those with bachelor's and master's degrees reported negative evaluations at rates of 22.9% (8/35) and 33.3% (8/24) of their respective totals. Bachelor's and master's degree holders significantly surpass junior college degree holders in their positive assessments of school teaching and clinical practices; the proportion of junior college degree holders who gave positive evaluations (choosing "completely satisfied," "a little to satisfy," and "slightly satisfied" options) is 36.4% (8/22) of their total, while bachelor's and master's degree holders selected positive evaluation options at rates of 77.1% (27/35) and 66.7% (16/24), respectively. Thus, bachelor's and master's degree holders evidently hold more favorable views of school teaching and clinical operations compared with those with junior college degrees. For further details, refer to Figure 3. Similarly, discrepancies based on educational backgrounds were also apparent in other phases, such as root canal disinfection and temporary sealing, root canal filling, postoperative supplies organization, doctor-patient and nursing coordination, as depicted in Figure 3.

Figure 3. The perceptions of respondents with different educational backgrounds regarding the integration of school teaching and clinical practices across 5 stages of the randomized controlled trial (RCT).



In Figure 3, junior college degree holders who selected negative evaluation options such as "completely unsatisfied," "most cannot satisfy," and "slightly unsatisfied" constituted 50% (11/22), 45.5% (10/22), 59.1% (13/22), 54.6% (12/22), and 63.6% (14/22) of their demographic, respectively. In contrast, those who opted for positive evaluations like "completely satisfied," "a little to satisfy," and "slightly satisfied" represented 36.4% (8/22), 40.9% (9/22), 27.3% (6/22), 31.8% (7/22), and 22.7% (5/22) of the same group. On the other hand, bachelor's and master's degree students who chose negative evaluation options made up 26.2% (16/61), 16.4% (10/61), 24.6% (15/61), 21.3% (13/61), and 32.8% (20/61) of their respective totals, while those who selected positive evaluation options accounted for 70.5% (43/61), 77.0% (47/61), 70.5% (43/61), 68.9% (42/61), and 62.3% (38/61) of their groups.

In the phases of root canal disinfection and temporary sealing operations, it is apparent that the "40~" age group perceives a significantly higher alignment between school teaching and clinical practices compared with the "21-29" age group (P<.05). Among participants aged 40 and above, 50% (5/10) selected "slightly satisfied," followed by 30% (3/10) who chose "a little to satisfy" and 20% (2/10) who opted for "completely satisfied." Conversely, within the 21–29 age group, 23.5% (12/51) of participants indicated "slightly unsatisfied", which trailed only behind the 25.4% (13/51) who chose the "a little to satisfied" option. For further details, refer to Figure 4. The 40+ age group (yellow) exclusively opted for positive attitudes (choosing "completely satisfied," "a little to satisfy," and "slightly satisfied" options), whereas 41.2% (21/51) of the 21 - 29 age group (blue) selected negative attitudes (selecting "completely unsatisfied," "most cannot satisfy," and "slightly unsatisfied" options).

On the issue of the practical difficulty of clinical treatment, respondents' evaluations were homogeneous, revealing no significant categorical disparities. With respect to the overall difficulty of clinical treatment, 82% of respondents (68/83) selected the options of "slightly difficult" and "normal," exhibiting a skewness value of 0.701. In the assessment of the difficulty of each phase of RCT, respondents generally perceived the difficulty of each stage to lie between "normal" and "slightly easy," with skewness values lower than the overall difficulty skewness value. Detailed findings are illustrated in Figure 5.

Regarding perception issues in certain phases, the responses of the participants have exhibited significant variances. For instance, in the equipment adjustment phase, the congruence of the participants from the Shanghai area is generally higher than that of the participants from other regions, and they also perceive the clinical practice difficulty of this phase to be relatively low (P<.05). For the composition of the difference in the equipment adjustment phase, refer to Figure 5A: the graph of the sample of the Shanghai area shows a positively skewed distribution, with a skewness value of 0.385; the graph of the sample from other areas exhibits a negatively skewed distribution, with a skewness value of −1.013. In other phases such as indications judgment, equipment adjustment, root canal preparation, root canal disinfection and temporary sealing, root canal filling, postoperative supplies organization, and the phase of doctor-patient and nursing coordination in the clinic, the graphical representation of the sample from the Shanghai area also reveals a significantly positive bias compared with those from other areas, as detailed in Figure 6.

In the survey assessing the disparities between clinical work and on-campus learning, 69.9%(58/83) of the respondents opined that the scenario of judging the indications for RCT in the clinic is more intricate, 68.7% (57/83) of the respondents

held the view that there is a marked divergence in the modes and methods of doctor-patient communication in the clinic vis-a-vis the related instruction in the academic setting, and with regard to the configuration and application of equipment, 51.8% (43/83) of the respondents contended that there is a significant variance between the clinical RCT work and the on-campus pedagogy. See for details.

When asked about the teaching methods encountered in the school, 78% (64/83) of the respondents received cavity preparation and filling training in the school, 75.6% (62/83) of the respondents reported that video demonstration was a prevalent teaching method, and 61% (50/83) of the respondents were exposed to the practice of taking x-ray imaging in the school. However, 25.6% (21/83) of the respondents were exposed to dental modeling software practice and digital assessment and review, and only 3.7% (3/83) of the respondents were exposed to VR or AR in the school learning process.

56.6% (47/83) of the respondents believed that there is too much didactic teaching in the school teaching process, while the proportion of practical practice is insufficient. 48.2% (40/83) of the respondents advocated for an enhancement in the identification and treatment teaching of different pulp conditions. 45.8% (38/83) of the respondents expressed the desire for an augmentation in the identification and treatment instruction of various periodontal conditions. In addition, respondents who believed that the school should focus on the content of patient psychological care/doctor-patient communication and appointment schedules in the clinic also reached 43.4% (36/83). See Figure 7 for details.

Finally, this study collated and categorized the self-statements of the participants in the “other” option. In the clinical practice of RCT, there is a noted difficulty in identifying or using/disbursing tools or medications: 20% (3/15) of the respondents think it is difficult to use the Gutta Percha Obturation Guns in the root canal filling phase. 33.33% (5/15) of the dental practitioners believed residual pulpitis is a common complication of RCT. In addition, 20% (3/15) of the respondents emphasized that in the on-campus instruction, the importance of the practice instructors cannot be understated, of which 66.67% (2/3) are college degree holders.

Figure 4. In the context of root canal disinfection and temporary sealing phase, respondents of different age groups expressed their views on the integration of school teaching and clinical operations.

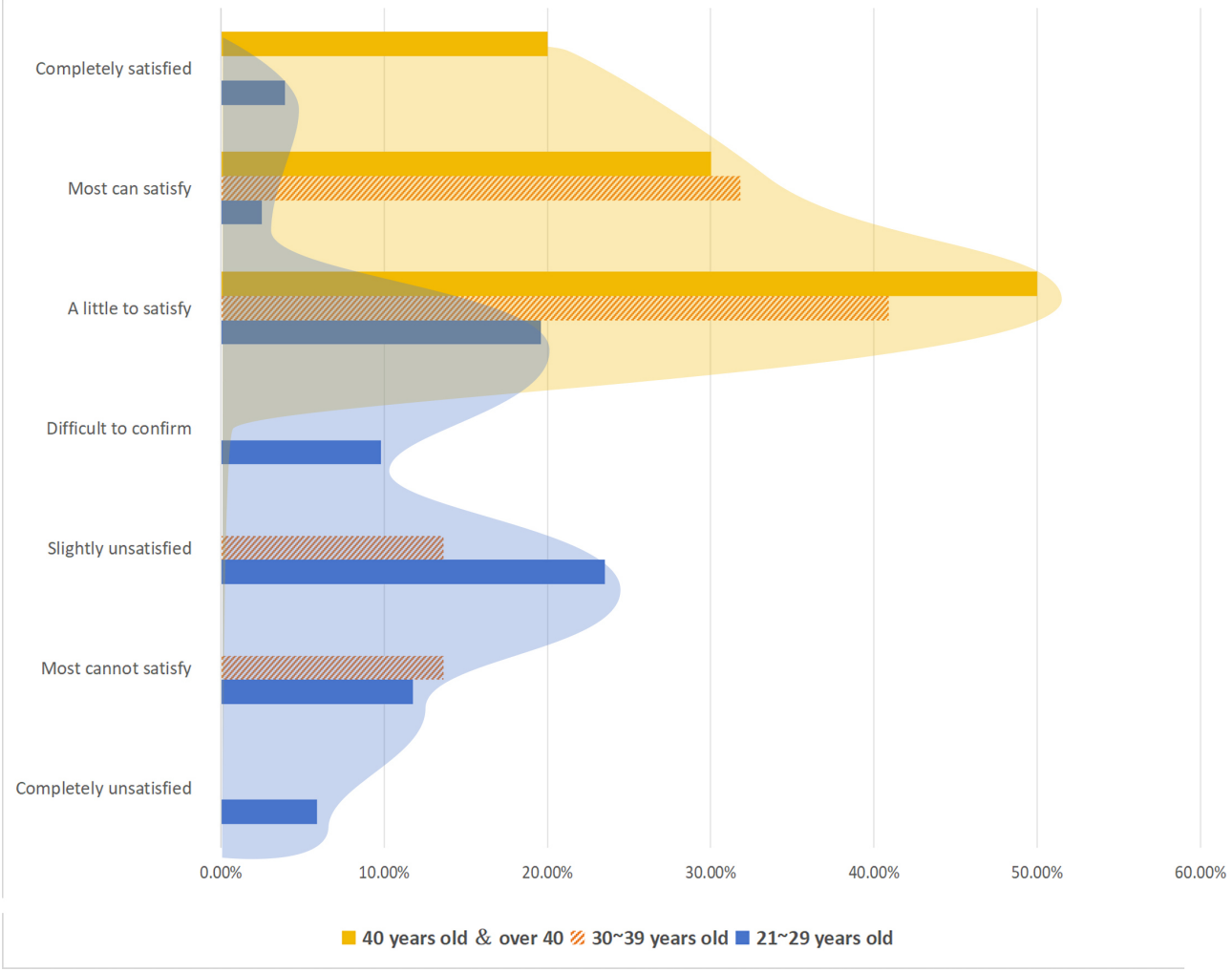


Figure 5. The trend in respondents' answers to the overall difficulty of root canal treatment and the difficulty of each stage is depicted as follows: the solid line (representing the overall difficulty of a randomized controlled trial [RCT]) has a skewness value of 0.701, which is larger than the skewness values of the dashed lines (representing the difficulty of each individual stage).

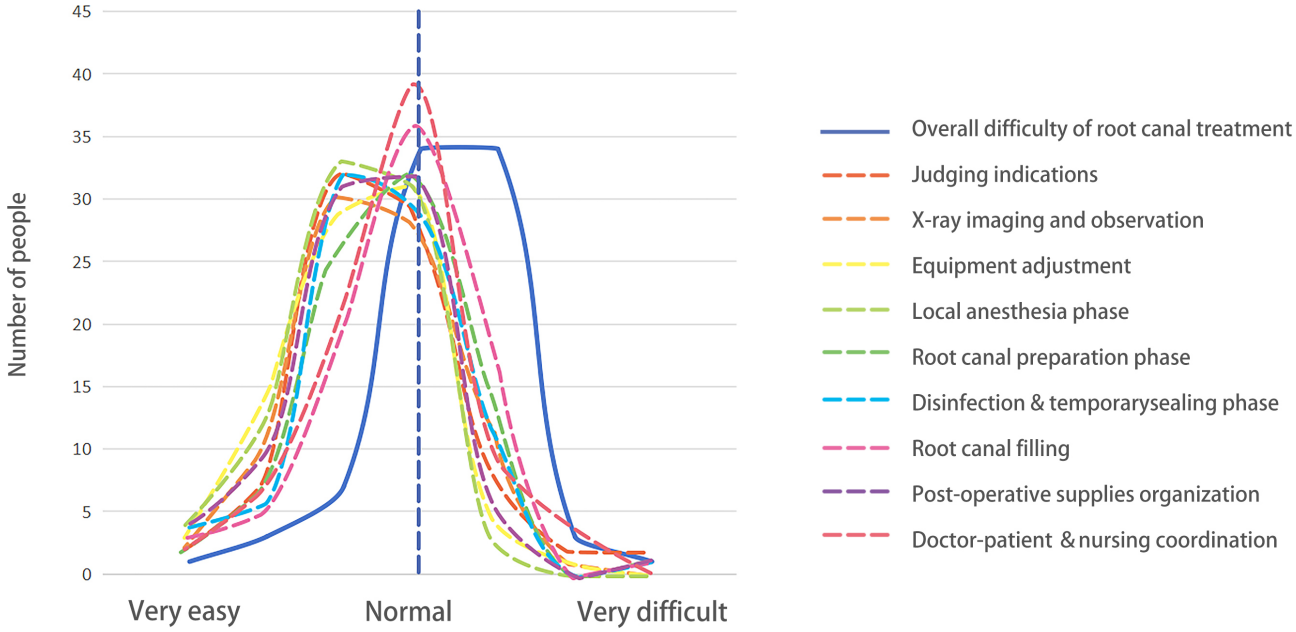


Figure 6. The trend in responses to questions about the variability in certain phases of root canal treatment shows that the skewness values for the Shanghai population samples, from left to right across the 7 graphs (A-G), are 0.385, 0.105, 0.567, 0.323, 0.499, 0.578, and 0.758, respectively. In contrast, the corresponding skewness values for the population samples from other regions are −1.013, −0.339, −0.306, −0.942, −0.383, −0.211, and −0.215.

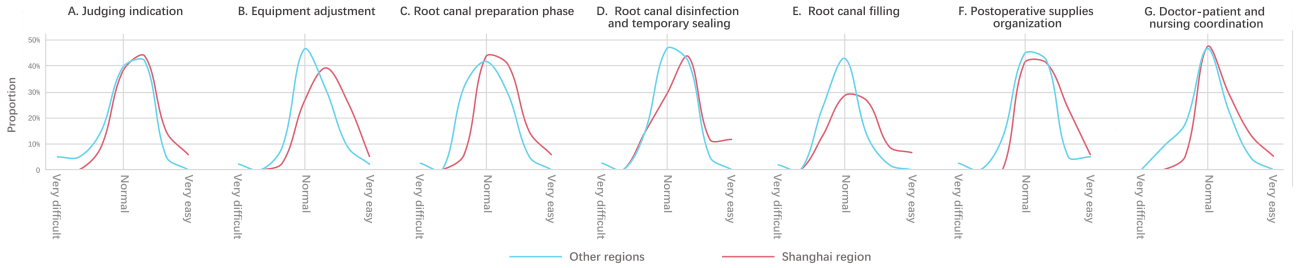
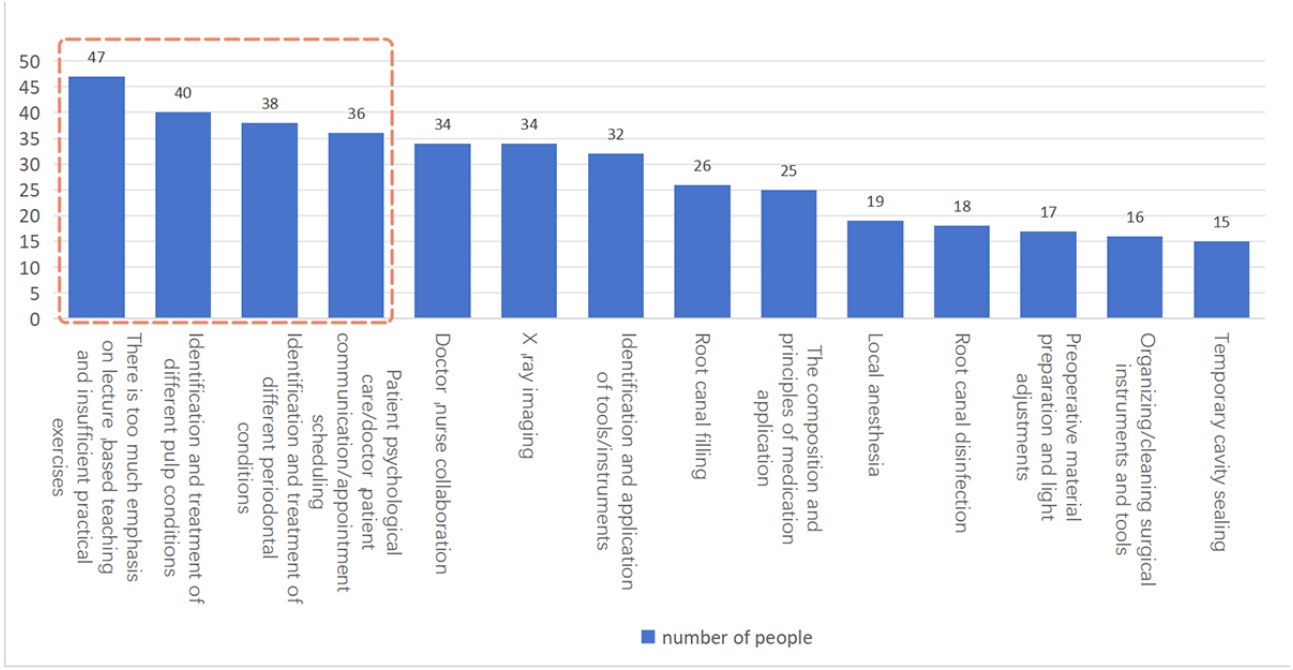


Figure 7. Respondents considered an improved aspect of the school: the red box denotes the 4 aspects of root canal treatment teaching that a significant number of participants suggest should be improved.



Discussion

Principal Findings

The results of this study revealed multiple factors affecting the satisfaction of root canal therapy teaching in the Southeastern Region of China, including educational background, regional differences, age stratification, and insufficient periodontal disease teaching content. These findings not only reflect the intuitive feedback of clinical medical staff on teaching effectiveness but also provide important references for optimizing dental education.

Education Stratification and Uneven Distribution of Teaching Resources

This study shows that medical staff with a specialized degree evaluate the integration of teaching and clinical practice significantly lower than those with undergraduate and master's degrees. This difference may be closely related to the uneven distribution of teaching resources. In China, most specialized colleges focus their dental courses on basic operation training, while systematic teaching of complex techniques such as root canal therapy is mostly concentrated at the undergraduate level and above. Similar phenomena were also mentioned in the studies of Hua [16] and Jiang et al [17], which pointed out that the number of endodontic class hours in Chinese specialized colleges is only 50% - 60% that of undergraduate colleges, and there is a lack of standardized textbook support. It can be seen that China's specialized education urgently needs to increase the depth of teaching in root canal therapy to narrow the capability gap between different educational groups.

Further analysis showed that the teaching content of specialized education is insufficient for determining the indications of root canal therapy and dealing with complex root canal anatomical shapes, leading to a lack of systematic thinking when graduates face complex clinical cases [16]. It is worth noting that nearly half of China's oral medical practitioners have a specialized background [18,19], and their professional abilities directly affect the quality of primary medical care. However, there is very limited research on specialized dental teaching and professional skills. We searched Google Scholar with "pulpology," "root canal treatment," "teaching," and "students" as core keywords. In the related Chinese literature in the past 6 years (2017 - 2024), only Hua [16] and Jiang et al [17] mentioned the problem of oral education for people with a specialized degree, while 92.3% (24/26) of the studies focused on people with undergraduate degrees and above. This reflects a serious imbalance between educational research and practical needs.

Regional Differences Reflect Inequality in Educational Resources

Respondents in the Shanghai area had a significantly higher level of satisfaction with teaching quality compared with other regions, a result highly correlated with regional disparities in the distribution of oral medical resources. A study on national oral health resource surveys also showed issues of inequality in the distribution of the national oral labor force and institutions [19]. Beijing has the highest proportion of dentists, while Tibet has the lowest [15]. In addition to the differences caused by economics between provinces and cities, there are also differences within cities. For example, although the total number of dental care personnel in Shanghai is relatively adequate, its distribution is unfair, with fewer dental care personnel employed in suburban areas [18]. Furthermore, clinical internship bases in the Shanghai area are mostly top-level hospitals, offering students more opportunities to engage with cutting-edge technologies compared with intern hospitals in other regions.

This disparity essentially reflects the synergistic effect of "education-healthcare" resources. In economically developed regions, clinical needs drive the update of teaching content, forming a virtuous cycle. The application of advanced technology in clinical settings leads to the transformation of teaching cases, which in turn enhances student capabilities. While in less developed areas, limited by equipment and teachers, the teaching content lags behind clinical practice [20], resulting in students facing the dilemma of "not taught in school, but needed in clinical practice."

Age Factor's Influence on the Satisfaction of Root Canal Teaching

This study also found significant differences in the perception of certain skills based on age, such as the discrepancy in the root canal disinfection and temporary sealing operation. As shown in Figure 3, the perception of the difficulty of clinical root canal treatment varies with age. The "40 years old and above" group had a significantly higher level of satisfaction with the integration of teaching and clinical practice than the "21–29 years old" group. The "40 years old and above" group generally had a more positive perception of school teaching, while nearly half of the "21–29 years old" group was not very satisfied with school teaching. The reasons for this diversity could be multifaceted. On one hand, over the past 20 years, there may have been changes in the teaching content and methods related to clinical root canal treatment in China's dental specialty programs. On the other hand, older respondents, due to their longer work experience, may have memory biases about their school teaching and, because of their extensive work experience, may be less sensitive to the perceived difficulty of root canal procedures. In addition, there may be a disconnect between teaching content and clinical practice. Some teaching may not keep pace with clinical changes or emphasize certain operational aspects in school teaching, causing some students to struggle with the clinical practice's complexity. Further investigation is needed to clarify these situations.

Interdisciplinary Impact of the Periodontal Disease Teaching Gap

Nearly half of the respondents called for an increase in teaching on differential diagnosis between periodontal and pulp diseases, exposing the current lack of interdisciplinary integration in the curriculum. In other countries, the issue of periodontal disease teaching is also prominent. In the United States and Europe, there is a problem of poor consistency in periodontal disease teaching [21,22]. The conclusions of these previous surveys align with the conclusions of this study, indicating that periodontal disease teaching is a challenge in both oral education and practice. Periodontal disease courses in Chinese schools

receive the least amount of time [13]. Due to insufficient class hours, students lack proficiency in key skills related to periodontal diseases, leading to a tendency to overlook periodontal factors in clinical operations and increasing treatment risks.

New Media Technology

Despite the widely recognized potential of VR or AR technology in dental education [22], only 3.7% of respondents in this study had exposure to such technology. This situation reveals the slow digital transformation in China's oral education. The reasons for this are two-fold: first, most institutions lack the financial support to purchase VR equipment on a large scale; second, teachers receive inadequate training on the pedagogical suitability of new technologies.

Other Considerations

In terms of difficulty perception, the results of this study reflected an interesting phenomenon, respondents generally believed that the overall difficulty of root canal treatment was higher than the difficulty of each individual step. As shown in Figure 4, this indicates a lack of "holistic thinking" training in teaching. Currently, most institutions adopt a step-by-step teaching approach (such as practicing root canal preparation and obturation separately), but clinical operations require integrated capabilities across multiple dimensions, such as "case assessment, treatment plan design, instrument selection, and emergency management." Students lack awareness of the "cascade effects of operational errors" during their school years (such as strategies for managing perforations caused by excessive root canal preparation). Drawing on experiences from other disciplines, Jiangxi University of Traditional Chinese Medicine has adopted "full-process situational simulation" as a teaching method, effectively improving the effectiveness of clinical skills training. It is recommended that China's dental specialty teaching include "clinical scenario comprehensive training" to train students' decision-making abilities through high-fidelity models and case libraries, narrowing the gap between "item-by-item proficiency" and "overall competence."

Due to the wide age range of the samples in this study, there may be differences in teaching methods across different periods. However, after excluding the older sample population, we found that age does not interfere with the impact of other factors (such as education level, region, etc.) on the satisfaction of root canal treatment teaching. Therefore, it suggests that age is not the primary factor, and the disconnection between root canal treatment teaching and clinical practice still exists to some extent.

Strengths and Limitations

This study adopts the perspective of in-service oral health care workers to retrospectively examine issues in root canal treatment teaching, providing a better understanding of whether school teachings can be effectively applied in clinical practice. The random sampling in this study, without limiting educational background, well reflects the situation and needs of clinical oral health care workers. Moreover, this study uses a Likert 7-point scale, which provides more accurate sensitivity assessments compared with the commonly used 5-point scale.

The main limitation of this study is that the data was primarily collected from the Shanghai area and surrounding cities in the Southeastern part of China, without distinguishing between urban and rural areas, thus preventing further analysis on urban-rural disparities. In addition, due to the limitation of the sample size, the generalizability of the findings from this survey needs to be further improved, necessitating further research to evaluate the effectiveness of school teaching and the details of clinical operations and the difficulty of RCTs.

Conclusions

This cross-sectional investigation systematically identifies structural discrepancies between preclinical RCT education and clinical workflows in China, with experiential, demographic (eg, training backgrounds), and geographic factors significantly modulating educational-clinical alignment. The findings reveal structural disparities in factors such as clinical experience, educational qualifications, and hospital location. These disparities influence clinical treatment plan decisions, potentially leading to overtreatment or inappropriate management, thus exacerbating patient discomfort, extending treatment duration, and increasing treatment burden. Optimizing classroom content (VR or AR, thermal obturation systems, etc) can expedite students' transition to clinical practice and bolster their confidence in the diagnostic and therapeutic process. By correlating clinicians' educational satisfaction metrics with workplace competency demands, this study establishes a novel evaluative framework for dental pedagogy reform, emphasizing competency-based curricula, standardized clinical protocols, and technology-enhanced immersive learning. Subsequent studies can further expand the sample sizes through multicenter sampling across diverse regions to evaluate the changes in academic–clinical consistency.

Acknowledgments

The authors sincerely thank Prof. Zhou Jin and Dr. Hibrising for their unparalleled care and support during the surveys and all dental practitioners who helped us to finish the data collection.

Data Availability

The datasheets of the study are available after communication with the corresponding author.

Authors' Contributions

XM contributed to the conception and design of the study, data collection, analysis and interpretation of the results, drafting of the manuscript, and review and editing of the manuscript. JH contributed to the conception and design of the study, guided data analysis, and participated in drafting, reviewing, and editing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary material.
[DOCX File, 23 KB - mededu_v11i1e65534_app1.docx]

References

1. Tibúrcio-Machado CS, Michelon C, Zanatta FB, et al. The global prevalence of apical periodontitis: a systematic review and meta-analysis. Int Endod J 2021 May;54(5):712-735. [doi: 10.1111/iej.13467] [Medline: 33378579]
2. Haug SR, Linde BR, Christensen HQ, Vilhjalmsson VH, Bårdsen A. An investigation into security, self-confidence and gender differences related to undergraduate education in Endodontics. Int Endod J 2021 May;54(5):802-811. [doi: 10.1111/iej.13455] [Medline: 33253460]
3. Dahlström L, Lindwall O, Rystedt H, Reit C. "Working in the dark": Swedish general dental practitioners on the complexity of root canal treatment. Int Endod J 2017 Jul;50(7):636-645. [doi: 10.1111/iej.12675] [Medline: 27374421]
4. Albino J, Dye BA, D'Souza RN. The decades ahead for dental education. J Dent Educ 2022 Jun;86(6):635-636. [doi: 10.1002/jdd.12932] [Medline: 35611789]
5. Field J, Stone S, Orsini C, et al. Curriculum content and assessment of pre-clinical dental skills: A survey of undergraduate dental education in Europe. Eur J Dent Educ 2018 May;22(2):122-127. [doi: 10.1111/eje.12276] [Medline: 28636116]
6. Scott J. Dental education in Europe: the challenges of variety. J Dent Educ 2003 Jan;67(1):69-78. [Medline: 12540108]
7. Fu D, Yao L, Zhu H, et al. The landscape of endodontic education research area: A bibliometric analysis. J Dent Educ 2023 May;87(5):711-720. [doi: 10.1002/jdd.13170] [Medline: 36646984]
8. Li C, Zheng J, Guo C, et al. An introduction to clinical practice guideline for Chinese undergraduates in stomatology. Eur J Dent Educ 2014 May;18(2):110-114. [doi: 10.1111/eje.12064] [Medline: 24118682]
9. Wang YH, Zhao Q, Tan Z. Current differences in dental education between Chinese and Western models. Eur J Dent Educ 2017 Nov;21(4):e43-e49. [doi: 10.1111/eje.12216] [Medline: 27339198]
10. Sonntag D, Bärwald R, Hülsmann M, Stachniss V. Pre-clinical endodontics: a survey amongst German dental schools. Int Endod J 2008 Oct;41(10):863-868. [doi: 10.1111/j.1365-2591.2008.01438.x] [Medline: 18699788]
11. Sacha SR, Sonntag D, Burmeister U, Rüttermann S, Gerhardt-Szép S. A multicentric survey to evaluate preclinical education in Endodontology in German-speaking countries. Int Endod J 2021 Oct;54(10):1957-1964. [doi: 10.1111/iej.13584] [Medline: 34081783]
12. Segura-Egea JJ, Zarza-Rebollo A, Jiménez-Sánchez MC, et al. Evaluation of undergraduate Endodontic teaching in dental schools within Spain. Int Endod J 2021 Mar;54(3):454-463. [doi: 10.1111/iej.13430] [Medline: 33063865]
13. Chen Y, Deng J, Li B, et al. Curriculum setting and students' feedback of pre-clinical training in different dental schools in China-A national-wide survey. Eur J Dent Educ 2022 Feb;26(1):28-35. [doi: 10.1111/eje.12669] [Medline: 33511722]
14. Yan X, Zhang X, Shen Y, et al. Career choice and future plan of Chinese 8-year stomatology medical doctor program students. J Chin Med Assoc 2015 Sep;78(9):555-561. [doi: 10.1016/j.jcma.2015.06.006] [Medline: 26298259]
15. Liu DL, Xie YF, Shu R. Statistical analysis of current oral health care and dental education resources in China. Chin J Dent Res 2019;22(1):37-43. [doi: 10.3290/j.cjdr.a41773] [Medline: 30746531]
16. Hua Z. Differences and characteristics of the dental degree between China and USA[J]. Med Educ Manag 2017;3(1):23-27. [doi: 10.3969/j.issn.2096-045X.2017.01.006]
17. Jiang H, Shen L, Zhang Y, Yang J. Attitudes towards and use of dental dams by final-year dental students in Chongqing, China: a cross-sectional study. BMJ Open 2022 Jul;12(7):e059148. [doi: 10.1136/bmjopen-2021-059148]
18. Gu Q, Lu HX, Feng XP. Status of the dental health care workforce in Shanghai, China. Int Dent J 2012 Dec;62(6):331-336. [doi: 10.1111/j.1875-595x.2012.00132.x] [Medline: 23252591]
19. Sun XY, Yuan C, Wang XZ, et al. Report of the national investigation of resources for oral health in China. Chin J Dent Res 2018;21(4):285-297. [doi: 10.3290/j.cjdr.a41087] [Medline: 30264045]
20. Bo H, Zhang DH, Zuo TM, et al. Survey and analysis of the current state of residency training in medical-school-affiliated hospitals in China. BMC Med Educ 2014 Jun 2;14:111. [doi: 10.1186/1472-6920-14-111] [Medline: 24885865]
21. Gürsoy M, Wilensky A, Claffey N, et al. Periodontal education and assessment in the undergraduate dental curriculum-A questionnaire-based survey in European countries. Eur J Dent Educ 2018 Aug;22(3):e488-e499. [doi: 10.1111/eje.12330] [Medline: 29460375]

22. Dhoble D, Raina D, Maheshwari D. Dental intern's perception of difficulties in performing root canal treatment: A cross sectional study. Int J Appl Dent Sci 2023 Jan 1;9(1):90-96. [doi: 10.22271/oral.2023.v9.i1b.1661]

Abbreviations

RCT: root canal treatment

Edited by T Gladman; submitted 19.08.24; peer-reviewed by AV Mahuli, MC Skelton, SS Oberoi; revised version received 25.04.25; accepted 20.05.25; published 29.07.25.

Please cite as:
Ma X, Huang J
Alignment Between Classroom Education and Clinical Practice of Root Canal Treatment Among Dental Practitioners in China: Cross-Sectional Study
JMIR Med Educ 2025;11:e65534
URL: https://mededu.jmir.org/2025/1/e65534
doi:10.2196/65534

Enhancing Preclinical Training for Removable Partial Dentures Through Participatory 3D Simulation: Development and Usability Study

Yikchi Siu, BDS; Hefei Bai, BDS, PhD; Jung-Min Yoon, BDS; Hongqiang Ye*, DDS, PhD; Yunsong Liu*, DDS, PhD; Yongsheng Zhou, DDS, PhD

Department of Prosthodontics, Peking University School and Hospital of Stomatology, National Center of Stomatology, National Clinical Research Center for Oral Diseases, National Engineering Research Center of Oral Biomaterials and Digital Medical Devices, No. 22 Zhongguancun South Avenue, Haidian District, Beijing, China

*these authors contributed equally

Corresponding Author:
Yunsong Liu, DDS, PhD
Department of Prosthodontics, Peking University School and Hospital of Stomatology, National Center of Stomatology, National Clinical Research Center for Oral Diseases, National Engineering Research Center of Oral Biomaterials and Digital Medical Devices, No. 22 Zhongguancun South Avenue, Haidian District, Beijing, China

Abstract

Background: The integration of digital technology in dental education has been recognized for its potential to address the challenges in training removable partial denture (RPD) design. RPD framework design is crucial to long-term success in the treatment of dentition defects, but traditional training methods often fall short of adequately preparing students for real-world applications.

Objective: This study aimed to evaluate the efficacy of a 3D simulation–based preclinical training software for RPDs in enhancing learning outcomes among first-year stomatology master's students, while also assessing user perceptions among students and faculty.

Methods: RTS (Yikchi Siu) is a preclinical training software that simulates the clinical process of treating patients with partial edentulism. In this study, 26 newly enrolled master's degree students in stomatology who volunteered to participate were randomly divided into a control group (n=13) and a training group (n=13). The training group used the RTS for 2 credit hours (90 min) of self-study, while the control group received theoretical lessons and case practice from an instructor. After 2 hours, both groups completed the theoretical knowledge and drawing tests for RPD simultaneously. Test results were evaluated and graded by 2 experts in prosthodontics. Both users and teachers filled out a questionnaire afterward about their training experience.

Results: Participants in the training group obtained better final grades compared to controls (theoretical test: 88.8, SD 2.3; 85.7, SD 3.3, respectively; P=.01; drawing test: 89.8, SD 4.5; 85.1, SD 4.3, respectively; P=.01). The training group had a shorter completion time in the drawing test (12.6, SD 19 min; 17.7, SD 3 min, respectively; P<.001) but there were no significant differences in the completion times in the theoretical test (23.2, SD 2.2 min; 24.9, SD 2.8 min, respectively; P=.14). Students and faculty generally had a favorable opinion of the RTS.

Conclusions: The effectiveness of the RTS for newly enrolled master's degree students in stomatology to understand and apply their knowledge of RPD framework design was validated; the system was well received by both students and faculty members, who reported that it improved the effectiveness and convenience of teaching.

(JMIR Med Educ 2025;11:e71743) doi:10.2196/71743

KEYWORDS

removable partial dentures; preclinical training; simulation; 3D; participatory

Introduction

As life expectancy and the size of the aging population continue to increase, so does the prevalence of patients with partial edentulism [1]. Removable partial dentures (RPDs) are versatile and cost-effective alternatives to fixed and implant restorations, and thus an increasing number of patients are choosing them to restore their dentition [2-4].

In the conventional training paradigm, students predominantly use 2D graphic representations and textual exercises to design RPD frameworks [5]. However, this is very different from how it is done in actual clinical practice [6,7]. In fact, RPD framework design is an ongoing challenge within the context

of prosthodontics training [8]. This is largely attributable to the intricate nature of RPD components and the considerable diversity of design proposals [9,10]. The design process requires the synthesis of theoretical knowledge and clinical practice; that is, a high degree of clinical decision-making ability is necessary [7,11].

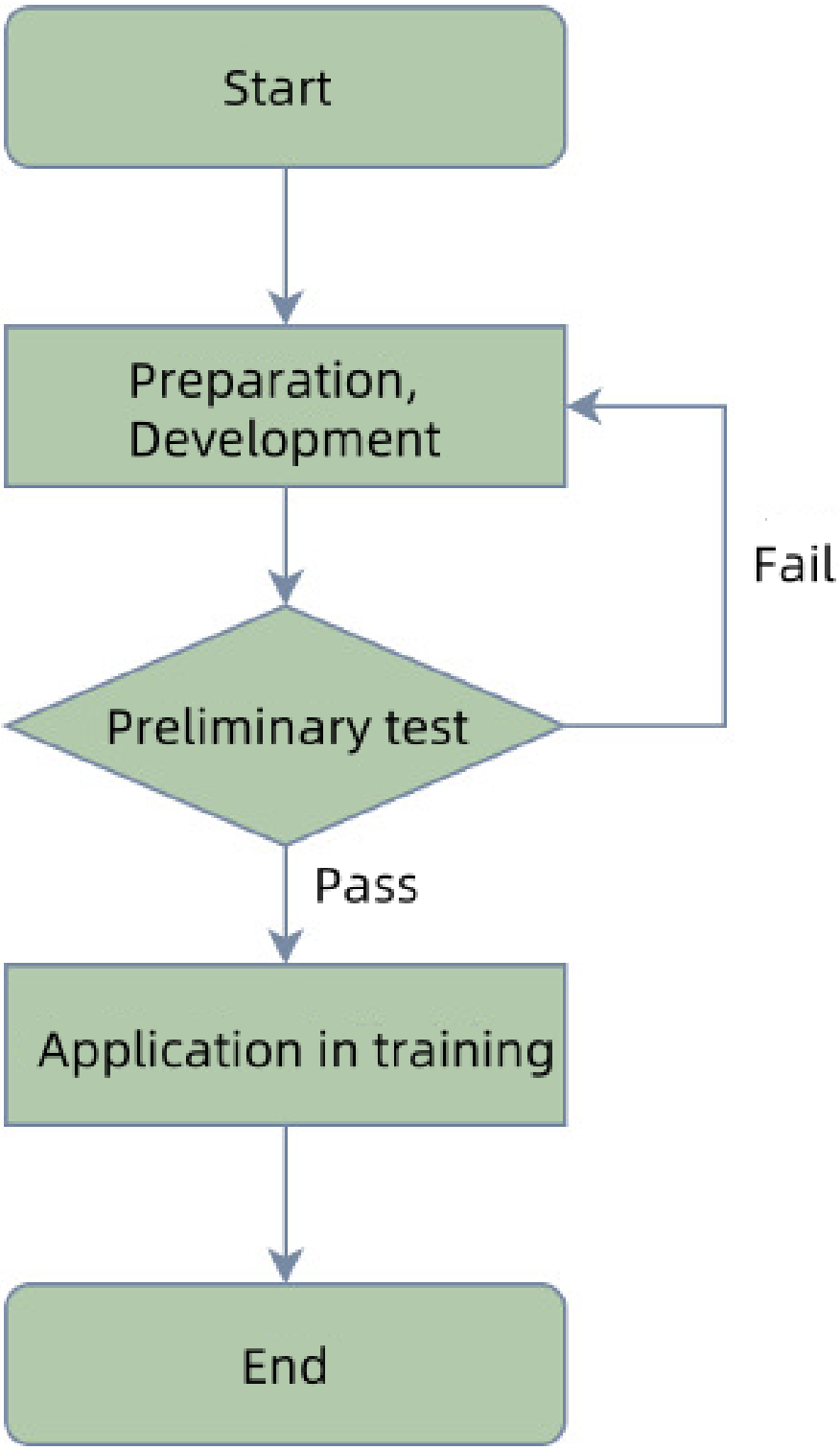
The advent of simulation technology in dentistry has led to immersive training systems including some based on force feedback and mixed reality [12]. However, there is a lack of studies on nonimmersive simulation approaches. In this context, to enhance the efficiency of training in the design of RPD frameworks, the use of training systems that incorporate 3D casts has increasingly been recognized as an innovative and reliable approach [13]. One reason for this is the popularity of digital RPD design software, which enables users to export 3D casts in standard tessellation language format for use in training. The use of such casts facilitates a more comprehensive understanding of the design principles and components of the RPD framework [14-16]. A substantial body of evidence indicates that integrating such systems into teaching curricula can significantly enhance student interest and learning outcomes as well as the efficacy of teaching [5,16,17]. This study describes novel 3D RPD simulation software designed for preclinical training in RPD framework design and provides a preliminary evaluation of its efficacy. We hypothesized that RTS (Yikchi Siu) would improve the test scores of RPD design and reduce completion time compared to traditional training methods.

Methods

RPD Training System

RTS is a participatory training program developed to enhance the RPD design skills of users without preclinical training. We designed and tested it in 3 phases: software preparation and development, preliminary tests, and application in training (Figure 1). During the development phase (August 2023 to July 2024), a prototype was prepared, designed, and developed. In the test phase (August 2024), 5 students (3 master's students and 1 doctoral student) enrolled at the Peking University School of Stomatology were recruited to test the system, and based on the outcomes, a panel of experts was asked to suggest modifications to the RTS; the system was modified accordingly. Finally, in the training phase (September 2024 to October 2024), the RTS was introduced to 26 newly enrolled postgraduate students who had not received preclinical training, and the system was quantitatively and qualitatively assessed by analyzing their test scores and feelings about its use. The study design followed the recommended framework for software development [18] and the criteria for the validation process [19].

Figure 1. Flowchart of RTS development.

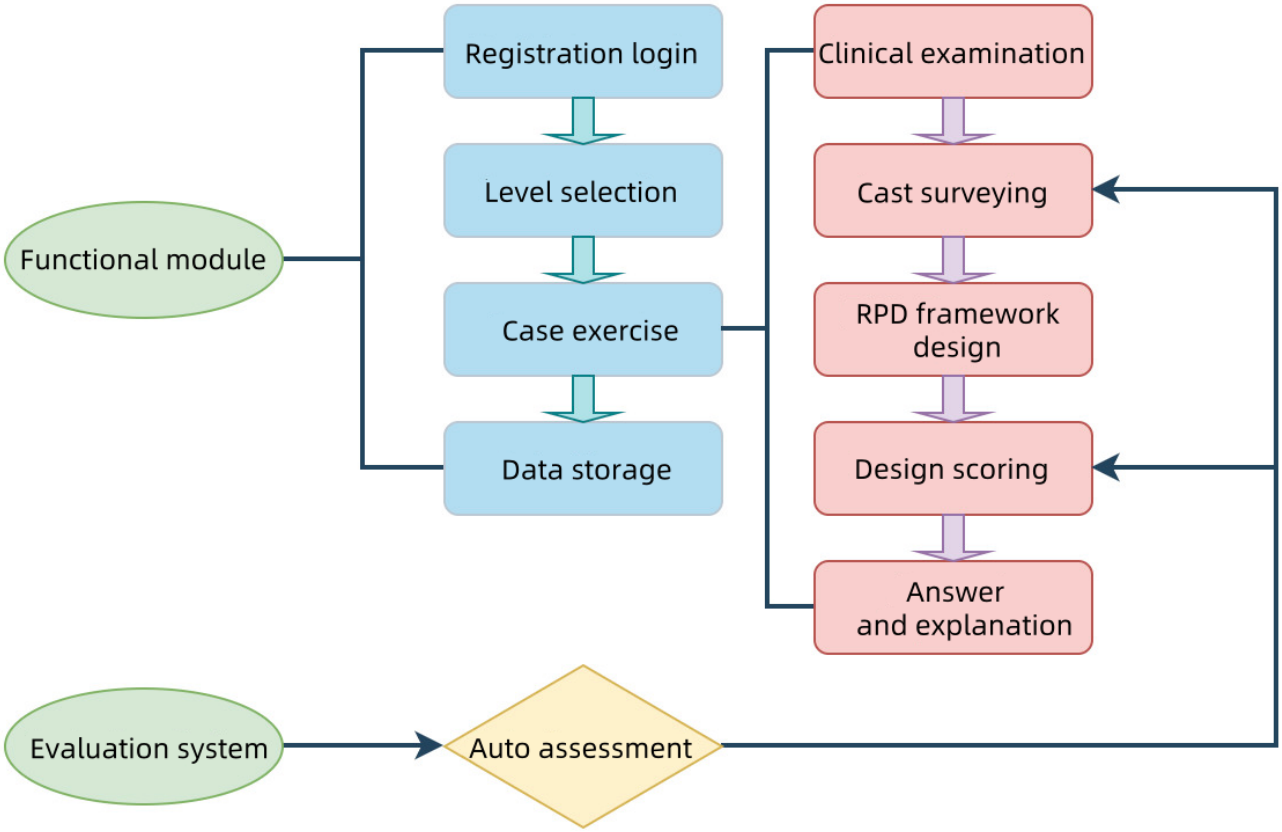


RTS was developed using Visual Studio Code and is a standalone program for Windows 10 (Microsoft Corporation). The software includes 3D digital casts created from clinical plaster casts from 19 dental patients. These cases were carefully selected and prepared by the software developers and then reviewed and approved for relevance and accuracy by 2 prosthodontics experts. The digital casts were created using the 3Shape cast scanner, replicating real-world clinical conditions.

The training content is divided into 5 topics: clinical examination, cast surveying, RPD framework design, design scoring, and answer and explanation (Figure 2). The RTS includes a functional module and an evaluation system. The former facilitates user engagement, and the latter involves assessment, design, and evaluation of user outcomes. The auto assessment component quantitatively assesses user performance, providing a comprehensive educational tool for preclinical dental training. Users view the 3D cast and perform digital surveying with mouse clicks. If the surveying perspective is

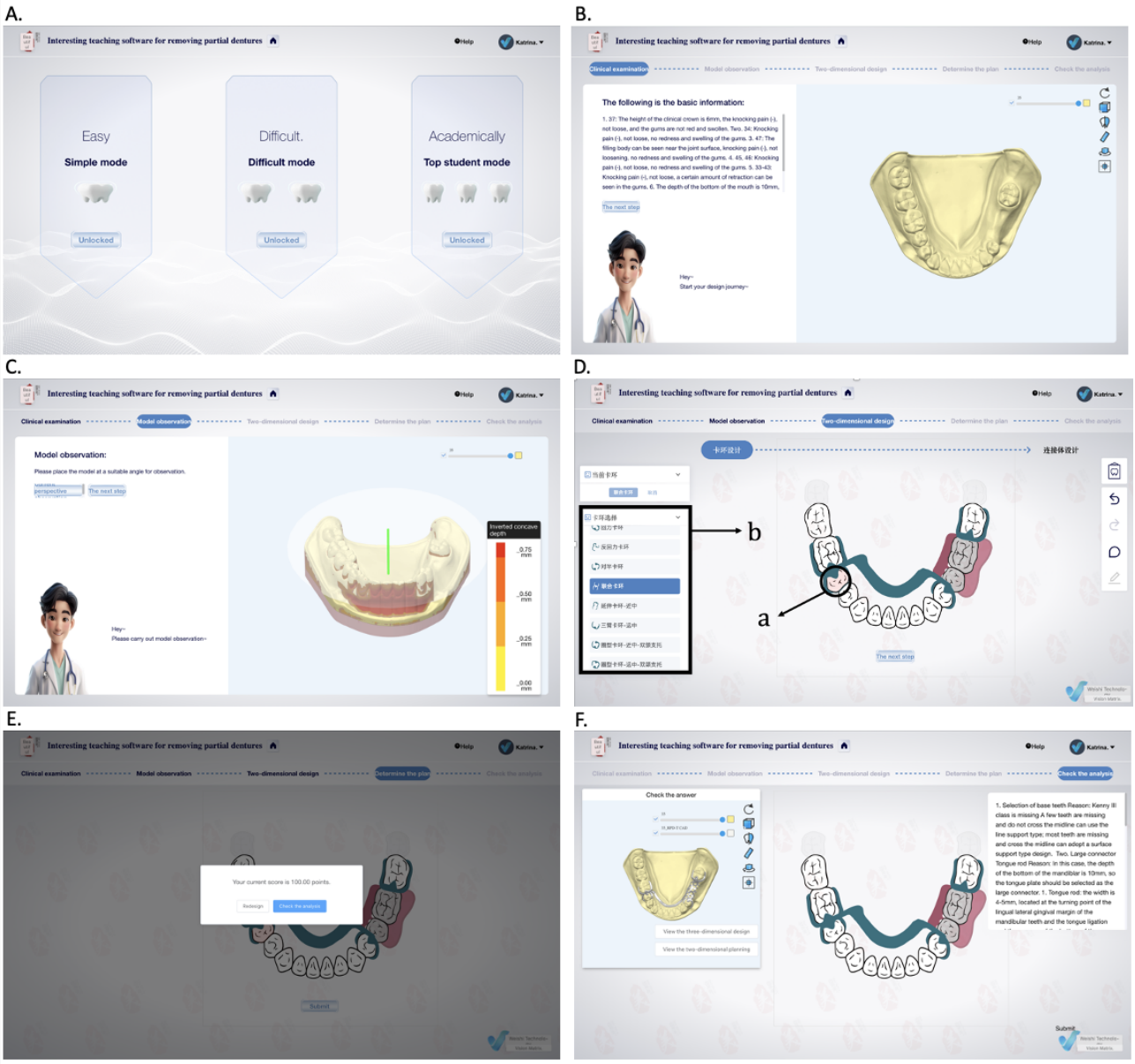
incorrect, an error message appears, and the user cannot proceed with the RPD framework design until the cast is oriented correctly. Once the surveying angle is correct, a confirmation message is displayed. Users then design an RPD framework blueprint based on the clinical examination and cast survey data.

Figure 2. Framework of RTS. RPD: removable partial denture.



After completing the design, users submit it for automatic scoring by the RTS evaluation system, which provides immediate feedback, including a standardized design for comparison. Faculty members can also review and manually grade the designs in the background (Figure 3). Figure 3 shows an example of a Kennedy classification III case exercise. The user logs on to the webpage and selects an available training mode (Figure 3A). Each mode contains 6 to 7 practice cases, all of which must be successfully completed before the next mode will be unlocked. The user clicks and rotates the 3D cast of a dental defect to view it, determine the Kennedy type, and obtain information about the intraoral clinical examination (Figure 3B). Then the user adjusts the appropriate viewing angle to survey the inverted concave area (Figure 3C). Next, the user designs the framework for RPD (a: red indicates the selected tooth as an abutment; b: selection of an appropriate clasp design), including features such as abutment teeth, clasps, rests, and major and minor connectors, to complete the design for the case (Figure 3D). The design is submitted and the system automatically scores it (Figure 3E). A system score of 60 and above will display the standardized design plan and an explanation; otherwise, the case must be designed again (Figure 3F). To start the RPD case practice, users must open the provided URL [20], register, and log in. Then the system automatically saves their practice records.

Figure 3. Screenshots from the RTS.



Application

Sample Size Calculation

A pilot study will be conducted with 5 newly admitted graduate students per group. Sample sizes will be calculated independently for both primary end points (drawing test scores and test time), with the larger of the 2 values being adopted to guarantee adequate power for all outcome assessments. Using pilot data of drawing test time (control group: 17.7, SD 2.7 min; training group: 12.6, SD 1.8 min) with α=.05 (2-tailed) and 80% power, the minimum sample size was 7 per group (total 14). Considering educational intervention complexity, we enrolled 13 per group (total 26).

Quantitative Assessment

For quantitative assessment, 26 newly enrolled graduate students who had not received RPD preclinical training volunteered to assess the system. In addition, the students were recruited and randomly allocated via sealed envelope drawing into 2 groups, the control group (n=13) and the training group (n=13).

Both groups first underwent a preclinical training course for RPDs. The instructor provided an explanation of the functionality of the software to ensure that each participant had the requisite knowledge and proficiency to use it. Then the training group was instructed to use the RTS to learn and practice the RPD framework design independently. By contrast, the graduate students in the control group were required to learn the theoretical aspects of RPD and complete the associated classroom exercises under the guidance of an instructor. During the period, the control group was not permitted to use RTS for learning purposes. At the conclusion of the experiment, both groups were administered a quiz assessing their theoretical knowledge of the RPD framework design and their design ability. The test included 10 clinical cases of dentition defects which were developed by 2 experts in prosthodontics, including 3 cases each of Kennedy Class I and II, and 2 cases each of

Class III and IV. The assessment was conducted by 2 prosthodontic experts using a 100-point grading scale established based on RPD design principles and requirements from textbooks [21,22] and the National Dental Practitioner Training Manual (Table 1). The final scores were calculated as the mean values of the ratings given by 2 independent experts for each student group's assessments.

Table . Removable partial denture design drawing evaluation rubric.

Category (subcriteria)	Maximum score	Remarks
I. Basic design principle (20)		
Fulcrum line design	6	• Consistency with task description
Force distribution	6	• Balanced occlusal loading
Antirotation or sinking measures	4	• Indirect retainers or rests
Lever arm	2	• Minimize cantilever
Stress breaker	2	• If applicable
II. Component design (20)		
Direct retainers	10	• Clasp type selection
Indirect retainers	5	• Position rationality
Major connectors	5	• Type selection
III. Clinical applicability (10)		
Kennedy classification	5	• Correct I-IV
Esthetic considerations	2.5	• Consistency with case description
Functional considerations	2.5	• Consistency with case description
IV. Drawing (50)		
Missing tooth position	20	• Accurate identification and correct marking
Color	10	• Blue metal components and red resin bases
Neatness	10	• Neat and accurate
Component drawing	20	• Component positioning and drawing • Minor connector spacing and drawing • Finish line

Qualitative Assessment

To evaluate the subjective efficacy of the RTS, a questionnaire was administered to students and staff to assess user experiences as well as instructor and student satisfaction with the system (Multimedia Appendix 1). The questionnaire included 10 questions, with respondents indicating their level of satisfaction or agreement with statements ranked on a 5-point Likert item (1=strongly dissatisfied or strongly disagree to 5=strongly satisfied or strongly agree) [12]. The questionnaire includes 6 student-specific questions and 4 teacher-specific questions. An open-ended question was also included to elicit feedback on potential shortcomings.

Statistical Analysis

This study used a combination of descriptive and inferential statistics to evaluate the scores and time spent on theoretical and drawing tests. First, all data were summarized descriptively using means and SD to illustrate the central tendency and dispersion of each group. Normality was assessed using the Shapiro-Wilk test, which was prioritized for its robustness with small sample sizes (n<50). Homogeneity of variance was evaluated via *F* tests to validate the assumptions for subsequent parametric tests. Results confirmed that all 4 datasets met both normality and homoscedasticity assumptions. Independent samples *t* tests were used for 2-group analyses, reporting *t* values, degrees of freedom (*df*), and *P* values. Mann-Whitney *U* tests were concurrently performed to ensure robustness, reporting *U*, *z*, and *P* values. Effect sizes were quantified using Cohen *d* to complement statistical significance. To examine linear agreement between 2 experts' scores, interrater reliability was assessed through Pearson correlation coefficient and intraclass correlation coefficient (ICC) under a 2-way

random-effects absolute agreement model, accounting for both systematic differences and random error between raters. Statistical evaluation was performed using the SPSS analysis program (IBM SPSS Statistics 26.0, IBM Corp). GraphPad PRISM 10.0 software (GraphPad Software) was used to create graphs.

Ethical Considerations

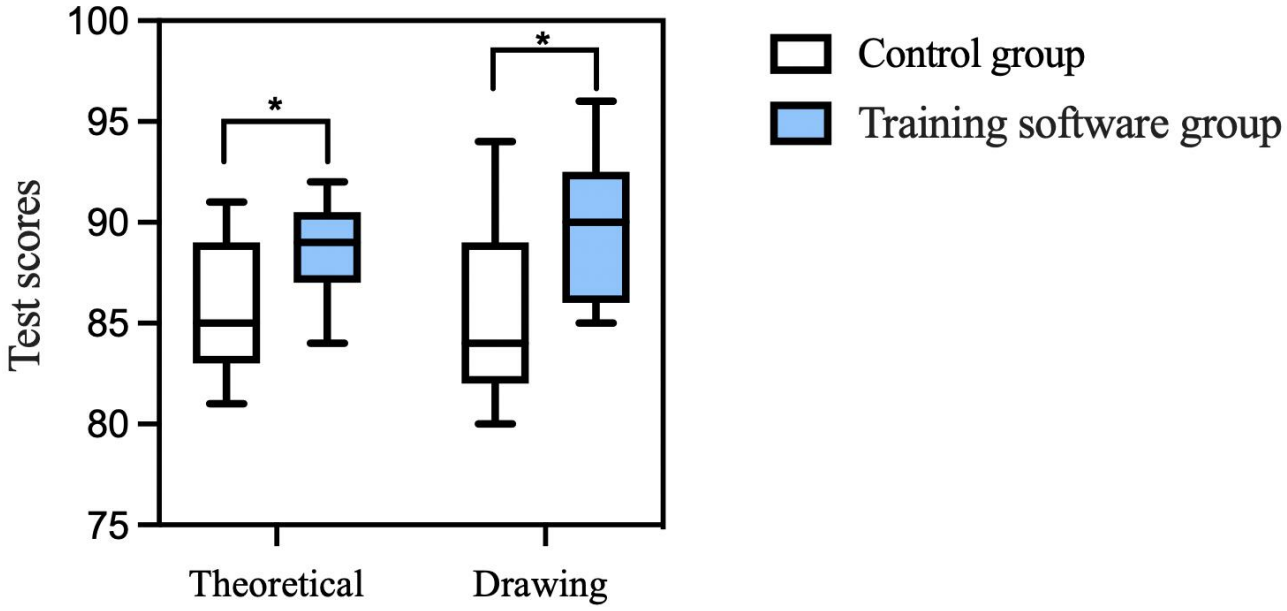
This study was approved by the ethics committee of Peking University School and Hospital of Stomatology (PKUSSIRB-202273003-免). Participants understood the study details and their rights, including their right to withdraw at any point without consequences. Participants' privacy is secured in all written and published data arising from this study. Names of participants or other identifying information are not and will not be used in reports or published papers.

Results

Quantitative Assessment: Test Scores

As shown in Figure 4, the theoretical test score in the training group was 88.8 (SD 2.3), while the drawing test score was 89.8 (SD 4.5). These values were 85.7 (SD 3.3) and 85.1 (SD 4.3), respectively, in the control group. Thus, the training group had significantly better scores for both tests (P=.01). Moreover, the Mann-Whitney U test was further applied for evaluation, and the P value remained less than .05, both groups with a large effect size (Cohen d = 1.1). The nonparametric test confirmed this result on the theoretical test (U=33.0, z=–2.5, P=.01) and the drawing test (U=33.0, z=–2.5, P=.01) (Figure 4). A strong positive correlation was observed between the raters' evaluations (Pearson r=0.70, P<.001; ICC=0.84, 95% CI 0.7-0.9).

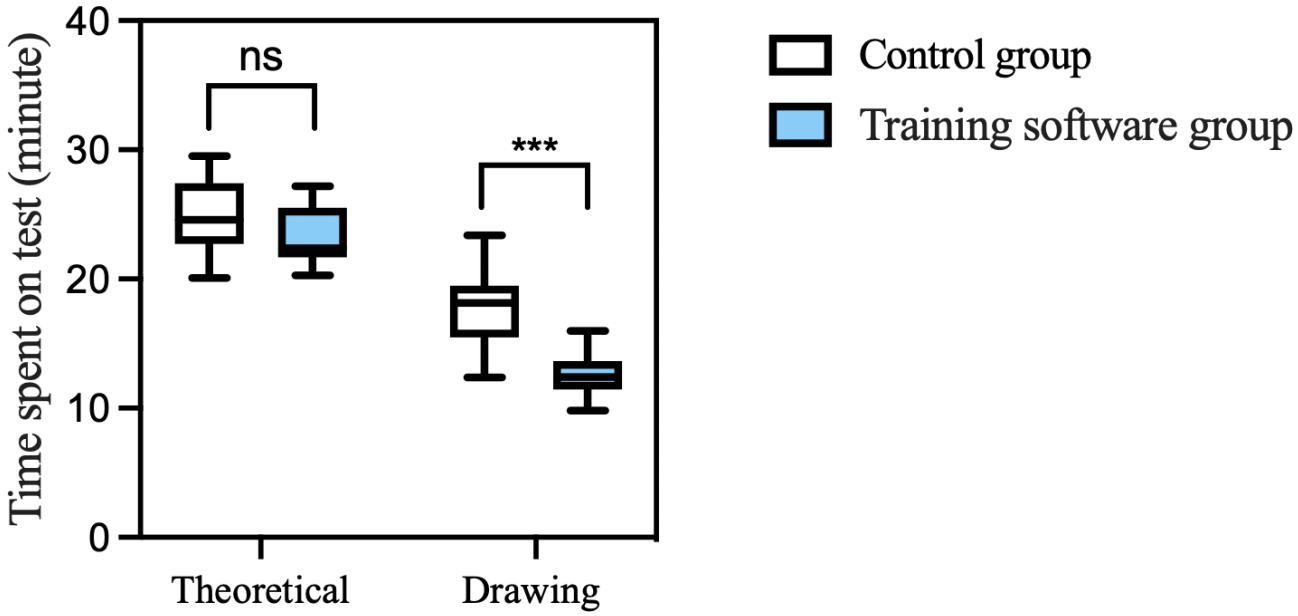
Figure 4. Comparison of test scores between the control and training groups. *P<.05.



Quantitative Assessment: Test Time

As shown in Figure 5, the mean completion times were 23.2 (SD 2.2) minutes in the training group and 24.9 (SD 2.8) minutes in the control group for the theoretical test; they were 12.6 (SD 1.9) minutes and 17.7 (SD 3) minutes for the drawing test, respectively, with a significant difference in the latter test only (P<.001).

Figure 5. Time spent on tests. ***P<.001, ns: not significant.



Response to Questionnaire

As shown in Table 2, users found the RPD framework design to be easier and more accessible when using the RTS. In addition, all users indicated that the program was convenient and straightforward to use. Users indicated a greater preference for the RTS over traditional training methods and expressed great satisfaction with the responsive features. Faculty members also had positive views, reporting that the system significantly reduced preparation time for teaching and improved teaching effectiveness.

Table . Responses of the training group (n=13) to the questionnaire.

Questionnaire item	Very dissatisfied or strongly disagree, n (%), 1 Point	Dissatisfied or disagree, n (%), 2 Points	Neutral, n (%), 3 Points	Satisfied or agree, n (%), 4 Points	Very satisfied or strongly agree, n (%), 5 Points	Mean (SD)
The software program encourages users to practice in designing RPD[a] frameworks.	0 (0)	0 (0)	0 (0)	1 (8)	12 (92)	4.9 (0.3)
The software program is highly responsive and allows for real-time feedback.	0 (0)	0 (0)	0 (0)	2 (15)	11 (85)	4.8 (0.4)
The software program is easy to use.	0 (0)	0 (0)	0 (0)	0 (0)	13 (100)	5 (0)
The software program effectively helps users to learn and practice RPD design, enhancing decision-making and critical skills.	0 (0)	0 (0)	1 (8)	1 (8)	11 (85)	4.8 (0.6)
The software program allows users to learn more about surveying 3D casts and digitally drawing designs.	0 (0)	0 (0)	0 (0)	0 (0)	13 (100)	5 (0)
I would like to use the RTS for skills training in the future.	0 (0)	0 (0)	1 (8)	3 (23)	9 (69)	4.6 (0.6)

[a]RPD: removable partial denture.

Discussion

Principal Findings

Based on the above experiments, it can be concluded that the RTS significantly improved both theoretical and drawing test scores related to RPD design among newly enrolled graduate students. Moreover, the system substantially reduced the time required for students to complete RPD design drawings. The positive feedback from both students and faculty members further suggests that RTS shows great promise as a preclinical training software for RPD design.

With the accelerated development of digital technology and the emergence of new educational models, online responsive training methods are playing an increasingly pivotal role in the field of medical education [23-25]. Such methods have notable benefits in terms of reduced teaching costs, providing an immersive experience, enabling resource sharing, and offering personalized learning and feedback, to name a few [26,27]. In the field of dental education, RPD design is a key ongoing challenge and area of focus. Students frequently have difficulty grasping the concepts related to RPD components and lack access to case-based instruction [5,28]. This is due to the inherent limitations of traditional teaching methods, which often lack the necessary casts and RPD frameworks needed for effective preclinical training, as well as the inability to practice on clinical cases [16,29,30]. The AiDental system was previously designed to teach RPD design, and has demonstrated effective teaching outcomes for RPD design instruction in dental education. However, Mahrous et al [31] have revealed 2 persistent limitations in the study: a suboptimal user interface design that may reduce learner engagement and insufficient analytical feedback on student performance. The design of Musawi et al [32] contains a specific clinical RPD process but uses 2D graphics, which lack a sense of immersion and interactivity. The RTS overcomes these limitations.

Using RTS, students can treat dentition defects and design RPD frameworks in a simulated clinical setting. The absence of a surveyor or plaster casts during the surveying and design processes ensures that the training content is more closely aligned with the clinical environment, thereby creating a more realistic training experience. This approach allows the trainer to integrate theoretical knowledge with clinical practice effectively, facilitating a smoother transition from theory to practice.

Regarding its efficacy, the training group achieved better results than did controls in both the theory and drawing tests. This means that the program helped users to understand and master RPD design theory better, increasing the design speed. This

may be because the program allows users to create personalized study plans based on their individual learning abilities and mastery levels, which significantly improves learning efficiency.

There were no statistically significant differences in the times taken by the 2 groups to perform the theory test, while the training group spent significantly less time completing the drawing test. This indicates that the program facilitated a more comprehensive understanding, making students more proficient in RPD design. The immediate feedback on, and comprehensive analysis of, incorrect responses provides a rapid-response mechanism for challenging content, enabling users to identify knowledge gaps accurately in a time-efficient manner. Moreover, the case practice and personalized guidance module of the program facilitates one-on-one training for students, in contrast with the traditional classroom setting.

Finally, both students and teachers viewed the software program positively. Students gave high ratings for user-friendliness, indicating that it allowed them to grasp essential functions quickly. They expressed a high level of interest in the RTS and perceived it to be more engaging and participatory than conventional paper-based RPD design. In addition, users were more inclined to dedicate time to the RTS to gain insight into RPD design. In fact, all of the users (n=13) indicated that the program allowed them to learn how to survey casts digitally, which is particularly beneficial in the context of the increasing use of CAD (Computer-Aided Design) software and 3D casting. Furthermore, users indicated a desire to do more digital training over traditional training in the future.

Teachers reported that the system notably reduced the time required for lesson planning. Instructor preparation time decreased from 90 minutes (traditional methods) to 45 minutes using RTS. They also felt that it made teaching easier and resulted in improved outcomes. Finally, they indicated a willingness to integrate such systems into future lesson plans as a potential option for self-study and after-school practice with students.

While the current system demonstrates promising educational outcomes, several refinements could further enhance its efficacy. Incorporating advanced haptic feedback technology could improve students' understanding of RPD biomechanics by simulating real-world force distribution during clasp design and framework adaptation. Recent studies have shown that force-feedback systems significantly enhance skill acquisition in preclinical dental training, particularly in prosthodontic procedures requiring precise tactile sensitivity [33,34]. In addition, integrating augmented reality for surveying and undercut analysis could bridge the gap between simulation and clinical application, as augmented reality has proven effective in improving spatial awareness in dental education [35]. To maximize impact, future iterations should explore embedding this tool into official dental curricula as a standardized preclinical training module. Structured integration, as seen with successful simulation platforms in medical education [23], would facilitate widespread adoption. Multi-institutional collaborations could also validate their scalability, similar to the collaborative frameworks used in the Association for Dental Education in Europe guidelines for simulation-based training [36]. By addressing these refinements and curricular integration pathways, the system has the potential to become a benchmark for RPD education.

Limitations and Future Work

The limitations of this study include the relatively small sample size, the recruitment of participants from a single school, and the absence of a randomization process. In addition, there may have been unidentified differences in the intrinsic and current learning abilities of the students. To enhance the rigor and reliability of the study, a randomization process could be used to balance the groups and mitigate the impact of confounding variables, thereby yielding higher-quality evidence. There is a paucity of data on the efficacy of such training modules for dental students. Consequently, future studies should collect more data to validate the long-term effectiveness and clinical transferability of these methods in real-world preclinical training settings, thereby facilitating their broader adoption [37,38]. Due to the inability to perform real-time adjustments for nonstandard anatomical structures, the model surveying function in RTS demonstrates poorer adaptability for complex cases. Manual operation with traditional surveyors enhances spatial awareness and biomechanical understanding, which digital tools may not fully replicate. So, it is important to note that the surveying function in the RTS cannot replace the use of an actual physical surveyor for preclinical training. This feature is designed solely to provide students with preliminary practice in understanding model survey angles. Comprehensive mastery of model surveying techniques still requires training with a physical surveyor.

Simulation can be an indispensable tool for training students so that they develop the necessary skills for authentic real-world scenarios. Our results show that the RTS effectively enhances the efficacy of RPD design instruction. It offers a flexible training module that can be adapted to evolving trends in dental education.

Conclusions

The novel RTS effectively improves the teaching and practice of RPD design. This study has described the system and offered a preliminary validation of its effectiveness. The system was viewed positively by both students and teaching staff. It enhanced the design competence of students not previously exposed to standardized preservice training in prosthodontics. It also made teaching easier and led to better teaching and learning outcomes. Thus, it can be considered an effective tool for teaching and learning RPD design.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (82170929), the National Center of Stomatology Suitable Technology Promotion Project (2023NCSHTP04), and the Young Beijing Scholars Program (2024 - 089).

Authors' Contributions

YL, HY, YS, and YZ were involved in the design and development of the RTS. YS was responsible for the organization of this manuscript and authored the paper. YX, JMY, and HB collected the data and revised the manuscript. YL, HY, and HB conceived the study and revised the manuscript. All of the authors approved the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
User satisfaction questionnaire for RTS.
[DOCX File, 18 KB - mededu_v11i1e71743_app1.docx]

References

1. Kimble R, Papacosta AO, Lennon LT, et al. The relationships of dentition, use of dental prothesis and oral health problems with frailty, disability and diet quality: results from population-based studies of older adults from the UK and USA. J Nutr Health Aging 2023;27(8):663-672. [doi: 10.1007/s12603-023-1951-8] [Medline: 37702340]
2. Shah S, Kapadia UH, Patel P, et al. Removable partial denture outcomes in an academic dental setting. J Dental Health Oral Res 2024 Nov;5(3):1-6. [doi: 10.46889/JDHOR.2024.5310]
3. Qiu J, Liu W, Wu D, Qiao F, Sui L. Fit accuracy in the rest region of RPDs fabricated by digital technologies and conventional lost-wax casting: a systematic review and meta-analysis. BMC Oral Health 2023 Sep 15;23(1):667. [doi: 10.1186/s12903-023-03348-6] [Medline: 37715159]
4. Fueki K, Inamochi Y, Yoshida-Kohno E, Wakabayashi N. Cost-effectiveness analysis of prosthetic treatment with thermoplastic resin removable partial dentures. J Prosthodont Res 2021 Feb 24;65(1):52-55. [doi: 10.2186/jpr.JPOR_2019_418] [Medline: 32938866]
5. Mahrous A, El-Kerdani T. Teaching the design and fabrication of RPD frameworks with a digital workflow: a preclinical dental exercise. MedEdPORTAL 2020 Oct 30;16:11041. [doi: 10.15766/mep_2374-8265.11041] [Medline: 33150206]
6. Wang X, Ma D, Zhong S, et al. A digital workflow for designing and manufacturing metal frameworks and removable partial dentures: a novel dental technique. J Prosthodont 2024 Apr 3. [doi: 10.1111/jopr.13845] [Medline: 38566576]
7. Soltanzadeh P, Suprono MS, Kattadiyil MT, Goodacre C, Gregorius W. An in vitro investigation of accuracy and fit of conventional and CAD/CAM removable partial denture frameworks. J Prosthodont 2019 Jun;28(5):547-555. [doi: 10.1111/jopr.12997] [Medline: 30407685]
8. Almufleh B, Arellanob A, Tamimi F. Patient-reported outcomes and framework fit accuracy of removable partial dentures fabricated using digital techniques: a systematic review and meta-analysis. J Prosthodont 2024 Aug;33(7):626-636. [doi: 10.1111/jopr.13786] [Medline: 37930081]
9. Patel J, Jablonski RY, Hodson TM. Removable partial dentures: Part 1. Br Dent J 2024 Oct;237(7):537-542. [doi: 10.1038/s41415-024-7893-7] [Medline: 39394297]
10. Bonnet G, Lance C, Bessadet M, et al. Teaching removable partial denture design: 'METACIEL', a novel digital procedure. Int J Med Educ 2018 Jan 26;9:24-25. [doi: 10.5116/ijme.5a5b.2b0d]
11. Fidler A, Kučič AC, Gašperšič R, Kuralt M. Measurements of gingival recession changes: comparison of the conventional clinical and direct digital approach. J Dent (Shiraz) 2022 Jun;121:103974. [doi: 10.1016/j.jdent.2022.103974]
12. Li Y, Ye H, Wu S, et al. Mixed reality and haptic–based dental simulator for tooth preparation: research, development, and preliminary evaluation. JMIR Serious Games 2022 Mar 9;10(1):e30653. [doi: 10.2196/30653] [Medline: 35262501]
13. Liu K, Xu Y, Ma C, et al. Efficacy of a virtual 3D simulation–based digital training module for building dental technology students' long-term competency in removable partial denture design: prospective cohort study. JMIR Serious Games 2024 Apr 5;12:e46789. [doi: 10.2196/46789] [Medline: 38596827]
14. Arnold C, Hey J, Schweyen R, Setz JM. Accuracy of CAD-CAM-fabricated removable partial dentures. J Prosthet Dent 2018 Apr;119(4):586-592. [doi: 10.1016/j.prosdent.2017.04.017] [Medline: 28709674]
15. Rezaie F, Farshbaf M, Dahri M, et al. 3D printing of dental prostheses: current and emerging applications. J Compos Sci 2023 Feb;7(2):80. [doi: 10.3390/jcs7020080] [Medline: 38645939]
16. Mahrous A, Schneider GB, Holloway JA, Dawson DV. Enhancing student learning in removable partial denture design by using virtual three-dimensional models versus traditional two-dimensional drawings: a comparative study. J Prosthodont 2019 Oct;28(8):927-933. [doi: 10.1111/jopr.13099] [Medline: 31343801]
17. Ward SM, Balolia KL, Wilson LAB. A preliminary analysis of the effectiveness of online practical laboratory delivery using 3D models for higher education courses in biological anthropology. Evo Edu Outreach 2023 Jul;16(1):12. [doi: 10.1186/s12052-023-00190-w]
18. Cederbladh J, Cicchetti A, Suryadevara J. Early validation and verification of system behaviour in model-based systems engineering: a systematic literature review. ACM Trans Softw Eng Methodol 2024 Mar 31;33(3):1-67. [doi: 10.1145/3631976]

19. Santana SR, Antonelli LR, Thomas PJ. Best practices for requirements validation process. In: Communications in Computer and Information Science: Springer Science and Business Media Deutschland GmbH; 2022:139-156. [doi: 10.1007/978-3-031-05903-2_10]
20. Welcome to the removable partial denture interactive teaching software [Web page in Chinese]. Vision Matrix. URL: http://8.130.173.181/ [accessed 2024-11-06]
21. Phoenix RD, Cagna DR, DeFreest CF. Stewart's Clinical Removable Partial Prosthodontics, 4th edition: Quintessence Publishing; 2015.
22. Carr AB, Brown DT. McCracken's Removable Partial Prosthodontics, 13th edition: Elsevier; 2016.
23. Pottle J. Virtual reality and the transformation of medical education. Future Healthc J 2019 Oct;6(3):181-185. [doi: 10.7861/fhj.2019-0036] [Medline: 31660522]
24. Ogundiya O, Rahman TJ, Valnarov-Boulter I, Young TM. Looking back on digital medical education over the last 25 years and looking to the future: narrative review. J Med Internet Res 2024 Dec 19;26:e60312. [doi: 10.2196/60312] [Medline: 39700490]
25. McGee RG, Wark S, Mwangi F, et al. Digital learning of clinical skills and its impact on medical students' academic performance: a systematic review. BMC Med Educ 2024 Dec 18;24(1):1477. [doi: 10.1186/s12909-024-06471-2] [Medline: 39696150]
26. Li Z, Li F, Fu Q, et al. Large language models and medical education: a paradigm shift in educator roles. Smart Learn Environ 2024 Jun;11(1):26. [doi: 10.1186/s40561-024-00313-w]
27. Zhang SL, Ren SJ, Zhu DM, et al. Which novel teaching strategy is most recommended in medical education? A systematic review and network meta-analysis. BMC Med Educ 2024 Nov;24(1):1342. [doi: 10.1186/s12909-024-06291-4]
28. Liu Z, Na G, Liu L, Tian S, Shan Y. Investigation and implementation of case-based learning in the sino-foreign joint program of preventive medicine. BMC Med Educ 2024 Nov 28;24(1):1390. [doi: 10.1186/s12909-024-06372-4] [Medline: 39609841]
29. Liu F, Yu H, Wei W, Qin C. I-feed: a robotic platform of an assistive feeding robot for the disabled elderly population. Technol Health Care 2020;28(4):425-429. [doi: 10.3233/THC-202320] [Medline: 32538890]
30. Altintas L, Sahiner M. Transforming medical education: the impact of innovations in technology and medical devices. Expert Rev Med Devices 2024 Sep;21(9):797-809. [doi: 10.1080/17434440.2024.2400153] [Medline: 39235206]
31. Mahrous A, Botsko DL, Elgreatly A, Tsujimoto A, Qian F, Schneider GB. The use of artificial intelligence and game-based learning in removable partial denture design: a comparative study. J Dent Educ 2023 Aug;87(8):1188-1199. [doi: 10.1002/jdd.13225] [Medline: 37186466]
32. Musawi A, Omar H, Allinson R, Al-Wakeel H, Aubrey P, Rahhal M. Interactive tutorial for enhancing removable partial design skills for second year dental students. J Dent Educ 2023 Jun;87 Suppl 1(S1):884-887. [doi: 10.1002/jdd.13136] [Medline: 36315980]
33. Patrascu A, Michel J, Templin C. Recanalization of in-stent chronic total occlusion using intravascular lithotripsy and Firehawk® rapamycin target eluting coronary stents: a case report. Cardiol J 2021;28(6):991-992. [doi: 10.5603/CJ.2021.0143] [Medline: 34985122]
34. Borgogna NC, Griffin KR, Grubbs JB, Kraus SW. Understanding differences in problematic pornography use: considerations for gender and sexual orientation. J Sex Med 2022 Aug;19(8):1290-1302. [doi: 10.1016/j.jsxm.2022.05.144] [Medline: 35753890]
35. Chen J, Salerno D, Breslin N, et al. Concomitant tacrolimus and ketorolac therapy in pediatric liver transplant recipients: teaching old dogma new tricks. Clin Transplant 2021 Jan;35(1):e14141. [doi: 10.1111/ctr.14141] [Medline: 33145821]
36. Wang C, He W, Li B, Yuan Y. Author response to Letter to the Editor: "Are inflammation-based models feasible tools in predicting the outcome of patients with hepatocellular carcinoma?". Liver Int 2020 Jun;40(6):1499-1500. [doi: 10.1111/liv.14397] [Medline: 32020771]
37. Koolivand H, Shooreshi MM, Safari-Faramani R, et al. Comparison of the effectiveness of virtual reality-based education and conventional teaching methods in dental education: a systematic review. BMC Med Educ 2024 Jan;24(1):8. [doi: 10.1186/s12909-023-04954-2]
38. Daud A, Matoug-Elwerfelli M, Khalid A, Ali K. The impact of virtual reality haptic simulators in pre-clinical restorative dentistry: a qualitative enquiry into dental students' perceptions. BMC Oral Health 2024 Aug 23;24(1):988. [doi: 10.1186/s12903-024-04704-w] [Medline: 39180025]

Abbreviations

ICC: intraclass correlation coefficient
RPD: removable partial denture

Edited by P Kanzow; submitted 09.02.25; peer-reviewed by JH Kim, N Mungoli, R Schnell; revised version received 05.08.25; accepted 11.08.25; published 19.09.25.

Please cite as:
Siu Y, Bai H, Yoon JM, Ye H, Liu Y, Zhou Y
Enhancing Preclinical Training for Removable Partial Dentures Through Participatory 3D Simulation: Development and Usability Study
JMIR Med Educ 2025;11:e71743
URL: https://mededu.jmir.org/2025/1/e71743
doi:10.2196/71743

A Large-Scale Multispecialty Evaluation of Web-Based Simulation in Medical Microbiology Laboratory Education: Randomized Controlled Trial

Lei Xu[1], PhD; Xichuan Deng[2], BSc; Tingting Chen[2], PhD; Nan Lu[1], PhD; Yuran Wang[1], BSc; Jia Liu[2], MSc; Yanan Guo[1], PhD; Zeng Tu[1], PhD; Yuxin Nie[3]; Yeganeh Hosseini[4]; Yonglin He[1], PhD

[1]Department of Pathogenic Biology, School of Basic Medicine, Chongqing Medical University, 1 Medical College Road, Yuzhong District, Chongqing, China

[2]Pathogen Biology and Immunology Laboratory, Experimental Teaching and Management Center, Chongqing Medical University, Chongqing, China

[3]The Second Clinical College, Chongqing Medical University, Chongqing, China

[4]College of International Education, Chongqing Medical University, Chongqing, China

Corresponding Author:
Yonglin He, PhD
Department of Pathogenic Biology, School of Basic Medicine, Chongqing Medical University, 1 Medical College Road, Yuzhong District, Chongqing, China

Abstract

Background: Traditional laboratory teaching of pathogenic cocci faces challenges in biosafety and standardization across medical specialties. While virtual simulation shows promise, evidence from large-scale, multidisciplinary studies remains limited.

Objective: The study aims to evaluate the effectiveness of integrating virtual simulation with traditional laboratory practice in enhancing medical microbiology education, focusing on the identification of biosafety level 2 pathogenic cocci. The study assessed improvements in student performance, theoretical understanding, laboratory safety, and overall satisfaction, while achieving standardization and cost reduction across multiple medical specialties.

Methods: This randomized controlled trial involved 1282 medical students from 9 specialties. The experimental group (n=653) received virtual simulation training—featuring interactivity and intelligent feedback—prior to traditional laboratory practice, while the control group (n=629) did not receive such training. Our virtual system focused on biosafety level 2 pathogenic cocci identification with dynamic specimen generation.

Results: The experimental group showed significantly improved performance across specialties (P<.05 for each specialty), particularly in clinical medicine, in which the experimental group score was 89.88 (SD 13.09) and the control group score was 68.34 (SD 17.23; P<.001). The students reported that virtual simulation enhanced their theoretical understanding (1268/1282, 98.9%) and laboratory safety (1164/1282, 90.8%) while helping them achieve standardization (790/1282, 61.6%,) and cost reduction (957/1282, 74.6%). Overall student satisfaction reached 97.2% (1246/1282), with distinct learning patterns observed across specialties. The test scores were significantly higher in the experimental group, with a mean of 80.82 (SD 17.10), compared to the control group, with a mean of 67.45 (SD 16.81).

Conclusions: This large-scale study demonstrates that integrating virtual simulation with traditional methods effectively enhances medical microbiology education, providing a standardized, safe, and cost-effective approach for teaching high-risk pathogenic experiments.

(JMIR Med Educ 2025;11:e72495) doi:10.2196/72495

KEYWORDS

medical education; virtual simulation; pathogenic cocci; blended learning; multispecialty evaluation; laboratory safety; cost-effectiveness

Introduction

Pathogenic cocci remain significant pathogens in clinical practice, and experiments that include these microorganisms are crucial for medical education [1-3]. The increasing prevalence of antibiotic resistance and the inherent biosafety risks associated with handling pathogenic microorganisms present substantial challenges in educational settings [4-6]. Traditional laboratory teaching, while essential, faces limitations in ensuring consistent safety standards and providing standardized learning experiences across diverse medical specialties. Virtual simulation technology has emerged as a transformative tool in medical education [7-10]. In microbiology

education specifically, this technology offers unique advantages in visualizing microscopic processes and safely simulating high-risk procedures [11,12]. Recent studies have demonstrated its potential in enhancing spatial understanding and mastery of physiological concepts [13-15]. However, most existing research has focused on single-specialty applications or small-scale implementations, leaving a significant knowledge gap regarding the effectiveness of virtual simulations across multiple medical disciplines [16-18]. To address these challenges, our teaching team developed a comprehensive system called Virtual Simulation Experiment for Pathogenic Cocci in Pus Specimens. This innovative system aims to revolutionize teaching on the subject of pathogenic cocci through 3 key features: dynamic specimen generation, integrated biosafety level 2 (BSL-2) safety protocols, and real-time performance tracking.

This study addresses critical questions in medical laboratory education. Specifically, this study aims to evaluate the effectiveness of virtual simulation across 9 medical specialties, representing one of the largest multidisciplinary investigations in this field. We seek to assess the impact on learning outcomes, safety enhancement, and cost-effectiveness in teaching related to pathogenic cocci experiments, while analyzing specialty-specific learning patterns and their implications for customizing virtual simulation approaches. By addressing these objectives, this research provides comprehensive insights into the integration of virtual simulation in medical microbiology education.

Methods

Virtual Simulation System

Experimental Content

Based on Kern's 6-step curriculum development model [19], for the pathogenic cocci virtual simulation teaching system constructed for this study, we first identified core issues through teaching evaluation. We specifically identified a lack of standardization in BSL-2 pathogen handling and inadequacies in cocci identification skills (ie, problem identification). By analyzing biosafety standards and the needs of the teaching faculty, we established a hierarchical competency framework covering Gram staining, culture isolation, biochemical testing, and integrated diagnosis (ie, needs assessment). The system content strictly aligned with 3 key teaching objectives: the basic modules included standardized operating procedures compliant with aseptic techniques (objective formulation); the intermediate modules integrated a dynamic specimen generation system to enable diversified training (educational strategies); and the advanced modules incorporated an embedded assessment system to monitor diagnostic decision-making capabilities in real time (implementation and evaluation). All experimental procedures are embedded with BSL-2 protection protocols, and the precision of the virtual simulation meets the operational standards of BSL-2 laboratories (quality control during implementation).

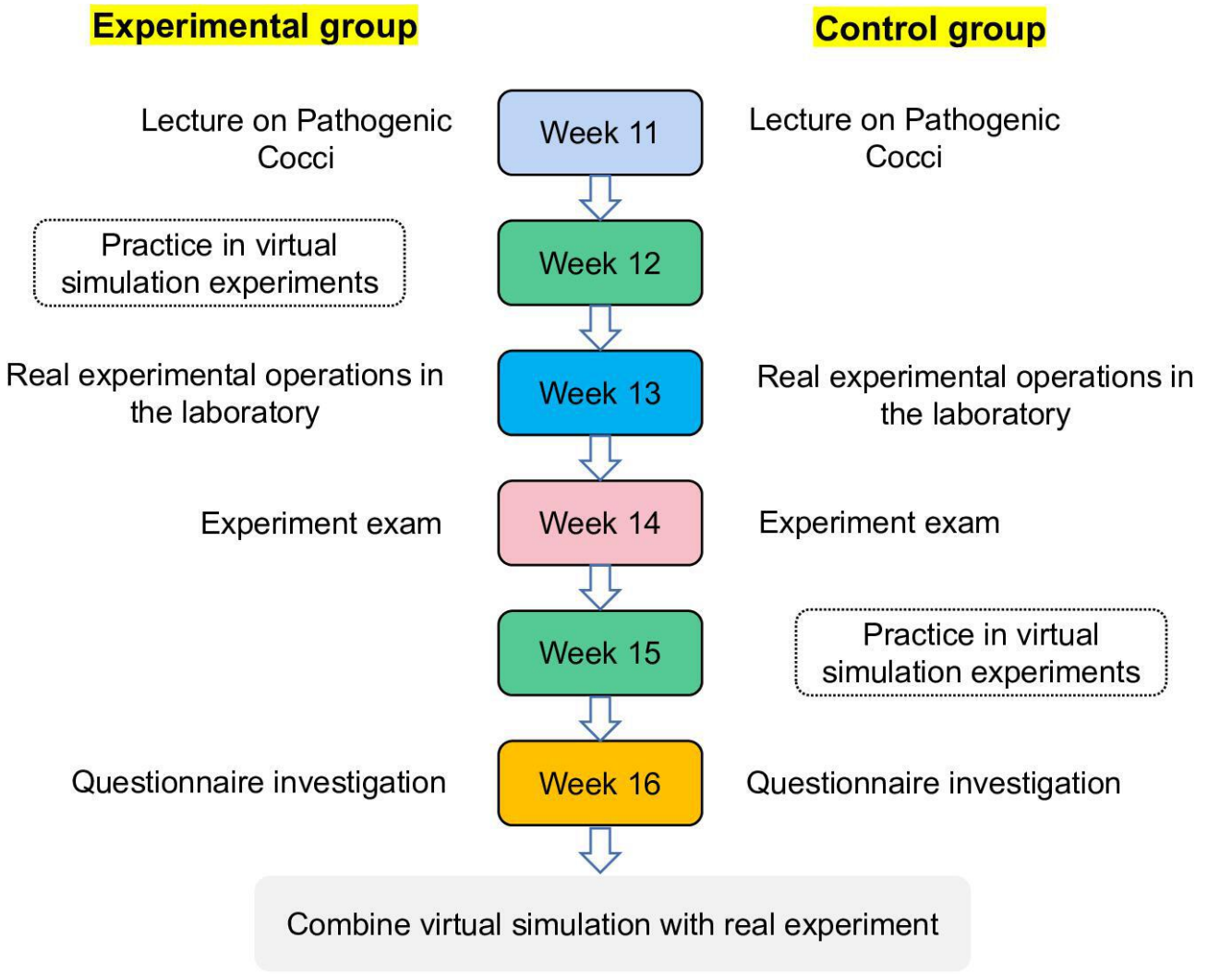
System Features

The design of this virtual simulation platform fully embodies the closed-loop concept of the Kern model: interactive virtual microscopy technology addresses the pain point of limited specimen types in traditional teaching (problem orientation); the real-time operation feedback system corresponds to the error-correction needs in the hierarchical teaching objectives (objective alignment); and the automated assessment module not only enables immediate evaluation of operational accuracy (formative assessment) but also generates personalized learning curves, providing data support for continuous curriculum improvement (summative assessment). Compared with traditional experimental methods, the platform, through intelligent error-correction guidance and a virtual consumables system (optimization of educational strategies), ensures full coverage of operational standards (needs response) while reducing teaching costs (verification of implementation effectiveness), thus completely realizing a closed-loop curriculum development process from problem identification to effect evaluation.

Study Design and Implementation

This study used a randomized controlled trial design with a mixed methods evaluation spanning from September 2023 to January 2024. A total of 1282 medical students from 9 specialties (clinical medicine, traditional Chinese medicine, pediatrics, nursing, medical imaging, preventive medicine, clinical pharmacy, acupuncture, moxibustion, and traditional Chinese pharmacy) participated in the study, with equal, random assignment to experimental (n=653) and control (n=629) groups. The teaching implementation followed a systematic schedule across 6 weeks (detailed in Figure 1), integrating theoretical lectures, virtual simulation exercises, practical laboratory sessions, and comprehensive assessments of both groups. This design enabled direct comparison of learning effectiveness between traditional and virtual simulation–enhanced teaching approaches.

Figure 1. Course schedule.



Teaching Process

The implementation of this teaching system integrated both virtual simulation and traditional laboratory approaches. The teaching process followed a systematic structure comprising preclass preparation, in-class activities, and postclass assessment. All participating students received standardized learning materials and safety protocols, with the experimental group using the virtual simulation platform under instructor guidance while the control group followed traditional laboratory teaching methods. Both groups engaged in comprehensive learning activities designed to achieve the specified learning objectives, with emphasis on laboratory safety awareness and standardized experimental procedures.

Data Collection

A comprehensive evaluation framework was established to assess the effectiveness of the teaching system. The assessment strategy incorporated 3 primary components: performance assessment through standardized practical tests, learning experience evaluation via structured questionnaires, and comparative cost analysis of resource use. Key evaluation indicators encompassed experimental operation proficiency, theoretical knowledge mastery, student satisfaction levels, teaching resource efficiency, and safety protocol compliance. This multidimensional assessment approach enabled thorough evaluation of both learning outcomes and teaching effectiveness.

Statistical Analysis

The collected data underwent rigorous statistical analysis using SPSS (version 25.0; SPSS Inc). Statistical methods included independent 2-tailed *t* tests for comparing group differences in continuous variables and χ^2 tests for analyzing categorical data. All statistical analyses were conducted with a significance level set at $P<.05$. This analytical approach ensured robust evaluation of the teaching system's effectiveness while maintaining scientific rigor in data interpretation.

Ethical Considerations

This study was approved by the ethics committee of Chongqing Medical University (approval number 2023027). The study was conducted in accordance with local legislation and institutional requirements. We obtained informed consent from participants through an online questionnaire system. Participants were provided with a research information statement detailing the study's purpose, participation requirements, potential risks and benefits, data protection measures, and a declaration that participation was voluntary. Participants were required to confirm they had read and understood this information and agreed to participate before being included in the study. No

financial or material compensation was provided to participants. All data were fully anonymized prior to analysis.

Results

Virtual Laboratory System Development

The developed virtual laboratory system (Multimedia Appendix 1) successfully integrated 3 core modules: knowledge review, a learning module, and an assessment module. The experimental procedure flow (Multimedia Appendix 2) was effectively translated into the virtual environment. The 3-module design received high approval from 1214 of 1282 (94.7%) participants across all medical specialties (Multimedia Appendix 3). Figure 2 shows the CONSORT (Consolidated Standards of Reporting Trials) diagram of participation flow.

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) flow chart of participation.

Learning Performance Analysis

Comparative analysis revealed significant differences in test scores between the experimental and control groups (Table 1). The experimental group demonstrated higher performance across multiple medical fields, with particularly notable improvements in clinical medicine, nursing, traditional Chinese pharmacology, and the science of acupuncture and moxibustion (overall $P<.001$).

Table . Experiment test scores of students from different majors.

Subject	Control group score, mean (SD)	Experiment group score, mean (SD)	*P* value
Clinical medicine	68.34 (17.23)	89.88 (13.09)	<.001
Traditional Chinese medicine	60.54 (17.01)	67.97 (13.14)	.02
Pediatrics	76.29 (18.83)	84.84 (15.26)	.003
Nursing	64.25 (12.56)	73.11 (16.82)	<.001
Medical imaging	70.39 (19.39)	79.94 (18.50)	.01
Preventive medicine	68.83 (15.13)	78.56 (17.72)	.01
Clinical pharmacy	69.41 (21.12)	86.53 (16.05)	.003
Science of acupuncture and moxibustion	60.11 (11.32)	71.47 (15.66)	<.001
Traditional Chinese pharmacology	65.33 (11.28)	82.96 (14.16)	<.001
Overall	67.45 (16.81)	80.82 (17.10)	<.001

System Advantages Analysis

Student perception analysis (Table 2) highlighted key advantages of the virtual system: 90.8% of students reported enhanced safety, 79.6% appreciated the unlimited repetition capability, and 74.6% mentioned cost reduction. The standardized experimental conditions (61.6%) and personalized teaching features (50.6%) also received positive recognition.

Table . Distribution of responses among medical students (n=1282) on key advantages of virtual simulation experiments.

Option	Responses, n (%)
Enhanced safety	1164 (90.8)
Standardized experimental conditions	790 (61.6)
Lower costs	957 (74.6)
Unlimited repetition	1020 (79.6)
Simulating abnormal situations	701 (54.7)
Personalized teaching	649 (50.6)
Intelligent assessment system	607 (47.3)

Interface Design and Implementation

The interface design evaluation (Multimedia Appendix 4) showed that 658 of 1282 (51.3%) students found it well designed and clear. Regarding difficulty levels (Multimedia Appendix 5), 1168 (91.1%) students found the experiment appropriately challenging, and 1156 (90.2%) students reported that the first-person perspective enhanced their learning experience (Table 3), with the 3-module design receiving high approval (Multimedia Appendix 3).

Table . First-person perspective: impact on biosafety level 2 lab experience and understanding of microbial techniques among medical students (n=1282).

Medical specialty	Significantly enhanced, n (%)	Little impact, n (%)	Uncertain, n (%)
Total	1156 (90.2)	62 (4.8)	64 (5.0)
Clinical medicine	340 (87.9)	21 (5.4)	26 (6.7)
Traditional Chinese medicine	89 (88.1)	8 (7.9)	4 (4.0)
Pediatrics	136 (90.1)	7 (4.6)	8 (5.3)
Nursing	243 (93.1)	10 (3.8)	8 (3.1)
Medical imaging	98 (89.9)	5 (4.6)	6 (5.5)
Preventive medicine	70 (93.3)	3 (4.0)	2 (2.7)
Clinical pharmacy	47 (88.7)	2 (3.8)	4 (7.5)
Science of acupuncture and moxibustion	92 (92.9)	2 (2.0)	5 (5.1)
Traditional Chinese pharmacology	41 (89.1)	4 (8.7)	1 (2.2)

Learning Outcomes and Satisfaction

The results showed that 1268 of 1282 (98.9%) students reported improved comprehension (Table 4). Student acceptance data (Table 5) indicated that 1145 (89.3%) were willing to adopt virtual simulation as a supplementary learning tool. Overall satisfaction (Table 6) reached 97.2%, with 39.6% being “very satisfied” and 57.6% “generally satisfied.”

Table . Enhancement of knowledge and understanding through virtual simulation experiments among medical students (n=1282).

Medical specialty	Significantly deepened understanding, n (%)	Somewhat helpful, n (%)	Little to no help, n (%)
Total	383 (29.9)	885 (69.0)	14 (1.1)
Clinical medicine	109 (28.2)	269 (69.5)	9 (2.3)
Traditional Chinese medicine	26 (25.7)	75 (74.3)	0 (0.0)
Pediatrics	47 (31.1)	103 (68.2)	1 (0.7)
Nursing	78 (29.9)	183 (70.1)	0 (0.0)
Medical imaging	32 (29.4)	77 (70.6)	0 (0.0)
Preventive medicine	21 (28.0)	54 (72.0)	0 (0.0)
Clinical pharmacy	21 (39.6)	30 (56.6)	2 (3.8)
Science of acupuncture and moxibustion	37 (37.4)	60 (60.6)	2 (2.0)
Traditional Chinese pharmacology	12 (26.1)	34 (73.9)	0 (0.0)

Table . Survey on willingness to use virtual simulation experiments as a supplement to pathogenic biology experiments among medical students (n=1282).

Medical specialty	Willing, n (%)	Unwilling, n (%)	Indifferent, n (%)
Total	1145 (89.3)	73 (5.7)	64 (5.0)
Clinical medicine	341 (88.1)	26 (6.7)	20 (5.2)
Traditional Chinese medicine	87 (86.1)	10 (9.9)	4 (4.0)
Pediatrics	138 (91.4)	4 (2.6)	9 (6.0)
Nursing	235 (90.0)	17 (6.5)	9 (3.5)
Medical imaging	97 (89.0)	6 (5.5)	6 (5.5)
Preventive medicine	70 (93.3)	1 (1.3)	4 (5.3)
Clinical pharmacy	49 (92.5)	3 (5.7)	1 (1.9)
Science of acupuncture and moxibustion	87 (87.9)	5 (5.1)	7 (7.1)
Traditional Chinese pharmacology	41 (89.1)	1 (2.2)	4 (8.7)

Table . Overall satisfaction assessment of the virtual simulation experiment among medical students (n=1282).

Medical specialty	Highly satisfied, n (%)	Generally satisfied, n (%)	Dissatisfied, n (%)
Total	508 (39.6)	738 (57.6)	36 (2.8)
Clinic medicine	164 (42.4)	209 (54.0)	14 (3.6)
Traditional Chinese medicine	29 (28.7)	70 (69.3)	2 (2.0)
Pediatrics	62 (41.1)	84 (55.6)	5 (3.3)
Nursing	94 (36.0)	161 (61.7)	6 (2.3)
Medical imaging	41 (37.6)	66 (60.6)	2 (1.8)
Preventive medicine	31 (41.3)	43 (57.3)	1 (1.3)
Clinical pharmacy	23 (43.4)	27 (50.9)	3 (5.7)
Science of acupuncture and moxibustion	43 (43.4)	53 (53.5)	3 (3.0)
Traditional Chinese pharmacology	21 (45.7)	25 (54.3)	0 (0.0)

Discussion

Principal Findings

Virtual simulation technology offers innovative and effective solutions for teaching related to laboratory experiments [20-23]. Our program was developed to address a critical training gap identified in institutional needs assessments: over 80% of students sought enhanced training in pathogenic bacteria, yet 90% of faculty opposed live cocci experiments due to biosafety concerns.

The 3-module design (the knowledge review, learning module, and assessment module) successfully reconciled this “must-learn vs cannot-do” paradox, achieving 94.7% student approval. This structure aligns with cognitive learning theory [20] while specifically addressing the staged requirements of microbiology experiments [21]. Quantitative assessments confirmed significant improvements across specialties (P<.05), including a 98.9% enhancement in theoretical understanding and 90.8% increase in safety awareness—effectively overcoming key limitations of traditional teaching through (1) standardized experimental procedures, (2) 60% cost reduction over 5 years, and (3) integrated BSL-2 safety training.

These findings are corroborated by prior research [11,24-26], particularly regarding virtual laboratories’ advantages in safety and accessibility. The program’s cost-effectiveness and scalability (serving large student populations with easy updates) will allow the further optimization of teaching models [7-9] and expansion of remote education possibilities.

Nevertheless, real experiments retain irreplaceable advantages in cultivating practical operational skills and sensory experiences [27]. Feedback from 1.1% of students indicating limited assistance from virtual experiments highlights the persistent limitations in substituting for hands-on operations in real environments [4-6]. Consequently, virtual simulation experiments and real experiments each possess distinct advantages [13-15,28]. The optimal teaching strategy should therefore involve an organic integration of both approaches [16-18,29].

Based on tracking student feedback, we identified the main issues with the virtual experiment system as lag, lack of speed adjustment options, poor learning module integration, incomplete results analysis, and inadequate error simulation compared to real experiments. To address these problems, we developed a series of optimization strategies. First, server performance should be enhanced and the program’s code should be optimized to resolve lag issues. Second, the introduction of an intelligent time management system would allow students to adjust the experiment’s progress. In terms of instructional design, the learning modules could be improved by increasing interactive guidance and feedback. Simultaneously, incorporating data visualization tools would strengthen results analysis, and refining the simulation fidelity would allow it to better reflect real experimental conditions. Additionally, we designed a multidimensional evaluation system to continuously collect and analyze student feedback for iterative system optimization. Lastly, it is important to strengthen teacher training, which is crucial for guiding student learning through virtual experiments. Through comprehensive implementation of these strategies [30], we expect to significantly enhance the technical implementation, instructional design, and learning experience of virtual experiments, thereby substantially improving their educational effectiveness.

To ensure analytical rigor, this study used a complete case analysis method, including only data from students who completed both tests and questionnaires. While this approach enhanced data quality and analytical consistency, it also reduced the effective sample size, potentially limiting the generalizability of the results. Future research might consider advanced techniques such as multiple imputation to handle missing data and extending the scope of the study to different educational levels to enhance its representativeness. Subsequent studies should adopt a mixed methods approach, integrating quantitative analysis with qualitative research, and use a longitudinal design to assess the long-term educational effectiveness of virtual experiments. These research directions would deepen understanding of virtual experiments’ educational effectiveness, promote teaching innovation, and provide crucial guidance for future education practices.

Conclusion

The virtual simulation experiment system described here serves as an effective complement to real experiments, offering significant advantages in terms of safety, repeatability, and standardization. Our study demonstrates its effectiveness through improved student performance and a high satisfaction rate (97.2%). The 3-module design effectively enhanced students' understanding (98.9% reporting improvement) and engagement with the experimental procedures. While virtual experiments show distinct advantages in visualization and allow repeated practice, they work best when integrated with traditional hands-on experiments to provide a comprehensive learning experience. This combination leverages the strengths of both approaches—the safety and flexibility of virtual simulation with the irreplaceable tactile experience of real laboratory practice. Future development should focus on technical optimization and enhanced integration with traditional teaching methods to further improve educational outcomes.

Acknowledgments

This work was supported by the Chongqing Education Teaching Reform Research Projects (213120); Education and Teaching Research Projects of Chongqing Medical University (JY200318 and JY20240101); and the Program for Youth Innovation in Future Medicine, Chongqing Medical University (W0021). The authors thank all the students who participated in the study.

Authors' Contributions

LX contributed to investigation, study design, funding acquisition, and writing (original draft). XD contributed to study design, conceptualization, and data curation. TC contributed to study design and data curation. NL contributed to methodology and software. JL and YG contributed to study design and writing (original draft). ZT contributed to study design, resources, and validation. YW and YN contributed to writing (original draft and editing) and software. Y Hosseini contributed to writing (original draft). Y He contributed to conceptualization, data curation, funding acquisition, supervision, and writing (original draft).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Ideas for identifying pathogenic cocci.
[PNG File, 126 KB - mededu_v11i1e72495_app1.png]

Multimedia Appendix 2
Virtual experiment process.
[PNG File, 147 KB - mededu_v11i1e72495_app2.png]

Multimedia Appendix 3
Evaluation of virtual experiment's 3-module design: review, learning, and assessment.
[XLSX File, 9 KB - mededu_v11i1e72495_app3.xlsx]

Multimedia Appendix 4
Evaluation of the virtual simulation experiment interface design and usability.
[XLSX File, 9 KB - mededu_v11i1e72495_app4.xlsx]

Multimedia Appendix 5
Perceived difficulty levels of the virtual simulation experiments.
[XLSX File, 9 KB - mededu_v11i1e72495_app5.xlsx]

Checklist 1
CONSORT (Consolidated Standards of Reporting Trials) checklist.
[PDF File, 2067 KB - mededu_v11i1e72495_app6.pdf]

References

1. Gajdács M, Ábrók M, Lázár A, Burián K. Increasing relevance of gram-positive cocci in urinary tract infections: a 10-year analysis of their prevalence and resistance trends. Sci Rep 2020 Oct 19;10(1):17658 [FREE Full text] [doi: 10.1038/s41598-020-74834-y] [Medline: 33077890]

2. Zhu H, Xu J, Wang P, et al. The status of virtual simulation experiments in medical education in China: based on the National Virtual Simulation Experiment Teaching Center (iLAB-X). Med Educ Online 2023 Dec;28(1):2272387 [FREE Full text] [doi: 10.1080/10872981.2023.2272387] [Medline: 37883485]
3. Gebremariam NM, Bitew A, Tsige E, Woldesenbet D, Tola MA. A high level of antimicrobial resistance in gram-positive cocci isolates from different clinical samples among patients referred to Arsho Advanced Medical Laboratory, Addis Ababa, Ethiopia. Infect Drug Resist 2022;15:4203-4212 [FREE Full text] [doi: 10.2147/IDR.S372930] [Medline: 35946034]
4. Miller WR, Arias CA. ESKAPE pathogens: antimicrobial resistance, epidemiology, clinical impact and therapeutics. Nat Rev Microbiol 2024 Oct;22(10):598-616 [FREE Full text] [doi: 10.1038/s41579-024-01054-w] [Medline: 38831030]
5. GBD 2019 Antimicrobial Resistance Collaborators. Global mortality associated with 33 bacterial pathogens in 2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2022 Dec 17;400(10369):2221-2248 [FREE Full text] [doi: 10.1016/S0140-6736(22)02185-7] [Medline: 36423648]
6. Zhang G, Liu J, He Y, et al. Modifying Escherichia coli to mimic Shigella for medical microbiology laboratory teaching: a new strategy to improve biosafety in class. Front Cell Infect Microbiol 2023;13:1257361 [FREE Full text] [doi: 10.3389/fcimb.2023.1257361] [Medline: 37780843]
7. Guerrini F, Bertolino L, Safa A, et al. The use of technology-based simulation among medical students as a global innovative solution for training. Brain Sci 2024 Jun 23;14(7):627 [FREE Full text] [doi: 10.3390/brainsci14070627] [Medline: 39061368]
8. Spencer D, McKeown C, Tredwell D, et al. Student experiences with a molecular biotechnology course containing an interactive 3D immersive simulation and its impact on motivational beliefs. PLOS ONE 2024;19(7):e0306224 [FREE Full text] [doi: 10.1371/journal.pone.0306224] [Medline: 39052561]
9. Li WY, Li HZ, Zhan SY, Wang SF. Application of virtual simulation technology in epidemiology education: a systematic review. Zhonghua Liu Xing Bing Xue Za Zhi 2024 Jul 10;45(7):1014-1023 [FREE Full text] [doi: 10.3760/cma.j.cn112338-20240210-00072] [Medline: 39004975]
10. Yeung AWK, Parvanov ED, Hribersek M, et al. Digital teaching in medical education: scientific literature landscape review. JMIR Med Educ 2022 Feb 9;8(1):e32747 [FREE Full text] [doi: 10.2196/32747] [Medline: 35138260]
11. de Vries LE, May M. Virtual laboratory simulation in the education of laboratory technicians-motivation and study intensity. Biochem Mol Biol Educ 2019 May;47(3):257-262 [FREE Full text] [doi: 10.1002/bmb.21221] [Medline: 30748084]
12. Tsirulnikov D, Suart C, Abdullah R, Vulcu F, Mullarkey CE. Game on: immersive virtual laboratory simulation improves student learning outcomes & motivation. FEBS Open Bio 2023 Mar;13(3):396-407 [FREE Full text] [doi: 10.1002/2211-5463.13567] [Medline: 36723273]
13. Lu J, Yang X, Zhao W, Lin J. Effect analysis of a virtual simulation experimental platform in teaching pulpotomy. BMC Med Educ 2022 Nov 7;22(1):760 [FREE Full text] [doi: 10.1186/s12909-022-03836-3] [Medline: 36345029]
14. Gao F, Qiu J, Chen L, Li L, Ji M, Zhang R. Effects of virtual reality simulation on medical students' learning and motivation in human parasitology instruction: a quasi-experimental study. BMC Med Educ 2023 Sep 3;23(1):630 [FREE Full text] [doi: 10.1186/s12909-023-04589-3] [Medline: 37661271]
15. Padilha JM, Machado PP, Ribeiro A, Ramos J, Costa P. Clinical virtual simulation in nursing education: randomized controlled trial. J Med Internet Res 2019 Mar 18;21(3):e11529 [FREE Full text] [doi: 10.2196/11529] [Medline: 30882355]
16. Meng L, Liu X, Ni J, Shen P, Jiao F. An investigation for the efficacy of teaching model of combining virtual simulation and real experiment for clinical microbiology examination. Front Med (Lausanne) 2024;11:1255088 [FREE Full text] [doi: 10.3389/fmed.2024.1255088] [Medline: 38449889]
17. Trevethan-Cravioto S, Sierra-Fernández C, López-Meneses M, Azar-Manzur F, Jiménez-Garcés V, Gaspar-Hernández J. The rescue of medical education crippled by the COVID-19 pandemic. Arch Cardiol Mex 2023;93(Supl 6):22-27 [FREE Full text] [doi: 10.24875/ACM.22000225] [Medline: 38537221]
18. Sharma D, Bhaskar S. Addressing the Covid-19 burden on medical education and training: the role of telemedicine and tele-education during and beyond the pandemic. Front Public Health 2020;8:589669 [FREE Full text] [doi: 10.3389/fpubh.2020.589669] [Medline: 33330333]
19. Young C, Daly K, Hurtado A, et al. Applying Kern's model to the development and evaluation of medical student well-being programs. J Gen Intern Med 2023 Oct;38(13):3047-3050 [FREE Full text] [doi: 10.1007/s11606-023-08265-6] [Medline: 37340253]
20. Cook DA. Learning and cognitive styles in web-based learning: theory, evidence, and application. Acad Med 2005 Mar;80(3):266-278 [FREE Full text] [doi: 10.1097/00001888-200503000-00012] [Medline: 15734809]
21. Ibrahim W, Ibrahim W, Zoubeidi T, Marzouk S, Sweedan A, Amer H. An online management system for streamlining and enhancing the quality of learning outcomes assessment. Educ Inf Technol (Dordr) 2022;27(8):11325-11353 [FREE Full text] [doi: 10.1007/s10639-022-10918-8] [Medline: 35542311]
22. Plch L. Perception of technology-enhanced learning by medical students: an integrative review. Med Sci Educ 2020 Dec;30(4):1707-1720 [FREE Full text] [doi: 10.1007/s40670-020-01040-w] [Medline: 34457833]
23. Dustman WA, King-Keller S, Marquez RJ. Development of gamified, interactive, low-cost, flexible virtual microbiology labs that promote higher-order thinking during pandemic instruction. J Microbiol Biol Educ 2021;22(1):22.1.53 [FREE Full text] [doi: 10.1128/jmbe.v22i1.2439] [Medline: 33884094]

24. Baumann-Birkbeck L, Anoopkumar-Dukie S, Khan SA, Cheesman MJ, O'Donoghue M, Grant GD. Can a virtual microbiology simulation be as effective as the traditional wetlab for pharmacy student education? BMC Med Educ 2021 Nov 17;21(1):583 [FREE Full text] [doi: 10.1186/s12909-021-03000-3] [Medline: 34789233]
25. Mistry D, Brock CA, Lindsey T. The present and future of virtual reality in medical education: a narrative review. Cureus 2023 Dec;15(12):e51124 [FREE Full text] [doi: 10.7759/cureus.51124] [Medline: 38274907]
26. Tsai HP, Lin CW, Lin YJ, Yeh CS, Shan YS. Novel software for high-level virological testing: self-designed immersive virtual reality training approach. J Med Internet Res 2023 Jun 21;25:e44538 [FREE Full text] [doi: 10.2196/44538] [Medline: 37342081]
27. Wilcha RJ. Effectiveness of virtual medical teaching during the COVID-19 crisis: systematic review. JMIR Med Educ 2020 Nov 18;6(2):e20963 [FREE Full text] [doi: 10.2196/20963] [Medline: 33106227]
28. Ayer Miller V, Marks T, Thompson DK. Student performance and perceptions in a hybrid laboratory model: an exploratory study of interactive virtual simulations and in-person integration in a foundational microbiology course. J Microbiol Biol Educ 2025 Apr 24;26(1):e0020324 [FREE Full text] [doi: 10.1128/jmbe.00203-24] [Medline: 40047415]
29. Joji RM, Kumar AP, Almarabheh A, et al. Perception of online and face to face microbiology laboratory sessions among medical students and faculty at Arabian Gulf University: a mixed method study. BMC Med Educ 2022 May 30;22(1):411 [FREE Full text] [doi: 10.1186/s12909-022-03346-2] [Medline: 35637505]
30. O'Connor S, Kennedy S, Wang Y, Ali A, Cooke S, Booth RG. Theories informing technology enhanced learning in nursing and midwifery education: a systematic review and typological classification. Nurse Educ Today 2022 Nov;118:105518 [FREE Full text] [doi: 10.1016/j.nedt.2022.105518] [Medline: 36030581]

Abbreviations

BSL: biosafety level
CONSORT: Consolidated Standards of Reporting Trials

Edited by J Gentges; submitted 12.02.25; peer-reviewed by A Girma, J Husseini; revised version received 03.07.25; accepted 03.07.25; published 30.07.25.

Please cite as:
Xu L, Deng X, Chen T, Lu N, Wang Y, Liu J, Guo Y, Tu Z, Nie Y, Hosseini Y, He Y
A Large-Scale Multispecialty Evaluation of Web-Based Simulation in Medical Microbiology Laboratory Education: Randomized Controlled Trial
JMIR Med Educ 2025;11:e72495
URL: https://mededu.jmir.org/2025/1/e72495
doi:10.2196/72495

Open-Access Web-Based Gamification in Pharmacology Education for Medical Students: Quasi-Experimental Study

Lujain Aloum[1], MSc; Halah Ibrahim[1], MD, MEHP; Senthil Kumar Rajasekaran[1], MD, MMHPE; Eman Alefishat[1,2,3], PhD

[1]Department of Medical Sciences, College of Medicine and Health Sciences, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates

[2]Department of Biopharmaceutics and Clinical Pharmacy, Faculty of Pharmacy, University of Jordan, Amman, Jordan

[3]Department of Biomedical and Translational Sciences, Carle Illinois College of Medicine, University of Illinois Urbana-Champaign, 506 South Mathews Avenue, Urbana, IL, United States

Corresponding Author:
Eman Alefishat, PhD
Department of Medical Sciences, College of Medicine and Health Sciences, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates

Abstract

Background: Medical education continues to favor didactic lectures as the predominant method of instruction. However, in recent years, there has been a shift toward active learning methodologies such as gamification.

Objective: This study aimed to describe the implementation of 3 open-access, web-based pharmacology games tailored for medical students: *Cross DRUGs*, *Find the DRUG*, and *DRUGs Escape Room*. The study also evaluated the impact of gamification on knowledge retention, student engagement, and learning experience in pharmacology education.

Methods: We used a quasi-experimental design to examine the effects of gamification on knowledge retention by comparing pretest and posttest scores between the gamer and control groups. Each week, students self-selected into either the gamer group or the control group based on personal preference. All students were provided with online access to the same lecture slides. Students in the control group completed both the pretest and posttest but did not play any of the games. A survey was administered to assess students' perceptions of gamification as a learning tool.

Results: Of the 72 students enrolled in the course, 49 (68%) agreed to participate, with 40 (56%) students completing both the pretest and posttest and being included in our analysis. As participation could vary weekly, an individual student might have appeared in both groups across different weeks, resulting in 59 gamer sessions and 20 control sessions. The mean pretest scores were 6.05 (SD 2.31) for the control group and 6.20 (SD 2.13) for the gamer group. The mean posttest scores were 6.90 (SD 2.02) for the control group and 8.47 (SD 1.30) for the gamer group. The gamer group exhibited significantly improved posttest scores (P=.006), while the control group did not (P=.21). Most respondents (25/30, 83%) found the games enjoyable and agreed that the games effectively helped them understand pharmacological concepts (24/30, 80%). Additionally, 70% (21/30) of students believed they learned better from the gaming format than from didactic lectures. Most favored a blended approach that combines lectures with games or case studies.

Conclusions: Gamification can serve as an effective complementary teaching tool for helping medical students learn pharmacological concepts.

(JMIR Med Educ 2025;11:e73666) doi:10.2196/73666

KEYWORDS

pharmacology; gamification; open access; medical education; medical students

Introduction

Medical education continues to favor didactic lectures as the predominant method of instruction [1,2], even though this traditional approach fosters passive learning, reinforces teacher-centeredness [3], and yields lower rates of knowledge retention compared to other approaches [4]. In recent years, there has been a growing shift toward innovative and active learning methodologies, among which gamification has emerged as a promising strategy [5].

In its simplest terms, gamification involves integrating game elements into nongame contexts [6,7]. Gamification in education is grounded in several learning theories, including humanistic learning and adult learning theories [8,9]. It incorporates features such as goal setting, incremental challenges, immediate feedback, progression systems, and rewards to foster deeper

involvement in learning tasks [10,11]. Gamification has been shown to enhance motivation, improve academic achievement, and foster social interaction, thereby supporting its use as an effective teaching tool [11-13].

For Generation Z learners, considered to be "digital natives," gamification may be the most appropriate pedagogical method as it meets Generation Z's familiarity and immersion in digital platforms [14-17]. Furthermore, their extensive exposure to technology may have resulted in the development of constrained attention spans, a preference for visual and kinesthetic learning activities, and a need for immediate feedback [14]. This trend in learning preferences requires educators to transition from content providers to facilitators of learning who can skillfully leverage technology, including gamification, to boost student engagement and motivation [14,18,19].

Although many studies have demonstrated that gamification enhances the learning experience and increases knowledge retention [20-24], they have been criticized for publication bias and for consistently reporting positive outcomes in health professions education programs [25]. Methodological concerns have also been raised, with most studies lacking proper control groups, raising doubts about the reliability of the evidence supporting the impact of gamification on learning outcomes. A systematic review on gamification underscored the need for more rigorous research designs with defined control groups to accurately assess the benefits of gamification in health professions education [25]. This study aimed to describe the implementation of 3 open-access, web-based pharmacology games for medical students and to explore the impact of these games on knowledge retention and student experience.

Methods

Setting and Study Design

Khalifa University College of Medicine and Health Sciences is a 4-year postgraduate entry medical school in the United Arab Emirates. Many of the students do not follow the traditional premedical route and instead have engineering backgrounds. The pharmacology course is a mandatory 4-week course for all first-year medical students. It typically requires memorization of large amounts of information, which presents an academic challenge for students. Recognizing that a lecture-based delivery of the pharmacology course might not benefit students who may struggle with rote memorization of a long list of facts, we sought innovative teaching methods to enhance the learning experience.

We conducted a quasi-experimental study to examine the effects of gamification on student experience and knowledge retention during the first 3 weeks of a 4-week pharmacology course; the final exam week was excluded. Quasi-experimental designs are commonly used in medical education when randomization is not feasible due to the practical and ethical constraints of real-world educational settings [26]. We used a pretest-posttest, nonrandomized control group design. This approach is consistent with the framework of nonequivalent group designs where participants are not randomly assigned but are compared based on pretest and postintervention outcomes to infer the impact of the intervention [27,28].

Ethical Considerations

The CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist was used to guide our reporting [29]. The study was approved by the Institutional Review Board of Khalifa University of Science and Technology (H20-036). At the start of the course, each student received an email invitation to participate in the study along with an informed consent form, and all participants provided informed consent. The email described the study's purpose and explained that it was anonymous and confidential. Participation was voluntary, and no incentives were offered. Students were informed that they could withdraw from the study at any time without any consequences. Student participation and data deidentification were handled by a study coordinator who was not involved in student teaching or grading.

Participants and Group Allocation

Students were allowed to self-select into either the gamified (gamer) or nongamer (control) group each week based on their personal preference. While this self-selection limited random assignment, it is a pragmatic approach for studying educational interventions that mirror real-world classroom decision-making [30]. Given the weekly self-selection process, a single student could participate in both types of sessions across different weeks, resulting in 59 gamer sessions and 20 control sessions, reflecting learning sessions rather than unique participants.

Game Design

We chose 3 open-access, web-based game formats—crossword puzzle, word search, and escape room—because they align with adult learning theories by incorporating clear goals, incremental challenges, and immediate feedback [7]; require minimal technological infrastructure; and support asynchronous, self-directed learning suited to Generation Z's digital preferences. *Cross DRUGs* (Multimedia Appendix 1), based on a crossword game, was generated via an online crossword puzzle generator available on the Education website [31]. *Find the DRUG* (Multimedia Appendix 1) was based on the Hunting Words game and was produced using an online word search puzzle generator available on the Educolorir website [32]. *DRUGs Escape room* (Multimedia Appendix 1), created via Google Forms using the EduGame template, required students to solve a series of sequential, drug-related clues to progress through the game.

Each game tested students on the drugs covered during the preceding week of the pharmacology course, including the mechanisms of action, primary indications, and clinically relevant side effects. Two authors, who were faculty members in the pharmacology course, independently drafted each game's questions based on weekly learning outcomes. A third author, also a course instructor, reviewed the games for accuracy and curricular alignment. Games were single-player and time-limited, but students were allowed unlimited attempts. Students were given 48 hours to participate in the game, and only 1 attempt was required for inclusion. Upon completion, correct answers and brief explanations were provided via the learning management system.

Assessment

All students received the same curriculum. At the end of each week and before game access, all students completed a 9-item multiple-choice pretest. The pretest was conducted to assess the pharmacological knowledge of students regarding the content covered that week and to ensure the homogeneity of the knowledge acquired by both the control and gamer groups. Instructions on how to access the games were provided via email, along with a reminder encouraging students to play the game. At the end of the 48-hour period for game completion, a matched 9-item posttest was released to both groups. Students in both the gamer and control groups accessed the posttest at the same time. The pretest and posttest contained identical items to leverage spaced repetition principles and measure knowledge gains attributable to gameplay. Spaced repetition involves revisiting material at intervals to reinforce understanding over time [33]. The tests were formative and did not influence course grades. Students in the control group had online access to the lecture slides that were also available to the gamer group, while the latter engaged in gameplay.

Survey Instrument

At the end of the course, all students who participated as gamers were asked to complete an online questionnaire assessing enjoyment, perceived learning benefits, and preferences for gamification versus lectures using a 5-point Likert scale (1=strongly agree to 5=strongly disagree). The survey also evaluated the importance of incorporating a reward system into gamification and compared gamification to other teaching methodologies. The postgame survey was administered only to students in the gamer group, as it focused on perceptions of the gamified activities. Because the control group did not receive any additional instructional method beyond lecture slides, the questionnaire was not applicable. We calculated Cronbach α for internal consistency (overall α=0.88; reward-related items α=0.65) [34]. Deleting any question caused the α coefficient to decrease, suggesting that each question contributed to the overall reliability, as shown in Multimedia Appendix 2.

Data Analysis

We performed analyses in R (version 4.2.3; R Foundation for Statistical Computing). In line with prior studies on the impact of games on learning outcomes, score differences (posttest minus pretest) between control and gamers were used to assess the effectiveness of the games [35]. The Shapiro-Wilk test and Mann-Whitney test were performed using the built-in stats package, and Cronbach α was calculated using the psych package. To assess the normality of the pretest and posttest scores, the Shapiro-Wilk test and histogram inspection were conducted. Because the scores were not normally distributed, we compared groups with the Mann-Whitney test (Wilcoxon rank-sum test with continuity correction) and set statistical significance at $P<.05$.

Results

Of the 72 students enrolled in the course, 49 (68%) agreed to participate, with 40 (56%) students completing both the pretest and posttest and being included in our analysis (Figure 1). As students self-selected into the gamer or control group each week, there were 59 gamer sessions and 20 control sessions. The average number of attempts per gamer was 3.4 (SD 0.85). Participant demographics are presented in Multimedia Appendix 3. The self-identified mean age was 23.9 (SD 2.69) years for the control group and 24.0 (SD 1.79) years for the gamer group. Students had educational backgrounds in engineering or scientific disciplines (*others*), typically holding bachelor's degrees in chemistry, biology, biochemistry, or psychology. None of the students reported prior educational experience in pharmacy or pharmacology.

Figure 1 displays the average pretest and posttest scores for both groups. The mean pretest scores were 6.05 (SD 2.31) for the control group and 6.20 (SD 2.13) for the gamer group. The mean posttest scores were 6.90 (SD 2.02) for the control group and 8.47 (SD 1.30) for the gamer group. The gamer group demonstrated a significant improvement in posttest scores ($P=.006$), whereas the control group did not show a statistically significant change ($P=.21$).

Survey results demonstrated positive perceptions of gamification as a learning tool. Of the 30 students surveyed, 24 (80%) agreed that the games were an efficient way to understand pharmacological concepts, and 25 (83%) reported that the games were enjoyable. Most students (21/30, 70%) believed that they learned better from the gaming format than from didactic lectures, and 67% (20/30) agreed that awarding points, such as extra credits, for game-related activities would be beneficial. Less than half of the respondents (12/30, 40%) agreed with allocating prizes to game winners, while the remaining students were evenly divided between neutral and disagreement (Table 1). Surveyed students favored a teaching methodology that combines lectures with games or case studies, with the highest percentage of students (9/30, 30%) preferring lectures as their third preference (Figure 2).

Figure 1. Mean (SD) number of correct responses on the pretest and posttest for students who did not play the game (control group; n=20 sessions) and those who played the game (the gamer group; n=59 sessions). Each data point represents a learning session rather than a unique participant. The highest attainable score was 9. *P=.006.

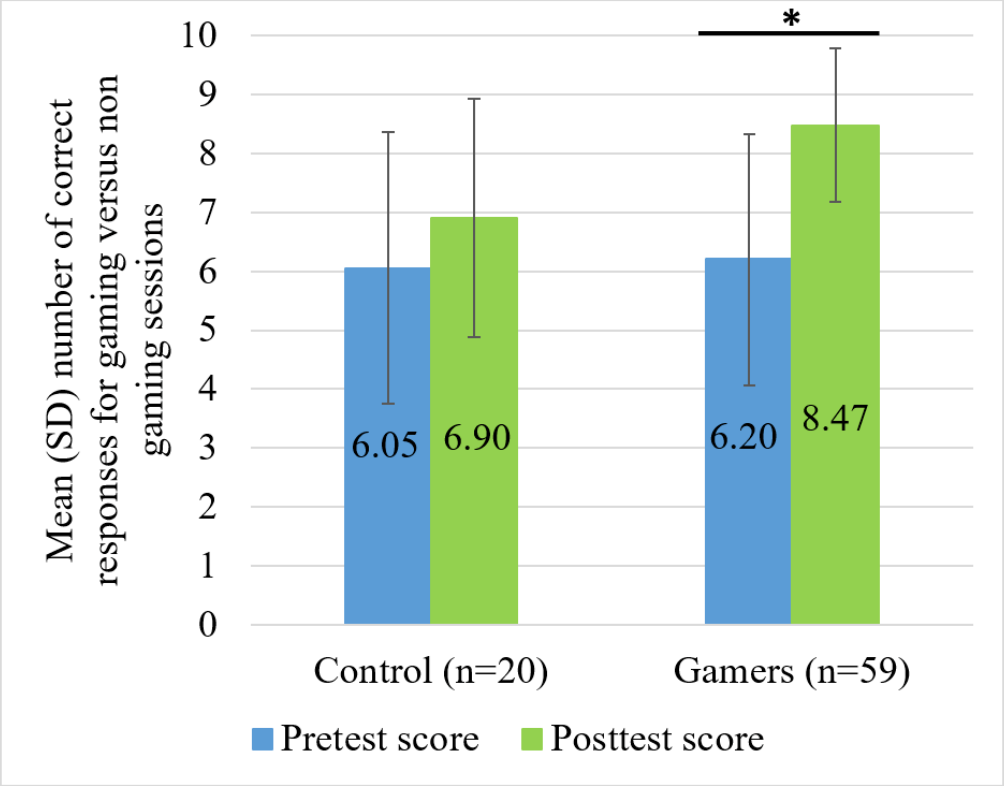
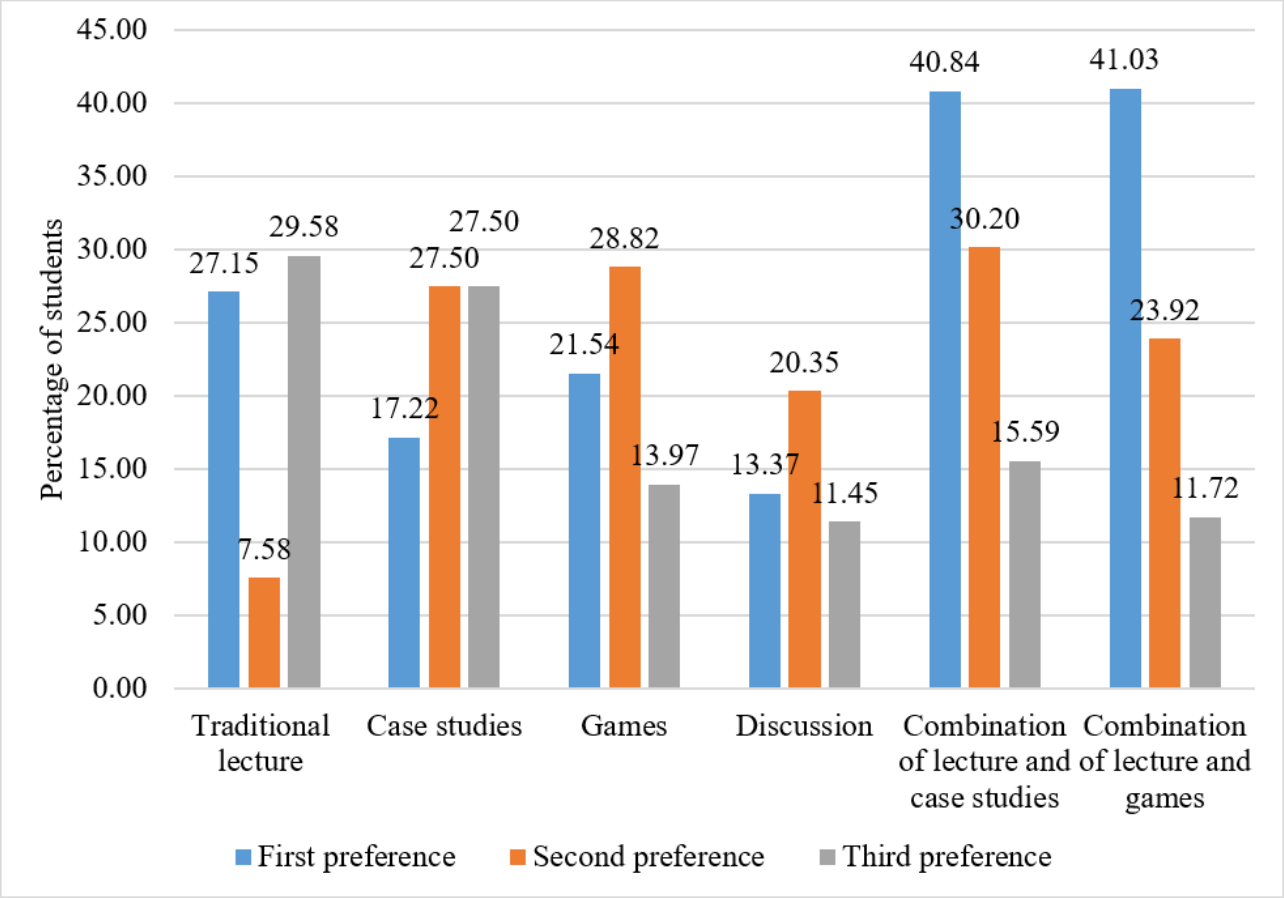


Table . Students' responses to 5 evaluation statements on gamified pharmacology learning activities (n=30). The statements assessed perceived effectiveness, enjoyment, learning preferences, and attitudes toward rewards.

Statements	Strongly agree or agree, n (%)	Neutral, n (%)	Strongly disagree or disagree, n (%)
The game was an effective way to learn pharmacological concepts	24 (80)	3 (10)	3 (10)
I enjoyed the game	25 (83)	3 (10)	2 (7)
I learn better in a game format than in a didactic lecture	21 (70)	3 (10)	6 (20)
I feel that prizes should be awarded to the winners of the games	12 (40)	9 (30)	9 (30)
Points (either as extra credit or incorporated into the overall grading scheme) should be associated with game activities	20 (67)	4 (13)	6 (20)

Figure 2. Students' preferences for different instructional formats in pharmacology education, presented as percentages for first, second, and third choices (n=30).



Discussion

We integrated 3 low-cost, easily implemented, open-access pharmacology games into a pharmacology course for medical students. Students who engaged in the game-based learning approach demonstrated better knowledge retention with a significant improvement in test scores compared to the control group. Most students also perceived the games as effective, enjoyable, and preferable compared to traditional lectures. Many students suggested that a blended instructional model combining didactic sessions with interactive games would optimize learning by balancing content delivery with engaging reinforcement.

Our decision to implement gamification in pharmacology education was partly driven by the academic profile of our students, many of whom came from engineering backgrounds. These students often prefer spontaneous, pragmatic, and concrete learning styles with hands-on, goal-oriented, and sequential tasks. Research has shown that engineering students tend to favor trial-and-error learning; practical application; and structured, step-by-step problem-solving approaches [36]. These characteristics made them well suited for gamified learning environments that emphasize active participation and immediate feedback. In line with this, we selected crossword puzzles, word search activities, and web-based escape rooms, as they require minimal setup and support self-directed, asynchronous learning [37,38]. By breaking up complex pharmacology concepts into brief game sessions with immediate feedback, students experienced incremental successes that built self-efficacy and intrinsic motivation [25,39]. Furthermore, the variety of game formats helped sustain attention and catered to diverse learning preferences within the student cohort.

Two recent systematic reviews in pharmacy education and higher education reported that the most common type of research methodology was pretest and posttest evaluation [35,40]. Among the 3 games we implemented, the crossword and word search games lack supporting evidence in the literature that uses pretest and posttest scores [24,38,41]. In contrast, escape room–based strategies have been more rigorously evaluated, frequently using pretest and posttest to quantify knowledge gains. Consistent with our findings, several studies reported substantial improvements in posttest scores among pharmacy and medical students following escape room activities, although many lacked control groups [20,42-44]. While different in design, other gamified learning tools have also demonstrated improved knowledge retention. For instance, second-year medical students who played "Who Wants to be a Physician"—a game inspired by the TV show "Who Wants to be a Millionaire"—achieved higher posttest scores compared to peers who attended traditional tutorial sessions [45]. Similarly, both preclinical and final-year medical students and residents who engaged in board or card games showed posttest score improvements [46-48]. Additionally, the "Pharmacotrophy" tournament, which incorporated Kahoot-based quizzes and in-person matches, significantly enhanced PharmD students' knowledge acquisition, likely due to its incorporation of fun elements and a relaxed, competitive environment [49]. These findings suggest that

gamified approaches can be effectively integrated across various stages of medical education, from preclinical training to residency.

A common methodological limitation across many gamification studies is the lack of a control group, which limits the ability to draw causal inferences [20,24,38,41-44,46-48,50]. Notably, a systematic review of pharmacy education found that just 2% of studies included a control arm [40]. Our study addressed this gap by incorporating a clearly defined control group each week, enabling us to objectively demonstrate that game-based learners achieved significantly greater knowledge gains than their nongamer peers.

Consistent with previous research, most students perceived the games as efficient, enjoyable, and of more educational value than conventional didactic lectures. Students in previous studies found gamified learning methods, such as the diabetes board game, virtual escape room, crossword puzzles, and "Who Wants to be a Physician," to be enjoyable ways to learn pharmacology, often citing increased engagement and confidence in the subject matter [20,24,38,41,45,46]. Similarly, use of the Kahoot platform increased medical student motivation and participation [51]. In our study, more than 80% of gamers agreed that the puzzles and escape room enhanced their understanding of pharmacological concepts and enjoyment of the subject. These findings underscore that well-designed gamification can meaningfully enrich medical education by providing opportunities for interactive learning.

Although students enjoyed the games, many favored a blended approach that combines interactive activities with conventional lectures. This preference suggests that while gamification can foster engagement, learners may not yet be ready to replace lectures entirely, especially within a traditionally structured curriculum. Our voluntary, supplemental design likely contributed to the positive findings by allowing motivated students to opt in without penalizing others. One study found that the impact of gamification depended on participant personality traits, suggesting that it was not beneficial for all learners [52], with some students considering games inefficient or tedious [53]. The moderate-enjoyment hypothesis further cautions that excessive game elements can overload cognitive resources, reducing learning gains beyond an optimal point [54]. Accordingly, we aimed to balance play and learning—using play to complement and reinforce complex information to increase learning efficiency, rather than as a stand-alone replacement for didactic teaching. This strategy aligns with recent recommendations to incorporate game-based learning as an optional strategy and only for tedious or difficult concepts [53]. A recent review supported the use of game-based learning as complementary tools, as it noted that the long-term applicability of these methods requires further exploration [5].

It is notable that the 3 games used in this study deliberately centered on a single game element—"rules and goals"—to provide clear objectives for each activity without introducing confounding game elements such as competition, narratives, or collaboration [55]. By focusing solely on structured challenges with immediate feedback, we could more confidently attribute the observed gains in retention and engagement to goal-oriented game elements [25]. We intentionally omitted competition, a common gamification feature [13,56-58], because it shifts the focus from learning to winning and has been associated with increased anxiety [58,59]. A recent randomized crossover study warned against using competitive gamification in medical education, finding no evidence of benefits on students' competence or internal motivation [60].

Our use of free, open-access platforms addressed barriers in resource-limited settings, where 50% of gamified training platforms for preclinical medical education required paid access [61]. This aligns with the United Nations Educational, Scientific and Cultural Organization "Education for All" initiative by promoting equitable educational resources worldwide [62]. Moreover, while most research on gamification as an educational tool in health professions has been conducted in the United States or Canada [25], by conducting this study in the Middle East—an underrepresented region in gamification research—we contributed valuable data to global medical education literature and demonstrated the approach's feasibility beyond North American contexts.

Further research is essential to build on our findings and to fully understand the potential of gamification in pharmacology education. Investigating the broader impact of open-access, web-based gamification will help validate and generalize these results, paving the way for more engaging and effective learning experiences for medical students. In future research, we would like to explore additional parameters such as academic achievement, long-term memory retention, and practical applicability. As a next step, we also propose developing a free mobile app. This tool has significant potential for learning due to its easy accessibility and affordability, especially because mobile apps are rarely used in pharmacology education [63].

Our study has several limitations. As a single-site, quasi-experimental study with a modest sample size, its generalizability may be limited. As students self-selected as gamers, it is possible that those who were more academically motivated chose the games as an additional study tool. There were also substantially more gamer instances than controls. To mitigate potential bias from unequal group sizes, we analyzed learning sessions rather than unique students, treating each week's participation as an independent observation reflecting exposure to the intervention. We used nonparametric statistics (Mann-Whitney U test), which do not assume equal variances or normally distributed data, thereby making our comparisons robust to group-size differences. Future studies should use randomized allocation or stratified sampling to ensure balanced groups.

In conclusion, integrating simple, low-cost web-based games into a pharmacology curriculum can enhance knowledge retention and learner engagement. As a complementary, optional strategy, gamification offers a feasible strategy to enrich traditional didactics and meets the needs of diverse medical student populations.

Acknowledgments

This work was initially conceived as part of a project submitted to the Leading Innovations in Health Care and Education course at the Harvard Macy Institute. The authors gratefully acknowledge the course directors and faculty for their guidance, support, and inspiration, which were instrumental in shaping the development of this work.

Data Availability

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

Conflicts of Interest

None declared.

Multimedia Appendix 1
Screenshots of the open-access digital pharmacology games developed for the study: *Cross DRUGs* (crossword puzzle), *Find the DRUG* (word search), and *DRUGs Escape Room* (interactive Google Form–based challenge).
[PDF File, 342 KB - mededu_v11i1e73666_app1.pdf]

Multimedia Appendix 2
Cronbach α tool for the level of satisfaction with gamification as a learning tool.
[PDF File, 297 KB - mededu_v11i1e73666_app2.pdf]

Multimedia Appendix 3
Demographics of the students who participated (gamers) and did not participate (control) in the games.
[PDF File, 418 KB - mededu_v11i1e73666_app3.pdf]

References

1. Alaagib NA, Musa OA, Saeed AM. Comparison of the effectiveness of lectures based on problems and traditional lectures in physiology teaching in Sudan. BMC Med Educ 2019 Sep 23;19(1):365. [doi: 10.1186/s12909-019-1799-0] [Medline: 31547817]
2. Wolff M, Wagner MJ, Poznanski S, Schiller J, Santen S. Not another boring lecture: engaging learners with active learning techniques. J Emerg Med 2015 Jan;48(1):85-93. [doi: 10.1016/j.jemermed.2014.09.010] [Medline: 25440868]
3. Luscombe C, Montgomery J. Exploring medical student learning in the large group teaching environment: examining current practice to inform curricular development. BMC Med Educ 2016 Jul 19;16(1):184. [doi: 10.1186/s12909-016-0698-x] [Medline: 27435852]
4. Subramanian A, Timberlake M, Mittakanti H, Lara M, Brandt ML. Novel educational approach for medical students: improved retention rates using interactive medical software compared with traditional lecture-based format. J Surg Educ 2012;69(2):253-256. [doi: 10.1016/j.jsurg.2011.12.007] [Medline: 22365876]
5. Xu M, Luo Y, Zhang Y, Xia R, Qian H, Zou X. Game-based learning in medical education. Front Public Health 2023;11:1113682. [doi: 10.3389/fpubh.2023.1113682] [Medline: 36935696]
6. Deterding S, Khaled R, Nacke LE, Dixon D. Gamification: toward a definition. Presented at: CHI 2011 Gamification Workshop Proceedings; May 7-12, 2011 URL: https://www.researchgate.net/publication/273947177_Gamification_Toward_a_definition [accessed 2025-11-09]
7. Gentry SV, Gauthier A, L'Estrade Ehrstrom B, et al. Serious gaming and gamification education in health professions: systematic review. J Med Internet Res 2019 Mar 28;21(3):e12994. [doi: 10.2196/12994] [Medline: 30920375]
8. Bigdeli S, Hosseinzadeh Z, Dehnad A, et al. Underpinning learning theories of medical educational games: a scoping review. Med J Islam Repub Iran 2023;37(1):26. [doi: 10.47176/mjiri.37.26] [Medline: 37180860]
9. Salehi AM, Mohammadi HA, Jenabi E, Khanlarzadeh E, Ashtari K. Quality of evidence and pedagogical strategy in using gamification in medical education literature: a systematic review. Simul Gaming 2023 Dec;54(6):598-620. [doi: 10.1177/10468781231195903]
10. Alnuaim A. The impact and acceptance of gamification by learners in a digital literacy course at the undergraduate level: randomized controlled trial. JMIR Serious Games 2024 Aug 23;12(1):e52017. [doi: 10.2196/52017] [Medline: 39177662]
11. Molina-Torres G, Rodriguez-Arrastia M, Alarcón R, et al. Game-based learning outcomes among physiotherapy students: comparative study. JMIR Serious Games 2021 Mar 24;9(1):e26007. [doi: 10.2196/26007] [Medline: 33759800]
12. Forni MF, Garcia-Neto W, Kowaltowski AJ, Marson GA. An active-learning methodology for teaching oxidative phosphorylation. Med Educ 2017 Nov;51(11):1169-1170. [doi: 10.1111/medu.13418] [Medline: 28857228]
13. Worm BS, Buch SV. Does competition work as a motivating factor in e-learning? A randomized controlled trial. PLoS One 2014;9(1):e85434. [doi: 10.1371/journal.pone.0085434] [Medline: 24465561]

14. Adamson MA, Chen H, Kackley R, Micheal A. For the love of the game: game- versus lecture-based learning with generation Z patients. J Psychosoc Nurs Ment Health Serv 2018 Feb 1;56(2):29-36. [doi: 10.3928/02793695-20171027-03] [Medline: 29117424]
15. Beattie M. Using gamification to motivate Gen Z. Alvaria Blog. 2022 Jun 9. URL: https://www.alvaria.com/blog/using-gamification-to-motivate-gen-z [accessed 2024-12-06]
16. Kuo CL, Chuang YH. Kahoot: applications and effects in education. Hu Li Za Zhi 2018 Dec;65(6):13-19. [doi: 10.6224/JN.201812_65(6).03] [Medline: 30488408]
17. Prensky M. Digital natives, digital immigrants part 1. On Horizon 2001 Sep;9(5):1-6. [doi: 10.1108/10748120110424816]
18. Chapman JR, Rich PJ. Does educational gamification improve students' motivation? If so, which game elements work best? J Educ Bus 2018 Oct 3;93(7):315-322. [doi: 10.1080/08832323.2018.1490687]
19. Seemiller C, Grace M. Generation Z: educating and engaging the next generation of students. About Campus 2017 Jul;22(3):21-26. [doi: 10.1002/abc.21293]
20. Barrickman A, McMillan A, Gálvez-Peralta M, Purnell K. Development and assessment of integrated virtual escape rooms to reinforce cardiology content and skills. Am J Pharm Educ 2023 Apr;87(3):ajpe8899. [doi: 10.5688/ajpe8899] [Medline: 36270662]
21. Jones JS, Tincher L, Odeng-Otu E, Herdman M. An educational board game to assist PharmD students in learning autonomic nervous system pharmacology. Am J Pharm Educ 2015 Oct 25;79(8):114. [doi: 10.5688/ajpe798114] [Medline: 26689278]
22. Kow AWC, Ang BLS, Chong CS, Tan WB, Menon KR. Innovative patient safety curriculum using iPAD game (PASSED) improved patient safety concepts in undergraduate medical students. World J Surg 2016 Nov;40(11):2571-2580. [doi: 10.1007/s00268-016-3623-x] [Medline: 27417109]
23. Saxena A, Nesbitt R, Pahwa P, Mills S. Crossword puzzles: active learning in undergraduate pathology and medical education. Arch Pathol Lab Med 2009 Sep 1;133(9):1457-1462. [doi: 10.5858/133.9.1457]
24. Shah S, Lynch LMJ, Macias-Moriarity LZ. Crossword puzzles as a tool to enhance learning about anti-ulcer agents. Am J Pharm Educ 2010 Sep;74(7):117. [doi: 10.5688/aj7407117]
25. van Gaalen AEJ, Brouwer J, Schönrock-Adema J, Bouwkamp-Timmer T, Jaarsma ADC, Georgiadis JR. Gamification of health professions education: a systematic review. Adv in Health Sci Educ 2021 May;26(2):683-711. [doi: 10.1007/s10459-020-10000-3]
26. Colliver JA, Kucera K, Verhulst SJ. Meta-analysis of quasi-experimental research: are systematic narrative reviews indicated? Med Educ 2008 Sep;42(9):858-865. [doi: 10.1111/j.1365-2923.2008.03144.x] [Medline: 18715482]
27. Park S, Kim H. Development and effect of prenatal education programs using virtual reality for pregnant women hospitalized with preterm labor: experimental study. J Med Internet Res 2025 Jun 30;27(1):e75585. [doi: 10.2196/75585] [Medline: 40587895]
28. Verma GS, Gopalakrishnan L, Ayadi AE, et al. Preliminary effectiveness of a postnatal mHealth and virtual social support intervention on newborn and infant health and feeding practices in Punjab, India: quasi-experimental pre-post pilot study. JMIR Pediatr Parent 2025 Jun 27;8(1):e65581. [doi: 10.2196/65581] [Medline: 40577696]
29. Eysenbach G, Stoner S, Drozd F. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
30. Shadish WR, Cook TD, Campbell DT. Experimental and Quasi-Experimental Designs for Generalized Causal Inference: A Norton Critical Edition: Houghton Mifflin; 2002.
31. Crossword puzzle worksheet generator. Education.com. URL: https://www.education.com/worksheet-generator/reading/crossword-puzzle/ [accessed 2024-06-12]
32. Word search generator [Web page in Portuguese]. Educolorir. URL: https://www.educolorir.com/gerador-de-sopa-de-letras [accessed 2024-06-12]
33. Helle N, Vikman MD, Dahl-Michelsen T, Lie SS. Health care and social work students' experiences with a virtual reality simulation learning activity: qualitative study. JMIR Med Educ 2023 Jan 1;9:e49372. [doi: 10.2196/49372]
34. De Vellis RF. Scale Development Theory and Applications, 4th edition: Sage Publication; 2017.
35. Vlachopoulos D, Makri A. The effect of games and simulations on higher education: a systematic literature review. Int J Educ Technol High Educ 2017 Dec;14(1):1-33. [doi: 10.1186/s41239-017-0062-1]
36. Lee CK, Sidhu MS. Engineering students learning preferences in UNITEN: comparative study and patterns of learning styles on JSTOR. Educ Technol Soc 2015;18(3):266-281 [FREE Full text]
37. dos Reis Lívero FA, da Silva GR, Amaral EC, et al. Playfulness in the classroom: gamification favor the learning of pharmacology. Educ Inf Technol 2021 Mar;26(2):2125-2141. [doi: 10.1007/s10639-020-10350-w]
38. Patrick S, Vishwakarma K, Giri VP, et al. The usefulness of crossword puzzle as a self-learning tool in pharmacology. J Adv Med Educ 2018 Oct;6(4):181-185 [FREE Full text] [Medline: 30349830]
39. Cantwell C, Saadat S, Sakaria S, Wiechmann W, Sudario G. Escape box and puzzle design as educational methods for engagement and satisfaction of medical student learners in emergency medicine: survey study. BMC Med Educ 2022 Jul 2;22(1):518. [doi: 10.1186/s12909-022-03585-3] [Medline: 35780126]
40. Hope DL, Grant GD, Rogers GD, King MA. Gamification in pharmacy education: a systematic quantitative literature review. Int J Pharm Pract 2023 Mar 13;31(1):15-31. [doi: 10.1093/ijpp/riac099]

41. Bawazeer G, Sales I, Albogami H, et al. Crossword puzzle as a learning tool to enhance learning about anticoagulant therapeutics. BMC Med Educ 2022 Apr 11;22(1):267. [doi: 10.1186/s12909-022-03348-0] [Medline: 35410242]
42. Hu J, Sonnleitner M, Weldon E, Kejriwal S, Brown B, Shah A. An escape room to teach first- and second-year medical students nephrology. Med Sci Educ 2024 Feb;34(1):71-76. [doi: 10.1007/s40670-023-01917-6] [Medline: 38510392]
43. Eukel HN, Frenzel JE, Cernusca D. Educational gaming for pharmacy students - design and evaluation of a diabetes-themed escape room. Am J Pharm Educ 2017 Sep;81(7):6265. [doi: 10.5688/ajpe8176265] [Medline: 29109566]
44. Frenzel JE, Cernusca D, Marg C, Schotters B, Eukel HN. Design‐based research: studying the effects of an escape room on students' knowledge and perceptions. J Am Coll Clin Pharm 2020 Nov;3(7):1326-1332. [doi: 10.1002/jac5.1290]
45. Gudadappanavar AM, Benni JM, Javali SB. Effectiveness of the game-based learning over traditional teaching-learning strategy to instruct pharmacology for phase II medical students. J Educ Health Promot 2021;10(1):91. [doi: 10.4103/jehp.jehp_624_20] [Medline: 34084838]
46. Twist KE, Ragsdale JW. Candy gland: a diabetes board game for medical students. MedEdPORTAL 2022;18:11294. [doi: 10.15766/mep_2374-8265.11294] [Medline: 36654983]
47. Ghelfenstein-Ferreira T, Beaumont AL, Dellière S, et al. An educational game evening for medical residents: a proof of concept to evaluate the impact on learning of the use of games. J Microbiol Biol Educ 2021;22(2):e00119-21. [doi: 10.1128/jmbe.00119-21] [Medline: 34594443]
48. Sannathimmappa MB, Nambiar V, Aravindakshan R. Learning out of the box: fostering intellectual curiosity and learning skills among the medical students through gamification. J Educ Health Promot 2022;11:79. [doi: 10.4103/jehp.jehp_683_21] [Medline: 35434144]
49. Delage C, Palayer M, Lerouet D, Besson VC. "Pharmacotrophy": a playful tournament for game- and team-based learning in pharmacology education - assessing its impact on students' performance. BMC Med Educ 2024 Mar 1;24(1):219. [doi: 10.1186/s12909-024-05157-z] [Medline: 38429772]
50. Matreja PS, Kaur J, Yadav L. Acceptability of the use of crossword puzzles as an assessment method in pharmacology. J Adv Med Educ 2021 Jun 1;9(3):154-159. [doi: 10.30476/JAMP.2021.90517.1413]
51. Shawwa L, Kamel F, Shawwa L, Kamel F. Assessing the knowledge and perceptions of medical students after using kahoot. Cureus 2023 Mar 28;15(3). [doi: 10.7759/CUREUS.36796] [Medline: 37012955]
52. Smiderle R, Rigo SJ, Marques LB, Peçanha de Miranda Coelho JA, Jaques PA. The impact of gamification on students' learning, engagement and behavior based on their personality traits. Smart Learn Environ 2020 Dec;7(1):1-11. [doi: 10.1186/s40561-019-0098-x]
53. Van Gaalen AEJ, Jaarsma ADC, Georgiadis JR. Medical students' perceptions of play and learning: qualitative study with focus groups and thematic analysis. JMIR Serious Games 2021 Jul 28;9(3):e25637. [doi: 10.2196/25637] [Medline: 34319237]
54. Conati C. Probabilistic assessment of user's emotions in educational games. Appl Artif Intell 2002 Aug;16(7-8):555-575. [doi: 10.1080/08839510290030390]
55. Maxim RI, Arnedo-Moreno J. Identifying key principles and commonalities in digital serious game design frameworks: scoping review. JMIR Serious Games 2025 Mar 5;13(1):e54075. [doi: 10.2196/54075] [Medline: 40053743]
56. Bou Nemer L, Kalin D, Fiorentino D, Garcia JJ, Estes CM. The labor games. Obstet Gynecol 2016 Oct;128(Suppl 1):1S-5S. [doi: 10.1097/AOG.0000000000001572] [Medline: 27662001]
57. Nevin CR, Westfall AO, Rodriguez JM, et al. Gamification as a tool for enhancing graduate medical education. Postgrad Med J 2014 Dec;90(1070):685-693. [doi: 10.1136/postgradmedj-2013-132486] [Medline: 25352673]
58. Van Nuland SE, Roach VA, Wilson TD, Belliveau DJ. Head to head: the role of academic competition in undergraduate anatomical education. Anat Sci Educ 2015 Sep;8(5):404-412. [doi: 10.1002/ase.1498] [Medline: 25319077]
59. Reeve J, Deci EL. Elements of the competitive situation that affect intrinsic motivation. Pers Soc Psychol Bull 1996 Jan;22(1):24-33. [doi: 10.1177/0146167296221003]
60. Kirsch J, Spreckelsen C. Caution with competitive gamification in medical education: unexpected results of a randomised cross-over study. BMC Med Educ 2023 Apr 19;23(1):259. [doi: 10.1186/s12909-023-04258-5] [Medline: 37072842]
61. McCoy L, Lewis JH, Dalton D. Gamification and multimedia for medical education: a landscape review. J Am Osteopath Assoc 2016 Jan;116(1):22-34. [doi: 10.7556/jaoa.2016.003] [Medline: 26745561]
62. Zainuddin Z, Chu SK, Shujahat M, Perera CJ. The impact of gamification on learning and instruction: a systematic review of empirical evidence. Educ Res Rev 2020 Jun;30:100326. [doi: 10.1016/j.edurev.2020.100326]
63. Zammarchi G, Del Zompo M, Squassina A, Pisanu C. Increasing engagement in pharmacology and pharmacogenetics education using games and online resources: the PharmacoloGenius mobile app. Drug Dev Res 2020 Dec;81(8):985-993. [doi: 10.1002/ddr.21714] [Medline: 32633017]

Abbreviations

CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
UNESCO: United Nations Educational, Scientific and Cultural Organization

Edited by D Chartash; submitted 16.03.25; peer-reviewed by JM Raimundo, MK Ghanta, S Karekar; revised version received 16.07.25; accepted 16.07.25; published 05.12.25.

Please cite as:
Aloum L, Ibrahim H, Rajasekaran SK, Alefishat E
Open-Access Web-Based Gamification in Pharmacology Education for Medical Students: Quasi-Experimental Study
JMIR Med Educ 2025;11:e73666
URL: https://mededu.jmir.org/2025/1/e73666
doi:10.2196/73666

Authors' Reply: Enhancing AI-Driven Medical Translations: Considerations for Language Concordance

Joyce Teng[1], MD, PhD; Roberto Andres Novoa[1,2*], MD; Maria Alexandrovna Aleshin[1*], MD; Jenna Lester[3*], MD; Kira Seiger[4*], MD, MBA; Fiatsogbe Dzuali[3*], MD; Roxana Daneshjou[1,5], MD, PhD

[1]Department of Dermatology, Stanford University, 700 Welch Rd, Stanford, United States
[2]Department of Pathology, Stanford University, Redwood City, CA, United States
[3]Department of Dermatology, University of California, San Francisco, CA, United States
[4]Department of Dermatology, University of Washington, Seattle, WA, United States
[5]Department of Biomedical Data Science, Stanford University, Stanford, CA, United States
[*]these authors contributed equally

Corresponding Author:
Joyce Teng, MD, PhD
Department of Dermatology, Stanford University, 700 Welch Rd, Stanford, United States

Related Articles:

https://mededu.jmir.org/2024/1/e51435

Comment in: https://mededu.jmir.org/2025/1/e70420

(JMIR Med Educ 2025;11:e71721) doi:10.2196/71721

KEYWORDS

ChatGPT; artificial intelligence; language; translation; health care disparity; natural language model; survey; patient education; accessibility; preference; human language; communication; language-concordant care

We appreciate the thoughtful insights shared by Quon and Zhou [1] regarding our study on the application of ChatGPT in translating patient education materials [2]. We wholly agree that the linguistically distinct languages, such as Mandarin, can present challenges in capturing all the nuances and achieving precise translations.

In response to the comment regarding the use of multiple prompts, we acknowledge the complexity and variability in artificial intelligence (AI)-generated translations. However, it is important to consider the practical limitations within a clinical setting. Asking providers to use various prompts in real time may not be feasible due to time constraints and the need for efficiency in patient care. We believe that focusing on a single, effective prompt can streamline the translation process while we explore avenues for improvement in the AI's capabilities. This could be a productive avenue for future research.

Addressing the concern regarding the reliance on board-certified dermatologists for post-translation review, we want to clarify that, in addition to being board-certified dermatologists, all reviewers were native speakers in the language they reviewed, including fluency in Mandarin at a college level. This proficiency allows for a confluence of both clinical and linguistic insights when evaluating translations, reinforcing the validity of our findings. We appreciate the importance of rigor in translation review and remain committed to enhancing the integrity of our translated materials.

Overall, while we recognize the areas where ChatGPT can improve, we also see its current utility as a valuable tool for expanding access to language-concordant care in clinical settings. Our study serves as a helpful step toward identifying and addressing the limitations of AI translations, and we welcome continued dialogue to refine these practices.

Conflicts of Interest

None declared.

References

1. Quon S, Zhou S. Enhancing AI-driven medical translations: considerations for language concordance. JMIR Med Educ 2025;11. [doi: 10.2196/70420]
2. Dzuali F, Seiger K, Novoa R, et al. ChatGPT may improve access to language-concordant care for patients with non-English language preferences. JMIR Med Educ 2024 Dec 10;10:e51435. [doi: 10.2196/51435] [Medline: 39657144]

Abbreviations

AI: artificial intelligence

Edited by T Leung; submitted 24.01.25; this is a non–peer-reviewed article;accepted 27.01.25; published 11.04.25.

Please cite as:
Teng J, Novoa RA, Aleshin MA, Lester J, Seiger K, Dzuali F, Daneshjou R
Authors' Reply: Enhancing AI-Driven Medical Translations: Considerations for Language Concordance
JMIR Med Educ 2025;11:e71721
URL: https://mededu.jmir.org/2025/1/e71721
doi:10.2196/71721

Author's Reply: Examining Multimodal AI Resources in Medical Education: The Role of Immersion, Motivation, and Fidelity in AI Narrative Learning

Tyler Bland, PhD

Department of Medical Education, University of Idaho, 875 Perimeter Drive MS 4061, Moscow, ID, United States

Corresponding Author:
Tyler Bland, PhD
Department of Medical Education, University of Idaho, 875 Perimeter Drive MS 4061, Moscow, ID, United States

Related Articles:

https://mededu.jmir.org/2025/1/e63865

https://mededu.jmir.org/2025/1/e72190

Abstract

(JMIR Med Educ 2025;11:e72336) doi:10.2196/72336

KEYWORDS

artificial intelligence; cinematic clinical narrative; cinemeducation; medical education; narrative learning; pharmacology; AI; medical students; preclinical education; long-term retention; AI tools; GPT-4; image; applicability; CCN

I extend my sincere appreciation for the thoughtful critique [1] of my study, "Enhancing Medical Student Engagement Through Cinematic Clinical Narratives: Multimodal Generative AI–Based Mixed Methods Study" [2]. The author's insights regarding engagement mechanisms, theoretical expansion, and methodological refinements offer valuable perspectives that contribute to the broader discourse on the pedagogical applications of generative artificial intelligence in medical education.

While the Cognitive Affective Model of Immersive Learning framework originated to explain learning with immersive virtual reality technologies [3], I concur that its underlying principles are applicable to my study. The debate over the role of media versus instructional methods in learning has been longstanding. While some argue that the medium itself shapes cognition, social structures, and cultural norms [4], others reject this notion, asserting that media are merely delivery mechanisms and that instructional methods alone drive learning outcomes [5]. The Cognitive Affective Model of Immersive Learning reframes this debate by emphasizing that it is not the medium (eg, immersive virtual reality) that inherently enhances learning, but rather how instructional methods leverage the unique affordances of that medium. In the context of cinematic clinical narratives (CCNs), the structured narrative and multimodal design capitalize on engagement mechanisms similar to those observed in immersive learning. Future research could further examine how instructional design within CCNs optimally harnesses these principles to promote knowledge retention and clinical application.

The author's recommendation of the integration of pretest and posttest methodologies is well-founded. While the published study employed posttest assessments to measure comprehension, incorporating pretest measures would facilitate a more granular evaluation of baseline knowledge and attitudinal shifts attributable to CCNs. Furthermore, longitudinal assessments could provide critical insights into the durability of knowledge retention and the sustained impact of CCNs over extended timeframes. I aim to incorporate these into future studies.

The author's call for broader contextual applications of CCNs beyond traditional classroom settings is well-taken. While the study examined CCN implementation within a structured learning environment, I am currently working on converting CCNs into self-contained short films that can be viewed online for self-directed learning. This adaptation aims to provide learners with greater flexibility while maintaining the engagement and narrative-driven structure of CCNs. Investigating how these self-contained films perform across varied instructional modalities could yield valuable insights into their scalability and applicability within diverse educational contexts.

Finally, I concur with the author's observation that medical students are increasingly turning to digital platforms such as social media for information and engagement. Medical educators should take note and examine the factors that make these

platforms so compelling. By understanding the draw of these digital environments, educators can incorporate similar characteristics into medical school learning materials to meet students where they are. Expanding CCN research to explore how elements such as interactivity, brevity, and personalization influence learner engagement could provide valuable insights into modernizing medical education. I am grateful for the astute observations and constructive recommendations by the author. These perspectives will undoubtedly inform my future research directions and further the integration of artificial intelligence–driven methodologies in my studies on medical education.

Conflicts of Interest

None declared.

References

1. Jacobs C. Examining Multimodal AI Resources in Medical Education: The Role of Immersion, Motivation, and Fidelity in AI Narrative Learning. JMIR Med Educ 2025;11:e72190. [doi: 10.2196/72190]
2. Bland T. Enhancing medical student engagement through cinematic clinical narratives: multimodal generative AI–based mixed methods study. JMIR Med Educ 2025 Jan 6;11(1):e63865. [doi: 10.2196/63865] [Medline: 39791333]
3. Makransky G, Petersen GB. The Cognitive Affective Model of Immersive Learning (CAMIL): a theoretical research-based model of learning in immersive virtual reality. Educ Psychol Rev 2021 Sep;33(3):937-958. [doi: 10.1007/s10648-020-09586-2]
4. McLuhan M. Understanding Media: The Extensions of Man: McGraw-Hill; 1964.
5. Clark RE. Media will never influence learning. ETR&D 1994 Jun;42(2):21-29. [doi: 10.1007/BF02299088]

Abbreviations

CCN: cinematic clinical narrative

Edited by S Nedunchezhiyan; submitted 07.02.25; this is a non–peer-reviewed article;accepted 14.02.25; published 18.03.25.

Please cite as:
Bland T
Author's Reply: Examining Multimodal AI Resources in Medical Education: The Role of Immersion, Motivation, and Fidelity in AI Narrative Learning
JMIR Med Educ 2025;11:e72336
URL: https://mededu.jmir.org/2025/1/e72336
doi:10.2196/72336

Examining Multimodal AI Resources in Medical Education: The Role of Immersion, Motivation, and Fidelity in AI Narrative Learning

Chris Jacobs, MB BChir, BSc, MRes, MD(Res)

Department of Psychology, University of Bath, Claverton Down, Bath, United Kingdom

Corresponding Author:
Chris Jacobs, MB BChir, BSc, MRes, MD(Res)
Department of Psychology, University of Bath, Claverton Down, Bath, United Kingdom

Related Articles:

https://mededu.jmir.org/2025/1/e63865

Comment in: https://mededu.jmir.org/2025/1/e72336

(JMIR Med Educ 2025;11:e72190) doi:10.2196/72190

KEYWORDS

artificial intelligence; cinematic clinical narrative; cinemeducation; medical education; narrative learning; AI; medical students; preclinical education; long-term retention; pharmacology; AI tools; GPT-4; image; applicability; CCN

I read with great interest the recent study by Bland [1], "Enhancing Medical Student Engagement Through Cinematic Clinical Narratives: Multimodal Generative AI–Based Mixed Methods Study," which explored the use of cinematic clinical narratives (CCNs) in medical education. The findings highlighted the potential of multimodal generative artificial intelligence (AI) to enhance engagement and knowledge retention among medical students. While the study effectively demonstrated novel use of AI to modernize case learning, further exploration of engagement mechanisms and broader learning theories may deepen our understanding of how these approaches can be optimized for educational impact.

Engagement in learning is multifaceted and can be linked to immersion [2] and intrinsic motivation [3]. As Bland observed, students exhibited heightened situational interest in CCNs, reinforcing the idea that immersive learning environments can enhance attention and recall. However, beyond the constructivist learning theory discussed in the study, additional models could be considered to expand the theoretical framework for understanding these results. One such model is the Cognitive Affective Model of Immersive Learning framework [4], which emphasizes the interplay between representational fidelity, cognitive load, and technological mediation in shaping learner experiences. Exploring the Cognitive Affective Model of Immersive Learning and similar frameworks could provide a more nuanced perspective on how technology interacts with learner motivation and engagement.

Another area for further inquiry is the role of pretest and posttest methodologies in evaluating learning outcomes. As the study rightly acknowledges, medical education research is often limited to single posttest designs [5], which may not capture students' initial levels of engagement or attitudes toward learning. Additionally, fidelity of experience is important for recall, as realistic and contextually accurate learning environments can enhance memory retention and application. Implementing pretest assessments could help quantify shifts in engagement, allowing researchers to distinguish between students who are inherently motivated and those whose interest is primarily triggered by the intervention itself. Such an approach could offer valuable insights into how CCNs influence different learner profiles.

Moreover, the conditions under which learning occurs significantly affect student engagement. The study employed a classroom-based intervention, which effectively bridged the gap between controlled laboratory settings and real-world educational environments. Future research might explore how CCNs perform across varied instructional contexts, including self-paced online learning and clinical simulation settings, to assess their adaptability and impact on different learning situations and scenarios.

Bland provides good evidence that multimodal AI resources hold promise in medical education, underscoring the need to adapt to the evolving preferences of modern medical students who increasingly turn to social media for information and engagement. Expanding this research to integrate additional learning theories and measures, with varied instructional settings will help establish best practices for how to use these innovative tools. I appreciate the author's novel approach and am eager to see how future developments in this field will shape medical education.

Conflicts of Interest

None declared.

References

1. Bland T. Enhancing medical student engagement through cinematic clinical narratives: multimodal generative AI–based mixed methods study. JMIR Med Educ 2025 Jan 6;11:e63865. [doi: 10.2196/63865] [Medline: 39791333]
2. Jacobs C, Maidwell-Smith A. Learning from 360-degree film in healthcare simulation: a mixed methods pilot. J Vis Commun Med 2022 Oct;45(4):223-233. [doi: 10.1080/17453054.2022.2097059] [Medline: 35938350]
3. Ryan RM, Deci EL. Intrinsic and extrinsic motivations: classic definitions and new directions. Contemp Educ Psychol 2000 Jan;25(1):54-67. [doi: 10.1006/ceps.1999.1020] [Medline: 10620381]
4. Makransky G, Petersen GB. The Cognitive Affective Model of Immersive Learning (CAMIL): a theoretical research-based model of learning in immersive virtual reality. Educ Psychol Rev 2021 Sep;33(3):937-958. [doi: 10.1007/s10648-020-09586-2]
5. Scerbo MW, Calhoun AW, Hui J. Research and hypothesis testing: moving from theory to experiment. In: Nestel D, Hui J, Kunkler K, Scerbo MW, Calhoun AW, editors. Healthcare Simulation Research: A Practical Guide: Springer International Publishing; 2019:161-167. [doi: 10.1007/978-3-030-26837-4_22]

Abbreviations

AI: artificial intelligence
CCN: cinematic clinical narrative

Edited by S Nedunchezhiyan; submitted 05.02.25; this is a non–peer-reviewed article;accepted 14.02.25; published 18.03.25.

Please cite as:
Jacobs C
Examining Multimodal AI Resources in Medical Education: The Role of Immersion, Motivation, and Fidelity in AI Narrative Learning
JMIR Med Educ 2025;11:e72190
URL: https://mededu.jmir.org/2025/1/e72190
doi:10.2196/72190

Enhancing AI-Driven Medical Translations: Considerations for Language Concordance

Stephanie Quon[1], BASc; Sarah Zhou[2], BSc

[1]Faculty of Medicine, University of British Columbia, 2194 Health Sciences Mall, Vancouver, BC, Canada

[2]Faculty of Science, University of British Columbia, Vancouver, BC, Canada

Corresponding Author:
Stephanie Quon, BASc
Faculty of Medicine, University of British Columbia, 2194 Health Sciences Mall, Vancouver, BC, Canada

Related Articles:

Companion article: https://mededu.jmir.org/2024/1/e51435

https://mededu.jmir.org/2025/1/e71721

(JMIR Med Educ 2025;11:e70420) doi:10.2196/70420

KEYWORDS

letter to the editor; ChatGPT; AI; artificial intelligence; language; translation; health care disparity; natural language model; survey; patient education; accessibility; preference; human language; communication; language-concordant care

We commend the recent publication by Dzuali et al [1], which explored the application of ChatGPT for translating patient education materials into multiple languages. This important study highlights a critical area where artificial intelligence (AI) can potentially bridge gaps in language-concordant care. To further this research, we would like to raise several points to enrich the discussion and understanding of the findings.

The study demonstrates that while ChatGPT provides clinically usable translations for Spanish and Russian, its performance with Mandarin is suboptimal. This inconsistency raises important questions regarding the linguistic complexities and structural differences between English and Mandarin, which may hinder the accuracy and appropriateness of translations. Previous research has shown that the nuanced sentence structures and specialized terminology in Mandarin pose challenges for AI models such as ChatGPT, suggesting the need for more refined approaches when using AI for translation in linguistically distinct languages [2].

Being familiar with the Mandarin language, we have firsthand experience with the challenges that come with translating between languages with distinct linguistic structures. Mandarin, with its nuanced sentence structures and specialized terminology, presents difficulties for large language models such as ChatGPT. These challenges are compounded by differences in grammar, idiomatic expressions, and cultural contexts, which may lead to inaccuracies and misunderstandings in translations. Therefore, this study could provide additional insight into how cultural context influences translation quality. Mandarin, for example, involves not only linguistic precision but also an understanding of cultural nuances that could affect comprehension [3]. Future studies could explore how AI models such as ChatGPT are trained to account for these contextual factors to ensure culturally appropriate translations.

Another area for potential exploration in this study is the testing of alternative prompts and the impact they may have on translation quality. While the study focuses on a single translation prompt—“Translate this into <target language>”—the variability of AI-generated translations could be better evaluated through a variety of prompts. Utilizing multiple prompts could reveal a broader range of performance outcomes, especially for linguistically complex languages such as Mandarin and Russian. Other studies have shown that different AI prompts can produce vastly different results [4].

Lastly, the study heavily relies on the involvement of board-certified dermatologists for posttranslation review, which is applicable to the context of dermatology-related information, but may not fully address the extent of errors and misinformation. While human oversight is essential, the study could benefit from a more robust evaluation of how different levels of human intervention—such as linguistic experts or specialists in medical translation—might improve translation accuracy [5]. Future research should explore how different combinations of AI-generated translations and human review from varied sources could optimize clinical usability.

Overall, while ChatGPT shows promise for improving access to language-concordant patient education, further refinement and validation are required. This study is an important milestone in starting this discussion surrounding AI-translation in medical contexts, and we commend the authors for their valuable contribution to advancing the field. They clearly demonstrate a meticulous approach, thoughtful analysis, and commitment to improving patient care through innovative solutions.

Conflicts of Interest

None declared.

References

1. Dzuali F, Seiger K, Novoa R, et al. ChatGPT may improve access to language-concordant care for patients with non-English language preferences. JMIR Med Educ 2024 Dec 10;10:e51435. [doi: 10.2196/51435] [Medline: 39657144]
2. Jiao W, et al. Is chatgpt A good translator? A preliminary study. arXiv. Preprint posted online on Oct 1, 2023. [doi: 10.48550/arXiv.2301.08745]
3. Duff P, et al. Learning Chinese: Linguistic, Sociocultural, and Narrative Perspectives: Walter de Gruyter; 2013, Vol. 5.
4. Oppenlaender J, Linder R, Silvennoinen J. Prompting AI art: an investigation into the creative skill of prompt engineering. Int J Hum Comput 2024:1-23. [doi: 10.1080/10447318.2024.2431761]
5. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]

Abbreviations

AI: artificial intelligence

Edited by T Leung; submitted 20.12.24; this is a non–peer-reviewed article;accepted 27.01.25; published 11.04.25.

Please cite as:
Quon S, Zhou S
Enhancing AI-Driven Medical Translations: Considerations for Language Concordance
JMIR Med Educ 2025;11:e70420
URL: https://mededu.jmir.org/2025/1/e70420
doi:10.2196/70420

Citation Accuracy Challenges Posed by Large Language Models

Manlin Zhang[1*], MPhil; Tianyu Zhao[2*], MSc

[1]Department of Obstetrics and Gynecology, Shengjing Hospital of China Medical University, Shenyang, China

[2]Department of Science and Technology Studies, University College London, Gower Street, London, United Kingdom

[*]all authors contributed equally

Corresponding Author:
Tianyu Zhao, MSc
Department of Science and Technology Studies, University College London, Gower Street, London, United Kingdom

Related Articles:

https://mededu.jmir.org/2025/1/e63400

Comment in: https://mededu.jmir.org/2025/1/e73698

(JMIR Med Educ 2025;11:e72998) doi:10.2196/72998

KEYWORDS

chatGPT; medical education; Saudi Arabia; perceptions; knowledge; medical students; faculty; chatbot; qualitative study; artificial intelligence; AI; AI-based tools; universities; thematic analysis; learning; satisfaction; LLM; large language model

Large language models (LLMs) such as DeepSeek, ChatGPT, and ChatGLM have significant limitations in generating citations, raising concerns about the quality and reliability of academic research. These models tend to produce citations that are correctly formatted but fictional in content, misleading users and undermining academic rigor. In the recent study titled "Perceptions and earliest experiences of medical students and faculty with ChatGPT in medical education: qualitative study," the section addressing concerns about ChatGPT deserves a deeper discussion [1].

There are several reasons for the citation issues in LLMs, which can be analyzed as follows. First, most LLMs cannot access paid subscription databases and therefore solely rely on open-access resources [2]. This limits the citations generated by LLMs to open-access journals, potentially omitting more significant research published in subscription-based journals. Second, LLMs are trained on vast amounts of text data and generate content by analyzing patterns and structures in text. However, they lack the ability to understand the content of the text or think critically, implying that they cannot judge the accuracy and reliability of information. Third, the algorithms underlying LLMs are often opaque, leaving users unable to understand the specific processes of information handling. This makes it difficult for users to determine the reliability of citations generated by LLMs and to effectively evaluate their results. Recent research also stated that half of generated search results lack citations, and only 75% of those with citations support the claims, posing trust concerns as user reliance grows[3].

Recently, an experiment conducted by the Journal of Clinical Anesthesia involved publishing a fictional article titled "Spinal Cord Ischemia After ESP Block" to test the spread and citation of a fabricated academic content. Surprisingly, the fictional article was widely cited, over 400 times, including in some journals with high impact factors[4], revealing a lack of rigor in academic citation practices, where many authors may not check the original literature and instead copy references directly. This incident sparked widespread discussion about academic citation practices, emphasizing the importance of critical thinking by scholars while citing materials.

The use of fictional citations by LLMs poses a multifaceted problem: it misleads users into drawing incorrect conclusions and making inappropriate decisions, undermines the rigor and credibility of academic research, and hinders the dissemination of knowledge by limiting access to accurate scientific information [5]. The issue of LLMs generating fictional citations is complex and requires the combined efforts of multiple stakeholders for resolution. Developers must continuously improve the LLM technology and algorithms, users must increase their awareness and critical evaluation skills while using LLMs, and academic institutions must strengthen the management and education in academic practices. Only through these efforts can we ensure that LLMs play a positive role in academic research and promote the dissemination and progress of knowledge.

Conflicts of Interest

None declared.

References

1. Abouammoh N, Alhasan K, Aljamaan F, et al. Perceptions and earliest experiences of medical students and faculty with ChatGPT in medical education: qualitative study. JMIR Med Educ 2025 Feb 20;11:e63400. [doi: 10.2196/63400] [Medline: 39977012]
2. Perianes-Rodríguez A, Olmeda-Gómez C. Effects of journal choice on the visibility of scientific publications: a comparison between subscription-based and full open access models. Scientometrics 2019 Dec;121(3):1737-1752. [doi: 10.1007/s11192-019-03265-y]
3. Peskoff D, Stewart B. Credible without credit: domain experts assess generative language models. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics 2023 Jul;2:427-438. [doi: 10.18653/v1/2023.acl-short.37]
4. Marcus A, Oransky I, De Cassai A. Please don't cite this editorial. J Clin Anesth 2025 Jan 8:111741. [doi: 10.1016/j.jclinane.2025.111741] [Medline: 39779384]
5. Rasul T, Nair S, Kalendra D, et al. The role of ChatGPT in higher education: benefits, challenges, and future research directions. JALT 2023 May 10;6(1):41-56. [doi: 10.37074/jalt.2023.6.1.29]

Abbreviations

LLM: large language model

Edited by S Nedunchezhiyan; submitted 23.02.25; this is a non–peer-reviewed article;accepted 12.03.25; published 02.04.25.

Please cite as:
Zhang M, Zhao T
Citation Accuracy Challenges Posed by Large Language Models
JMIR Med Educ 2025;11:e72998
URL: https://mededu.jmir.org/2025/1/e72998
doi:10.2196/72998

Authors' Reply: Citation Accuracy Challenges Posed by Large Language Models

Mohamad-Hani Temsah[1], MD; Ayman Al-Eyadhy[1], MD; Amr Jamal[2], MBBS; Khalid Alhasan[1], MBBS; Khalid H Malki[3], PhD

[1]Pediatric Department, College of Medicine, King Saud University, King Abdullah Road, Riyadh, Saudi Arabia

[2]Department of Family and Community Medicine, King Saud University Medical City, Riyadh, Saudi Arabia

[3]Research Chair of Voice, Swallowing, and Communication Disorders, Department of Otolaryngology-Head and Neck Surgery, College of Medicine, King Saud University, Riyadh, Saudi Arabia

Corresponding Author:
Mohamad-Hani Temsah, MD
Pediatric Department, College of Medicine, King Saud University, King Abdullah Road, Riyadh, Saudi Arabia

Related Articles:

https://mededu.jmir.org/2025/1/e72998

https://mededu.jmir.org/2025/1/e63400

(JMIR Med Educ 2025;11:e73698) doi:10.2196/73698

KEYWORDS

ChatGPT; Gemini; DeepSeek; medical education; AI; artificial intelligence; Saudi Arabia; perceptions; medical students; faculty; LLM; chatbot; qualitative study; thematic analysis; satisfaction; RAG retrieval-augmented generation

We appreciate the thoughtful critique of our manuscript "Perceptions and earliest experiences of medical students and faculty with ChatGPT in medical education: qualitative study" [1] by Zhao and Zhang [2]. Concerns over the generation of hallucinated citations by large language models (LLMs), such as OpenAI's ChatGPT, Google's Gemini, and Hangzhou's DeepSeek, warrant exploring advanced and novel methodologies to ensure citation accuracy and overall output integrity [3].

The LLMs have demonstrated a propensity to generate well‐formatted yet fictitious references—a limitation largely attributed to restricted access to subscription-based databases and their reliance on probabilistic text generation [4]. As LLMs evolve, future iterations may integrate more reliable retrieval-based architectures, enhancing their capacity to cite legitimate sources while reducing fabricated references [4,5]. However, until such improvements are systematically validated, scholars must remain cautious.

One suggested enhancement is using retrieval-augmented generation (RAG) [6]. This approach integrates up-to-date external information, substantially improving real-world applicability. However, even RAG-based systems can misinterpret or distort source content under high-trust conditions. To address this, the authors developed Hallucination-Aware Tuning (HAT) [6]. HAT trains dedicated detection models to generate labels and detailed descriptions of identified hallucinations. These descriptions are then used by GPT-4 to correct discrepancies. The combination of corrected and original outputs forms a preference dataset that, when used for Direct Preference Optimization training, yields LLMs with reduced hallucination rates and improved answer quality [6].

We also propose another solution aimed at fundamentally reducing citation errors: the development of "Reference-Accurate" academic LLM by major global publishers. Leading journals could develop their own specialized LLM, trained exclusively on rigorously verified academic literature from robust databases. This targeted training would ensure that every generated reference is accurate and directly traceable to published work. Ideally, these publisher-backed LLMs would be made freely available to promote open science.

Therefore, we recommend a dual approach that combines advanced RAG methodologies with publisher-developed academic LLMs. Comparative studies should be conducted to evaluate the citation accuracy, factual consistency, and overall performance of RAG-HAT-tuned models against these publisher-specific models. Collaborative efforts among academic institutions, publishers, and AI developers are essential to establish standardized protocols and reliable training datasets. Such partnerships would not only enhance the reliability of LLM-generated outputs but also foster greater trust in AI-assisted scholarly communication.

Moreover, the broader academic community bears responsibility for critically appraising AI-generated content. While LLMs can streamline information retrieval and synthesis, human oversight remains indispensable for safeguarding academic integrity. Rather than dismissing AI-driven tools due to their current flaws, we advocate for further research to ensure greater alignment

with evidence-based scholarship and authentic publications. Future LLM iterations may rapidly overcome these limitations, but until then, transparency, responsible usage, and ongoing improvements in AI training remain imperative.

In conclusion, while RAG augmented by HAT represents a potential advancement in reducing hallucinations, the development of specialized, reference-accurate academic LLMs by publishers may offer a promising pathway. By integrating both strategies and ensuring human oversight, the academic community can ensure that AI-driven tools reliably support the rigor and transparency essential to scholarly research.

Conflicts of Interest

None declared.

References

1. Abouammoh N, Alhasan K, Aljamaan F, et al. Perceptions and earliest experiences of medical students and faculty with ChatGPT in medical education: qualitative study. JMIR Med Educ 2025 Feb 20;11:e63400. [doi: 10.2196/63400] [Medline: 39977012]
2. Zhang M, Zhao T. Citation accuracy challenges posed by large language models. JMIR Med Educ 2025 [FREE Full text] [doi: 10.2196/72998]
3. Temsah A, Alhasan K, Altamimi I, et al. DeepSeek in healthcare: revealing opportunities and steering challenges of a new open-source artificial intelligence frontier. Cureus 2025 Feb;17(2):e79221. [doi: 10.7759/cureus.79221] [Medline: 39974299]
4. Aljamaan F, Temsah MH, Altamimi I, et al. Reference hallucination score for medical artificial intelligence chatbots: development and usability study. JMIR Med Inform 2024 Jul 31;12:e54345. [doi: 10.2196/54345] [Medline: 39083799]
5. Howard A, Hope W, Gerada A. ChatGPT and antimicrobial advice: the end of the consulting infection doctor? Lancet Infect Dis 2023 Apr;23(4):405-406. [doi: 10.1016/S1473-3099(23)00113-5] [Medline: 36822213]
6. Song J, Wang X, Zhu J, et al. RAG-HAT: a hallucination-aware tuning pipeline for LLM in retrieval-augmented generation. 2024 Presented at: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; Miami, Florida, US. [doi: 10.18653/v1/2024.emnlp-industry.113]

Abbreviations

HAT: Hallucination-Aware Tuning
LLM: large language model
RAG: retrieval-augmented generation

Edited by S Nedunchezhiyan; submitted 10.03.25; this is a non–peer-reviewed article;accepted 12.03.25; published 02.04.25.

Please cite as:
Temsah MH, Al-Eyadhy A, Jamal A, Alhasan K, Malki KH
Authors' Reply: Citation Accuracy Challenges Posed by Large Language Models
JMIR Med Educ 2025;11:e73698
URL: https://mededu.jmir.org/2025/1/e73698
doi:10.2196/73698

Guidelines for Rapport-Building in Telehealth Videoconferencing: Interprofessional e-Delphi Study

Paula D Koppel[1*], PhD, RNC, GNP-BC, AHN-BC, NBC-HWC; Jennie C De Gagne[1*], PhD, DNP, RN, NPD-BC, CNE; Michelle Webb[1*], DNP, RN, CHPCA; Denise M Nepveux[2*], PhD, OTR/L; Janelle Bludorn[2,3*], MS, PA-C; Aviva Emmons[4*], BSN, BA, RN, AMB-BC; Paige S Randall[1*], PhD, RN, CNE; Neil S Prose[5*], MD

[1]School of Nursing, Duke University, DUMC 3322, 307 Trent Drive, Durham, NC, United States

[2]School of Medicine, Duke University, Durham, NC, United States

[3]Duke Center for Interprofessional Education and Care, Duke University, Durham, NC, United States

[4]Duke Cancer Institute, Durham, NC, United States

[5]Pediatrics and Dermatology, Global Health Institute, Duke University, Durham, NC, United States

[*]all authors contributed equally

Corresponding Author:
Paula D Koppel, PhD, RNC, GNP-BC, AHN-BC, NBC-HWC
School of Nursing, Duke University, DUMC 3322, 307 Trent Drive, Durham, NC, United States

Abstract

Background: Telehealth training is increasingly incorporated into educational programs for health professions students and practicing clinicians. However, existing competencies and standards primarily address videoconferencing visit logistics, diagnostic modifications, and etiquette, often lacking comprehensive guidance on adapting interpersonal skills to convey empathy, cultural humility, and trust in web-based settings.

Objective: This study aimed to establish consensus on the knowledge, skills, and attitudes required for health professions students and clinicians to build rapport with patients in telehealth videoconferencing visits and to identify teaching strategies that best support these educational goals.

Methods: An e-Delphi study was conducted using a panel of 12 interprofessional experts in telehealth and telehealth education. Round 1 involved interviews, followed by anonymous surveys in rounds 2 - 4 to build consensus.

Results: All 12 experts participated in rounds 1 - 3. In total, 19 themes related to rapport-building and 77 specific curriculum items were identified, all achieving the established level of consensus.

Conclusions: Using a competency-based education framework, this study provides guidance for health professions educators, teaching clinicians, and students on how to adapt interpersonal skills for telehealth including detailed content related to knowledge, skills, attitudes, and teaching strategies. Future research is needed to test the feasibility, acceptability, and effectiveness of curricula based on these competencies and teaching strategies.

Trial Registration: OSF Registries tjmkx; https://osf.io/tjmkx

(JMIR Med Educ 2025;11:e76260) doi:10.2196/76260

KEYWORDS

health professions education; health care professional development; clinician-patient relationship; interpersonal skills; health care communication; empathy; web-based care

Introduction

Background

Telehealth, including the use of videoconferencing visits (VV), is quickly becoming an important and common vehicle for delivering health care [1,2]. Growth in the use of VV is anticipated, given that clients or patients (hereafter referred to collectively as patients) and clinicians are increasingly receptive to VV as a supplement to in-person care [2].

Although telehealth competencies and training are being incorporated into some health professions educational programs [3-7], a recent scoping review found that 53% of sources emphasized a need for additional education and practice in telehealth etiquette and interpersonal aspects of care [8]. Many health care professionals remain uncertain of ways to establish rapport during VV [9-11]. A mixed methods study by Elsevier Health that included 3000 nurses and doctors found that over half believed telehealth would negatively impact their ability to demonstrate empathy and requested further guidance on

interpersonal skills in VV [12]. Studies show that not all interpersonal skills are interchangeable between in-person and VV [2,13] and that adapting skills for VV often does not come naturally but can be learned [14,15].

Definition of Rapport and Importance

Rapport, a term often used colloquially, has been defined within the health sciences and psychology as a desirable state of interpersonal connection that requires positivity, mutual responsiveness, and behavioral coordination [16,17] as well as respect, acceptance, empathy, and a mutual commitment to engage [18]. When intentionally fostered, rapport facilitates the development of a therapeutic relationship with outcomes that include trust, improved care outcomes, and satisfaction for both the patient and clinician [10, 13, 19-23]. Connection or rapport within a web-based care environment has only begun to be explored. These studies suggest that many in-person relationship-based skills are effective, but adaptations are necessary and must be used intentionally by clinicians and patients [9,10,20,24].

Knowledge Gap and Need for Research

Recent evidence-based telehealth etiquette and interpersonal checklists [25,26], along with educational interventions for health care students [14] and professional development workshops for clinicians [15,27], show promise in preparing current and future clinicians to adapt their interpersonal skills. However, telehealth etiquette often emphasizes practical and professional behaviors for efficient VV (eg, technical preparedness, ensuring a professional environment, and communication techniques to ensure diagnostic and treatment accuracy) rather than adapting interpersonal skills to convey empathy, cultural humility, and trust or to build personal connection in a web-based environment. Additionally, evidence-based resources are needed for interprofessional educators to support curriculum development across health care disciplines [7].

Building rapport in VV requires specific knowledge, skills, and attitudes as well as teaching strategies designed to facilitate relationship-based care competencies. Building upon previous descriptive studies [9,10,13,20,28-31], this project aimed to use an e-Delphi study to explore what relationship-based care elements are important to include in a telehealth curriculum. This will provide a strong foundation for the future curricula development of interprofessional students and clinicians across disciplines while closing a crucial gap in our practice and pedagogical knowledge.

Research Aim and Questions

This study aimed to establish expert consensus on training health professions students and clinicians in rapport-building during VV. Two research questions guided the study: (1) what knowledge, skills, and attitudes are essential for health professions students and clinicians to build rapport with patients in telehealth VV? (2) What teaching strategies best facilitate the development of this knowledge and these skills and attitudes?

Methods

Study Design

An e-Delphi methodology was used to build consensus by collecting, analyzing, and sharing data results with an interprofessional panel of experts in telehealth and telehealth education. The Delphi technique, widely used in nursing and health care research, assumes that group opinion is more valid and reliable than individual opinion, making it effective for achieving consensus on topics with limited evidence or uncertain practice standards [32]. This iterative, structured method uses interviews and surveys to gather expert input over multiple rounds until consensus is reached. Experts remain anonymous to one another to prevent any individual from dominating the process. After each survey round, data are summarized and shared with experts, reflecting both their individual responses and the group's collective responses. This feedback allows experts to revise their input based on group insights. Rounds typically continue, often between 2 and 4 times, until an acceptable level of consensus is achieved [32]. The web-based Delphi technique (e-Delphi), increasingly adopted for its cost-effectiveness, facilitates efficient participation from geographically diverse experts [33].

Recruitment and Selection of Participants

An interprofessional panel of national and international experts was recruited to participate in this e-Delphi study. The panel included licensed nursing and medical professionals (registered nurses, medical doctors, and advanced practice providers), as well as speech and occupational therapists, all proficient in reading, writing, and conversing in English. Participants were required to have experience in at least 1 of the following activities: (1) conducted research on VV, (2) participated in telehealth advisory panels, (3) developed telehealth curricula, or (4) provided extensive instruction in telehealth, with a significant focus on adapting relational and communication skills for telehealth care.

Our interprofessional research team initially identified 21 experts meeting these criteria. These experts were invited to recommend additional participants, resulting in 7 further suggestions. Recruitment occurred over 3 rounds of targeted emails to assemble a diverse panel of telehealth experts with broad perspectives and experiences. To encourage participation across all rounds of the study, participants received a small financial incentive for their time upon completion.

Ethical Considerations

This educational research was deemed to pose minimal risk to the expert participants, as defined under the Common Rule; therefore, written consent was not conducted. A statement in the initial questionnaire clarified that participants' willingness to complete the interview and surveys constituted their verbal consent. The study received an exempt determination from the health system's institutional review board for clinical investigations (protocol ID: Pro00117125) prior to beginning the study. To maintain participant privacy, data were deidentified, and all study communications, interviews, and study data were conducted and stored on a secure

password-protected server with firewall protection and multifactor authentication. Additionally, the study was registered in the Open Science Framework on January 22, 2025 [34].

Data Collection Procedures

Four rounds of data collection were planned to build adequate consensus among the experts. In round 1, interviews were conducted with experts individually on a university-approved secure Microsoft Teams videoconferencing platform using an interview guide with semistructured open-ended questions developed by the research team. This guide, informed by the 2 original research questions, focused on the knowledge, skills, attitudes, and teaching strategies essential for building rapport in VV. Interviews were chosen as a method for the initial round to gather qualitative descriptive data based on the experts' experiences, given the limited published research on rapport-building in VV.

The specific knowledge, skills, attitudes, and teaching strategies identified using qualitative content analysis of the interview data in round 1 were compiled into a Qualtrics survey for round 2. The round 2 survey asked experts to prioritize the knowledge, skills, attitudes, and teaching strategies using a 4-point Likert scale by asking the extent of importance (4=very important, 3=important, 2=of some importance, and 1=not important). A 4-point scale was selected to enhance survey completion and encourage participants to choose between an item being important or unimportant [35]. Contextually, this aligns with an educator's need to decide whether to include or exclude content from their curriculum. Open-ended questions allowed the participants to provide comments related to their decisions and describe any need to revise or expand the content.

The round 3 survey included the same questions as round 2 and was designed to show experts their ratings compared to those of the other experts in the study. The experts were given an opportunity to rerate each of their original responses or keep them the same. Revision of round 2 items or recommendations for new items from open-ended questions in round 2 were also incorporated into the round 3 survey. Member checking using a final questionnaire was planned in round 4 to evaluate if the panel felt the results reflected their expertise and recommendations. All Qualtrics surveys were sent to participants via the institution's secure Microsoft Teams platform.

Expert participants received emails with survey links for all rounds of the study, including a questionnaire to collect their relevant professional background details prior to the first-round interview. Surveys were developed and administered using Qualtrics, a secure university-approved web-based survey platform.

Data Analysis and Statistical Considerations

Descriptive statistics were selected to report panel responses for each survey item (mean, SD, IQR, and consensus level) and attrition rates over the course of rounds. Since building a curriculum necessitates making a dichotomous decision to include or exclude topics or teaching strategies, for the purposes of calculating consensus, responses were collapsed into 2 categories: important (items rated very important or important) and unimportant (items rated less important or unimportant). The percentage of agreement between the experts was used to measure the level of consensus. Although curricula to teach rapport-building in VV may not require complete agreement, rapport is an important indicator of quality care; therefore, consensus was considered sufficient when an item was rated important or very important by at least 70% of the experts on the panel [32,36]. Statistical analysis was supported by using Microsoft Copilot within Microsoft Excel, leveraging its artificial intelligence (AI)–driven capabilities to generate insights and summaries.

A 5-member coding team conducted qualitative content analysis of the narrative interview data from round 1 and the open-ended responses for all later survey rounds. Codes were derived directly from the transcribed text data and kept close to the participants' descriptions [37], and the analysis was guided by a process outlined by Elo and Kyngäs [38]. NVivo software (version 14.0; QSR International Pty Ltd) was used to facilitate the qualitative analysis, including coding and development of a codebook. At least 20% of the data were reviewed by 2 or more team members, with conflicts resolved through discussion. After initial manual coding, the team used Copilot to generate preliminary summaries and thematic overviews of the interview data. Testing and refining prompts ensured that results aligned with the manually generated NVivo codebook. The study's rigor was enhanced by (1) team-based analysis by researchers with extensive qualitative experience; (2) judicious use of AI tools to validate findings [39]; (3) regular coding meetings to refine codes, categories, themes, and survey items [40]; (4) concurrent data collection and analysis [41]; (5) detailed memos to maintain an audit trail of analytical decisions [39]; and (6) member checking by asking participants to clarify interview responses, evaluate survey items, and reflect on the final analysis [42]. The ACCORD (Accurate Consensus Reporting Document) guided the reporting of results [43].

Results

Overview

In total, 12 of the 16 interprofessional telehealth experts emailed an invitation consented to participate in the study, representing a 75% response rate. Interviews were conducted with the panel of experts in January 2025. This was followed by 3 survey rounds from February to April 2025. All 12 experts completed the round 1 interview and the round 2 and round 3 surveys, representing a 100% participation rate. A total of 11 of the 12 (92%) experts participated in round 4 member checking.

Panel Characteristics and Expertise

Most participants had doctoral degrees (8/12, 67%) and more than 20 years of professional experience (8/12, 67%). A third of the experts had 11 or more years of experience using telehealth with a deep range of involvement in practice, education, research, and advisory capacities (see Table 1 for further characteristics). Of the 12 experts, 2 (17%) were international experts, with the balance residing in the United States. The panel's expertise included using telehealth in a variety of contexts, including adult and pediatric care, rural settings, palliative and hospice care, and mental health. Members of the expert panel have contributed to the development of

practice standards and educational curricula, established telehealth clinics, taught educational courses, conducted research, and published journal papers and books on telehealth. In addition, several experts in the study have developed conceptual models and conducted research specifically on relational aspects of telehealth. All participants were actively involved in teaching professional and interprofessional educational programs with a focus on telehealth.

Table . Expert participant characteristics (N=12).

Characteristic		Values, n (%)
Education level		
	Master's degree	2 (17)
	Doctoral degree	8 (67)
	Professional degree (MD)	2 (17)
Licensure		
	Nursing or midwifery	7 (58)
	Medicine	2 (17)
	Speech-language pathology	1 (8)
	Physician assistant	2 (17)
Length of professional experience (years)		
	11 - 12	4 (33)
	21 - 30	3 (25)
	31 - 40	4 (33)
	>40	1 (8)
Professional experiences with telehealth		
	Conducted visits or consultations	9 (75)
	Conducted research	10 (83)
	Participated on advisory panels	8 (67)
	Developed curriculum	8 (67)
	Provided instruction focused on rapport	11 (92)
Length of telehealth experience (years)		
	2 - 4	4 (33)
	5 - 10	4 (33)
	11 - 20	4 (33)

First Round Results

Overview

Data collected and coded from the 12 expert interviews were organized into educational areas of knowledge, skills, attitudes, and teaching strategies as directed by the research question. Within each of these areas, themes emerged that helped organize more specific items that the experts recommended for a curriculum focused on building rapport in VV. The themes and specific items related to each topic are described below. Since these educational areas, themes, and specific items were incorporated into the round 2 survey, they can also be reviewed in their entirety in Tables 2-5, along with their descriptive statistics and consensus ratings.

Table . Education areas within the knowledge domain (N=12).

Theme and item		Consensus[a], n (%)	Mean (SD)[b]	Median (IQR)
Basic knowledge of telehealth and its applications				
	Telehealth taxonomy	9 (75)	2.83 (0.835)	3 (2.25-3)
	Evidence-based uses of video visits	12 (100)	3.83 (0.389)	4 (4-4)
	Types of video visits (eg, urgent care, follow-up care, and medical management)	12 (100)	3.50 (0.522)	3.5 (3-4)
	Types of clinician roles and responsibilities (eg, clinical evaluation, preoperative education, and facilitating support group)	11 (92)	3.33 (0.651)	3 (3-4)
	Situations not appropriate for video visits	12 (100)	3.75 (0.452)	4 (3.25-4)
	Common challenges in video visits	12 (100)	3.92 (0.289)	4 (4-4)
	General adaptations necessary	12 (100)	3.58 (0.515)	4 (3-4)
How to create a web-based environment that supports rapport				
	Ground rules for video visits (eg, what to expect, what to do if internet connection lost, and ways to enhance the quality of the visit)	12 (100)	3.83 (0.389)	4 (4-4)
	Awareness of how generational, cultural, educational, geographic, and socioeconomic factors influence a patient's participation in video visits	12 (100)	3.75 (0.452)	4 (3.25-4)
	Importance of privacy (auditory or visual) and confidentiality in the patient's and clinician's spaces (eg, private spaces and data protection)	12 (100)	3.92 (0.289)	4 (4-4)
	Choice of video visit setting or background (eg, quiet and nondistracting)	12 (100)	3.50 (0.522)	3.5 (3-4)
	Adaptations related to patient population or type of visit (eg, children, non-English speakers, team-based consultation, mental health evaluation, and group-based care)	12 (100)	3.75 (0.452)	4 (3.25-4)
Functional use of videoconferencing technology platforms				
	Understanding of videoconferencing platform functions and set-up	12 (100)	3.33 (0.492)	3 (3-4)
	Problem-solving of most common technical problems	12 (100)	3.25 (0.452)	3 (3-3.75)
Rapport basics				

Theme and item		Consensus[a], n (%)	Mean (SD)[b]	Median (IQR)
	Awareness of attributes of rapport (eg, shared experience of comfort and engagement and being "in sync")	12 (100)	3.83 (0.389)	4 (4-4)
	Awareness of the outcomes of rapport for patients and clinicians (eg, trust and collaboration)	12 (100)	3.83 (0.389)	4 (4-4)
	Basic rapport-building strategies (eg, active listening and showing genuine interest, respect, and empathy)	12 (100)	4.00 (0.000)	4 (4-4)
	Importance of patients feeling heard and understood	12 (100)	4.00 (0.000)	4 (4-4)
	Increased importance of facial expression and body mannerisms in video visits	12 (100)	3.67 (0.492)	4 (3-4)
	Indicators that rapport has been established (eg, shared smiles, laughter, warmth, and sense of connection)	12 (100)	3.83 (0.389)	4 (4-4)
	Unique opportunities to cultivate rapport in video visits	12 (100)	3.67 (0.492)	4 (3-4)
	Unique rapport challenges in video visits (eg, internet lag interfering with the ability to feel "in sync" and visits becoming brief and transactional)	12 (100)	3.83 (0.389)	4 (4-4)

[a]This describes the percentage of experts who scored the item as either "important" or "very important."

[b]Group mean Likert score. Note that each specific item within the survey was rated as 4=very important, 3=important, 2=of less importance, or 1=unimportant.

Table . Education areas within the skills domain (N=12).

Theme and item	Consensus[a], n (%)	Mean (SD)[b]	Median (IQR)
Creating a safe and comfortable web-based environment			
Supporting patient's comfort with technology	11 (92)	3.58 (0.669)	4 (3-4)
Managing technical challenges	11 (92)	3.33 (0.651)	3 (3-4)
Ensuring patient's desired level of privacy during the video visit	12 (100)	3.83 (0.389)	4 (4-4)
Navigating technological and other video visit distractions to maintain rapport	12 (100)	3.67 (0.482)	4 (3-4)
Taking time to build rapport at the beginning of the video visit	12 (100)	3.92 (0.289)	4 (4-4)
Monitoring level of rapport throughout the video visit	12 (100)	3.50 (0.522)	3.5 (3-4)
Navigating the patient's wants and needs within video visit parameters with mutually agreeable care goals	12 (100)	3.83 (0.389)	4 (4-4)
Demonstrate knowing the patient with recall of important information (eg, medical and social)	12 (100)	3.58 (0.515)	4 (3-4)
Demonstrating cultural humility (eg, respecting differences and practicing self-awareness)	12 (100)	3.67 (0.492)	4 (3-4)
Managing time and flow so patients do not feel rushed	12 (100)	3.50 (0.522)	3.5 (3-4)
Avoiding abruptly ending visits, ensuring that patients' questions and needs are addressed	12 (100)	3.92 (0.289)	4 (4-4)
Navigating moments of disconnection (eg, anger, disagreement, and disengagement)	12 (100)	3.75 (0.452)	4 (3.25-4)
Adapting verbal communication techniques to facilitate rapport in video visits			
Keeping communication clear and simple	12 (100)	3.67 (0.492)	4 (3-4)
Managing conversation flow and pacing to reduce interruptions and "talk overs"	12 (100)	3.75 (0.452)	4 (3.25-4)
Using words as a replacement for physical touch to acknowledge emotions and demonstrate support (eg, "I wish I could give you a hug")	11 (92)	3.58 (0.669)	4 (3-4)

Theme and item		Consensus[a], n (%)	Mean (SD)[b]	Median (IQR)
	Increasing use of reflective practices (eg, restating what the patient has said, asking more questions, and using teach-backs) to validate or confirm accurate understanding	12 (100)	3.83 (0.389)	4 (4-4)
	"Narrating the visit" to describe what you are doing that might be difficult for the patient to interpret (eg, looking at your laboratory work and writing orders for medication)	12 (100)	3.75 (0.452)	4 (3.25-4)
Enhancing nonverbal communication techniques to facilitate rapport in video visits				
	Using body language (eg, facial expressions, hand gestures, and body posturing) to express feelings, emphasize suggestions, or show intentions (eg, eye contact, nodding, and leaning in to show attentiveness, smiling, and hand over heart to show empathy)	12 (100)	4.00 (0.000)	4 (4-4)
	Using platform technology to keep the patient engaged and enhance learning or understanding (eg, screen sharing images, placing links to information in chat, and "show and then tell")	11 (92)	3.50 (0.674)	4 (3-4)
Maintaining attentiveness and presence				
	Pausing or breaking between video visits to reset and cultivate presence	10 (83)	3.17 (0.718)	3 (3-4)
	Managing personal distractions (eg, phone and computer alerts)	12 (100)	3.58 (0.515)	4 (3-4)
	Setting up technology and lighting to ensure that both the clinician and the patient can view the other's body language (eg, "passport view")	12 (100)	3.58 (0.515)	4 (3-4)
	Setting up technology to enhance eye contact with the patient during documentation and medical record review (eg, close to the computer camera)	12 (100)	3.75 (0.452)	4 (3.25-4)
	Heightening attentiveness to visual and auditory signals that might reflect patient emotion	12 (100)	3.58 (0.516)	4 (3-4)
Using a web-based environment to humanize the care experience				
	Seeking cues in the patient's environment to know them as a person	10 (83)	3.25 (0.754)	3 (3-4)

Theme and item		Consensus[a], n (%)	Mean (SD)[b]	Median (IQR)
	Using cues in the patient's environment to build connection	10 (83)	3.25 (0.754)	3 (3-4)
	Using self-disclosure based on cues in the patient's environment to build connection	9 (75)	2.92 (0.900)	3 (2.25-3.75)
	Personalizing your professional web-based environment to facilitate connection	9 (75)	3.00 (0.739)	3 (2.25-3.75)
	Using idiosyncratic aspects of video visits as opportunities to build connection (eg, unexpected pet encounters or technology issues)	12 (100)	3.25 (0.452)	3 (3-3.75)
Other				
	Magnifying use of active listening techniques	12 (100)	3.83 (0.389)	4 (4-4)
	Heightening self-awareness to ensure that your verbal and nonverbal behaviors reflect empathy and caring	12 (100)	3.92 (0.289)	4 (4-4)

[a]This describes the percentage of experts who scored the item as either "important" or "very important."

[b]Group mean Likert score. Note that each specific item within the survey was rated as 4=very important, 3=important, 2=of less importance, or 1=unimportant.

Table . Education areas within the attitudes domain (N=12).

Theme and item		Consensus n (%)[a]	Mean (SD)[b]	Median (IQR)
Willingness to adapt to changing situations with ease				
	Willingness to make extra effort to connect in video visits	12 (100)	3.67 (0.492)	4 (3-4)
	Patience with unexpected situations and technological challenges	12 (100)	3.67 (0.492)	4 (3-4)
Respect for patient challenges associated with video visits				
	Respect associated with being in someone's home digitally	12 (100)	3.83 (0.389)	4 (4-4)
	Respecting patient's perception of whether video visits can meet their needs and goals	12 (100)	3.67 (0.492)	4 (3-4)
Other				
	Valuing video visits as a viable care delivery model	12 (100)	3.75 (0.452)	4 (3.25-4)
	Receptivity to learning new skills for video visits	12 (100)	3.67 (0.492)	4 (3-4)
	Valuing relational aspects of care with a desire to adapt interpersonal skills for video visits	12 (100)	3.75 (0.452)	4 (3.25-4)

[a]This describes the percentage of experts who scored the item as either "important" or "very important."

[b] Group mean Likert score. Note that each specific item within the survey was rated as 4=very important, 3=important, 2=of less importance, or 1=unimportant.

Table . Teaching strategies (N=12).

Theme and item		Consensus n (%)[a]	Mean (SD)[b]	Median (IQR)
Experiential learning methods				
	Simulation with standardized patients	11 (92)	3.58 (0.669)	4 (3-4)
	Role-playing with other learners	11 (92)	3.50 (0.674)	4 (3-4)
Real-time feedback				
	Real-time feedback from real and standardized patients	12 (100)	3.75 (0.452)	4 (3.25-4)
	Real-time feedback from educators, colleagues, and mentors	12 (100)	3.75 (0.452)	4 (3.25-4)
	Real-time feedback from artificial intelligence technology (eg, avatars and natural language processing tools)	9 (75)	3.00 (0.953)	3 (2.25-4)
Self-paced learning				
	Asynchronous learning (eg, recorded presentations)	10 (83)	2.92 (0.793)	3 (3-3)
	Videos demonstrating effective and ineffective rapport-building skills	11 (92)	3.50 (0.674)	4 (3-4)
	Module-based learning	9 (75)	3.08 (0.793)	3 (2.25-4)
Small group methods				
	Case studies	11 (92)	3.33 (0.888)	3.5 (3-4)
	Discussion groups	10 (83)	3.00 (0.853)	3 (3-3.75)
	Lecture with breakout practice sessions	10 (83)	3.08 (0.900)	3 (3-4)
Affective learning				
	Opportunities for self-reflection	10 (83)	3.33 (0.985)	4 (3-4)
	Role-playing that includes having the clinician be a patient	10 (83)	3.33 (0.779)	3.5 (3-4)
Other				
	Peer learning (eg, observing, interacting with preceptors, mentors, and role models)	12 (100)	3.58 (0.515)	4 (3-4)
	Integration of telehealth adaptations into standard curricula (ie, when learning about developing in-person rapport, students would also learn about appropriate adaptations for telehealth)	12 (100)	3.92 (0.289)	4 (4-4)

[a]This describes the percentage of experts who scored the item as either “important” or “very important.”

[b] Group mean Likert score. Note each specific item within the survey was rated as 4=very important, 3=important, 2=of less importance, or 1=unimportant.

Knowledge

Expert participants described the importance of ensuring that students and clinicians had basic knowledge of telehealth. This included the evidence-based uses of VV, types of VV, roles and responsibilities of clinicians, situations not appropriate for VV, common challenges, and required adaptations. Providing information to ensure a functional understanding of videoconferencing technology and how to address common technical problems was also identified as basic telehealth

knowledge, as was instruction on specific ways to create a web-based environment supporting rapport. This included creating ground rules for VV and understanding how generational, cultural, educational, geographic, and socioeconomic factors influence a patient's participation in VV. The importance of ensuring privacy in not only the patient's but also the clinician's environment was emphasized as essential to developing rapport and trust. Finally, the experts recommended that basic knowledge of the antecedents (eg, feeling safe and comfortable, privacy, and attentiveness), attributes (eg, shared positive experiences, shared smiles, engagement, and sense of connection), and outcomes (eg, trust, confidence, and collaboration) of rapport were important to teach, along with the unique opportunities and challenges associated with rapport-building in VV.

Skills

The data highlighted essential skills for students and clinicians, including creating a safe and comfortable environment, adapting verbal and nonverbal communication techniques for VV, maintaining attentiveness and presence, and using the web-based environment to humanize the care experience. Experts provided detailed recommendations for each category that helped enrich the survey items developed in round 2. For example, they emphasized the importance of establishing rapport at the beginning of a video visit and monitoring it throughout. Managing the time and flow of the VV to avoid rushing patients and preventing abrupt endings were identified as critical skills. Demonstrating cultural humility and navigating moments of interpersonal disconnection were also identified as crucial for maintaining rapport.

Experts highlighted the need to pace verbal communication to reduce interruptions or "talk overs" and to "narrate the visit" by explaining the clinician's actions that might be unclear or not visible to patients. Using words to replace physical touch to acknowledge emotions and demonstrate care was emphasized. Enhancing active listening techniques, including body language, to reflect empathy and caring was described as particularly important. Examples included placing a hand over the heart to show empathy or leaning into the camera to demonstrate attentiveness.

Participants also stressed the importance of encouraging attentiveness by setting up technology to improve eye contact, sense of presence, and visualization of emotions. Finally, experts noted opportunities within a VV to humanize the experience and build rapport by seeking cues in the patient's environment to know them as a person and identify common interests. Using unique aspects of a VV, such as unexpected pet encounters or technology issues (eg, sharing a laugh about challenges of technology rather than focusing on frustration), to build interpersonal connection was also described as useful.

Attitudes

The importance of cultivating affirming attitudes was mentioned throughout the interview data. This included a willingness to adapt to changing situations with ease, such as making extra efforts to connect in VV and having patience with unexpected situations and technological challenges. Respect associated with entering a patient's home environment digitally was highlighted, as well as respecting the patient's perception of whether a VV could meet their needs. Other attitudes that the experts felt needed to be nurtured included valuing VV as a viable care model and adapting relational aspects of care for VV. Being receptive and excited to learn new skills for VV was also emphasized as an important attitude for educators to foster.

Teaching Strategies

The importance of integrating telehealth adaptations into the standard educational curriculum was frequently discussed in the interviews, and a variety of teaching strategies were recommended. Methods supporting experiential and affective learning, such as simulation with standardized patients and opportunities for self-reflection, were identified. Self-paced modules, videos demonstrating effective and ineffective rapport-building skills, and small group methods like case studies and discussion groups were identified as useful strategies. The opportunity to receive real-time feedback and learn by observing and interacting with preceptors, mentors, and role models was emphasized.

Second Round Results

All 75 items that emerged from the qualitative content analysis in round 1 were presented to the experts in a survey format. Participants rated each item's level of importance from a Likert scale score of 4=very important to 1=not important. The percentage of agreement, the group's mean Likert score, SD, and IQR for each item follow and are organized by the educational areas: knowledge (Table 2), skills (Table 3), attitudes (Table 4), and teaching strategies (Table 5). All 75 items presented to the experts for rating reached the a priori level of at least 70% consensus in round 2. In total, 71% (53/75) of the items achieved 100% consensus, indicating that all 12 experts felt that these items should be included in a curriculum focused on teaching rapport-building in VV. There were very high levels of consensus among the items related to knowledge and attitude. A total of 20 of 22 (91%) of the items in the knowledge domain had 100% consensus, and all 7 items related to attitudes had 100% consensus. In total, 4 of 15 teaching strategy items reached 100% consensus. Table 6 illustrates the strong levels of consensus achieved in round 2.

Table . Level of consensus by educational area and teaching strategy.

Educational areas (total items)	Items with 100% consensus, n (%)	Items with 92% consensus, n (%)	Items with 83% consensus, n (%)	Items with 75% consensus, n (%)
Knowledge (n=22)	20 (91)	1 (5)	0 (0)	1 (5)
Skills (n=31)	22 (71)	4 (13)	3 (10)	2 (7)
Attitudes (n=7)	7 (100)	0 (0)	0 (0)	0 (0)
Teaching strategies (n=15)	4 (27)	4 (27)	5 (33)	2 (13)

The group's mean Likert scores ranged from 2.83 to 4.00 (SD 0-0.985) for each item. The high mean of most of the items also reflects strong agreement on the items evaluated in round 2. Although there were no disputed items, 5 items only reached a 75% consensus level (Table 6). The high levels of consensus resulted in most items having relatively low SD and IQR ratings.

Three items (taxonomy, asynchronous learning, and discussion groups) in particular had lower means, with higher SDs. The median of these items was 3 (IQR 2.25-3), 3 (IQR 3-3), and 3 (IQR 3-3.75), respectively. This suggests that a core group of the experts closely agreed, but a few participants felt differently about these 3 items. The median for real-time feedback with AI was 3 (IQR 2.25-4), and the median for modular learning was 3 (IQR 2.25-4). Both also had high SDs (Table 5). This suggests that these items were more complex or controversial with a broader range of opinions, even though they still met the consensus threshold.

Open-ended responses from several of the experts in this round suggested the need for 2 additional items. This included the importance of incorporating information on ways a clinician could adapt teaching methods in VV and providing opportunities for practicing VV that involved an interprofessional team.

Third Round Results

Since all 75 items in round 2 reached the established consensus threshold, round 3 focused on gathering feedback from the expert panel on the 2 new items developed from the open-ended survey responses in round 2. Consistent with round 2, these items were presented in a Qualtrics survey, with experts asked to rate their importance on a 4-point Likert scale. Both items met the consensus threshold. The first item, "importance of adapting teaching methods to patient learning preferences and experience with videoconferencing," had 100% (12/12) consensus, a mean Likert score of 3.58 (SD 0.52), and a median of 4 (IQR 3-4). The second item evaluated in round 3, "opportunities to practice VV involving an interprofessional team," achieved 92% (11/12) consensus, a mean Likert score of 3.17 (SD 0.58), and a median of 3 (IQR 3-3.75). Open-ended feedback in round 3 suggested distinguishing educational content on patient learning preferences from content on the patient experience with VV, as these are distinct topics. Experts also noted that some patient learning preferences may not be viable options in VV.

Fourth Round Results

The focus of round 4 was to share the study's overall results and provide the experts an opportunity to indicate if these accurately reflected their experiences and perspectives. The participants' responses affirmed the results, and as one expert shared, "Building rapport through video-mediated communication can be challenging, and the items highlighted tangible and effective steps to mitigate this." Some of the experts' responses helped explain areas where there was less than 100% consensus. One expert indicated that in the populations they worked with, patients may not be comfortable with clinicians stating, "I wish I could offer you a hug" or having clinicians humanize a VV by looking for things in their environment. These comments reflected the importance of cultural humility, understanding the context of the VV, and individual patient preferences and perspectives. One expert expressed surprise that having students role-play being a patient did not receive a higher level of consensus sharing, "Empathy is a muscle. If you don't consistently remind yourself what it feels like to be the patient, you are more likely to forget and deliver more callous, depersonalized care."

All the participants indicated that they felt the results could inform interprofessional practice and teaching. As one expert stated,

> *These results can be a great jumping-off point for developing more focused training and education around telehealth communication. Since there seems to be a general awareness of the importance of rapport, but a lack of clarity on the specific skills needed, there's an opportunity to design tools, workshops, or even curricula that help providers build those skills more intentionally.*

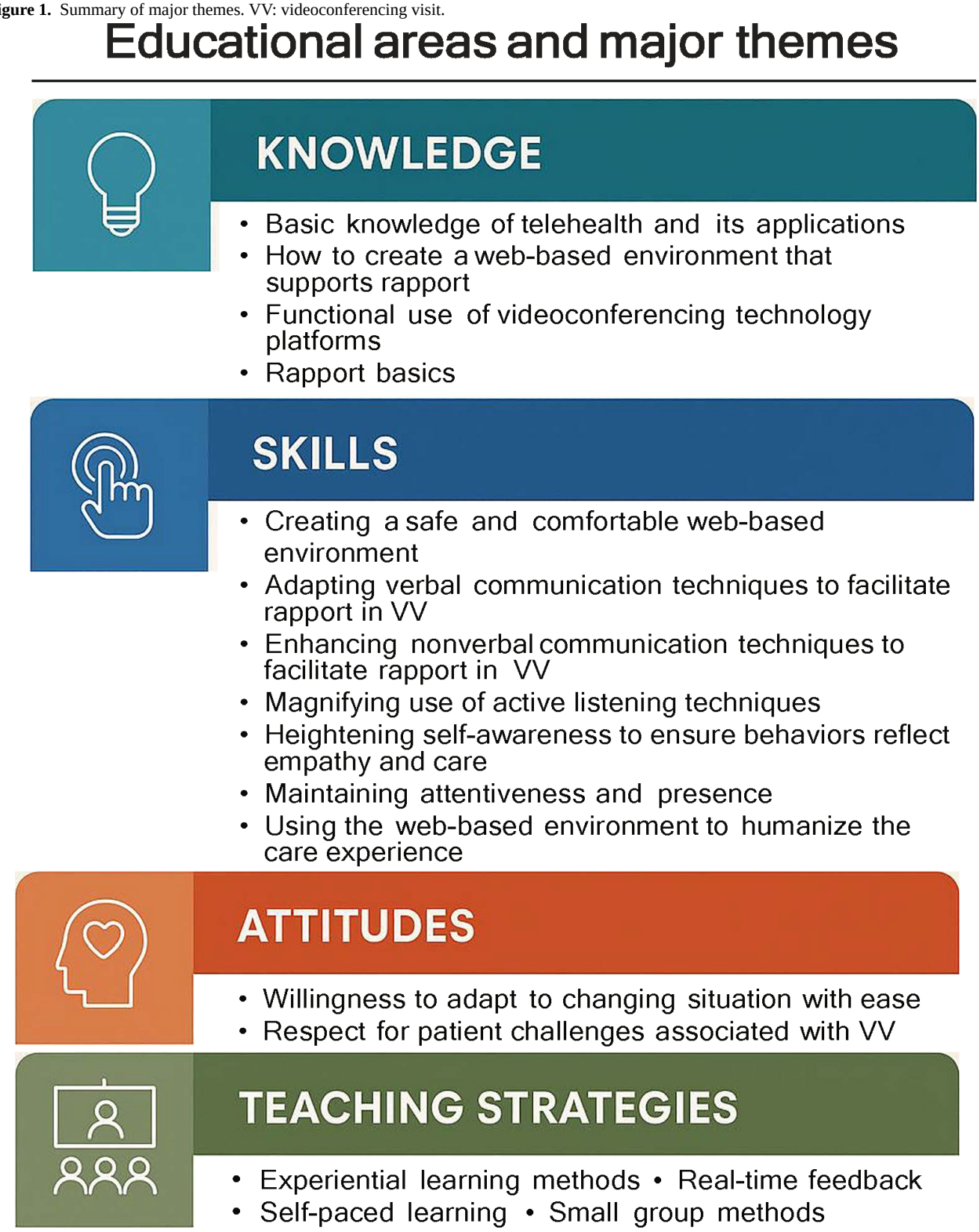
Another shared, "This could really push institutions to start thinking beyond the technology and more about the patient experience."

Discussion

Principal Findings

Teaching and learning how to establish and maintain rapport in VV were identified as important educational gaps. The aim of this study was to build consensus around the knowledge, skills, and attitudes required for health professions students and clinicians to build rapport with patients in VV as well as to identify what teaching strategies best support these educational objectives. This e-Delphi approach was an effective and efficient method for consolidating the expertise of 12 interprofessional telehealth clinicians and educational experts (Figure 1). The high retention rate throughout the study may indicate the importance and relevance of this specific topic to telehealth experts.

Figure 1. Summary of major themes. VV: videoconferencing visit.



The experts quickly agreed on the importance of 19 themes and 77 items to guide curriculum development. The depth and breadth of these themes and items demonstrate the importance of enhancing relationship skills beyond those described in etiquette guidelines or competencies. Although the number of themes and items could be perceived as excessive or impractical to incorporate into telehealth education, evidence that clinicians lack confidence in their relationship-based skills demonstrates the need for a more robust curriculum. The study results provide a comprehensive and useful reference for building a curriculum based on developing knowledge, skills, and attitudes. Like other complex competencies, these may need to be integrated throughout a health profession curriculum or offered in self-paced modules for professional development.

The number of themes and items related to adapting relationship-based care and communication skills for VV demonstrates the importance of this element in the curriculum and aligns with existing research [13,14,20,29]. The expert panel also emphasized the importance of learning to create interpersonal connections at the beginning of a VV and monitor the level of rapport throughout the visit. Additionally, not rushing or ending the VV abruptly and taking time to clearly understand the patient's wants and needs with mutually agreeable care goals were highlighted as important to building rapport. These suggestions align with telehealth best practices reported in oncology and palliative care [29,31,44], as well as practices identified as important in ambulatory care [10,13,45,46].

Similarly, fostering a sense of presence and attentiveness during the VV was a key focus. Many of these items have been reported in other research as important, including how to set up technology to enhance telepresence and body language communication (eg, eye contact, facial expressions, and body posture) and paying close attention to visual and auditory signals that might reflect patient emotion [20,26,47]. Best practices related to camera positioning and eye gaze originally recommended by the American Telemedicine Association [48] have been shown to influence interpersonal comfort and self-disclosure in web-based interactions and may facilitate trust and collaboration [49].

The curricular themes and items also addressed navigating barriers that impact building rapport in VV. Rettinger and Kuhn [50] described internet connectivity issues and a lack of skills in navigating or problem-solving technological challenges in VV as the barriers most frequently described in the 56 studies included in their scoping review. What was highlighted by the experts in this e-Delphi study is the importance of managing these challenges to maximize clinician-patient rapport. In addition to problem-solving skills, the overall attitude of the clinician in managing these challenges was highlighted. The experts felt that patience and light-heartedness were essential attitudes, even suggesting that technological challenges could present opportunities to build rapport around a shared experience. The high level of importance expressed by the expert panel for teaching rapport basics and managing VV challenges affirms the need for clinicians to be proficient and comfortable in both interpersonal and technological skills.

A recent scoping review on rapport in oncology ambulatory care found that more studies focused on nurses' attitudes than knowledge and skills [51]. Attitudes were also identified as critical by the experts in this study, who indicated that without positive attitudes, motivating adult learners to acquire new knowledge and skills becomes challenging. They articulated the need to cultivate positive attitudes toward telehealth and a willingness to adapt interpersonal skills accordingly. This was also identified as an important barrier in the scoping review of Rettinger and Kuhn [50].

Affective competency is reflected in students' ability to conceptualize and internalize subject matter related to attitudes, emotions, or biases in ways that influence their values, decisions, and behaviors [52,53]. Teaching in this domain of learning requires that educators place a greater emphasis on their students becoming caring, emotionally and culturally intelligent, and ethical professionals rather than on their acquisition of factual knowledge [20,53,54]. Achieving this goal necessitates creating a safe learning environment and using teaching strategies that enhance student self-awareness and encourage exploration of diverse perspectives and needs [53,55]. This highlights the importance of building a curriculum and using teaching strategies that address the affective domain of learning and aligns with the importance the experts placed on attitudes and reflective teaching strategies.

Research suggests that opportunities to practice conducting VV with standardized patients and in simulation laboratories have promoted positive learning outcomes [56]. This aligns with the results of this study, where real-time feedback and experiential learning methods reached higher levels of consensus. While module-based learning is evidence-based and necessary, given that this is how many health educational and professional programs are currently delivered [57], the experts in this study emphasized the importance of a learning environment that enabled hands-on practice.

Our results are relevant to various health professions that engage in telehealth. The Interprofessional Education Collaborative (IPEC) Core Competencies for Interprofessional Collaborative Practice [58] provide a framework for interprofessional collaborative practice centered upon 4 competency domains: communication, teams and teamwork, roles and responsibilities, and values and ethics. Each of the 4 competency domains is well-aligned with our panel's insights into effective education for telehealth rapport-building. Specifically, the IPEC competencies in communication and teams and teamwork resonate with our expert panel's emphasis on clear, culturally humble, and empathic communication adapted to the web-based setting. IPEC competency statements regarding communication that is responsive, respectful, and compassionate, as well as practicing active listening, directly relate to skills our expert panel prioritized, such as narrating visits, managing disruptions, and adapting both verbal and nonverbal cues to build rapport. The roles and responsibilities IPEC competency domain align with the panel's suggestion to teach interprofessional learners how to understand one another's scopes of practice and leverage each team member's unique contributions to enhance web-based care experiences. Finally, the values emphasized throughout the IPEC values and ethics domain, which include respect, cultural humility, and patient-centeredness, mirror the attitudinal attributes our experts identified as critical to building trust and interpersonal connection in video visits. The high level of consensus across the Delphi rounds and the incorporation of strategies like team-based practice opportunities, patient privacy, and rapport measurement underscore the practical application of IPEC competencies when designing meaningful telehealth education for various health professions.

Implications and Future Research

The research questions and analysis used a competency-based educational framework often applied by educators when developing competency-based education [59]. Organizing the findings by cognitive or mental skills (knowledge), manual

psychomotor skills (skills), and affective skills (attitudes) was done intentionally to make it relevant and easy for health profession educators to incorporate the results into their existing competency-based telehealth programs. When paired with existing checklists [25,26,46] and professional competency guidelines [3,4,58], these results can support the creation of stand-alone programs on relationship-based care in telehealth or module-based programs that can be integrated into broader telehealth curricula.

Future research is needed to test the feasibility, acceptability, and effectiveness of building curricula based on these competencies. Pilot testing of the proposed curriculum model would be an important next step in evaluating learning outcomes. This could be followed by a controlled study to examine whether these outcomes translate into measurable improvements in perceived rapport scores. In addition, gaining a better understanding of the patient perspective on rapport-building strategies is an important area for further study, along with exploring how these strategies may vary based on the VV context, purpose, and the type of professional conducting the VV.

Strengths and Limitations

The strengths of this study include the collection of data from interprofessional health care clinicians and educators. The use of the ACCORD [43] and multiple strategies to enhance the trustworthiness bolstered methodological rigor. The analysis incorporated a broad range of participants' responses to ensure credibility. Offering participants opportunities in rounds 2 - 4 to review the findings and provide feedback or suggestions also enhanced the study's dependability. The context of the participants was also thoroughly described to allow readers to determine how the findings can be applied in other programs and settings. Although the goal of a Delphi study is to build consensus, there is a risk of forcing conformity [32]. By interviewing participants individually in the first round, this study provided each expert an opportunity to fully share their views and ensured the participants remained anonymous to one another. The rapid and high levels of consensus for most of the items suggest that the analysis of the interview data and subsequent development of survey items were dependable, largely representing the experts' opinions.

This study has several limitations. Given the limited existing research on building rapport in VV, the reliance on expert interviews in round 1 constrained the participant pool. Although purposive criteria guided recruitment, the panel did not encompass all professions engaged in the use and instruction of VV. As such, the panel's composition may not fully reflect the broader population of telehealth clinicians and educators, particularly those in early-career stages. In addition, most participants were based in the United States, which may limit the applicability of findings in international or cross-cultural contexts. This may stem from social and cultural differences in communication norms. In addition, variations in health care systems, technological proficiency, device access, and internet availability across countries may further limit the generalizability of the findings. Finally, although there are no strict guidelines for Delphi panel size, the relatively small sample may not have captured the full range of perspectives. These factors may have increased the risk of selection bias and could affect the validity and generalizability of the study findings [32].

Conclusions

It is widely recognized that a strong rapport between patients and clinicians is essential in VV. Yet, despite this understanding, evidence-based practical guidance on adapting interpersonal skills for web-based care remains limited. This study offers a tangible first step toward addressing this gap. Its results provide detailed guidance on the knowledge, skills, and attitudes necessary for health professions students and clinicians to build rapport in VV settings. In addition to outlining these competencies, the study presents a clear structure, actionable content, and recommended teaching strategies to support curriculum development in health professions education. While web-based care brings numerous benefits, failing to deliberately adapt relationship-based skills for VV risks reducing patient interactions to impersonal, task-focused exchanges. To fully realize the potential of telehealth, it is critical to invest in evidence-based approaches that preserve and promote meaningful interpersonal connections in the web-based environment—for the benefit of both patients and clinicians.

Acknowledgments

This project was supported by small grants, including an award from Duke Academy for Health Professions Education and Academic Development and an Emerging Pedagogies Seed Grant from the Center for Applied Research and Design in Transformative Education, part of Duke University Learning Innovation & Lifetime Education.

Authors' Contributions

PDK, JCDG, and MW designed the study protocol. PDK interviewed all study participants. PDK, DMN, AE, JCDG, and MW coded and analyzed the data. PDK drafted the initial manuscript, and all other authors contributed to critical reviews and revisions of the manuscript. All authors have approved the final manuscript and agreed to publication.

Conflicts of Interest

None declared.

References

1. Fact sheet: Telehealth. American Hospital Association. URL: https://www.aha.org/fact-sheets/2025-02-07-fact-sheet-telehealth [accessed 2025-02-02]
2. Shaver J. The state of telehealth before and after the COVID-19 pandemic. Prim Care 2022 Dec;49(4):517-530. [doi: 10.1016/j.pop.2022.04.002] [Medline: 36357058]
3. Noronha C, Lo MC, Nikiforova T, et al. Telehealth competencies in medical education: new frontiers in faculty development and learner assessments. J Gen Intern Med 2022 Sep;37(12):3168-3173. [doi: 10.1007/s11606-022-07564-8] [Medline: 35474505]
4. Anglea C C, Murray M, Mastal M, Clelland S, editors. Scope and Standards of Practice for Professional Telehealth Nursing, 6th edition: American Academy of Ambulatory Care Nursing; 2018:1-55.
5. Accreditation standards for physician assistant education. Sixth edition, first draft. : Accreditation Review Commission on Education for the Physician Assistant; 2024 Sep URL: https://www.arc-pa.org/wp-content/uploads/2024/10/Standards-6-ed-FIRST-DRAFT-09-2024.pdf [accessed 2024-04-08]
6. Hollander JE, Davis TM, Doarn C, et al. Recommendations from the first national academic consortium of telehealth. Popul Health Manag 2018 Aug;21(4):271-277. [doi: 10.1089/pop.2017.0080] [Medline: 28976250]
7. Chike-Harris KE, Durham C, Logan A, Smith G, DuBose-Morris R. Integration of telehealth education into the health care provider curriculum: a review. Telemed J E Health 2021 Feb;27(2):137-149. [doi: 10.1089/tmj.2019.0261] [Medline: 32250196]
8. Pittmann R, Adair White BA, Danaher-Garcia N, Thompson A. Where's the etiquette? Telehealth etiquette in health professions education and practice: a scoping review. Internet J Allied Health Sci and Pract 2024;22(4):22.
9. English W, Robinson J, Gott M. How are the vibes? Patient and family experiences of rapport during telehealth calls in palliative care. Patient Exp J 2023;10(2):75-85. [doi: 10.35680/2372-0247.1786]
10. Koppel PD, De Gagne JC, Docherty S, Smith S, Prose NS, Jabaley T. Exploring nurse and patient experiences of developing rapport during oncology ambulatory care videoconferencing visits: qualitative descriptive study. J Med Internet Res 2022 Sep 8;24(9):e39920. [doi: 10.2196/39920] [Medline: 36074558]
11. Rettinger L, Putz P, Aichinger L, et al. Telehealth education in allied health care and nursing: web-based cross-sectional survey of students' perceived knowledge, skills, attitudes, and experience. JMIR Med Educ 2024 Mar 21;10:e51112. [doi: 10.2196/51112] [Medline: 38512310]
12. Goodchild L, Mulligan A, Shearing Green E, Mueller T. Clinician of the Future: Elsevier; 2022.
13. Duffy LV, Evans R, Bennett V, Jones R. Exploring therapeutic relational connection in virtual healthcare: insights from nurse practitioner practice with young adults living with chronic illness. J Adv Nurs 2025 Jul;81(7):3962-3971. [doi: 10.1111/jan.16654] [Medline: 39651647]
14. Gustin TS, Kott K, Rutledge C. Telehealth etiquette training: a guideline for preparing interprofessional teams for successful encounters. Nurse Educ 2020;45(2):88-92. [doi: 10.1097/NNE.0000000000000680] [Medline: 31022072]
15. Pittmann R, Danaher-Garcia N, Adair White BA, Thompson A. The impact of a professional development workshop on healthcare professionals' knowledge and readiness to use telehealth etiquette in virtual care. J Telemed Telecare 2024 Oct 17;17:2024. [doi: 10.1177/1357633X241285938] [Medline: 39415640]
16. Tickle-Degnen L, Rosenthal R. The nature of rapport and its nonverbal correlates. Psychol Inq 1990 Oct;1(4):285-293. [doi: 10.1207/s15327965pli0104_1]
17. Tickle-Degnen L. Nonverbal behavior and its functions in the ecosystem of rapport. In: Patterson V, editor. The Sage Handbook of Nonverbal Communication Sage 2006:381-399. [doi: 10.4135/9781412976152.n20]
18. Epstein RM, Street RL. Patient-Centered Communication in Cancer Care: Promoting Healing and Reducing Suffering: National Cancer Institute, NIH Publication No. 07-6225; 2007:1-203.
19. English W, Robinson J, Gott M. Health professionals' experiences of rapport during telehealth encounters in community palliative care: an interpretive description study. Palliat Med 2023 Jul;37(7):975-983. [doi: 10.1177/02692163231172243] [Medline: 37129344]
20. Duffy LV, Evans R, Bennett V, Hady JM, Palaniappan P. Therapeutic relational connection in telehealth: concept analysis. J Med Internet Res 2023 Jun 22;25:e43303. [doi: 10.2196/43303] [Medline: 37347526]
21. English W, Gott M, Robinson J. The meaning of rapport for patients, families, and healthcare professionals: a scoping review. Patient Educ Couns 2022 Jan;105(1):2-14. [doi: 10.1016/j.pec.2021.06.003] [Medline: 34154861]
22. Haverfield MC, Tierney A, Schwartz R, et al. Can patient-provider interpersonal interventions achieve the quadruple aim of healthcare? A systematic review. J Gen Intern Med 2020 Jul;35(7):2107-2117. [doi: 10.1007/s11606-019-05525-2] [Medline: 31919725]
23. Kelley JM, Kraft-Todd G, Schapira L, Kossowsky J, Riess H. The influence of the patient-clinician relationship on healthcare outcomes: a systematic review and meta-analysis of randomized controlled trials. PLoS ONE 2014;9(4):e94207. [doi: 10.1371/journal.pone.0094207] [Medline: 24718585]
24. Groom LL, Brody AA, Squires AP. Defining telepresence as experienced in telehealth encounters: a dimensional analysis. J Nurs Scholarsh 2021 Nov;53(6):709-717. [doi: 10.1111/jnu.12684] [Medline: 34060218]

25. Pittmann R, Danaher-Garcia N, Adair White BA, Thompson A. Development and validation of the Telehealth Etiquette Competency Checklist: a Delphi study. J Telemed Telecare 2024 Sep 23:1357633X241279494. [doi: 10.1177/1357633X241279494] [Medline: 39311041]
26. Henry BW, Billingsly D, Block DE, Ehrmann J. Development of the teaching interpersonal skills for telehealth checklist. Eval Health Prof 2022 Sep;45(3):260-269. [doi: 10.1177/0163278721992831] [Medline: 33557609]
27. Ostrovsky DA, Heflin MT, Bowers MT, et al. Development, implementation, and assessment of an online modular telehealth curriculum for health professions students. Adv Med Educ Pract 2024;15:743-753. [doi: 10.2147/AMEP.S468833] [Medline: 39099682]
28. Antoine MD, Cherba M, Grosjean S, Boet S, Waldolf R. How to support patient-provider communication during telemedicine consultations? A scoping review of challenges and existing tools. Univ Ottawa J Med 2023;12(1):25-33. [doi: 10.18192/uojm.v12i1.6320]
29. Banerjee SC, Staley JM, Howell F, et al. Communicating effectively via tele-oncology (Comskil TeleOnc): a guide for best practices for communication skills in virtual cancer care. J Cancer Educ 2022 Oct;37(5):1343-1348. [doi: 10.1007/s13187-021-01959-7] [Medline: 33544315]
30. Castro MJA, Zaig S, Nissim R, et al. Telehealth outpatient palliative care in the COVID-19 pandemic: patient experience qualitative study. BMJ Support Palliat Care :spcare-2023-004189. [doi: 10.1136/spcare-2023-004189]
31. Chua IS, Jackson V, Kamdar M. Webside manner during the COVID-19 pandemic: maintaining human connection during virtual visits. J Palliat Med 2020 Nov;23(11):1507-1509. [doi: 10.1089/jpm.2020.0298] [Medline: 32525744]
32. Keeney S, Hasson F, McKenna HP. The Delphi Technique in Nursing and Health Research: Wiley-Blackwell; 2011.
33. Gill FJ, Leslie GD, Grech C, Latour JM. Using a web-based survey tool to undertake a Delphi study: application for nurse education research. Nurse Educ Today 2013 Nov;33(11):1322-1328. [doi: 10.1016/j.nedt.2013.02.016] [Medline: 23510701]
34. Koppel PD. Educational guidelines for building rapport in telehealth video visits: an e-delphi study. OSF Registries. 2025 Jan 22. URL: https://osf.io/tjmkx [accessed 2025-07-24]
35. DeVellis RF. Scale Development: Theory and Applications: SAGE; 2017.
36. Sorensen J, Michaëlis C, Olsen JMM, Krasnik A, Bozorgmehr K, Ziegler S. Diversity competence training for health professionals in Europe: a modified Delphi study investigating relevant content for short or online courses. BMC Med Educ 2023 Aug 21;23(1):590. [doi: 10.1186/s12909-023-04563-z]
37. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
38. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008 Apr;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
39. Prescott MR, Yeager S, Ham L, et al. Comparing the efficacy and efficiency of human and generative AI: qualitative thematic analyses. JMIR AI 2024 Aug 2;3:e54482. [doi: 10.2196/54482] [Medline: 39094113]
40. Elo S, Kääriäinen M, Kanste O, Pölkki T, Utriainen K, Kyngäs H. Qualitative content analysis: a focus on trustworthiness. SAGE Open 2014 Feb;4(1):2158244014522633. [doi: 10.1177/2158244014522633]
41. Morse JM, Barrett M, Mayan M, Olson K, Spiers J. Verification strategies for establishing reliability and validity in qualitative research. Int J Qual Methods 2002 Jun;1(2):13-22. [doi: 10.1177/160940690200100202]
42. Marshall C, Rossman GB. Designing Qualitative Research, 6th edition: SAGE Publications; 2016:1-352.
43. Gattrell WT, Logullo P, van Zuuren EJ, et al. ACCORD (ACcurate COnsensus Reporting Document): a reporting guideline for consensus methods in biomedicine developed via a modified Delphi. PLoS Med 2024 Jan;21(1):e1004326. [doi: 10.1371/journal.pmed.1004326] [Medline: 38261576]
44. Webb M, Hurley SL, Gentry J, Brown M, Ayoub C. Best practices for using telehealth in hospice and palliative care. J Hosp Palliat Nurs 2021 Jun 1;23(3):277-285. [doi: 10.1097/NJH.0000000000000753] [Medline: 33911060]
45. Millstein JH, Chaiyachati KH. Creating virtual presence during a pandemic. J Patient Exp 2020 Jun;7(3):285-286. [doi: 10.1177/2374373520930447] [Medline: 32821781]
46. Modic MB, Neuendorf K, Windover AK. Enhancing your webside manner: optimizing opportunities for relationship-centered care in virtual visits. J Patient Exp 2020 Dec;7(6):869-877. [doi: 10.1177/2374373520968975] [Medline: 33457513]
47. Schweickert PA, Rutledge CM, editors. Telehealth Essentials for Advanced Practice Nursing: SLACK Incorporated; 2020.
48. Ben-Arieh D, Charness N, Duckett K, Krupinski E, Leistner G, Strawderman L. A concise guide for telemedicine practitioners: human factors quick guide eye contact. : American Telemedicine Association; 2016 Feb.
49. Fitzpatrick JA, Tickle JJ. Eye contact matters: exploring its influence on interpersonal comfort and self-disclosure in virtual interactions. Couns Psychol Q :1-19. [doi: 10.1080/09515070.2024.2436964]
50. Rettinger L, Kuhn S. Barriers to video call-based telehealth in allied health professions and nursing: scoping review and mapping process. J Med Internet Res 2023 Aug 1;25:e46715. [doi: 10.2196/46715] [Medline: 37526957]
51. Koppel PD, Park HYK, Ledbetter LS, Wang EJ, Rink LC, De Gagne JC. Rapport between nurses and adult patients with cancer in ambulatory oncology care settings: a scoping review. Int J Nurs Stud 2024 Jan;149:104611. [doi: 10.1016/j.ijnurstu.2023.104611] [Medline: 37879272]
52. Myers SA, Goodboy AK. Reconsidering the conceptualization and operationalization of affective learning. Commun Educ 2015 Oct 2;64(4):493-497. [doi: 10.1080/03634523.2015.1058489]

53. Valiga TM. Attending to affective domain learning: essential to prepare the kind of graduates the public needs. J Nurs Educ 2014 May 1;53(5):247-247. [doi: 10.3928/01484834-20140422-10] [Medline: 24802228]
54. Oermann MH, Shellenbarger T, Gaberson KB. Clinical Teaching Strategies in Nursing, 6th edition: Springer Publishing; 2023.
55. Donlan P. Developing affective domain learning in health professions education. J Allied Health 2018;47(4):289-295. [Medline: 30508841]
56. Smith TS, Watts P, Moss JA. Using simulation to teach telehealth nursing competencies. J Nurs Educ 2018 Oct 1;57(10):624-627. [doi: 10.3928/01484834-20180921-10] [Medline: 30277549]
57. De Gagne JC, Randall PS, Koppel PD, Cho E, Blackwood ER, Kang HS. Online learning in nursing education: a 21st century bibliometric analysis. Nurse Educ Today 2025 Aug;151:106740. [doi: 10.1016/j.nedt.2025.106740] [Medline: 40222324]
58. IPEC core competencies for interprofessional collaborative practice: version 3. : Interprofessional Education Collaborative; 2023.
59. Cutcliffe JR, Sloan G. Towards a consensus of a competency framework for clinical supervision in nursing: knowledge, attitudes, and skills. Clin Superv 2014 Jul 3;33(2):182-203. [doi: 10.1080/07325223.2014.981494]

Abbreviations

ACCORD: Accurate Consensus Reporting Document
AI: artificial intelligence
IPEC: Interprofessional Education Collaborative
VV: videoconferencing visit

Edited by A Bahattab; submitted 19.04.25; peer-reviewed by K Garber, Y Yazdani; revised version received 10.06.25; accepted 17.06.25; published 07.08.25.

Please cite as:
Koppel PD, De Gagne JC, Webb M, Nepveux DM, Bludorn J, Emmons A, Randall PS, Prose NS
Guidelines for Rapport-Building in Telehealth Videoconferencing: Interprofessional e-Delphi Study
JMIR Med Educ 2025;11:e76260
URL: https://mededu.jmir.org/2025/1/e76260
doi:10.2196/76260

The Evolution of Medical Student Competencies and Attitudes in Digital Health Between 2016 and 2022: Comparative Cross-Sectional Study

Paula Veikkolainen[1], MD, MSc; Timo Tuovinen[1,2], MD, PhD; Petri Kulmala[2,3], MD, PhD; Erika Jarva[4], MSc, PhD; Jonna Juntunen[4], MSc; Anna-Maria Tuomikoski[5], MSc, PhD; Merja Männistö[6], MSc, PhD; Teemu Pihlajasalo[7], MD; Jarmo Reponen[1,2], MD, PhD

[1]FinnTelemedicum, Research Unit of Health Sciences and Technology, Faculty of Medicine, University of Oulu, Aapistie 5a, Oulu, Finland

[2]Medical Research Center, the Wellbeing Services County of North Ostrobothnia and the University of Oulu, Oulu, Finland

[3]Education Development and Service Unit, Faculty of Medicine, University of Oulu, Oulu, Finland

[4]Research Unit of Health Sciences and Technology, Faculty of Medicine, University of Oulu, Oulu, Finland

[5]Oulu University Hospital, Wellbeing Services County of North Ostrobothnia, Oulu, Finland

[6]School of Health and Social Studies, JAMK University of Applied Sciences, Jyväskylä, Finland

[7]Mehiläinen Vaasa, Mehiläinen Health Services, Vaasa, Finland

Corresponding Author:
Paula Veikkolainen, MD, MSc
FinnTelemedicum, Research Unit of Health Sciences and Technology, Faculty of Medicine, University of Oulu, Aapistie 5a, Oulu, Finland

Abstract

Background: Modern health care systems worldwide are facing challenges, and digitalization is viewed as a way to strengthen health care globally. As health care systems become more digital, it is essential to assess health care professionals' competencies and skills to ensure they can adapt to new practices, policies, and workflows effectively.

Objective: The aim of this study was to analyze how the attitudes, skills, and knowledge of medical students concerning digital health have shifted from 2016 to 2022 in connection with the development of the national health care information system architecture using the clinical adoption meta-model framework.

Methods: The study population consisted of 5th-year medical students from the University of Oulu in Finland during 2016, 2021, and 2022. A survey questionnaire was administered comprising 7 background questions and 16 statements rated on a 5-point Likert scale assessing students' attitudes toward digital health and their self-perceived digital capabilities. The results were recategorized into a dichotomous scale. The statistical analysis used Pearson χ^2 test. The Benjamini-Hochberg procedure was used for multiple variable correction.

Results: The study included 215 medical students (n=45 in 2016, n=106 in 2021, and n=64 in 2022) with an overall response rate of 53% (43% in 2016, 74% in 2021, and 42% in 2022). Throughout 2016, 2021, and 2022, medical students maintained positive attitudes toward using patient-generated information and digital applications in patient care. Their self-perceived knowledge of the national patient portal significantly improved, with agreement increasing by 35 percentage points from 2016 to 2021 (P<.001) and this trend continued in 2022 (P<.001). However, their perceived skills in using electronic medical records did not show significant changes. Additionally, students' perceptions of the impact of digitalization on health promotion improved markedly from 2016 to 2021 (with agreement rising from 53% to 78%, P=.002) but declined notably again by 2022.

Conclusions: Medical students' attitudes and self-perceived competencies have shifted over the years, potentially influenced by the national health information system architecture developments. However, these positive changes have not followed a completely linear trajectory. To address these gaps, educational institutions and policy makers should integrate more digital health topics into medical curricula and provide practical experience with digital technologies to keep professionals up-to-date with the evolving health care environment.

(JMIR Med Educ 2025;11:e67423) doi:10.2196/67423

KEYWORDS

digital health; eHealth; telemedicine; medical informatics; professional competence; medical education; digitalization; digital health; digital; technology; medical student; cross-sectional study; health care systems; health care; health care professional; health care information system; survey; questionnaire; healthcare digitalization; digital competence; innovation

Introduction

Modern health care systems around the world are facing challenges due to the aging population, increasing prevalence of chronic diseases, and other lifestyle-associated conditions [1,2]. At the same time, countries are grappling with shortages of the health care workforce, especially in remote and rural areas [3]. Different exposures to health risks create health inequalities between individuals with higher and lower education and income levels [4,5]. These factors put pressure on health care systems to shift their focus toward promoting health and preventing diseases through patient engagement and self-management.

According to the World Health Organization (WHO), the term "eHealth" focuses on using information and communication technologies in health care, while "digital health" serves as a broader umbrella term that also encompasses advanced computer sciences such as artificial intelligence. Furthermore, the term "mHealth," a subset of eHealth, is defined as "the use of mobile wireless technologies for health" [2]. Regardless, digital transformation is seen as an essential component and enabler to enhance the quality, accessibility, and affordability of health services [2,6]. In 2021, the WHO published the Global Strategy on Digital Health 2020 - 2025, seeking to assist nations in strengthening their health care systems through digitalization [7]. The European Union (EU) has named digital solutions as one of the key enablers to deliver health and care services more effectively to patients [8,9]. The COVID-19 pandemic, starting in 2020, and the resulting lockdowns accelerated the technological leap by forcing health care institutions worldwide to swiftly develop and implement digital strategies [10-13].

In 2022, Finland was ranked as the top country in the annual Digital Economy and Society Index report, which monitors the digital progress of the European Union Member States [14]. Accordingly, Finland has a long history of enforcing digitalization in health care [15,16]. One example of this development is the introduction and implementation of the Finnish nationwide, centralized shared electronic data system service, called the Kanta Services, which comprises several service entities such as an electronic patient portal, prescription database, and patient data repository. The implementation of these services has taken place in several stages throughout the 2010s [17-19].

In Finland, the key competence requirements for a graduating doctor have been established at a national level [20]. These requirements are based on international literature, including the UK's Generic Professional Capabilities framework and the International Association for Medical Education guidelines [21-23]. The basic education for medical professionals in Finland consists of 2 years of pre-clinical studies followed by 4 years of clinical training. In the EU, the profession of medical doctor is recognized as a qualification in all member states on the basis of harmonized minimum training requirements [24].

As health care systems become more digital, it's essential to assess health care professionals' competencies and skills to ensure they can adapt to new practices, policies, and workflows effectively [25]. According to current literature, medical students globally have positive attitudes toward learning about digital health and consider the introduction of digital health topics into the medical curricula to be important [26-34]. However, more reserved perceptions toward digital health have also been reported among students [35].

Many nations have made efforts to develop and implement national digital health strategies, including initiatives to incorporate digital health education at local and national levels [36-40]. Furthermore, the European Medical Students' Association has implemented policies emphasizing the inclusion of digital health in medical education curriculum to ensure future doctors in Europe possess crucial digital skills [41], and the recent Digital Health Competencies in Medical Education framework outlines the essential digital competencies for medical education on a global level [42]. Efforts have been made to modernize basic medical education in Finland as well; an example of this is the national MEDigi project (2018 - 2021) funded by the Finnish Ministry of Education and Culture [43]. In addition to digitizing the teaching of medicine and dentistry, the project aimed to ensure a high level of competence in the use of digital health care tools among medical students and to establish national eHealth competence themes [44].

In our previous study, we aimed to compare the attitudes of medical and nursing students toward digital health. Based on the study results, the differences between the 2 student groups were small, and overall, the students' attitudes toward digital health were positive [33]. Now, we seek to deepen our understanding of the changes in medical students' attitudes, skills, and knowledge regarding digital health in relation to the underlying development of the national health care information system architecture. For this purpose, we used the clinical adoption meta-model (CAMM) framework. This framework is developed to describe the health information system adoption over time, and it incorporates 4 dimensions: availability, use, behavior changes, and outcome changes [45]. In this study, the aim is to focus on the third dimension of the model, namely, to describe the behavior changes (attitudes and competencies) of medical students in connection with the development of the national health care information system architecture over time.

Methods

Ethical Considerations

The research was conducted in accordance with the instructions of the Finnish Advisory Board on Research Integrity [46], and in compliance with EU data protection regulations [47] as well

as the established research practices of the University of Oulu and the Faculty of Medicine. Therefore, no approval from the ethics committee was required. Full consideration was given to matters related to data protection in accordance with the ethical principles applicable to research subjects. Participation in the study was voluntary, and students were asked for their consent to collect and use data for the purpose of the study. The students were informed of the purpose of the study and their right to withdraw and prohibit the use of their data at any time. No incentives were offered for participation.

Study Design

The study population for this comparative cross-sectional study consisted of 5th-year medical students at the University of Oulu who participated in a compulsory 1-day digital health course held in spring 2016, 2021, and 2022. The aim of this 1-day course was to provide essential knowledge on digital health and its applications from the perspective of health care professionals. There were 105, 144, and 154 medical students who enrolled in the courses in 2016, 2021, and 2022, respectively. The students were invited to participate in the study via email in 2016, as well as through the course's Moodle environment in 2021 and 2022.

This study adheres to the EQUATOR CROSS (A Consensus-Based Checklist for Reporting of Survey Studies) guidelines for survey research [48]. The checklist was used to ensure comprehensive reporting of study design, data collection, analysis, and interpretation.

Study Questionnaire

After completing the digital health course, a web-based survey was conducted on the students' perceptions of digital health using a Webropol survey tool. The survey was compiled in 2016, and it was developed based on literature and expert reviews [49]. The survey was piloted prior to its use. The pilot group consisted of 2 fifth-year medical students and 2 medical teachers, both of whom had backgrounds in teaching digital health.

The Finnish-language survey questionnaire consisted of 16 statements (Q1-Q16) measured on a 5-point Likert scale ("Fully disagree" to "Fully agree"), surveying students' attitudes to digital health and their self-perceived digital competencies. The survey used in 2021 and 2022 was adapted from the 2016 survey by changing the term "medical doctor" to "health care professional," as nursing students also participated in the 1-day course in 2021 and 2022. An English translation of the latter questionnaire is presented in Multimedia Appendix 1. The statements were related to 5 themes concerning digital health: (1) the usage of patient-generated information and the role of digital applications in patient care; (2) health information systems; (3) digitalization of the working environment; (4) the changing role of patients and professionals; and (5) the culture of experimentation and readiness to participate in innovation activities.

Data Analysis

The 5-point Likert-scale responses "Fully agree" and "Somewhat agree" were combined to form the category "Agree." Similarly, the responses "Fully disagree," "Somewhat disagree," and "Neither agree or disagree" were combined to form the category "Disagree" using Excel (Microsoft Corp.).

The data analysis was conducted using a Pearson χ^2 test to examine relationships between the medical student's digital health attitudes and competencies in 2016, 2021, and 2022. We used χ^2 Calculator by Social Science Statistics to perform the statistical analysis and Effect Size Calculator (effect type: phi) by Statistics Kingdom to calculate effect sizes [50,51]. A *P* value less than .05 was considered statistically significant [52]. Cohen interpretation of the ϕ is as follows: small effect ϕ=0.1, medium effect ϕ=0.3, and large effect ϕ=0.5 [53]. Based on consultation with a statistician, significance values were corrected for multiple comparisons using the Benjamini-Hochberg procedure to control the false discovery rate (FDR =0.25) [54]. According to the Benjamini-Hochberg procedure, *P* values ≤.03 were found to be statistically significant. The results of data analysis, including the χ^2 statistics (χ^2), degrees of freedom (*df*), *P* values, phi coefficients (ϕ), and Q-values, are reported in Multimedia Appendix 2.

Results

Demographic Characteristics of the Study Sample

The sample size of the study included a total of 215 medical students (n=45 in 2016, n=106 in 2021, and n=64 in 2022), with an overall response rate of 53% (43% in 2016, 74% in 2021, and 42% in 2022). No participants were excluded due to incomplete or invalid responses. There were no missing values or other deviations in the questionnaire results. Details regarding gender distribution and age distribution are presented in Table 1.

Table . Demographic characteristics of the study sample (N=215), including gender distribution and age distribution by decade of birth.

Characteristics	Medical students in 2016 (n=45), n (%)	Medical students in 2021 (n=106), n (%)	Medical students in 2022 (n=64), n (%)
Gender distribution			
Women	27 (60)	56 (53)	34 (53)
Men	18 (40)	50 (47)	27 (42)
Other	—[a]	—	3 (5)
Age distribution (birth decade)			
1960	3 (7)	—	—
1970	3 (7)	—	1 (1.5)
1980	20 (44)	13 (12)	7 (11)
1990	19 (42)	93 (88)	55 (86)
2000	—	—	1 (1.5)

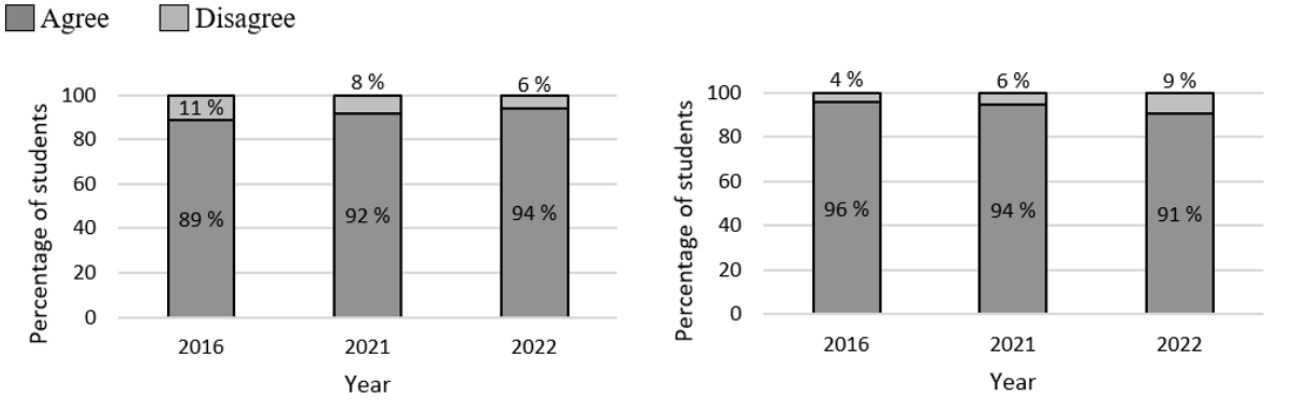
[a]Not applicable

The average working experience in health care sector was 1.1 (SD 1.3) years in 2021 and 0.9 (SD 1.1) years in 2022. In the 2016 survey, 96% of the students had worked as assistants to medical doctors. The work experience of the students in a field corresponding to their prior education was on average 1.1 (SD 2.5) years in 2021 and 0.9 (SD 2.1) years in 2022. While the 2016 dataset likely includes career changers (persons born in 1960 and 1970) [55], we lack sufficient information about their backgrounds to draw definitive conclusions about this.

The Usage of Patient-Generated Information and the Role of Applications in Patient Care

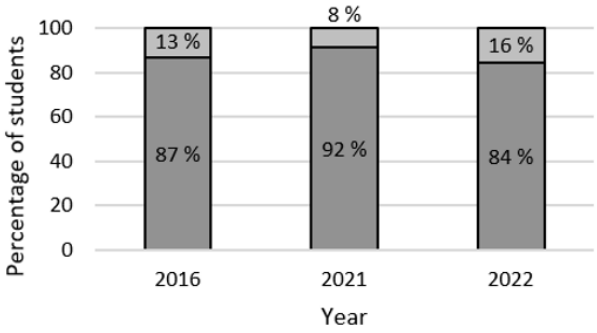
Overall, there were no major changes in medical students' attitudes toward using patient-generated information and the role of digital applications in patient care between 2016, 2021, and 2022 (Figure 1). Students consistently held positive views, asserting that digital applications benefit patients' health and motivation. They also emphasized the importance of health care professionals being proficient in using these applications.

Figure 1. Percentages of students agreeing and disagreeing with each statement related to the theme "The usage of patient-generated information and the role of applications in patient care." A statistical analysis was performed using a χ^2 test.

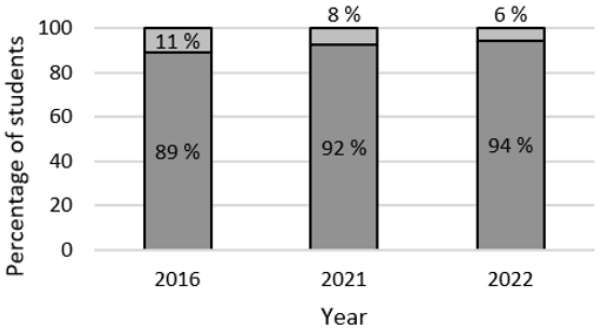


Q1) The usage of health applications by patients contributes positively to their healthcare.

Q2) It is important for healthcare professionals to be able to utilise digital applications in patient care.



Q3) It is important for healthcare professionals to be able to utilise patient-generated health data in patient care.

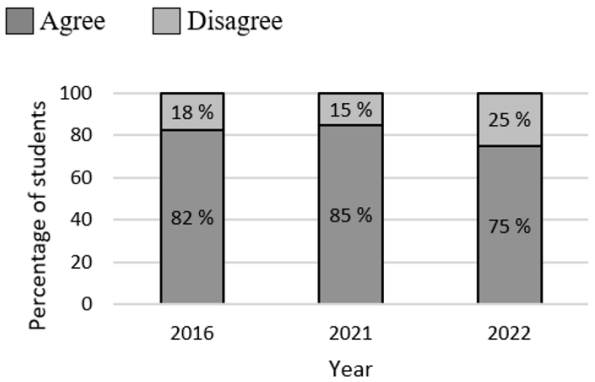


Q4) The engagement of the patient in their treatment and care (e.g., by using electronic self-monitoring and self-care systems) leads to better patient motivation and improved health outcomes.

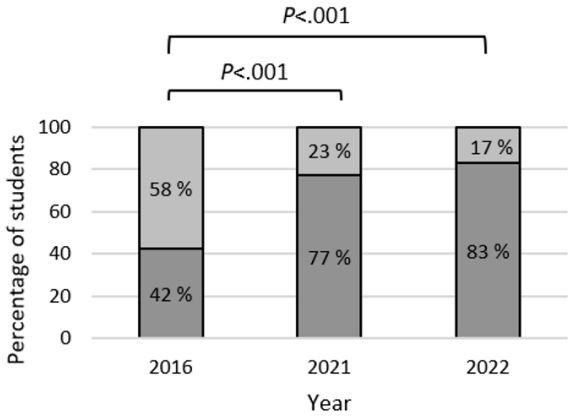
Health Information Systems

There was a significant improvement in the students' self-perceived knowledge of the information contained in the national patient portal (My Kanta Pages), with agreement increasing by 35 percentage points from 2016 to 2021, and this trend continued in 2022 (Figure 2).

Figure 2. Percentages of students agreeing and disagreeing with each statement related to the theme "Health information systems." A statistical analysis was performed using a χ^2 test.



Q5) I know how to use the tools within electronic medical record systems to facilitate my daily work as a healthcare professional.



Q6) I know what kind of information the national patient portal (My Kanta Pages) incorporates and what features it offers to the patient and the healthcare professional.

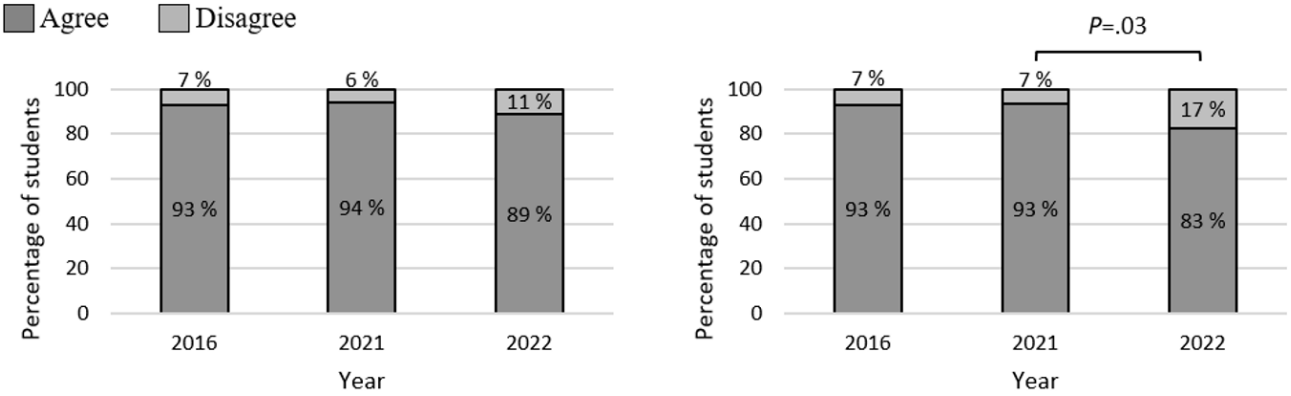
However, there was no significant change in students' self-perceived skills in using electronic medical records over

this period. If anything, the students perceived their skill set to be weaker in 2022 compared to 2016 and 2021, though this difference was not statistically significant.

Digitalization of the Working Environment

The survey data indicates that medical students consistently recognize the importance of digitalization in enhancing health care professionals' working methods (Figure 3). From 2021 to 2022, there was a significant shift in students' perceptions of how digitalization would affect their working lives in the coming years. In 2021, only 7% of the students disagreed that digitalization significantly changes health care professionals' practical work, consistent with 2016. By 2022, this disagreement rose to 17%, demonstrating greater uncertainty.

Figure 3. Percentages of students agreeing and disagreeing with each statement related to the theme "Digitalization of the working environment." A statistical analysis was performed using a χ^2 test.



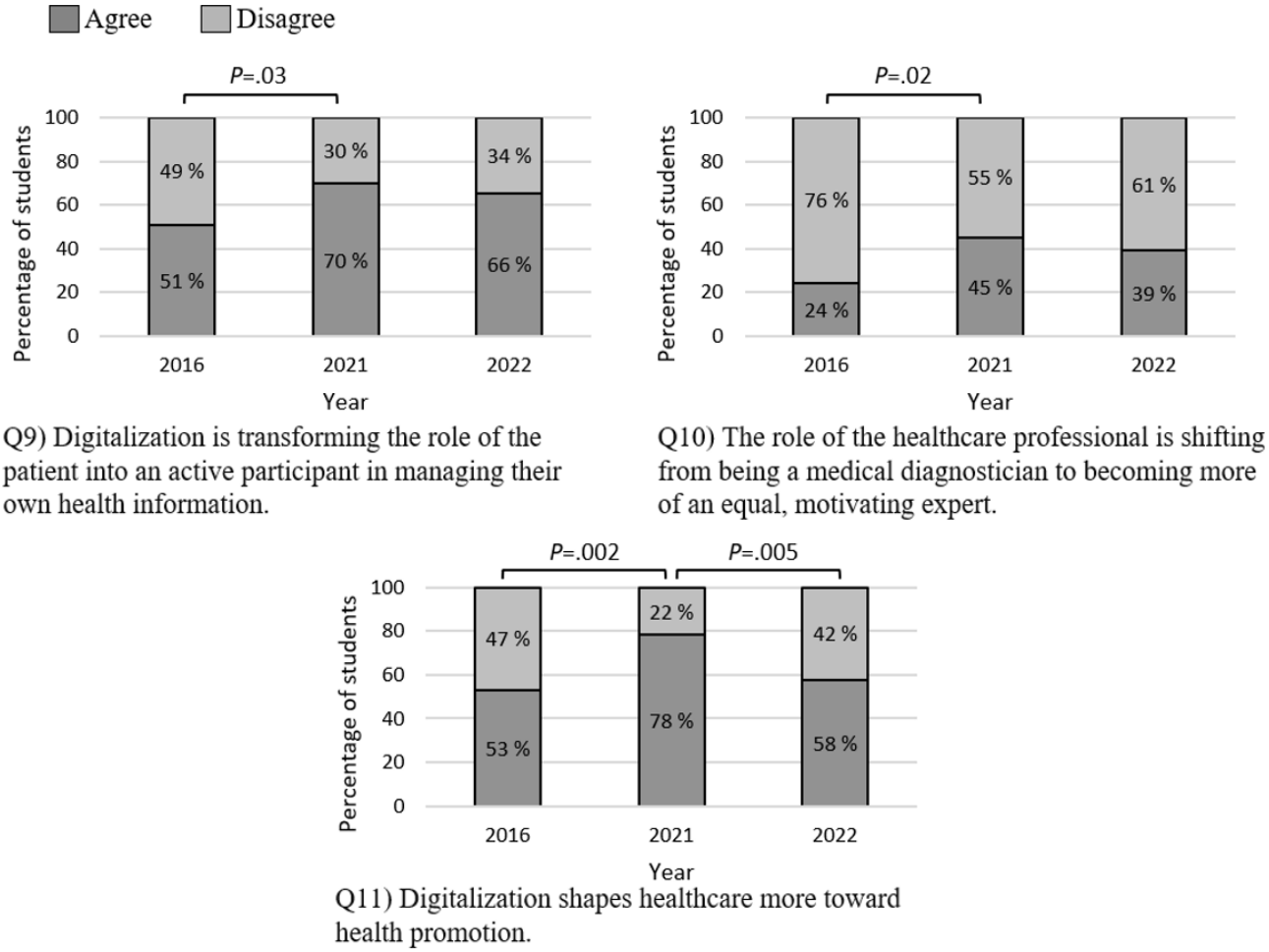
Q7) It is important for healthcare professionals to be able to improve their working methods and/or practices within their work community through digitalization.

Q8) The digitalization of healthcare is expected to significantly affect the practical work of healthcare professionals in the coming years.

The Changing Role of Patients and Professionals

Between 2016 and 2021, medical students' attitudes about the roles of patients and professionals evolved, reflecting increased patient involvement in managing their health information and more equitable relationship between patients and professionals (Figure 4). Additionally, the students' perceptions of the impact of digitalization on health promotion improved significantly from 2016 to 2021 (with agreement rising from 53% to 78%) but declined notably again by 2022.

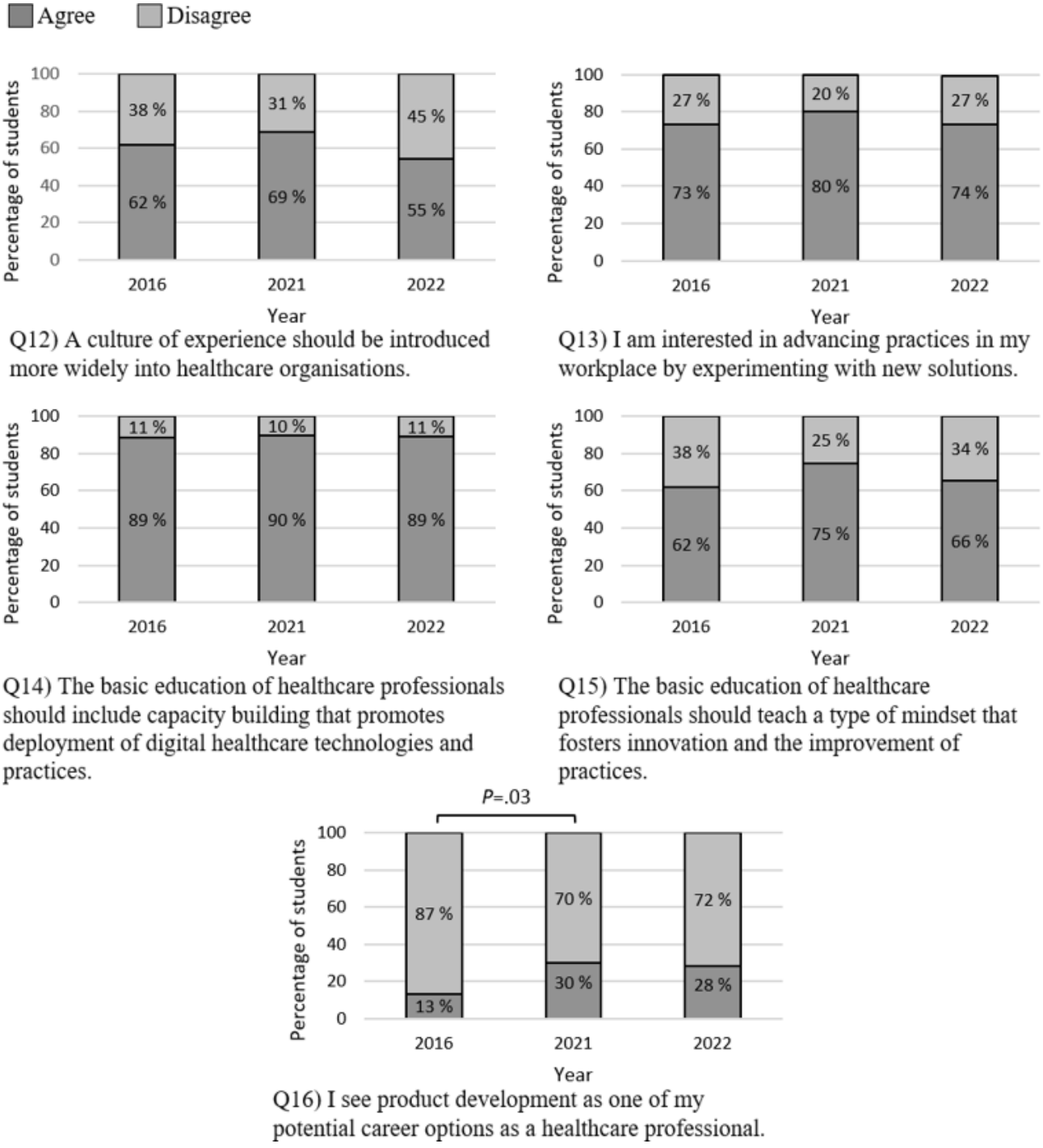
Figure 4. Percentages of students agreeing and disagreeing with each statement related to the theme "The changing role of patients and professionals." A statistical analysis was performed using a χ^2 test.



The Culture of Experimentation and Readiness to Participate in Innovation Activities

The participating students consistently valued the inclusion of digital health topics in basic medical education across 2016, 2021, and 2022, with no significant changes in attitudes (Figure 5). In addition, alternative career options, such as product development, became increasingly appealing, with the percentage of students considering this path more than doubling from 2016 to 2021 and remaining relatively stable in 2022.

Figure 5. Percentages of students agreeing and disagreeing with each statement related to the theme "The culture of experimentation and readiness to participate in innovation activities." A statistical analysis was performed using a χ^2 test.



Discussion

Principal Findings

Overall, we observed a positive shift in the participating medical students' attitudes and an improvement in their self-perceived digital competencies between 2016 and 2022. This coincides with advancements in the national health information system, providing an excellent opportunity to assess the outcomes from the CAMM framework viewpoint. Interestingly, we observed that the changes in attitudes were not consistently linear; while there was a positive trend overall, the 2022 results showed signs of stagnation or decline, with the exception of the students' increasing knowledge of the national patient portal.

To our knowledge, there has been only 1 umbrella review describing health care providers' attitudes to patient portals using CAMM [56]. This study primarily focused on the adoption of patient portals, revealing predominantly reserved attitudes and concerns among professionals, such as increased workload, insufficient training and resources, accuracy of information, and issues related to patient privacy. Our study contributes to this by extensively mapping the attitudes and digital competencies of future health care professionals, reflecting potential changes connected to the advancements in the national

health information system architecture across three different time points. In our work, students' self-perception of knowledge of the national patient portal had significantly improved. Our data also indicated that, as early as 2016, medical students valued the importance of using patient-generated information and patient engagement via self-care systems relatively high, with 87% and 89% of the students, respectively. This might indicate that while there have been initial concerns among professionals, future professionals are increasingly recognizing the value of digital tools and patient engagement.

In our study, medical students perceived the use of digital applications and patient-generated data in patient care as a positive factor: there were no statistically significant differences in the participating students' attitudes between 2016, 2021, and 2022 (see Figure 1). From a system adoption point of view, this finding suggests that the widespread use of digital applications had already prompted students to recognize their importance as early as 2016. EU legislation has emphasized the role of medical and health digital applications as a part of patient empowerment and has also recognized potential risk aspects [1,57]. Both the WHO and the European Commission have participated in establishing the European mHealth Knowledge and Innovation Hub as part of the Horizon 2020 project [58]. The initiative aims to support the integration of mHealth services into the national health systems of European countries.

Our results indicate a significant improvement in medical students' knowledge of the national patient portal since 2016 (see Figure 2, Q6). We believe that this development may be linked to the introduction and implementation of new national health information exchange services in the 2010s and beyond. For example, the implementation of the national patient data repository in public health care in Finland was only completed in late 2015, less than a year before our first data checkpoint [15,19]. Similar nationwide systems for the exchange of health information between patients and professionals can be found in several countries [59], but studies on students' competencies in the use of these systems are scarce in the literature. Additionally, we discovered that students in 2022 rated their ability to use electronic patient record systems slightly lower than students in both 2021 and 2016. Although this finding was not statistically significant, we know that requirements for recording patient data have become more demanding, and the functionalities of the systems have increased [60]. These trends may have influenced the students' perceptions in 2022.

Prior research indicates that medical students have concerns regarding the impact of digitalization on the patient-doctor relationship [28,35]. Our study reveals a shift in attitudes related to the roles of patients and professionals. We discovered a statistically significant difference between 2016 and 2021 in the attitudes of participating students toward patients' roles in managing their health information and collaborating with professionals. Additionally, students in 2021 recognized the role of digitalization in shaping health care toward health promotion. These findings align with both national and EU health strategies, which emphasize electronic services to support the active role of citizens in maintaining their own well-being [1,61]. In Finland, the national health exchange services were complemented by a personal health record repository service (Kanta Personal Health Record or Kanta PHR), which entered its first-phase production in 2018 [19]. The service allows citizens to input, store, and share their well-being data with professionals. At the EU level, the EU is establishing a European Health Data Space ecosystem aiming to empower individuals of member states with control over their health data [9].

Overall, we saw slightly more reserved attitudes toward digital health in 2022 compared to 2021. The assertion to presentation 'The digitalization of healthcare is expected to significantly affect the practical work of healthcare professionals in the coming years' was notably less supported in 2022 compared to 2021 (see Figure 3, Q8). This trend suggests that electronic health services and tools have become fully integrated into the health care system. It may even indicate that, from the students' perspective, digitalization has reached its apex in health care. This shift could be linked to the advancements in digital health education in the basic medical training in Finland, such as the completion of the national MEDigi project in 2021 and the introduction of its eHealth competence areas in 2020 [43,44].

Another noteworthy discovery was the significant increase in the number of students who agreed that "Digitalization shapes health care more toward health promotion" between 2021 and 2022, followed by a notable decrease in agreement between 2021 and 2022 (see Figure 4, Q11). This change in attitudes could be partly linked to the "care debt" and prolonged waiting times for treatment that emerged during the global health emergency caused by the COVID-19 pandemic in 2020 - 2023 [62,63]. It is also possible that after over 2 years of remote and blended education, students in 2022 may be exhibiting signs of digital fatigue [64,65]. As a result, the enthusiasm for digitalization driven by the COVID-19 pandemic may have started to decline by 2022. Nevertheless, these trends require further research to fully comprehend the underlying causes behind the phenomenon. The more reserved attitudes of the participating students and their somewhat lower self-assessed skills suggest the necessity for an increased focus on digital training for future health care professionals. This is crucial to ensure their competencies align with the broad health strategies in Europe and on a global scale.

The medical students who participated in this study in 2016, 2021, and 2022 were aligned in their belief that the basic education of health care professionals should include capacity building promoting the deployment of digital health solutions (see Figure 5, Q14). This consensus resonates with findings from previous studies [26-34]. The students' attitudes shifted positively toward alternative career paths such as product development (see Figure 5, Q16). This change was statistically significant between 2016 and 2021: in 2016, only 13% of students agreed with this claim, whereas almost a third of the students agreed with it in 2021. Although the difference between 2016 and 2022 was not statistically significant, the trend was consistent with percentages of 13% and 28%, respectively. This is an important finding, as end-user involvement is considered a critical success factor in information technology projects [66-69]. A Finnish study revealed that younger physicians were more eager to participate in health information system development compared to their older counterparts [70]. Additionally, research indicates that interdisciplinary

collaborations between health care and engineering professionals can foster innovation and new practices, underscoring the importance of interprofessional education [71,72]. However, introducing digital health topics and innovation activities into medical curricula has proven to be challenging in practice due to crowded curricula designs and competing interests [29,73-75].

Strengths and Limitations

Our overall response rate was 53%. Our study sample was collected from one of the 5 Finnish medical universities at 3 different time points. The results are likely to be applicable to other Finnish medical faculties given the similar surrounding health care system and relatively homogeneous education system [20,43]. Furthermore, these findings may also have relevance to other countries, particularly those with similar health care systems, health care information system architecture, and medical curriculum design. However, confirming this would necessitate further research.

There are also several limitations to this study. Firstly, we had a relatively restricted sample size from 1 education institution, which may affect the generalizability of the findings. While cross-sectional studies are useful for examining associations, causal relationships cannot be established, which should be considered when interpreting the findings. We relied on self-assessment to evaluate competencies in this study, which means we were unable to measure absolute changes in skills and knowledge. Because of minor variations in the collection of demographic characteristics among the study sample, a statistical comparison of the students' demographic characteristics could not be conducted. Furthermore, the dichotomization of variables may potentially hinder the reflection of results in real-world situations.

Future research should aim to include larger and more diverse samples from multiple institutions and possibly from other health care professions to enhance the generalizability of the findings. Additionally, longitudinal studies could provide more insight into causal relationships and how the attitudes and capabilities of the students may shift after entering working life. More information on the best teaching strategies for digital health topics is needed to facilitate optimal learning outcomes for health care professionals.

Conclusions

There has been a shift in self-perceived digital competence and attitudes of medical students over the years, potentially influenced by the development of national health information system architecture. However, this change has not followed a completely linear trajectory, and students' attitudes toward digital health have become somewhat more reserved in certain areas over time, particularly regarding the extent to which digitalization alters health care processes and its role in health promotion.

To address these gaps, medical educational institutions and policy makers should consider integrating more digital health topics into curricula, offer practical experience with digital health technologies, and implement effective teaching strategies for digital health. Given the rapid evolutions of the health care field, it is crucial to ensure the professionals keep up with the changes of this dynamic working environment. The integration of digital health education should be carefully considered and evaluated to ensure it meets the needs of both students and health care systems.

Acknowledgments

We would like to thank all the students who participated in this study as well as well as Lingsoft Language Services (Ltd) for English language editing of the manuscript.

Authors' Contributions

PV, TT, PK, EJ, JJ, AT, MM, TP, and JR were responsible for the conception and design of the study as well as the acquisition and interpretation of the data. PV conducted the statistical analyses and drafted the tables and figures. PV wrote the original draft conceptualization. TT, PK, EJ, JJ, AT, MM, TP, and JR reviewed and edited the manuscript. TP was responsible for the conception and design of the survey questions and the study as well as commenting on the manuscript. The final version of the manuscript was approved by all contributing authors.

Conflicts of Interest

The employer of PV, PK, TT, and JR received support for salaries from the MEDigi project funded by Finnish Ministry of Education and Culture (MEDigi OKM/270/523/2017). Otherwise, the authors declare no other known competing financial interests or personal relationships that may have influenced the work reported in this paper.

Multimedia Appendix 1
An English translation of the study questionnaire.
[PDF File, 119 KB - mededu_v11i1e67423_app1.pdf]

Multimedia Appendix 2
The results of the data analysis detailing the chi-square statistics (χ^2), degrees of freedom (df), P-values, phi coefficients (ϕ), and Q-values.

[PDF File, 134 KB - mededu_v11i1e67423_app2.pdf]

References

1. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on enabling the digital transformation of health and care in the Digital Single Market; empowering citizens and building a healthier society.: European Commission; 2018. URL: https://digital-strategy.ec.europa.eu/en/library/staff-working-document-enabling-digital-transformation-health-and-care-digital-single-market
2. Recommendations on Digital Interventions for Health System Strengthening - WHO Guideline: World Health Organization; 2019. URL: https://www.who.int/publications/i/item/9789241550505 [accessed 2025-04-06]
3. Woods L, Martin P, Khor J, Guthrie L, Sullivan C. The right care in the right place: a scoping review of digital health education and training for rural healthcare workers. BMC Health Serv Res 2024 Sep 2;24(1):1011. [doi: 10.1186/s12913-024-11313-4] [Medline: 39223581]
4. Lindberg MH, Chen G, Olsen JA, Abelsen B. Combining education and income into a socioeconomic position score for use in studies of health inequalities. BMC Public Health 2022 May 13;22(1):969. [doi: 10.1186/s12889-022-13366-8] [Medline: 35562797]
5. Marmot M, Friel S, Bell R, Houweling TAJ, Taylor S. Closing the gap in a generation: health equity through action on the social determinants of health. Lancet 2008 Nov;372(9650):1661-1669. [doi: 10.1016/S0140-6736(08)61690-6]
6. Agarwal R, Gao G, Desroches C, Jha A. The digital transformation of healthcare: current status and the road ahead. Inf Syst Res 2010 Dec 1;21:796-809. [doi: 10.1287/isre.1100.0327]
7. Digital Health and Innovation (DHI) WHO Team. Global Strategy on Digital Health 2020-2025: World Health Organization; 2021. URL: https://www.who.int/publications/i/item/9789240020924 [accessed 2025-04-06]
8. Publications Office of the European Union. Assessing the impact of digital transformation of health services: report of the expert panel on effective ways of investing in health (EXPH). European Commission. 2019. URL: https://health.ec.europa.eu/system/files/2019-11/022_digitaltransformation_en_0.pdf [accessed 2023-06-08]
9. Proposal for a Regulation of the European Parliament and of the Council on the European Health Data Space.: European Commission; 2022. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52022PC0197 [accessed 2023-11-28]
10. Liaw ST, Kuziemsky C, Farin H. Editorial: special issue on “The primary care Informatics response to COVID-19”. Int J Med Inform 2022 Apr;160:104690. [doi: 10.1016/j.ijmedinf.2022.104690] [Medline: 35067452]
11. Baudier P, Kondrateva G, Ammi C, Chang V, Schiavone F. Digital transformation of healthcare during the COVID-19 pandemic: patients’ teleconsultation acceptance and trusting beliefs. Technovation 2023 Feb;120:102547. [doi: 10.1016/j.technovation.2022.102547]
12. Budd J, Miller BS, Manning EM, et al. Digital technologies in the public-health response to COVID-19. Nat Med 2020 Aug;26(8):1183-1192. [doi: 10.1038/s41591-020-1011-4] [Medline: 32770165]
13. Wong BLH, Maaß L, Vodden A, et al. The dawn of digital public health in Europe: implications for public health policy and practice. Lancet Reg Health Eur 2022 Mar;14:100316. [doi: 10.1016/j.lanepe.2022.100316] [Medline: 35132399]
14. Digital Economy and Society Index (DESI) 2022: Thematic Chapters.: European Commission; 2022. URL: https://digital-strategy.ec.europa.eu/en/library/digital-economy-and-society-index-desi-2022 [accessed 2023-06-08]
15. Vehko T, Ruotsalainen S, Hyppönen H, editors. E-health and e-welfare of Finland: check point 2018 [Finnish]. : Finnish Institute for Health and Welfare; 2019 URL: https://www.julkari.fi/handle/10024/138244 [accessed 2023-11-11]
16. Vehko T, editor. E-Health and E-Welfare of Finland: Check Point 2022 [Finnish]: Finnish Institute for Health and Welfare; 2022. URL: https://urn.fi/URN:ISBN:978-952-343-891-0 [accessed 2023-11-05]
17. Jormanainen V, Reponen J. CAF and CAMM analyses on the first 10 years of national Kanta services in Finland. FinJeHeW 2020;12(4):302-315. [doi: 10.23996/fjhw.98548]
18. Jormanainen V, Vehko T, Lindgren M, Keskimäki I, Kaila M. Implementation, adoption and use of the Kanta Services in Finland 2010-2022. Stud Health Technol Inform 2023 May 18;302:227-231. [doi: 10.3233/SHTI230108] [Medline: 37203652]
19. Jormanainen V. Large-scale implementation and adoption of the Finnish national Kanta services in 2010–2017: a prospective, longitudinal, indicator-based study. Finnish J eHealth eWelfare 2018;10(4):381-395. [doi: 10.23996/fjhw.74511]
20. Merenmies J, Jääskeläinen J, Kortekangas-Savolainen O, Kulmala P, Nikkari S. Valmistuvan lääkärin osaamistavoitteet 10.6.2020 [Finnish]. URL: https://www.helsinki.fi/assets/drupal/2021-06/valmistuvan_laakarin_osaamistavoitteet.pdf [accessed 2025-05-29]
21. Outcomes for graduates 2018. General Medical Council. 2020. URL: https://www.gmc-uk.org/-/media/documents/dc11326-outcomes-for-graduates-2018_pdf-75040796.pdf [accessed 2023-12-14]
22. Harden RM. AMEE Guide No. 21: curriculum mapping: a tool for transparent and authentic teaching and learning. Med Teach 2001 Mar;23(2):123-137. [doi: 10.1080/01421590120036547] [Medline: 11371288]
23. Harden RM. AMEE Guide No. 14: outcome-based education: part 1 – an introduction to outcome-based education. Med Teach 1999 Jan;21(1):7-14. [doi: 10.1080/01421599979969]

24. EUR-Lex Access to European Union law. Directive 2005/36/EC of the European Parliament and of the Council of 7 September 2005 on the recognition of professional qualifications. : European Commission URL: https://eur-lex.europa.eu/eli/dir/2005/36/oj/eng [accessed 2025-04-06]
25. Konttila J, Siira H, Kyngäs H, et al. Healthcare professionals' competence in digitalisation: a systematic review. J Clin Nurs 2019 Mar;28(5-6):745-761. [doi: 10.1111/jocn.14710] [Medline: 30376199]
26. Ghaddaripouri K, Mousavi Baigi SF, Abbaszadeh A, Mazaheri Habibi MR. Attitude, awareness, and knowledge of telemedicine among medical students: a systematic review of cross-sectional studies. Health Sci Rep 2023 Mar;6(3):e1156. [doi: 10.1002/hsr2.1156] [Medline: 36992712]
27. El Kheir DYM, AlMasmoom NS, Eskander MK, et al. Perception of Saudi undergraduate medical students on telemedicine training and its implementation. J Family Community Med 2023;30(3):231-238. [doi: 10.4103/jfcm.jfcm_41_23] [Medline: 37675211]
28. Wernhart A, Gahbauer S, Haluza D. eHealth and telemedicine: practices and beliefs among healthcare professionals and medical students at a medical university. PLoS ONE 2019;14(2):e0213067. [doi: 10.1371/journal.pone.0213067] [Medline: 30818348]
29. Vossen K, Rethans JJ, van Kuijk SMJ, van der Vleuten CP, Kubben PL. Understanding medical students' attitudes toward learning eHealth: questionnaire study. JMIR Med Educ 2020 Oct 1;6(2):e17030. [doi: 10.2196/17030] [Medline: 33001034]
30. Seemann RJ, Mielke AM, Glauert DL, et al. Implementation of a digital health module for undergraduate medical students: a comparative study on knowledge and attitudes. THC 2022 Jun 16;31(1):157-164. [doi: 10.3233/THC-220138]
31. Ma M, Li Y, Gao L, et al. The need for digital health education among next-generation health workers in China: a cross-sectional survey on digital health education. BMC Med Educ 2023 Jul 31;23(1):541. [doi: 10.1186/s12909-023-04407-w] [Medline: 37525126]
32. Lotrean LM, Sabo SA. Digital health training, attitudes and intentions to use it among Romanian medical students: a study performed during COVID-19 pandemic. Healthcare (Basel) 2023 Jun 13;11(12):1731. [doi: 10.3390/healthcare11121731] [Medline: 37372849]
33. Veikkolainen P, Tuovinen T, Jarva E, et al. eHealth competence building for future doctors and nurses - attitudes and capabilities. Int J Med Inform 2023 Jan;169:104912. [doi: 10.1016/j.ijmedinf.2022.104912] [Medline: 36356432]
34. Machleid F, Kaczmarczyk R, Johann D, et al. Perceptions of digital health education among European medical students: mixed methods survey. J Med Internet Res 2020 Aug 14;22(8):e19827. [doi: 10.2196/19827] [Medline: 32667899]
35. Baumgartner M, Sauer C, Blagec K, Dorffner G. Digital health understanding and preparedness of medical students: a cross-sectional study. Med Educ Online 2022 Dec;27(1):2114851. [doi: 10.1080/10872981.2022.2114851] [Medline: 36036219]
36. Palm K, Brantnell A, Peolsson M, Özbek N, Hedström G. National eHealth strategies: a comparative study of nine OECD health systems. BMC Health Serv Res 2025 Feb 18;25(1):269. [doi: 10.1186/s12913-025-12411-7] [Medline: 39966936]
37. National Digital Health Strategy 2023-2028.: Australian Digital Health Agency; 2023. URL: https://www.digitalhealth.gov.au/national-digital-health-strategy/downloads [accessed 2025-03-30]
38. Feuille Route 2023-2027 - Supporting health through technology.: Ministry for Health - Digital Health Delegation; 2023. URL: https://gnius.esante.gouv.fr/sites/default/files/2025-01/Digital%20Health%20Roadmap%20France.pdf [accessed 2025-03-30]
39. eHealth Strategy for Ireland. : Department of Health (Ireland); 2024. URL: https://www.gov.ie/en/publication/0d21e-digital-for-care-a-digital-health-framework-for-ireland-2024-2030/ [accessed 2025-03-30]
40. Izquierdo-Condoy JS, Arias-Intriago M, Nati-Castillo HA, et al. Exploring smartphone use and its applicability in academic training of medical students in Latin America: a multicenter cross-sectional study. BMC Med Educ 2024 Nov 30;24(1):1401. [doi: 10.1186/s12909-024-06334-w] [Medline: 39616324]
41. Mosch L, Machleid F, Maltzahn F, et al. Digital health in the medical curriculum: addressing the needs of the future health workforce. European Medical Students' Association. 2019. URL: https://emsa-europe.eu/wp-content/uploads/2021/06/Policy-2019-04-Digital-Health-in-the-Medical-Curriculum_-Addressing-the-Needs-of-the-Future-Health-Workforce.pdf [accessed 2025-04-03]
42. Car J, Ong QC, Erlikh Fox T, et al. The digital health competencies in medical education framework: an international consensus statement based on a Delphi study. JAMA Netw Open 2025 Jan 2;8(1):e2453131. [doi: 10.1001/jamanetworkopen.2024.53131] [Medline: 39888625]
43. Levy A, Reponen J. Digital Transformation of Medical Education MEDigi Project Report: University of Oulu; 2021. URL: https://urn.fi/URN:ISBN:9789526232454 [accessed 2025-07-14]
44. Tuovinen T, Reponen J, Isoviita VM, et al. Sähköisten terveyspalveluiden opetus lääketieteessä [Finnish]. Duodecim 2021;137(17):1807-1813 [FREE Full text]
45. Price M, Lau F. The clinical adoption meta-model: a temporal meta-model describing the clinical adoption of health information systems. BMC Med Inform Decis Mak 2014 May 29;14(1):43. [doi: 10.1186/1472-6947-14-43] [Medline: 24884588]
46. Responsible conduct of research and procedures for handling allegations of misconduct in Finland.: Finnish National Board on Research Integrity; 2012. URL: https://tenk.fi/en/advice-and-materials/RCR-Guidelines-2012 [accessed 2023-11-05]

47. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016. EUR-LexAccess to European Union law.: European Union; 2016. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv%3AOJ.L_.2016.119.01.0001.01.ENG&toc=OJ%3AL%3A2016%3A119%3ATOC [accessed 2023-11-05]
48. Sharma A, Minh Duc NT, Luu Lam Thang T, et al. A Consensus-Based Checklist for Reporting of Survey Studies (CROSS). J Gen Intern Med 2021 Oct;36(10):3179-3187. [doi: 10.1007/s11606-021-06737-1] [Medline: 33886027]
49. Pihlajasalo T. Opiskelijoiden osaaminen ja asenteet sähköisiä terveyspalveluja ja terveydenhuollon digitalisoitumista kohtaan ennen ja jälkeen opetuspilotin [Students' skills and attitudes towards e-health services and the digitalisation of healthcare before and after the pilot]. : University of Oulu; 2017.
50. Chi square calculator for 2x2. Social Science Statistics. 2025. URL: https://www.socscistatistics.com/tests/chisquare/ [accessed 2025-04-06]
51. Effect size calculator. Statistic Kingdom. 2025. URL: https://www.statskingdom.com/effect-size-calculator.html [accessed 2025-04-06]
52. Munro BH. Statistical Methods for Health Care Research: Lippincott & Wilkins; 2005.
53. Cohen J. Statistical Power Analysis for the Behavioral Sciences: Lawrence Erlbaum Associates; 1988.
54. McDonald JH. Multiple Comparisons Handbook of Biological Statistics.: Sparky House Publishing; 2014. URL: https://www.biostathandbook.com/multiplecomparisons.html [accessed 2025-07-14]
55. Huikko-Tarvainen S, Tuovinen T, Kulmala P. Win–win practice: Finnish medical students' active role as doctors in the healthcare workforce regardless of background variables. JWAM 2025 Mar 18. [doi: 10.1108/JWAM-06-2024-0077]
56. Antonio MG, Petrovskaya O, Lau F. The state of evidence in patient portals: umbrella review. J Med Internet Res 2020 Nov 11;22(11):e23851. [doi: 10.2196/23851] [Medline: 33174851]
57. van der Storm SL, Jansen M, Meijer HAW, Barsom EZ, Schijven MP. Apps in healthcare and medical research; European legislation and practical tips every healthcare provider should know. Int J Med Inform 2023 Sep;177:105141. [doi: 10.1016/j.ijmedinf.2023.105141] [Medline: 37419042]
58. WHO - ITU mhealth hub in EU. CORDIS - EU research results.: European Union; 2022. URL: https://cordis.europa.eu/project/id/737427 [accessed 2023-11-02]
59. Essén A, Scandurra I, Gerrits R, et al. Patient access to electronic health records: differences across ten countries. Health Policy Technol 2018 Mar;7(1):44-56. [doi: 10.1016/j.hlpt.2017.11.003]
60. Haverinen J, Keränen N, Tuovinen T, Ruotanen R, Reponen J. National development and regional differences in eHealth maturity in Finnish public health care: survey study. JMIR Med Inform 2022 Aug 12;10(8):e35612. [doi: 10.2196/35612] [Medline: 35969462]
61. Information to support well-being and service renewal eHealth and eSocial Strategy 2020.: Ministry of Social Affairs and Health; 2015. URL: http://urn.fi/URN:ISBN:978-952-00-3575-4 [accessed 2025-07-14]
62. Living conditions and quality of life - COVID-19 and older people: impact on their lives, support and care. Eurofound.: Publications Office of the European Union; 2022. URL: https://www.eurofound.europa.eu/en/publications/2022/covid-19-and-older-people-impact-their-lives-support-and-care [accessed 2023-09-08]
63. Uimonen M, Kuitunen I, Paloneva J, Launonen AP, Ponkilainen V, Mattila VM. The impact of the COVID-19 pandemic on waiting times for elective surgery patients: a multicenter study. PLoS ONE 2021;16(7):e0253875. [doi: 10.1371/journal.pone.0253875] [Medline: 34228727]
64. Romero-Rodríguez JM, Hinojo-Lucena F, Kopecký K, García-González A. Digital fatigue in university students as a consequence of online learning during the Covid-19 pandemic. Educacion XX1 2023 Jun 14;26:165-184. [doi: 10.5944/educxx1.34530]
65. Lin Y, Yu Z. An integrated bibliometric analysis and systematic review modelling students' technostress in higher education. Behav Inf Technol 2025 Feb 25;44(4):631-655. [doi: 10.1080/0144929X.2024.2332458]
66. Høstgaard AM, Bertelsen P, Nøhr C. Methods to identify, study and understand end-user participation in HIT development. BMC Med Inform Decis Mak 2011 Sep 28;11:57. [doi: 10.1186/1472-6947-11-57] [Medline: 21955493]
67. Subramanyam R, Weisstein FL, Krishnan MS. User participation in software development projects. Commun ACM 2010 Mar;53(3):137-141. [doi: 10.1145/1666420.1666455]
68. He J, King WR. The role of user participation in information systems development: implications from a meta-analysis. J Manag Inf Syst 2008 Jul;25(1):301-331. [doi: 10.2753/MIS0742-1222250111]
69. Ayat M, Imran M, Ullah A, Kang CW. Current trends analysis and prioritization of success factors: a systematic literature review of ICT projects. Int J Manag Proj Bus 2021 Apr 6;14(3):652-679. [doi: 10.1108/IJMPB-02-2020-0075]
70. Martikainen S, Kaipio J, Lääveri T. End-user participation in health information systems (HIS) development: physicians' and nurses' experiences. Int J Med Inform 2020 May;137:104117. [doi: 10.1016/j.ijmedinf.2020.104117] [Medline: 32179254]
71. Kim HN. A conceptual framework for interdisciplinary education in engineering and nursing health informatics. Nurse Educ Today 2019 Mar;74:91-93. [doi: 10.1016/j.nedt.2018.12.010] [Medline: 30639937]
72. Zhou Y, Li Z, Li Y. Interdisciplinary collaboration between nursing and engineering in health care: a scoping review. Int J Nurs Stud 2021 May;117:103900. [doi: 10.1016/j.ijnurstu.2021.103900] [Medline: 33677250]

73. Keep M, Janssen A, McGregor D, et al. Mapping eHealth education: review of eHealth content in health and medical degrees at a Metropolitan Tertiary Institute in Australia. JMIR Med Educ 2021 Aug 19;7(3):e16440. [doi: 10.2196/16440]
74. Edirippulige S, Brooks P, Carati C, et al. It's important, but not important enough: eHealth as a curriculum priority in medical education in Australia. J Telemed Telecare 2018 Dec;24(10):697-702. [doi: 10.1177/1357633X18793282]
75. Arias J, Scott KW, Zaldivar JR, et al. Innovation-oriented medical school curricula: review of the literature. Cureus 2021 Oct;13(10):e18498. [doi: 10.7759/cureus.18498] [Medline: 34754659]

Abbreviations

CAMM: clinical adoption meta-model
EU: European Union
WHO: World Health Organization

Edited by D Chartash; submitted 11.10.24; peer-reviewed by E Ortiz-Prado, P Putz; revised version received 12.05.25; accepted 12.05.25; published 31.07.25.

Please cite as:
Veikkolainen P, Tuovinen T, Kulmala P, Jarva E, Juntunen J, Tuomikoski AM, Männistö M, Pihlajasalo T, Reponen J
The Evolution of Medical Student Competencies and Attitudes in Digital Health Between 2016 and 2022: Comparative Cross-Sectional Study
JMIR Med Educ 2025;11:e67423
URL: https://mededu.jmir.org/2025/1/e67423
doi:10.2196/67423

Original Paper

Refining Established Practices for Research Question Definition to Foster Interdisciplinary Research Skills in a Digital Age: Consensus Study With Nominal Group Technique

Jana Sedlakova[1,2], PhD; Mina Stanikić[1,2,3], MD, PhD; Felix Gille[1,2], PhD; Jürgen Bernard[1,4], Prof Dr; Andrea B Horn[1,5,6], Dr rer nat; Markus Wolf[1,6], Dr phil; Christina Haag[1,2,3], PhD; Joel Floris[1,7], Dr; Gabriela Morgenshtern[1,4], MSc; Gerold Schneider[1,8], Prof Dr; Aleksandra Zumbrunn Wojczyńska[1,9], Dr; Corine Mouton Dorey[1,10], Dr med, Dr phil; Dominik Alois Ettlin[1,9], PD, Dr med; Daniel Gero[1,11], PD, MD, PhD; Thomas Friemel[1,12], Prof Dr; Ziyuan Lu[1,7], Dr med; Kimon Papadopoulos[1,13], MPH; Sonja Schläpfer[1,14], MSc; Ning Wang[1], PhD; Viktor von Wyl[1,2,3], Prof Dr

[1]Digital Society Initiative, University of Zurich, Zurich, Switzerland

[2]Institute of Implementation Science in Healthcare, Faculty of Medicine, University of Zurich, Zurich, Switzerland

[3]Epidemiology, Biostatistics and Prevention Institute, Faculty of Medicine, University of Zurich, Zurich, Switzerland

[4]Department of Informatics, Faculty of Business, Economics and Informatics, University of Zurich, Zurich, Switzerland

[5]Center for Gerontology, University of Zurich, Zurich, Switzerland

[6]Department of Psychology, Faculty of Arts and Social Sciences, University of Zurich, Zurich, Switzerland

[7]Institute of Evolutionary Medicine, Faculty of Medicine, University of Zurich, Zurich, Switzerland

[8]Department of Computational Linguistics, Faculty of Business, Economics and Informatics, University of Zurich, Zurich, Switzerland

[9]Center of Dental Medicine, Faculty of Medicine, University of Zurich, Zurich, Switzerland

[10]Institute of Biomedical Ethics and History of Medicine, Faculty of Medicine, University of Zurich, Zurich, Switzerland

[11]Department of Surgery and Transplantation, University Hospital of Zurich, Zurich, Switzerland

[12]Department of Communication and Media Research, Faculty of Arts and Social Sciences, University of Zurich, Zurich, Switzerland

[13]Institute of Implementation Science in Healthcare, Faculy of Medicine, University of Zurich, Zurich, Switzerland

[14]Institute for Complementary and Integrative Medicine, University Hospital of Zurich, Zurich, Switzerland

Corresponding Author:
Jana Sedlakova, PhD
Digital Society Initiative
University of Zurich
Raemistrasse 69
Zurich, 8001
Switzerland
Phone: 41 0786753991
Email: sedlakova@ifi.uzh.ch

Abstract

Background: The increased use of digital data in health research demands interdisciplinary collaborations to address its methodological complexities and challenges. This often entails merging the linear deductive approach of health research with the explorative iterative approach of data science. However, there is a lack of structured teaching courses and guidance on how to effectively and constructively bridge different disciplines and research approaches.

Objective: This study aimed to provide a set of tools and recommendations designed to facilitate interdisciplinary education and collaboration. Target groups are lecturers who can use these tools to design interdisciplinary courses, supervisors who guide PhD and master's students in their interdisciplinary projects, and principal investigators who design and organize workshops to initiate and guide interdisciplinary projects.

Methods: Our study was conducted in 3 steps: (1) developing a common terminology, (2) identifying established workflows for research question formulation, and (3) examining adaptations of existing study workflows combining methods from health research and data science. We also formulated recommendations for a pragmatic implementation of our findings. We conducted

a literature search and organized 3 interdisciplinary expert workshops with researchers at the University of Zurich. For the workshops and the subsequent manuscript writing process, we adopted a consensus study methodology.

Results: We developed a set of tools to facilitate interdisciplinary education and collaboration. These tools focused on 2 key dimensions— content and curriculum and methods and teaching style—and can be applied in various educational and research settings. We developed a glossary to establish a shared understanding of common terminologies and concepts. We delineated the established study workflow for research question formulation, emphasizing the "what" and the "how," while summarizing the necessary tools to facilitate the process. We propose 3 clusters of contextual and methodological adaptations to this workflow to better integrate data science practices: (1) acknowledging real-life constraints and limitations in research scope; (2) allowing more iterative, data-driven approaches to research question formulation; and (3) strengthening research quality through reproducibility principles and adherence to the findable, accessible, interoperable, and reusable (FAIR) data principles.

Conclusions: Research question formulation remains a relevant and useful research step in projects using digital data. We recommend initiating new interdisciplinary collaborations by establishing terminologies as well as using the concepts of research tasks to foster a shared understanding. Our tools and recommendations can support academic educators in training health professionals and researchers for interdisciplinary digital health projects.

(JMIR Med Educ 2025;11:e56369) doi:10.2196/56369

KEYWORDS

research question; digitalization; digital data; data science; health research; interdisciplinary

Introduction

Background

Health research increasingly leverages the abundance of data from our "digital lives," including mobility data, social media data, or data from wearables [1,2]. Such digital data are commonly "unstructured" because it may not conform to a tabular format (eg, images, videos, sound, and free text) and often require specific expertise for harvesting; transforming; preprocessing; and creating meaningful insights into health, disease, and treatment [1,3-5]. Moreover, such digital data are often originally generated for nonresearch purposes and without addressing a specific research question [6]. In turn, they may lack standard quality attributes found in digital data collected for specific research purposes, such as depth, completeness, or consistency, which present methodological complexities to meaningfully use these data [1]. Therefore, reusing these digital unstructured data for health research requires diverse expertise, skills, and interdisciplinary collaboration between health domain experts (eg, clinicians and health scientists) and data scientists as method experts (eg, from data science, computer science, or statistics) [5,7,8].

Such interdisciplinary collaborations are often faced with challenges due to the seemingly conflicting research approaches between the disciplines. In addition to differences in terminologies and concept definitions, the prevailing emphasis of linear deductive approaches in health research contrasts with the often more explorative and iterative approaches used in data science [7]. In health research, it is customary to predefine key elements of the scientific process, including a research question and related hypothesis, in a protocol and scientific report (eg, STROBE [Strengthening the Reporting of Observational Studies in Epidemiology] or PRISMA [Preferred Reporting Items for Systematic Reviews and Meta-Analyses] guidelines) [9-12]. These standard practices are deeply influenced by the tradition of clinical trials and treatment development, which place a strong emphasis on measurement validity, robustness, scientific rigor, and safety [13], as errors in study conduct or treatment could place study participants at risk. By contrast, data science generally tends to emphasize exploration, pattern discovery, or hypothesis generation as well as more iterative and inductive analysis approaches [1,14]. Some health researchers may perceive this greater emphasis on iterative approaches as lacking scientific rigor or focus on specific research questions.

For young researchers, interdisciplinary digital health collaborations might be particularly challenging because they need to balance traditional scientific methods with more iterative data-driven techniques. This dual demand highlights the importance of fostering interdisciplinary skills in education, enabling students to balance the rigorous demands of hypothesis-driven research with the iterative and inductive approaches of data science. Addressing these complexities represents an educational challenge for both established and young researchers.

Despite broad recognition of their importance, both practical and teaching or educational guidance on how to manage and overcome the challenges of interdisciplinary digital health collaborations are scarce. Such guidance is also important for educational purposes to foster skills for interdisciplinary collaboration among both young and established researchers as well as health professionals. Continuous education for experienced researchers is equally important to keep them updated with evolving methods and foster effective collaboration across disciplines.

This Study

To address this need, our study focuses on skill development to successfully navigate interdisciplinary collaborations and education in health-related research fields. We reviewed established workflows for research question formulation and investigated whether and how established workflows in health research may require adaptations to accommodate inductive and exploratory data science practices and novel analysis techniques. The study findings were translated into a set of tools and recommendations designed to facilitate interdisciplinary education and collaboration. These tools focus on 2 key

dimensions—*content and curriculum* and *methods and teaching style*—and can be applied in various educational and research settings. Lecturers can use them to design interdisciplinary courses, supervisors can guide PhD and master's students in their interdisciplinary projects, and principal investigators can design and organize workshops to initiate and guide interdisciplinary projects. By implementing these tools, educators and researchers can create more cohesive and productive educational resources for interdisciplinary collaborations. In the following sections, we offer our insights and a more detailed outline of how our study findings can inform both the *content* and *methods* dimensions, using an existing interdisciplinary course as an example.

The aims and findings of our study are intended to be globally relevant and applicable to all researchers using digital data in the context of health research and health care. Importantly, they also provide academic educators with a clear workflow and practical recommendations for discussing and addressing the challenges of interdisciplinary collaboration. As the focus is on research question formulation, a fundamental aspect of the research process, these recommendations are especially valuable for educational purposes, helping educators guide researchers and students through this essential phase of research projects.

To achieve our aims, we chose a consensus study approach that is appropriate to harmonize and bridge insights from experts from diverse research disciplines. Moreover, we focused our effort on the different approaches of research question formulation as the guiding example for this study because it represents a central step in guiding the research process and subsequent study design decisions. This process also served as an illustrative example to highlight the differences in research approaches between health research and data science.

Methods

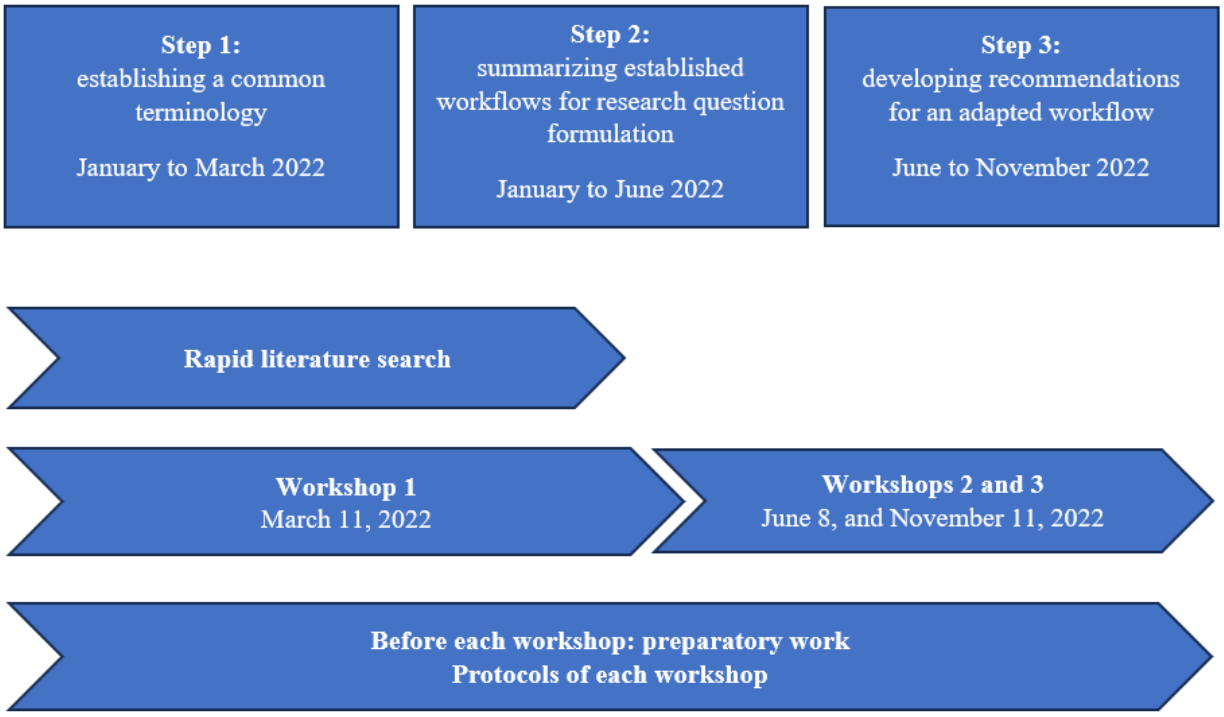
Consensus Methods

We used the nominal group technique with expert groups to gather insights from a diverse range of experts. This approach aimed to foster interdisciplinary skills and knowledge and achieve consensus on adapting research question development.

This study was structured by the following three high-level steps (Figure 1):

1. To create a common terminology to facilitate interdisciplinary and transdisciplinary collaborations that are required for research projects reusing digital data (ie, repurposing data originally generated for nonresearch purposes)
2. To describe the "established workflow" for research question formulation in health research on the basis of existing literature
3. To formulate suggestions and recommendations for adapting the "established workflow"

Figure 1. Study flow.



To inform steps 1 and 2, a rapid literature review was performed to identify established concepts for defining a research question in health research and data science as well as in other fields (refer to the Preparatory Research: Literature Search for the "Established Workflow" and an Example Scenario section). Expert inputs were gathered in a series of three 1.5-hour expert workshops. To foster a focused discussion in the workshops, participants were asked to complete preworkshop tasks. These inputs were summarized by JS and VvW and presented at the beginning of each workshop to discuss potential disagreements and allow participants to explain or comment on their and others' inputs. The consensus and agreement of each objective

were reached by an iterative, deliberative process. This included expert inputs before workshops, discussions during the workshop, and finally, expert feedback on and approval of the consolidated findings. These findings were synthesized, formulated, and shared by JS and VvW after each workshop. Furthermore, each participant was actively involved in the manuscript writing. These methods facilitated the systematic collection of input from participants in group and individual settings, enabling a comprehensive understanding of experts' knowledge and consolidating diverse perspectives. Workshops were recorded after receiving consent from the team. Workshop minutes, including the results of the preworkshop tasks, were sent for approval to the expert group. When necessary, individual researchers were contacted after workshops for clarification on specific issues raised during the workshops. The documentation and reporting of the workshop and the Accurate Consensus Reporting Document (ACCORD) checklist [15] for the consensus methodology are available in Multimedia Appendices 1 and 2.

The first 2 steps were accomplished in workshops 1 and 2. Building on these results, a third workshop was dedicated to identifying the need for the adaptation of established research practices in the health field to streamline collaboration with data scientists and to better integrate and communicate the need for research principles and standards, including open science and reproducibility.

Participants

The consensus meetings in the form of expert groups were led by JS and VvW, who led a previous project focusing on challenges and best practices of digital data, which inspired this study. Furthermore, JS is a scientific manager of the scientific community whose members were recruited for the consensus exercise. VvW's expertise lies in epidemiology and digital health research, and JS's expertise is mainly in digital ethics considering health research and health care. The workshop participants were recruited by JS and VvW among the diverse members of the Digital Society Initiative (DSI) Health Community at the University of Zurich. The members from the DSI were selected because it is a competence center for digital transformation that fosters interdisciplinary collaborations and projects studying the interplay and implications of digital transformations in society. Participants were included if they had experience with projects using digital data or planned to be involved in such projects. The workshop was promoted on the DSI website, through newsletters, and via word-of-mouth within the community. A total of 21 researchers from different disciplines and from all career stages participated in the workshops. This number of participants enabled to have an expert group with sufficient diversity to foster discussions and include insights from diverse disciplines. Of the 21 researchers, 13 (62%) represented health research, 3 (14%) represented data science, and 7 (33%) represented the social sciences and humanities.

Preparatory Research: Literature Search for the "Established Workflow" and an Example Scenario

A rapid literature search was conducted to inform the planning of the workshops and to develop a project roadmap (by JS and VvW). To gather information on the established workflow for research question formulation (steps 1 and 2), we searched the literature for publications, reviews, and course guidelines written either in English or German in PubMed and Google Scholar databases (search terms are provided in Multimedia Appendix 1). The search was further complemented by retrieval of guidelines from universities in Switzerland, Germany, the United States, and the United Kingdom, for which we searched on selected university websites. In addition, coauthors contributed materials they were familiar with or had previously used for teaching or research purposes. On the basis of this literature, we proposed the initial model for the "established workflow" that combines existing well-established frameworks and practices. To guide the discussions of our workshops, we developed an example scenario of digital data reuse for health research, which was communicated to participants before the workshops (Multimedia Appendix 3).

Ethical Considerations

The study followed the recommended procedures of the ethics committee of the Medical Faculty of the University of Zurich by completing the Data Protection/Ethics Self-Assessment Tool and received an exempt status. Participants were informed about the study's scope and goals as well as the nature of their involvement. They provided consent before the workshop and were informed that they could withdraw from the study at any time without providing a reason. The participants did not receive any compensation for their participation. The only personal information collected for the study was sociodemographic data, which were anonymized.

Results

Establishing a Common Terminology

Anticipating that a lack of harmonization concerning terminologies and concepts may hinder an effective interdisciplinary workshop collaboration, we aimed to establish a shared understanding of common terminologies. To this end, the workshop leaders (JS and VvW) developed a glossary before the first workshop, which was discussed and further refined by collecting written feedback from the participants after workshop 1 (Table 1).

The workshop discussions concerning the glossary centered around discipline-specific interpretations of concepts such as "research task," "research objectives," "research aims," and "research goals," whose interpretations were dependent on the embedding in different research methodologies, such as qualitative or quantitative research approaches. A central discussion centered around the recognition of different "research tasks," that is, high-level research aims from a methodological viewpoint, including, for example, exploration, confirmation, prediction, methods development, or theory development. For prediction and classification tasks, participants mentioned 2 subcategories of analyses, which are supervised learning methods that rely on labeled data and outcomes and include the broad class of (multivariable) regression models. By contrast, unsupervised methods (eg, neural networks) aim to find new data structures and features without the need for prior labeling and are often developed in a less linear, inductive manner.

Table 1. Common terminology for interdisciplinary research projects using digital data.

Terms	Definitions
Confirmatory research	• Hypothesis-driven research, experimental research, or research aiming at testing and confirming a hypothesis in a broader context of a theory. This research is also referred to as hypothetico-deductive research in some disciplines.
Exploratory research	• Data-driven research that aims at exploring new patterns and associations to formulate hypotheses.
Hypothesis	• A tentative, hypothetical prediction of the nature and direction of relationships between sets of data, phrased as a declarative statement. It is an assumption about scientific laws, causation, or empirical regularities. A hypothesis should be testable or falsifiable. This refers to quantitative evidence-based health research.
Unstructured data	• Raw data that are not in a predefined structure (eg, tables) or data that may be structured but still require substantial preprocessing or feature extraction (eg, continuous sensor data).
Principles and criteria of good research and research practice	• A set of values and norms for good conduct of research, including validity, scientific integrity, objectivity, and ethical study conduct.
Research aim	• The research aim is the overall, general, and long-term intention of a research project. The research aim describes the "what" of the research—where we aspire to be at the end.
Research design	• Research design describes the general outline of data collection (eg, cross-sectional and longitudinal studies) and analytical methods (eg, randomization, observational, and with or without control group) to answer the RQ[a]. It describes the "how" of research.
Research objective	• The specific goal linked to a RQ [16].
Research problem	• The research problem describes the rationale for a study, for example, by highlighting the societal or medical needs. It describes the "why"—the specific needs a study wants to address.
RQ	• A clear and concise question determining the research aim, objective, design, methodology, data collection, and analysis. The RQ narrows the aim and objective of the research. The process of defining a good RQ is dynamic and iterative. The RQ is refined through the different steps of the research cycle. We define the RQ in the context of quantitative evidence-based health research.
Research task	• A research task describes a high-level classification of aims or tasks in research, including descriptive research, exploratory research, confirmatory research, prediction and classification, theory development, or methods development.
Reuse of digital data	• The process of harvesting, transforming, and using structured or unstructured digital information that was initially generated for purposes other than research.
Theory or model	• A systematic, structured explanation or representation of facts, phenomena, or processes that sets the ground for research design, formulation of hypotheses, and predictions.
Tools to specify RQ	• Frameworks and tools that facilitate the development of specific aspects of defining the RQ or study design.
Types of RQ	• The type of RQ determines the main approach for achieving the research aim. Usually, there is a difference between quantitative and qualitative RQs that reflect quantitative and qualitative approaches. • Quantitative approaches use statistical and mathematical methods to address precise questions, typically using a deductive approach with a strong emphasis on the framework and structure. • By contrast, qualitative approaches use, for example, open-ended responses, focus groups, and interview-based techniques and focus on individual experiences and singularities. It seeks to determine or discover a process or define experiences. RQs tend to be inductive, flexible, adaptable, and nondirectional [17].

[a]RQ: research question.

The group further discussed the central role of hypotheses and linear, highly structured research approaches in health research, for example, in confirmatory research tasks (confirming a hypothesis, eg, by use of randomized controlled experiments or trials) and, to some extent, research focusing on predictions tasks (developing prediction models or classifiers to predict future events or out-of-sample attributes). In health research, it is generally recommended that the development of prediction methods follows a protocol that includes careful selection of predictors and (external) validation of the final model [18]. At the same time, it was also pointed out that some quantitative research tasks, such as methods development (ie, the development and validation of data analysis methods) or exploratory research (ie, detection of patterns and associations

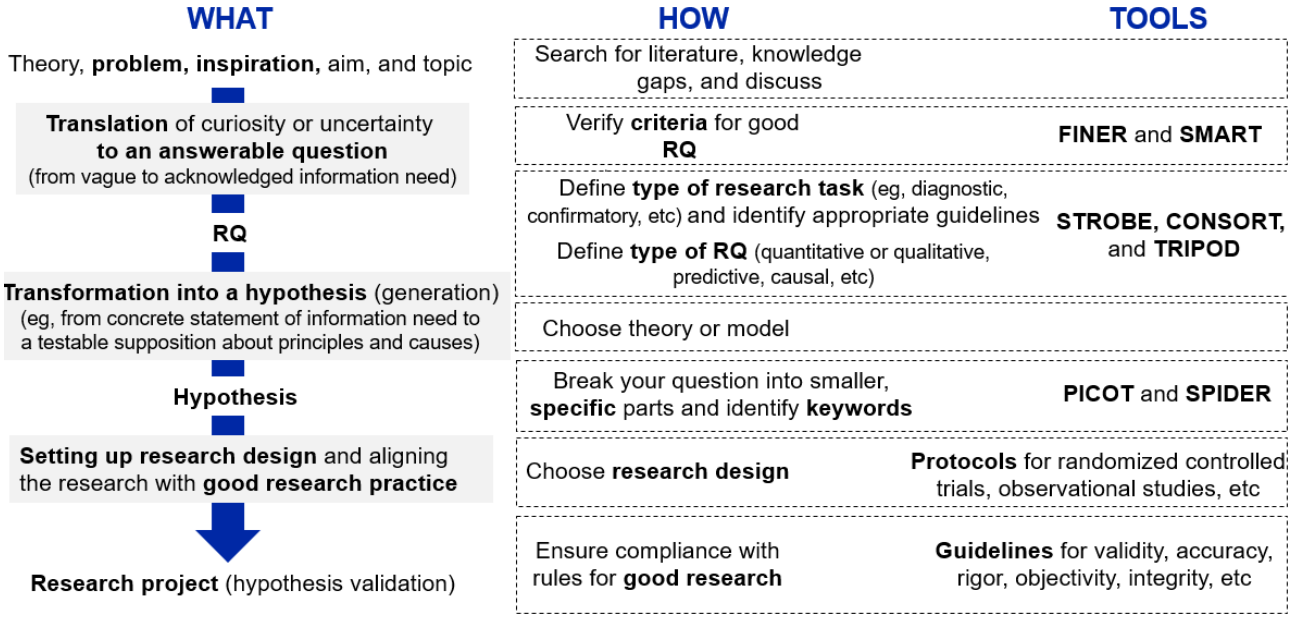
to generate new hypotheses), as well as qualitative research, generally depend much less on the specification of hypotheses. Workshop participants with a qualitative background argued that hypotheses can be implicitly involved in the research project. In qualitative research, it is common for the research question to evolve due to the necessity to critically reflect and adjust the study focus in each research step. As a result, the overall research process in qualitative research and some quantitative tasks, such as methods development, can be more iterative and dialectical when compared to deductive or confirmatory health research.

These discussions led to a key insight that interdisciplinary collaborations may be streamlined through the identification and discussion of the most appropriate "research task" early on, which can help guide subsequent discussions about the research question and the role of hypotheses in a common direction.

Summarizing Established Workflows for Research Question Formulation

The first 2 workshops were dedicated to better understanding how different disciplines approach the initial steps of a research project, including research question definition and study design choices. Informed by our rapid literature review, Figure 2 illustrates a summary workflow for established research design practices in health research. The vertical axis of Figure 2 illustrates the recommended steps for defining a research question (the "what"), starting from finding inspiration to developing a hypothesis, designing an appropriate study, and validating the hypothesis. Aligned with these definition steps, Figure 2 displays established practices (the "how") to execute the recommended steps. The third column references various frameworks and checklists aiding the implementation of each recommended step (the "tools").

Figure 2. Workflow of the recommended practices for defining a good research question. CONSORT: Consolidated Standards of Reporting Trials; FINER: Feasible, Interesting, Novel, Ethical, and Relevant; PICOT: Population, Intervention, Comparison, Outcome, and Time; RQ: research question; SMART: Specific, Measurable, Achievable, Realistic, and Timely; SPIDER: Sample, Phenomenon of Interest, Design, Evaluation, Research Type; STROBE: Strengthening the Reporting of Observational Studies in Epidemiology; TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis.



The workflow of established practices generally starts with identifying a meaningful problem or question to be addressed in a study. The inspiration often emerges from a real-world challenge or knowledge gaps, but it can also be derived from existing theories or be triggered by discussions among colleagues. Workshop members also mentioned the influential role of funding criteria (ie, to increase chances for funding success) or topic-specific funding calls. This inspiration, curiosity, or uncertainty then needs to be translated into an answerable question [19-21]. Although we found little guidance in the literature on how to operationalize this step, it is often recommended to check the research question against the FINER (feasible, interesting, novel, ethical, and relevant) and SMART (specific, measurable, achievable, realistic, and timely) quality attributes to ensure its suitability for testing in a research study [16,20,22,23].

The wording of the research question itself may already imply a specific research task (eg, exploratory, confirmatory, or qualitative research) [21]. We differentiate between research aim and research objective. Research aim is the overall goal of the research, whereas research objective is the specific goal linked to a research question [16]. Having clarity on the research task will also facilitate the identification of appropriate reporting guidelines, such as STROBE (for observational studies), CONSORT (Consolidated Standards of Reporting Trials; for randomized controlled studies), or TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis; for the development of prediction models). These reporting guidelines primarily intend to guide the communication of study results but can also be useful in converting the research question into a study design.

Ultimately, decisions regarding the study design should be guided by the research question, while also considering practical

limitations and available means and resources [23]. Frameworks such as PICOT (population, intervention, comparison, outcome, and time) and similar tools (eg, SPIDER [sample, phenomenon of interest, design, evaluation, research type] for qualitative research) [9,20,24] provide useful starting points for defining the study design. We include both the FINER and PICOT tools and their equivalents to ensure the best possible quality of the research question. Some research studies have shown that using only PICOT might be suboptimal [23]. The PICOT framework is frequently applied in health research, as PICOT is already defined above [21]. Further high-level study design decisions concern the study duration and measurement frequency (longitudinal vs cross-sectional studies), the allocation of study participants into comparator groups (randomization vs "as is" in observational research), as well as numerous practical aspects concerning study execution (eg, sample size and methods of data collection and analysis) [21,24]. Study design and study execution also need to adhere to the principles and criteria of good research and research practice to achieve valid, reliable, and accurate results [25]. Moreover, the research must comply with the standards of objectivity, reproducibility, and research integrity [26,27].

Overall, the workshop discussions confirmed that the workflow summary (Figure 2) represents a useful starting point for interdisciplinary collaborations to illustrate the established practices and to explore conceptual differences between health research, data science, and other scientific disciplines.

Developing Recommendations for an Adapted Workflow

Building on the proposed example scenario for using digital data in health research (Multimedia Appendix 3) and the established workflow description from step 1, the participants then discussed 2 types of workflow adaptations to better reflect practices and approaches from data science (Figure 2). These included the following: (1) structural modifications by changing the sequence of workflow steps (ie, introducing additional steps that should become standard in a novel workflow—the "what") and (2) the need for introducing additional contextual constraints or novel quality criteria (ie, modifications that do not change the workflow but may impact their execution—the "how").

Modifications to the "What": The Steps in Research Questions Workflow

Overall, the workshop participants perceived that the general sequence of the established workflow (the "what") still applies to studies using (structured and unstructured) digital data. Complementary steps with their potential pitfalls were proposed to better reflect the additional challenges of working with digital unstructured data (Multimedia Appendix 4.

First, for unstructured data, preprocessing and feature extraction should be allocated a distinct workflow step to emphasize the need for thorough consideration during study planning and execution, to ensure that the data are usable, credible, and useful for the research question at hand [28-31]. On one hand, the assessment of data quality and validity is more challenging. On the other hand, preprocessing and feature extraction through machine learning require additional assumptions and may lead to predictions and derived parameters with uncertain distributional characteristics (eg, normal ranges) or propagation of algorithmic errors and biases.

Second, the selection of appropriate analysis methods to address the research question as a new workflow step would underscore the importance of scientific rigor [1,31-38]. For example, deciding between pretrained deep learning models requires preliminary investigations about the model features and the training database, which goes beyond the choice of more standard statistical techniques (eg, regression models) [39,40].

Finally, the general importance of efforts to render science reproducible and transparent was identified as a new step in the workflow.

Modifications of the "How" of the Research Question Workflow

The workshop group identified potential contextual and methodological changes to research practices (the "how"; Table 2). These proposed changes can be grouped into 3 clusters.

Table 2. Proposed changes to the "how" parts of the research question formulation workflow.

Change number	Contextual constraints and quality criteria	Description	Steps this is applicable to
I	Consider real-life incentives and constraints in defining research problems	• The decision about RQ[a] can be strongly influenced by other nonacademic factors such as the availability of funding or data.	RQ
II	Acknowledge feasibility and resource constraints	• The choice of research design and data analysis tools involves costs that must be considered, particularly to ensure compliance with scientific integrity.	Research design
III	Declare limitations in RQ scope	• Each RQ has limitations; it is important to define what RQ can and cannot answer.	RQ
IV	Allow for and document iterations in RQ development and analysis	• Proper documentation is important for ensuring transparency and helps with evaluating and tracking the decisions regarding the iterations in RQ.	RQ
V	Acknowledge and respond to the increasing need for interdisciplinary expertise	• For the feasibility of the RQ, it is important to consider the needed expertise and skills. This becomes particularly important in research involving digital unstructured data as it requires an interdisciplinary set of skills.	All steps
VI	Enhance reproducibility	• Reproducibility in data science means obtaining consistent results using the same input data and methods. On a higher level, reproducibility in science also refers to the ability to duplicate findings if the same methods are used [41]. • Reproducibility in science also refers to the concept of making data; computational steps, methods, and codes; and conditions of analysis transparent and available, so that others can verify the findings.	All steps
VII	Enhance replicability	• Replicability refers to applying the same methods from a different study on different data. Observed differences in findings should be explicable by data-specific differences between studies.	All steps
VIII	Enhance robustness	• Robustness refers to analyses that apply the same database but use different methods. Observed differences in findings should be explicable by method-specific differences between studies. Within the same study, robustness is often evaluated by sensitivity analyses that use the same data but vary methods (eg, by applying different model parameters).	All steps
IX	Critically assess generalizability	• Generalizability means that the study results or outcomes are also applicable in other study settings and samples.	All steps

[a]RQ: research question.

The first cluster includes the acknowledgment of what we have labeled as real-life constraints (change numbers I and II) and limitations in the scope of research questions (change number III). Appropriately addressing such real-life constraints can be fostered by greater transparency and experience exchange.

The second cluster of proposed contextual changes pertains to enabling more interdisciplinary and iterative workflows (change numbers IV and V). Reasons for iterative approaches include more complex choices of analysis methods, the need for verifying the validity and robustness of model results, or the need to manually search for the best model parametrization. These challenges also require a greater emphasis on interdisciplinary collaborations that combine subject-domain knowledge and data science expertise.

The final cluster reflects the need for strengthening research quality criteria to foster open science, better reproducibility, and greater transparency (change numbers VI-IX). As analytical methods and databases become more complex, there is also an increasing need for transparency; adequate documentation; as well as publicly available analysis protocols, software codes, data, and analysis files. Studies should critically examine their findings under changing data or method combinations, thus exploring reproducibility, robustness, replicability, or generalizability (the 3RG criteria) and enhancing the overall quality of research. An important means to achieve these goals are open science and Findable, Accessible, Interoperable, and Reusable (FAIR) data principles [42].

Recommendations Toward a Pragmatic Approach of Teaching and Conducting Research Question Formulation

The workshop discussions produced a set of specific recommendations to promote approaches for defining good

research questions for reusing digital data. These recommendations are also well suited for educational use, helping to navigate the challenges of interdisciplinary collaboration and to foster interdisciplinary skills.

Iterative Research Question Formulation

As a principle, data collection, preprocessing, and analysis methods should follow the research question; researchers should not lose sight of the research aim, objective, and question. Defining a good research question is a fundamental and universal first step of science, which ideally should not be preceded by the choice of data or methods. However, the linear process of defining a research question common to health research may need several iterations to ensure that the complexity and feasibility of reusing and integrating digital data are accounted for.

The lecture instructors, supervisors, and principal investigators of interdisciplinary projects can apply this recommendation by emphasizing the importance of research question formulation in interdisciplinary projects. Furthermore, they can facilitate a discussion or create exercises for students to practice how the linear process of research question definition changes into a more iterative process when collaborating with other disciplines.

Reconciling Linear And Iterative Approaches: Continuum of Research Tasks

To reconcile the apparent conceptual differences between health research and data science approach research projects, we propose to reframe the scientific process as a continuum of knowledge accumulation over the course of multiple studies. Such a continuum can consist of several different research tasks (projects) combining deductive and inductive research approaches. Not all research tasks will involve explicit research questions or hypotheses. However, systematic reflections on how study results can inform new hypotheses and research questions and how they could be tested in future studies could become an integral part of a study, for example, as a last step in exploratory analyses.

Lecture instructors, supervisors, and principal investigators of interdisciplinary projects can use this recommendation to emphasize the continuum of research tasks. They can create exercises consisting of different research tasks where students practice combining deductive and inductive approaches in research design. These exercises can guide students to recognize that research is not always a straightforward process of hypothesis testing but may involve exploratory tasks that inform future studies. Instructors can also encourage students to reflect systematically on their research results, guiding them to think about how current findings can shape future hypotheses and research directions. This reflection can be incorporated into project work, where students work on iterative research tasks, examining how knowledge accumulates across studies and how inductive and deductive methods interact throughout this process. This practice prepares students to handle the nonlinear nature of interdisciplinary research, especially when bridging health research and data science.

Research Quality Criteria

The complexities involved in digital data preprocessing and analysis require careful design decisions and thorough reporting to ensure adherence to research quality standards. The reuse of existing, digital unstructured data and the need for extensive preprocessing may obfuscate or compound issues of external and internal validity [14]. Moreover, the use of machine learning techniques such as deep neural networks may generate "unexplainable" predictions or classifications that challenge the transparency and open science paradigms. The verification of "whether the data measure what they are supposed to measure (in the context of the research question)" [14] remains crucial and deserves appropriate attention, but it may become more difficult to achieve. We recommend that researchers systematically scrutinize interim results to ensure that they are "on the good track." Such checks can, for example, include the replication of results from different studies. Furthermore, transparency in reporting and reproducibility are key to scientific rigor.

Lecture instructors and supervisors can emphasize the importance of maintaining research quality in interdisciplinary projects. They can design exercises where students practice making careful design decisions in their research projects, ensuring that issues of validity, transparency, and reproducibility are addressed throughout the process. One approach could be to guide students in developing protocols for systematic checks of their interim results. Instructors can also promote transparency by teaching students how to document their research processes thoroughly, facilitating reproducibility and open science principles. By applying these exercises, students learn to critically evaluate the quality of their research.

Take Active Measures to Foster Interdisciplinarity

We recommend reflecting these aspects appropriately in teaching and training of next-generation researchers as well as in establishing new interdisciplinary research groups or collaborations. Therefore, in teaching, it is important to also convey a realistic view of how research works in practice. Students should be sensitized to real-world challenges and the need for pragmatic decision-making, while still striving for the basic principles of "good research practices." The literature review and our own experiences suggest that students are mostly taught the "ideal model," and thus, they are often not well prepared for the realities of research. It seems preferable to discuss challenges openly and to expose students to ethical and practical dilemmas early on.

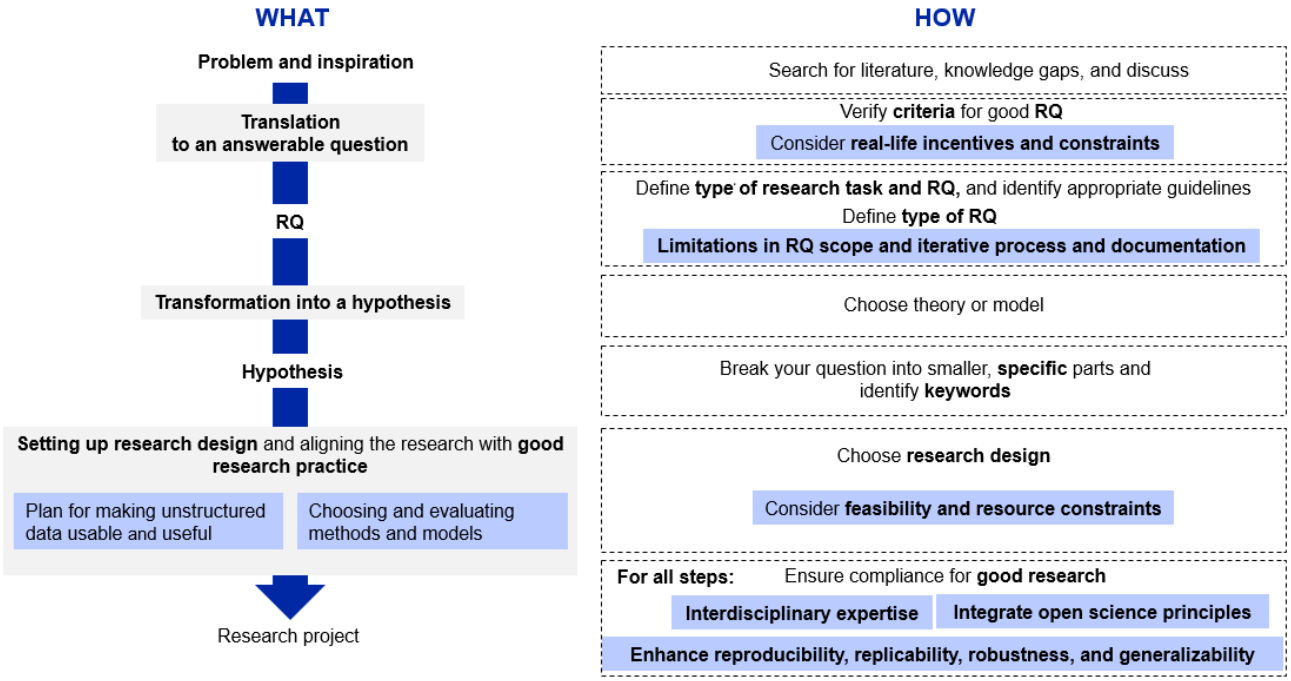
Lecture instructors and supervisors can sensitize students toward real-world challenges. They can prepare specific exercises where students can reflect on the problems that might arise from real-life constraints.

The added complexity and additional financial needs for education for interdisciplinary collaboration and open science should be acknowledged by funding agencies.

Specific Tools to Inform the Teaching of Interdisciplinary Courses on Real-World Data Analyses

Our study provides practical tools to guide the content and curricula of courses focused on interdisciplinary projects and collaborations. A more detailed description of the application of the study results to teaching is provided in Multimedia Appendix 5. The structure of our workshops (Figure 1) and the results of each workshop can be directly translated into the tools focusing on both dimensions of *content and curriculum* as well as *methods and teaching styles*. In terms of *content and curriculum*, the glossary with key concepts and terminology can be used to introduce students to interdisciplinary work. The workflow (Figure 2) combined with the glossary can serve as an interdiction to research practices for students with a different background, for example, humanities. Finally, our adapted workflow (Figure 3) sensitizes students for additional topics of transparency, FAIR data, reproducibility, and open science. Regarding *methods and teaching styles*, the sequence of workshops (Figure 1) and their results as outlined in the *content and curriculum section* can be directly translated into teaching phases, which build on top of each other. As illustrated by the example described in Multimedia Appendix 5, the structure of 3 teaching phases is useful and effective for teaching interdisciplinary research collaborations. A key insight from our workshop (modifications to the "how"—cluster 1) consisted of the need to acknowledge and address real-world challenges in study planning and execution. In our experience, case studies and illustrations of the scientific process of real-world examples are greatly appreciated by students.

Figure 3. Adapted workflow. The blue-colored boxes contain the added aspects in the "what" and "how" categories.



Discussion

Principal Findings

This study examined how interdisciplinary research collaborations between health research and data science can be streamlined by creating a shared conceptual understanding of terminologies and best practice workflows and by acknowledging or merging approaches from other disciplines. In a series of interactive workshops, our interdisciplinary group of coauthors concluded that the workflow of established practices for formulating a research question, generating hypotheses, and defining research designs remain valid. We argue that the reuse of digital data does not substantially change scientific activity, particularly the fundamental step of defining a good research question [1,43]. Achieving clarity on the research question benefits data analysis and interpretation by providing structure and informing the study design workflow. Moreover, a shared understanding of the research question and study workflow facilitates the inclusion of diverse domain knowledge to ensure research quality and result quality [6,14,44]. In line with this, the group noted general tendencies in research toward more open, transparent, and reproducible research, which are influenced by recommended data science practices. Along those lines, health research should increasingly foster good scientific practices that help to align the reuse of digital data with principles of reproducibility, robustness, generalizability, validity [1,32], transparency, and open science.

Our set of tools and recommendations can also be integrated into medical education by providing academic educators with a structured approach to teaching research question formulation in the context of using digital data in health research. By emphasizing the importance of both hypothesis-driven and data-driven research methods, educators can guide researchers in navigating the interdisciplinary challenges of health research and data science. The importance of creating a common terminology and discussion about scientific principles can further increase awareness about the challenges of interdisciplinary collaboration between health researchers and

data scientists. The proposed workflow and recommendations equip researchers with the tools to address the challenges of research question definition for interdisciplinary projects. The clear and practical steps provided by the workflow ensure that students not only grasp theoretical concepts but also apply them effectively in real-world scenarios, preparing them for collaborative, data-driven environments in health care and research.

For implementation, the set of tools and recommendations could be integrated into medical curricula and PhD programs through dedicated courses, workshops, or modules focusing on research methods and interdisciplinary collaboration for young researchers. Medical educators can adopt these recommendations to structure class discussions, assignments, and group projects, ensuring that students are exposed to both research approaches. In Multimedia Appendix 5, we provide an example of an interdisciplinary course implementing this set of tools. To evaluate the effectiveness of this implementation, a combination of qualitative and quantitative assessments can be used. Surveys and feedback from both students and educators can measure how well the workflow improves understanding and application of interdisciplinary research question formulation.

Our interdisciplinary effort recognized and discussed several potential obstacles toward bridging the approaches of established health research and data science. In the following sections, we repeat 4 key insights from our workshop interactions on how such obstacles can be overcome. First, we noted substantial differences in the use of terminologies across disciplines. For interdisciplinary collaborations, it is important to clarify key terms and concepts early on and to develop a shared understanding of the research aim and research question.

Second, in the early stages of the project, workshop participants expressed confusion about different types of analysis methods and their relationship with specific research tasks and high-level aims, such as prediction and classification, confirmatory research, or exploratory research. Agreeing on the high-level conceptual framework of “research tasks” helped structure the workshop discussions effectively. The discussion around the concepts of “research task” also fostered insights about commonalities and overlaps between concepts of data and health research. For example, many data science tasks can be classified as exploratory or prediction or classification tasks, which have conceptual counterparts in health research methodologies, each with corresponding reporting quality guidelines. Referring to specific research tasks rather than making global statements about data science or health research resonated well with the workshop participants and facilitated the discussions considerably.

Third, by introducing the concept of a “research task,” the group was also better able to examine the relationships among research aims, objectives, and tasks and how they are reflected in the workflow of established practices. Participants believed that exploratory or prediction or classification tasks, in particular, did not fit well into the workflow because such work is often not strictly hypothesis driven. However, 2 insights helped to align the workflow framework with the task concept: answering a research question may involve multiple research tasks in the same analysis, such as using prediction and classification tasks for data preprocessing, and later using these predictions in a confirmatory analysis, for example, as an exposure variable. Moreover, the scientific process can be viewed as a continuum of studies. From this perspective, the workflow of established practices can also be seen as a higher-level discovery cycle that spans across multiple studies. For example, an initial study may explore initial exploratory hypotheses or generate a first iteration of a prediction model, thus leading to new hypotheses. Indeed, exploratory and inductive methods can be useful to keep an open mind and become inspired by empirical data. In this way, the research tasks can be seen as a continuum—where data-driven research ends, hypothesis-driven research can start. Follow-up studies could then explore the hypotheses or validate the prediction model (whose structure can also be considered a hypothesis) in new data or in a confirmatory analysis. In combination, these multiple research tasks or study sequences are still likely to conform to the proposed workflow of recommended practices.

Finally, reusing unstructured and structured digital data brings new ethical challenges, such as privacy and consent issues, and problems with (public) trust and data diversity [45-49]. Traditional ethical assessments for data use in research and ethics review committees might not be well suited to address the challenges of digital data and might need adaptations [45,50]. Weighing the potential benefits and risks of using digital data becomes more complex. This problem is accentuated because the availability and production of digital data are often not based on a scientific decision, and rather, other factors such as political or social phenomena play a role [1]. While the need for novel ethical mechanisms to guide researchers is to be found in recently developed self-assessment tools for ethical data use [51,52], these new ethical mechanisms need further refinement to be widely adopted.

Strengths and Limitations

The strength of the expert groups was that participants represented a diverse group in terms of disciplines and career stages. However, it is possible that not all potentially relevant viewpoints were represented. A further strength was that the inputs from experts were collected systematically via different channels (eg, discussions, preworkshop tasks, and commenting on documents) throughout the consensus process. This allowed to harmonize and synthesize knowledge and insights from diverse disciplines. Experts also had several opportunities to review discussion outcomes and final summaries through workshop protocol and involvement in manuscript writing.

There are limitations regarding our proposed workflow. First, it represents an idealized process for defining a good research question, which is often challenged by funding and resource constraints or established norms. Some parts of the workflow might not be explicitly applicable to all types of research. The example scenario used to develop the workflow was based on hypothesis-driven deductive research, which often uses relational and causal research questions. We did not explicitly include inductive, qualitative approaches in health research, but we see the deductive and inductive research on a spectrum [53]. This limitation does not prevent the overall concept of the workflow

from being applied to other types of research, such as inductive, data-driven, or exploratory research.

Finally, although the literature review was conducted with great care and the expert group included several experienced researchers and faculty from different scientific disciplines, it was not possible to conduct a fully systematic search across all research disciplines due to resource constraints. Therefore, it is possible that some potentially relevant concepts and guidelines were not included.

Conclusions

In an age of digital transformation, established scientific practices with a strong focus on formulating research question design remain relevant and useful for gaining clarity about research aims. We recommend initiating new collaborations in the health domain with a review of terminologies and concepts to avoid misconceptions and problems further downstream in the research process. Our terminology and workflow may serve as tools to be used in medical education to support young and established researchers in interdisciplinary health research projects. To this end, we found the concept of "research tasks" particularly useful to foster a shared understanding among our collaborators. In addition, we recommend adapting the way the established workflow is taught to prospective researchers in health research and other disciplines, incorporating concepts from open science, the 3RG criteria, and the "science as a continuum" paradigm. We also call for funding agencies and publishers to incentivize and acknowledge investments in defining good research questions for complex novel data and analysis methods.

Data Availability

All data are available in the manuscript and multimedia appendices.

Authors' Contributions

JS contributed to conceptualization, data curation, investigation, methodology, project administration, resources, visualization, and writing the original draft. VvW contributed to the conceptualization, investigation, methodology, and writing the original draft. All the other authors contributed to writing, reviewing, and editing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Study and workshop protocols and preparatory tasks.
[PDF File (Adobe PDF File), 1792 KB - mededu_v11i1e56369_app1.pdf]

Multimedia Appendix 2
Accurate Consensus Reporting Document (ACCORD) checklist.
[PDF File (Adobe PDF File), 127 KB - mededu_v11i1e56369_app2.pdf]

Multimedia Appendix 3
Example scenario of digital data reuse to structure workshop discussions.
[PDF File (Adobe PDF File), 20 KB - mededu_v11i1e56369_app3.pdf]

Multimedia Appendix 4
Proposed modifications to the "what": additional steps to the workflow.
[PDF File (Adobe PDF File), 118 KB - mededu_v11i1e56369_app4.pdf]

Multimedia Appendix 5
Application of teaching tools: example from the course Interactive Data Science in Digital Health.
[DOCX File , 36 KB - mededu_v11i1e56369_app5.docx]

References

1. Caliebe A, Leverkus F, Antes G, Krawczak M. Does big data require a methodological change in medical research? BMC Med Res Methodol 2019 Jun 17;19(1):125 [FREE Full text] [doi: 10.1186/s12874-019-0774-0] [Medline: 31208367]
2. Chiavilli M, Campagnini S, Baretta T, Castagnoli C, Paperini A, Politi AM, et al. Design and implementation of a Stroke Rehabilitation Registry for the systematic assessment of processes and outcomes and the development of data-driven prediction models: the STRATEGY study protocol. Front Neurol 2022 Oct 10;13:919353 [FREE Full text] [doi: 10.3389/fneur.2022.919353] [Medline: 36299268]

3. Sim I. Mobile devices and health. N Engl J Med 2019 Sep 05;381(10):956-968. [doi: 10.1056/NEJMra1806949] [Medline: 31483966]
4. Batko K, Ślęzak A. The use of big data analytics in healthcare. J Big Data 2022;9(1):3 [FREE Full text] [doi: 10.1186/s40537-021-00553-4] [Medline: 35013701]
5. Sedlakova J, Daniore P, Horn Wintsch A, Wolf M, Stanikic M, Haag C, et al. Challenges and best practices for digital unstructured data enrichment in health research: a systematic narrative review. PLOS Digit Health 2023 Oct 11;2(10):e0000347 [FREE Full text] [doi: 10.1371/journal.pdig.0000347] [Medline: 37819910]
6. Kitchin R. Big data, new epistemologies and paradigm shifts. Big Data Soc 2014 Apr 01;1(1). [doi: 10.1177/2053951714528481]
7. Goldsmith J, Sun Y, Fried LP, Wing J, Miller GW, Berhane K. The emergence and future of public health data science. Public Health Rev 2021 Apr 26;42:1604023 [FREE Full text] [doi: 10.3389/phrs.2021.1604023] [Medline: 34692178]
8. Mirin N, Mattie H, Jackson L, Samad Z, Chunara R. Data science in public health: building next generation capacity. Harv Data Sci Rev 2022 Oct 27;4(4). [doi: 10.1162/99608f92.18da72db]
9. Cooke A, Smith D, Booth A. Beyond PICO: the SPIDER tool for qualitative evidence synthesis. Qual Health Res 2012 Oct;22(10):1435-1443. [doi: 10.1177/1049732312452938] [Medline: 22829486]
10. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008 Apr;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008] [Medline: 18313558]
11. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
12. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. PLoS Med 2007 Oct 16;4(10):e297 [FREE Full text] [doi: 10.1371/journal.pmed.0040297] [Medline: 17941715]
13. Donaldson L, Ricciardi W, Sheridan S, Tartaglia R. Textbook of Patient Safety and Clinical Risk Management. Cham, Switzerland: Springer International Publishing; 2020.
14. Hicks JL, Althoff T, Sosic R, Kuhar P, Bostjancic B, King AC, et al. Best practices for analyzing large-scale health data from wearables and smartphone apps. NPJ Digit Med 2019 Jun 3;2:45 [FREE Full text] [doi: 10.1038/s41746-019-0121-1] [Medline: 31304391]
15. Gattrell WT, Logullo P, van Zuuren EJ, Price A, Hughes EL, Blazey P, et al. ACCORD (ACcurate COnsensus Reporting Document): a reporting guideline for consensus methods in biomedicine developed via a modified Delphi. PLoS Med 2024 Jan 23;21(1):e1004326 [FREE Full text] [doi: 10.1371/journal.pmed.1004326] [Medline: 38261576]
16. Doody O, Bailey ME. Setting a research question, aim and objective. Nurse Res 2016 Mar;23(4):19-23. [doi: 10.7748/nr.23.4.19.s5] [Medline: 26997231]
17. Creswell JW. Qualitative Inquiry and Research Design: Choosing Among Five Approaches. 3rd ed. Thousand Oaks, CA: SAGE Publications; 2012.
18. Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: developing a prognostic model. BMJ 2009 Mar 31;338:b604. [doi: 10.1136/bmj.b604] [Medline: 19336487]
19. Lipowski EE. Developing great research questions. Am J Health Syst Pharm 2008 Sep 01;65(17):1667-1670. [doi: 10.2146/ajhp070276] [Medline: 18714115]
20. Aslam S, Emmanuel P. Formulating a researchable question: a critical step for facilitating good clinical research. Indian J Sex Transm Dis AIDS 2010 Jan;31(1):47-50 [FREE Full text] [doi: 10.4103/0253-7184.69003] [Medline: 21808439]
21. Tully MP. Research: articulating questions, generating hypotheses, and choosing study designs. Can J Hosp Pharm 2014 Jan;67(1):31-34 [FREE Full text] [doi: 10.4212/cjhp.v67i1.1320] [Medline: 24634524]
22. Fandino W. Formulating a good research question: pearls and pitfalls. Indian J Anaesth 2019 Aug;63(8):611-616 [FREE Full text] [doi: 10.4103/ija.IJA_198_19] [Medline: 31462805]
23. Abbade LP, Wang M, Sriganesh K, Jin Y, Mbuagbaw L, Thabane L. The framing of research questions using the PICOT format in randomized controlled trials of venous ulcer disease is suboptimal: a systematic survey. Wound Repair Regen 2017 Sep;25(5):892-900. [doi: 10.1111/wrr.12592] [Medline: 29080311]
24. Booth A. Clear and present questions: formulating questions for evidence based practice. Libr Hi Tech 2006;24(3):355-368. [doi: 10.1108/07378830610692127]
25. Mehta A, Malley B, Walkey A. Formulating the research question. In: Secondary Analysis of Electronic Health Records. Cham, Switzerland: Springer; 2016.
26. Sørensen MP, Ravn T, Marušić A, Elizondo AR, Kavouras P, Tijdink J, et al. Strengthening research integrity: which topic areas should organisations focus on? Humanit Soc Sci Commun 2021 Aug 12;8(1):1-15. [doi: 10.1057/s41599-021-00874-y]
27. Quiroga Gutierrez AC, Lindegger DJ, Taji Heravi A, Stojanov T, Sykora M, Elayan S, et al. Reproducibility and scientific integrity of big data research in urban public health and digital epidemiology: a call to action. Int J Environ Res Public Health 2023 Jan 13;20(2):1473 [FREE Full text] [doi: 10.3390/ijerph20021473] [Medline: 36674225]

28. Kandel S, Heer J, Plaisant C, Kennedy J, van Ham F, Riche NH, et al. Research directions in data wrangling: visualizations and transformations for usable and credible data. Inf Vis 2011 Sep 02;10(4):271-288. [doi: 10.1177/1473871611415994]
29. Munzner T. Visualization Analysis and Design. Boca Raton, FL: CRC Press; 2014.
30. Fekete JD. Visual analytics infrastructures: from data management to exploration. Computer 2013 Jul;46(7):22-29. [doi: 10.1109/MC.2013.120]
31. Arbesser C, Spechtenhauser F, Muhlbacher T, Piringer H. Visplause: visual data quality assessment of many time series using plausibility checks. IEEE Trans Visual Comput Graphics 2017 Jan;23(1):641-650. [doi: 10.1109/tvcg.2016.2598592]
32. Cerreta F, Ritzhaupt A, Metcalfe T, Askin S, Duarte J, Berntgen M, et al. Digital technologies for medicines: shaping a framework for success. Nat Rev Drug Discov 2020 Sep;19(9):573-574. [doi: 10.1038/d41573-020-00080-6] [Medline: 32398879]
33. Sedlmair M, Meyer M, Munzner T. Design study methodology: reflections from the trenches and the stacks. IEEE Trans Visual Comput Graphics 2012 Dec;18(12):2431-2440. [doi: 10.1109/tvcg.2012.213]
34. Muhlbacher T, Piringer H, Gratzl S, Sedlmair M, Streit M. Opening the black box: strategies for increased user involvement in existing algorithm implementations. IEEE Trans Visual Comput Graphics 2014 Dec 31;20(12):1643-1652. [doi: 10.1109/tvcg.2014.2346578]
35. Schreck T, Bernard J, von Landesberger T, Kohlhammer J. Visual cluster analysis of trajectory data with interactive Kohonen maps. Inf Vis 2009 Feb 12;8(1):14-29. [doi: 10.1057/ivs.2008.29]
36. Lipton ZC. The doctor just won't accept that!. arXiv Preprint posted online on November 20, 2017 [FREE Full text] [doi: 10.48550/arXiv.1711.08037]
37. Bernard J, von Landesberger T, Bremm S, Schreck T. Multi-scale visual quality assessment for cluster analysis with self-organizing maps. In: Proceedings of SPIE - The International Society for Optical Engineering. 2011 Jan Presented at: SPIE 2011; 2011; Bellingham, WA. [doi: 10.1117/12.872545]
38. Munzner T. A nested model for visualization design and validation. IEEE Trans Vis Comput Graph 2009 Nov;15(6):921-928. [doi: 10.1109/TVCG.2009.111] [Medline: 19834155]
39. de Hond AA, Leeuwenberg AM, Hooft L, Kant IM, Nijman SW, van Os HJ, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. NPJ Digit Med 2022 Jan 10;5(1):2 [FREE Full text] [doi: 10.1038/s41746-021-00549-7] [Medline: 35013569]
40. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ 2022 May 18;377:e070904 [FREE Full text] [doi: 10.1136/bmj-2022-070904] [Medline: 35584845]
41. Schwab S, Janiaud P, Dayan M, Amrhein V, Panczak R, Palagi PM, et al. Ten simple rules for good research practice. PLoS Comput Biol 2022 Jun 23;18(6):e1010139 [FREE Full text] [doi: 10.1371/journal.pcbi.1010139] [Medline: 35737655]
42. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3(1):160018 [FREE Full text] [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
43. Mazzocchi F. Could big data be the end of theory in science? A few remarks on the epistemology of data-driven science. EMBO Rep 2015 Oct 10;16(10):1250-1255 [FREE Full text] [doi: 10.15252/embr.201541001] [Medline: 26358953]
44. Lee EW, Yee AZ. Toward data sense-making in digital health communication research: why theory matters in the age of big data. Front Commun 2020 Feb 27;5. [doi: 10.3389/fcomm.2020.00011]
45. Ferretti A, Ienca M, Sheehan M, Blasimme A, Dove ES, Farsides B, et al. Ethics review of big data research: what should stay and what should be reformed? BMC Med Ethics 2021 Apr 30;22(1):51 [FREE Full text] [doi: 10.1186/s12910-021-00616-4] [Medline: 33931049]
46. Nebeker C, Torous J, Bartlett Ellis RJ. Building the case for actionable ethics in digital health research supported by artificial intelligence. BMC Med 2019 Jul 17;17(1):137 [FREE Full text] [doi: 10.1186/s12916-019-1377-7] [Medline: 31311535]
47. Facca D, Smith MJ, Shelley J, Lizotte D, Donelle L. Exploring the ethical issues in research using digital data collection strategies with minors: a scoping review. PLoS One 2020;15(8):e0237875 [FREE Full text] [doi: 10.1371/journal.pone.0237875] [Medline: 32853218]
48. Clark RA, Foote J, Versace VL, Brown A, Daniel M, Coffee NT, et al. The keeping on track study: exploring the activity levels and utilization of healthcare services of acute coronary syndrome (ACS) patients in the first 30-days after discharge from hospital. Med Sci (Basel) 2019 Apr 19;7(4):61 [FREE Full text] [doi: 10.3390/medsci7040061] [Medline: 31010168]
49. Arora A, Alderman JE, Palmer J, Ganapathi S, Laws E, McCradden MD, et al. The value of standards for health datasets in artificial intelligence-based applications. Nat Med 2023 Nov 26;29(11):2929-2938 [FREE Full text] [doi: 10.1038/s41591-023-02608-w] [Medline: 37884627]
50. Viberg Johansson J, Bentzen HB, Mascalzoni D. What ethical approaches are used by scientists when sharing health data? An interview study. BMC Med Ethics 2022 Apr 11;23(1):41 [FREE Full text] [doi: 10.1186/s12910-022-00779-8] [Medline: 35410285]
51. Wiltshire D, Alvanides S. Ensuring the ethical use of big data: lessons from secure data access. Heliyon 2022 Feb;8(2):e08981 [FREE Full text] [doi: 10.1016/j.heliyon.2022.e08981] [Medline: 35243099]

52. Bigdata-ethics-HeALth framework - BEHALF. Eidgenössische Technische Hochschule Zürich. URL: https://bioethics.ethz.ch/research/BEHALF.html [accessed 2022-11-28]
53. Young M, Varpio L, Uijtdehaage S, Paradis E. The spectrum of inductive and deductive research approaches using quantitative and qualitative data. Acad Med 2020 Jul;95(7):1122. [doi: 10.1097/ACM.0000000000003101] [Medline: 31833855]

Abbreviations

3RG: reproducibility, robustness, replicability, or generalizability
ACCORD: Accurate Consensus Reporting Document
CONSORT: Consolidated Standards of Reporting Trials
DSI: Digital Society Initiative
FAIR: findable, accessible, interoperable, and reusable
FINER: feasible, interesting, novel, ethical, and relevant
PICOT: population, intervention, comparison, outcome, and time
SMART: specific, measurable, achievable, realistic, and timely
SPIDER: sample, phenomenon of interest, design, evaluation, research type
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

Edited by B Lesselroth; submitted 15.01.24; peer-reviewed by G Richter, M Mahmic Kaknjo; comments to author 12.02.24; revised version received 16.05.24; accepted 23.11.24; published 23.01.25.

Please cite as:
Sedlakova J, Stanikić M, Gille F, Bernard J, Horn AB, Wolf M, Haag C, Floris J, Morgenshtern G, Schneider G, Zumbrunn Wojczyńska A, Mouton Dorey C, Ettlin DA, Gero D, Friemel T, Lu Z, Papadopoulos K, Schläpfer S, Wang N, von Wyl V
Refining Established Practices for Research Question Definition to Foster Interdisciplinary Research Skills in a Digital Age: Consensus Study With Nominal Group Technique
JMIR Med Educ 2025;11:e56369
URL: https://mededu.jmir.org/2025/1/e56369
doi:10.2196/56369
PMID:

Global Health care Professionals' Perceptions of Large Language Model Use In Practice: Cross-Sectional Survey Study

Ecem Ozkan[1], MD; Aysun Tekin[2], MD; Mahmut Can Ozkan[1], MD; Daniel Cabrera[3], MD; Alexander Niven[4], MD; Yue Dong[2], MD

[1]Department of Medicine, Jersey Shore University Medical Center, 1945 NJ-33, Neptune, NJ, United States

[2]Department of Anesthesiology, Mayo Clinic College of Medicine, Rochester, MN, United States

[3]Department of Emergency Medicine, Mayo Clinic College of Medicine, Rochester, MN, United States

[4]Department of Pulmonary and Critical Care Medicine, Mayo Clinic College of Medicine, Rochester, MN, United States

Corresponding Author:
Ecem Ozkan, MD
Department of Medicine, Jersey Shore University Medical Center, 1945 NJ-33, Neptune, NJ, United States

Abstract

Background: ChatGPT is a large language model-based chatbot developed by OpenAI. ChatGPT has many potential applications to health care, including enhanced diagnostic accuracy and efficiency, improved treatment planning, and better patient outcomes. However, health care professionals' perceptions of ChatGPT and similar artificial intelligence tools are not well known. Understanding these attitudes is important to inform the best approaches to exploring their use in medicine.

Objective: Our aim was to evaluate the health care professionals' awareness and perceptions regarding potential applications of ChatGPT in the medical field, including potential benefits and challenges of adoption.

Methods: We designed a 33-question online survey that was distributed among health care professionals via targeted emails and professional Twitter and LinkedIn accounts. The survey included a range of questions to define respondents' demographic characteristics, familiarity with ChatGPT, perceptions of this tool's usefulness and reliability, and opinions on its potential to improve patient care, research, and education efforts.

Results: One hundred and fifteen health care professionals from 21 countries responded to the survey, including physicians, nurses, researchers, and educators. Of these, 101 (87.8%) had heard of ChatGPT, mainly from peers, social media, and news, and 77 (76.2%) had used ChatGPT at least once. Participants found ChatGPT to be helpful for writing manuscripts (n=31, 45.6%), emails (n=25, 36.8%), and grants (n=12, 17.6%); accessing the latest research and evidence-based guidelines (n=21, 30.9%); providing suggestions on diagnosis or treatment (n=15, 22.1%); and improving patient communication (n=12, 17.6%). Respondents also felt that the ability of ChatGPT to access and summarize research articles (n=22, 46.8%), provide quick answers to clinical questions (n=15, 31.9%), and generate patient education materials (n=10, 21.3%) was helpful. However, there are concerns regarding the use of ChatGPT, for example, the accuracy of responses (n=14, 29.8%), limited applicability in specific practices (n=18, 38.3%), and legal and ethical considerations (n=6, 12.8%), mainly related to plagiarism or copyright violations. Participants stated that safety protocols such as data encryption (n=63, 62.4%) and access control (n=52, 51.5%) could assist in ensuring patient privacy and data security.

Conclusions: Our findings show that ChatGPT use is widespread among health care professionals in daily clinical, research, and educational activities. The majority of our participants found ChatGPT to be useful; however, there are concerns about patient privacy, data security, and its legal and ethical issues as well as the accuracy of its information. Further studies are required to understand the impact of ChatGPT and other large language models on clinical, educational, and research outcomes, and the concerns regarding its use must be addressed systematically and through appropriate methods.

(JMIR Med Educ 2025;11:e58801) doi:10.2196/58801

KEYWORDS

ChatGPT; LLM; global; health care professionals; large language model; language model; chatbot; AI; diagnostic accuracy; efficiency; treatment planning; patient outcome; patient care; survey; physicians; nurses; educators; patient communication; clinical; educational; utilization; artificial intelligence

Introduction

Large language model (LLM) refers to advanced artificial intelligence (AI) models designed for natural language processing tasks. LLMs are trained on vast amounts of text data and use deep learning techniques to understand and generate human-like language. They helped transform various fields, including medicine [1]. Some examples of most popular LLMs are LlaMA by Meta, Orca and Phi-1 by Microsoft, BLOOM, PaLM2 by Google, and GPT by OpenAI. ChatGPT, a chatbot powered by GPT-3/4 was released by OpenAI in November 2022, incorporating billions of parameters that enable it to comprehend and generate human-like text with the capability of context creation. Its intuitive interface and capacity for prompt engineering have enabled diverse applications across domains [2].

In medicine, recent studies have demonstrated ChatGPT's potential to support clinical decision-making, summarize complex medical data, and streamline documentation processes. For instance, ChatGPT has been evaluated for its ability to generate discharge summaries, assist in developing differential diagnoses, and simplify patient communication [3-5]. Its role in medical education has also been explored, demonstrating its utility in preparing students for licensing exams like the United States Medical Licensing Examination (USMLE) and enhancing self-directed learning through case-based scenarios [5-7]. ChatGPT was also shown to be capable of defining and answering clinical vignettes and achieved >60% of the threshold on the USMLE, which is the passing score for all three exams [8,9]. Additionally, its ability to provide personalized health education and assist in chronic disease management has been highlighted as a promising avenue for improving patient outcomes [4,10].

The integration of ChatGPT into health care settings is accelerating, with a growing body of literature examining its applications. Despite these advancements, significant challenges remain. Concerns about data privacy, ethical implications, and the accuracy of AI-generated content persist as barriers to widespread adoption [4,5,10]. Additionally, little is known regarding global health care professionals' perspectives and the extent and impact of ChatGPT's integration in health care settings [11,12]. Most studies to date, have been limited to localized settings or specific subgroups. Yet, successful and ethical integration of ChatGPT into health care workflows depends heavily on end-user acceptance, awareness of limitations, and perceptions regarding safety, usability, and value [5-7].

This study aimed to evaluate health care professionals' awareness and perceptions of ChatGPT, with a focus on its applications, challenges, and utility across clinical, educational, and research settings. We surveyed a diverse group of health care professionals—including physicians, nurses, researchers, and educators—from multiple countries and practice settings. Using a cross-sectional survey design, we collected data on their familiarity with ChatGPT, how and why they used it, and their concerns about its integration. Our a priori hypothesis was that while many health care professionals would recognize ChatGPT's potential benefits, such as improving efficiency, communication, and access to knowledge, they would also express concerns regarding ethical, legal, and accuracy-related issues.

This study offers timely insights for health care leaders, educators, and policymakers considering the responsible adoption of generative AI tools. By reflecting on global perspectives from frontline users, our findings may help shape discussions on how to balance innovation with safety and trust in clinical AI applications.

Methods

This study was conducted as a cross-sectional survey between April 20 and July 3, 2023 (Multimedia Appendix 1).

Survey Instrument Development and Validation

The questionnaire used in this study was developed de novo by the research team. The design process was informed by the research team's multidisciplinary experience in medicine, education, and digital health, as well as the evolving discourse around AI in health care. To assist with rapid prototyping, the research team used ChatGPT (OpenAI) to generate the first draft of the questionnaire. This initial draft provided a foundation for question phrasing and thematic organization. The final survey was iteratively refined by the study investigators to ensure clinical and contextual relevance.

To enhance clarity and assess feasibility, the questionnaire was piloted informally among five health care researchers affiliated with our institution. Their feedback informed improvements in question wording, branching logic, and estimated completion time (approximately 5 minutes). No formal psychometric validation was conducted.

The final survey included 33 questions and was distributed electronically using Research Electronic Data Capture (REDCap) (version 13.1.30; Vanderbilt University) [13]. The questionnaire was structured around six thematic domains: (1) respondent demographics and work environment, (2) awareness and familiarity with ChatGPT, (3) frequency and purpose of use, (4) perceived benefits and challenges of ChatGPT in daily practice, (5) views on ethical, legal, and data security concerns, and (6) future expectations and training needs. The questionnaire incorporated branching logic to adapt follow-up questions based on initial responses—for example, only respondents who reported using ChatGPT were asked about specific applications or frequency of use. A visual summary of the questionnaire flow and branching logic is provided in Multimedia Appendix 2. The final instrument has been reported in Multimedia Appendix 1.

Participants and Sampling Strategy

We used a convenience sampling approach. The questionnaire was distributed to health care professionals via targeted emails, and professional Twitter, LinkedIn, and Instagram accounts using a snowball technique [14]. No predefined inclusion or exclusion criteria were applied beyond the requirement of being a health care professional (eg, physician, nurse, educator, researcher). There were no regional or institutional restrictions.

As the survey was open and anonymous, we did not estimate a denominator or calculate a response rate. For the purposes of this study, we defined the application of ChatGPT in the medical field broadly to include its use in clinical care, research, medical education, and health care–related administrative tasks. This inclusive definition reflects the multifaceted roles that health care professionals fulfill and acknowledges that tools such as ChatGPT may support a wide range of activities beyond direct patient care, such as writing grants, academic correspondence, and synthesizing medical literature. Survey items were designed to capture this broad spectrum of use across domains relevant to daily professional practice.

Demographic information of participants was summarized. Among those familiar with ChatGPT, opinions on the tool and potential dissemination resources were assessed. For those who had not used it, barriers to usage were examined (Multimedia Appendix 2). Participants with experience using the ChatGPT were also asked about perceived challenges and approaches for enhancing usability. Summary statistics were provided as numbers and frequencies. Comparative analyses were conducted using the χ^2 test, with a two-sided P value <.05 considered statistically significant. JMP Pro (version 14.1.0 software; SAS Institute Inc.) was used for the analyses.

Ethical Considerations

The study protocol was evaluated by the Mayo Clinic institutional review board and it was determined that it was exempted under 45 CFR 46.102 of the Code of Federal Regulations (2/28/2023). No personally identifying information was collected, and all data were fully anonymous. Study participation was voluntary and survey completion was considered as consent. All survey responses were stored on secure, access-restricted servers in compliance with institutional data protection policies.

Results

Main Findings

A total of 115 health care professionals from 21 countries responded to the survey. Table 1 displays a summary of their demographic information (Figures 1–2).

Table . Baseline characteristics.

Variables	Participants (N=115), n (%)
Age (years)	
20 - 29	30 (26.1)
30 - 39	27 (23.5)
40 - 49	26 (22.6)
50 - 59	10 (8.7)
>60	22 (19.1)
Sex[a]	
Female	45 (39.5)
Male	68 (59.6)
Profession[a]	
Educator	16 (14.0)
NP/PA[b]	5 (4.4)
Physician	62 (54.4)
Researcher	25 (21.9)
RN[c]	5 (4.4)
Area/ Unit	
Internal medicine	20 (17.4)
Surgery	15 (13)
Emergency medicine	10 (8.7)
Psychiatry and Neurology	8 (7)
Anesthesiology/ICU[d]	10 (8.6)
Obstetrics and Gynecology	7 (6.1)
Radiology	6 (5.2)
Others[e]	39 (33.9)
Years since graduation	
<5	43 (37.4)
5 - 10	27 (23.5)
11 - 20	16 (13.9)
>20	29 (25.2)
Work length in hospital (years)[a]	
<5	66 (57.9)
5 - 10	11 (9.6)
11 - 20	19 (16.7)
>20	18 (15.8)
Country of work	
United States	53 (46.1)
Turkey	24 (20.9)
Tanzania	7 (6.1)
China	6 (5.2)
Croatia	3 (2.6)
Russia	2 (1.7)

Variables	Participants (N=115), n (%)
France	2 (1.7)
Canada	2 (1.7)
Italy	2 (1.7)
Saudi Arabia	2 (1.7)
Others[e]	12 (10.4)
Native language	
English	28 (24.3)
Turkish	32 (27.8)
Spanish	10 (8.7)
Chinese (Mandarin)	9 (7.8)
Arabic	5 (4.3)
Others[e]	31 (26.8)
Place of employment[f]	
Academic hospitals and medical centers	72 (64.2)
Community hospitals	9 (8.0)
Private hospitals	13 (11.6)
Public hospitals	15 (13.4)
Free clinics	6 (5.4)
Others[e]	6 (5.3)
Frequency of ChatGPT usage (n=68)	
Multiple times per day	14 (20.6)
Once per day	3 (4.4)
Three to five times per week	14 (20.6)
Less than three times a week	13 (19.1)
Only tried it few times	24 (35.3)

[a]Due to lack of responses, missing data are not included in the reported totals; as a result, some category counts may not sum to the overall sample size.
[b]NP/PA: nurse practitioner/physician assistant.
[c]RN: registered nurse.
[d]ICU: intensive care unit.
[e]For Others see Multimedia Appendix 3.
[f]The subcategories are not mutually exclusive.

Figure 1. ChatGPT usage based on participants' age, gender, and country of work.

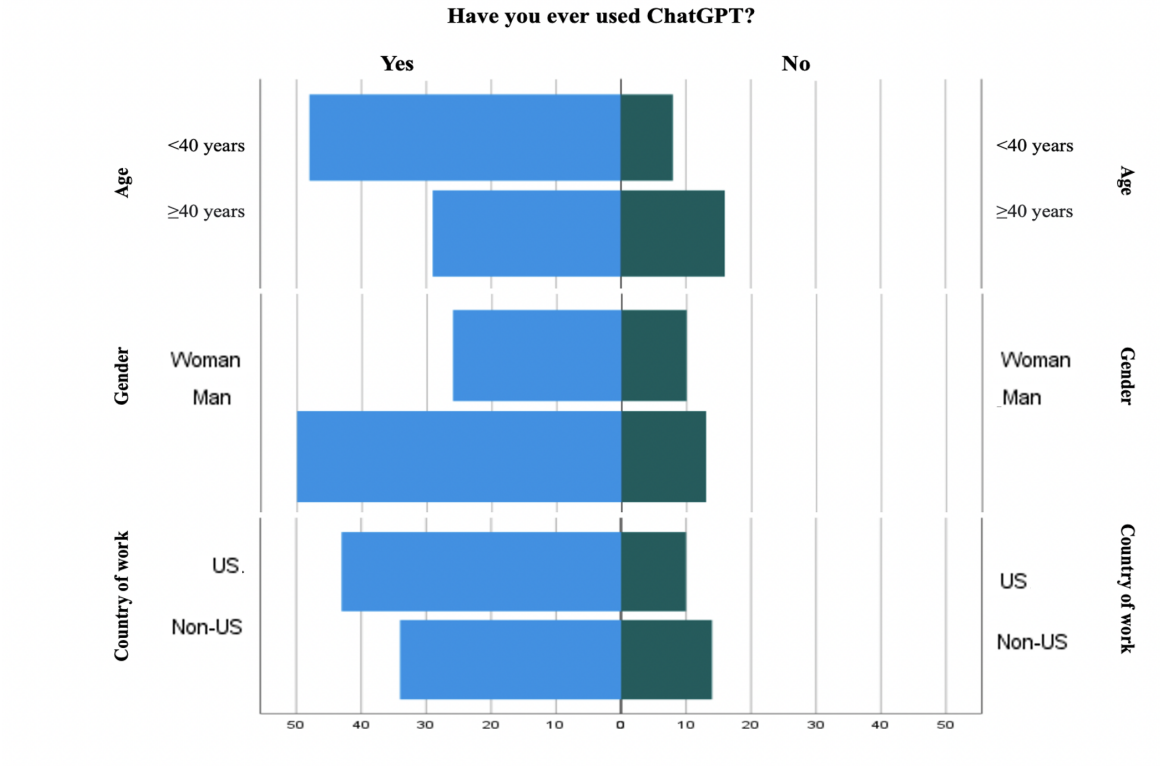
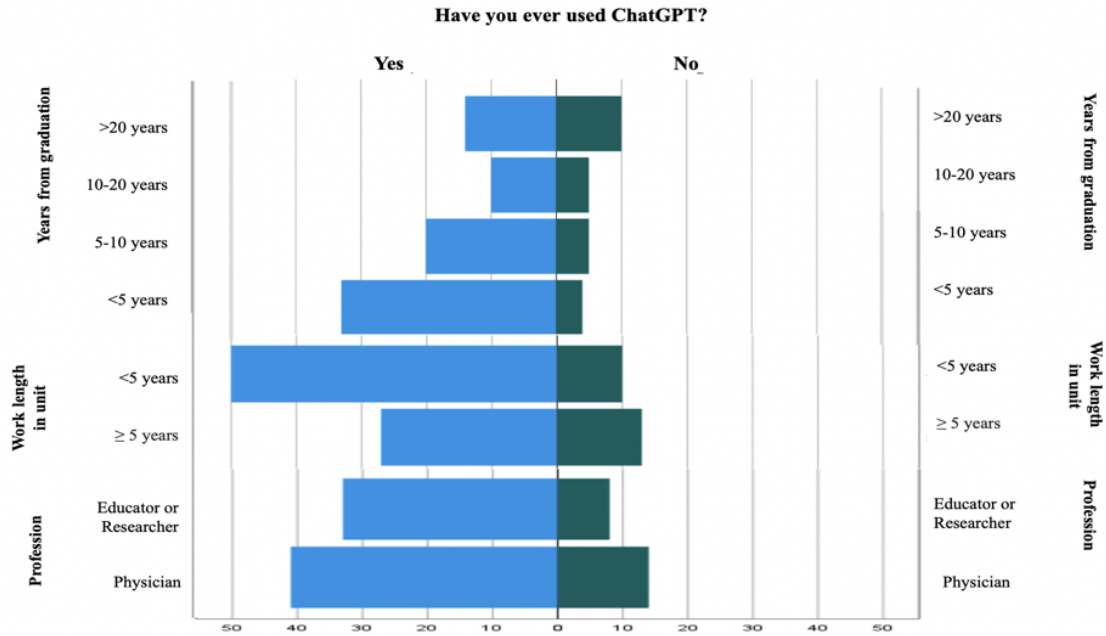


Figure 2. ChatGPT usage based on participants' years since graduation, length of work in the current unit, and profession.



Of the 115 participants, 101 (87.8%) had heard of ChatGPT, mainly from social media (n=33, 32.7%) and peers or colleagues (n=43, 42.6%). Of those, 77 (76.2%) had used ChatGPT before, with 18 (23.4%) using it multiple times per day and 23 (29.9%) having tried it only a few times. Moreover, 71 out of 77 (92.2%) participants used it in English. Among these, 50 were not native English speakers, and only 16/50 (32%) speakers used it both in English and their native language (Figure 3). Furthermore, variations in ChatGPT usage in daily practice were observed between participants using ChatGPT in English versus those who used it in their native language (Figure 4).

Figure 3. Ratio of native language use versus English use among participants while using ChatGPT.

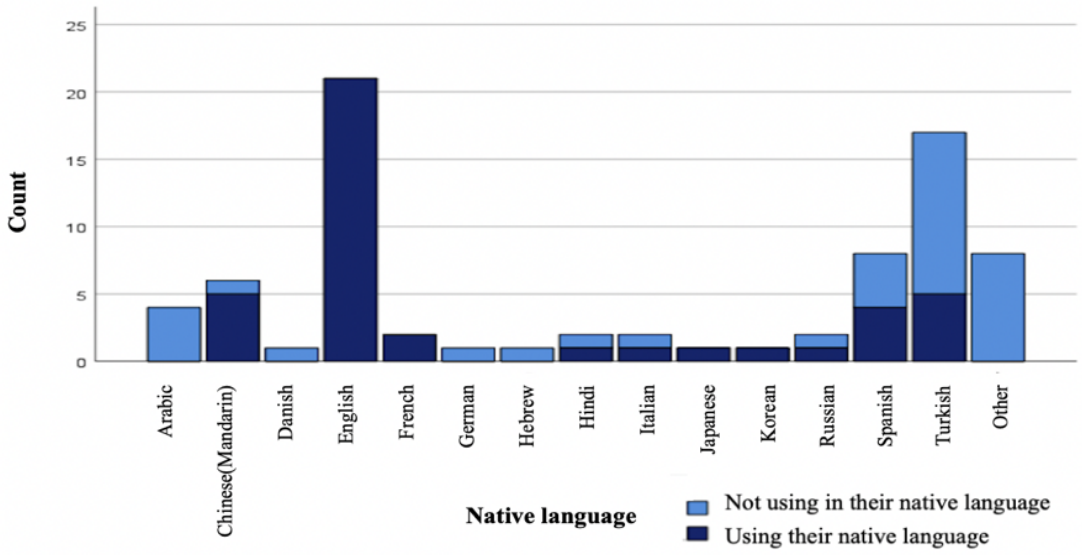
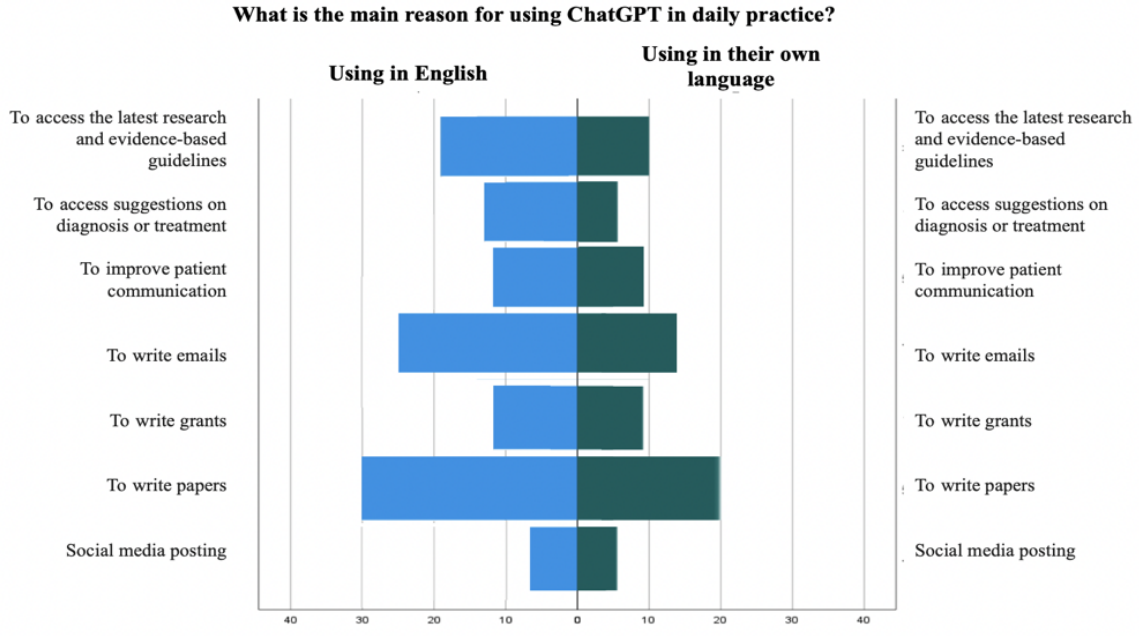


Figure 4. Main reasons for using ChatGPT in daily practice based on the language used by the participants.



The most common reasons to use ChatGPT included writing papers (n=29, 44.6%) and emails (n=25, 38.5%), and obtaining suggestions on diagnosis or treatment (n=14, 21.5%) (Table 2). Additional reasons for ChatGPT usage by health care professionals in daily practice are shared in Table 3.

Table . ChatGPT usefulness based on used features in daily practice.

ChatGPT features	Participants (n=68), n (%)
Usefulness in daily practice	
Not important	14 (20.6)
Slightly important	21 (30.9)
Moderately important	13 (19.1)
Important	13 (19.1)
Very important	7 (10.3)
ChatGPT's usefulness, 0 (most negative experience) to 10 (most positive experience)	
≥7	42 (61.8)
4-5-6	19 (27.9)
≤3	7 (10.3)
Most useful features	
To access and summarize research articles efficiently	22 (46.8)
To provide quick answers to clinical questions	15 (31.9)
To provide patient education materials	10 21.3
To write emails, grants, and papers	25 53.2

Table . Percentage of participants' main reasons for using ChatGPT in daily practice (multiple choice questions).

Main reason for using ChatGPT in daily practice	Participants (n=68), n (%)
Writing papers	31 (45.6)
Writing emails	25 (36.8)
To access the latest research and evidence-based guidelines	21 (30.9)
To access suggestions on diagnosis or treatment	15 (22.1)
To improve patient communication	12 (17.6)
To write grants	12 (17.6)

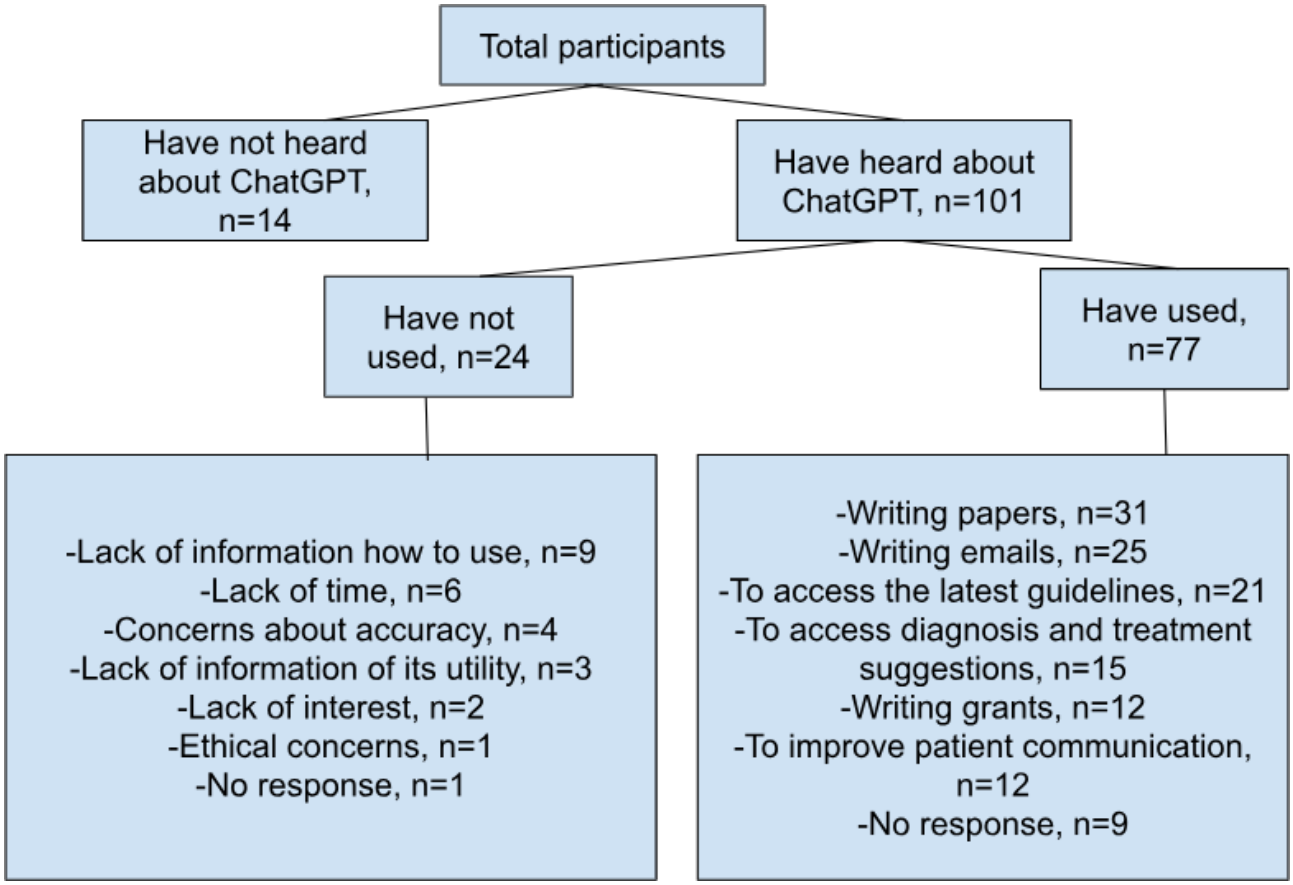
Incorporation of ChatGPT Into Daily Practice

Of the 77 participants who used ChatGPT, 36 (46.8%) used ChatGPT in their clinical practice, 58 (75.3%) used it for research, and 56 out of 77 (72.7%) used it for educational activities (Figure 5).

Among all respondents, 42/101 (43.6%) participants agreed that they would not be concerned if their clinician used ChatGPT while providing care to them if they were the patient, whereas 32 (32.7%) disagreed and preferred that their clinician not use ChatGPT during care.

The majority (n=79, 78.2%) of participants agreed that ChatGPT could be useful for medical or health care professional education. In nonclinical settings, participants stated that ChatGPT could help to reduce workload (n=57, 73.1%), improve efficiency by automating certain tasks (n=51, 65.4%), offer greater access and efficiently summarize research articles (n=52, 66.7%), create patient educational materials (n=49, 62.8%), provide quick answers to questions (n=48, 61.5%), and enhance the ability to write papers (n=37, 47.4%).

Figure 5. Factors contributing to use and nonuse of ChatGPT. The activities are not mutually exclusive and therefore, the total number of participants may exceed 115.



Challenges for Integrating ChatGPT Into Daily Practice

The main reasons for respondents not using ChatGPT included concerns about the accuracy of ChatGPT responses (n=14, 29.8%), limited applicability to their practice (n=18, 38.3%), legal and ethical considerations (n=6, 12.8%), limited diagnostic capabilities (n=4, 8.5%), lack of time (n=3, 6.4%), and lack of interest (n=2, 4.3%).

As one of the significant barriers is legal and ethical considerations, participants were asked to define plagiarism or copyright violations. Participants defined it as copying text or ideas from ChatGPT and using it for another source without citation (n=64, 63.4%), paraphrasing or summarizing content from ChatGPT and using it for another source without citation (n=41, 40.6%), using images from ChatGPT without permission (n=36, 35.6%), reusing or repurposing content from ChatGPT that was previously created for another purpose without permission (n=44, 43.6%).

In response to the legal and ethical challenges, participants proposed several solutions for integrating ChatGPT into daily practice. Participants stated that data encryption (n=63, 62.4%), access control (n=52, 51.5%), user authentication such as two-factor authentication (n=48, 47.5%), compliance with regulations such as Health Insurance Portability and Accountability Act or General Data Protection Regulation (n=62, 61.4%), transparency and informed consent (n=53, 52.5%), and regular training and awareness for health care professionals (n=58, 57.4%) are necessary to ensure patient privacy and data security.

Views on ChatGPT's success and other possible uses

When asked whether the participants knew ChatGPT had performed with ≥60% accuracy on the USMLE, 52 (51.5%) participants indicated they had heard this before. Additionally, 76 (68.5%) participants reported that they had not used any other AI platform.

Participants stated that ChatGPT can improve patient outcomes through personalized health education by providing tailored information and support (n=76, 75.2%); assisting with medication management through reminders and refill prescriptions, and provide information on side effects and interactions (n=55, 54.5%); telemedicine support for health care professionals to conduct virtual consultations, collect patient data, and provide decision support (n=48, 50%); aiding in symptom triage for patients (n=49, 48.5%); and offering mental health support by providing guidance on self-management techniques and coping strategies (n=49, 48.5%).

The distribution of responses based on different levels of postgraduate experience is reported in Multimedia Appendix 4. This distribution was largely balanced between the participants with fewer than 10 years and those with 10 or more years of experience.

Discussion

Principal Findings

This study offers a global perspective on how health care professional perceive and use ChatGPT in clinical, research, and educational context. Our findings demonstrate that awareness and adoption of ChatGPT are already widespread, with 76.2% of respondents having used the tool at least once. Participants primarily reported using ChatGPT for manuscript and email writing, grant application preparation, accessing research articles, clinical guideline support, diagnostic suggestions, and improving patient communication. Notably, more than three-quarters of participants agreed that ChatGPT holds potential utility in medical education, highlighting its ability to enhance learning experiences and facilitate task automation. Moreover, our study indicates that health care professionals endorse its use among colleagues. However, concerns about data privacy, ethical risks such as plagiarism, and the accuracy of AI-generated content remained as significant barriers to broader adoption. Proposed solutions included implementing safety protocols such as data encryption, access control, and regulatory compliance. In exploratory analyses comparing ChatGPT use, we did not identify significant differences across professional experience levels, which might be due to the limited sample size. Due to the wide range and uneven distribution of medical subspecialties represented, we were not able to conduct a formal comparison across specialties.

Implications of Findings

Our findings highlight the broad and flexible potential of ChatGPT in health care workflows. In clinical practice, ChatGPT is perceived as a tool that can enhance efficiency by automating routine documentation tasks, such as generating draft discharge summaries and patient letters. It also supports decision-making by offering fast access to evidence summaries and aids communication through the creation of patient-friendly materials [5,15]. In medical education, participants identified ChatGPT as a valuable educational supplement—one that could be incorporated into curricula to simulate real-world clinical scenarios and assist in preparing students for standardized exams like the USMLE [5,16]. It can also support personalized learning experiences tailored to individual needs and self-directed learning pathways. In research, ChatGPT was valued for its ability in grant writing, literature synthesis, and ideation, especially in the early stages of manuscript development or protocol design [5].

These findings underscore the need for structured training programs and ethical guidelines to support responsible integration of AI tools. Implementing human-in-the-loop systems, in which clinicians oversee and validate AI outputs, may enhance safety, and build user confidence while mitigating risks associated with biases or inaccuracies in AI-generated content [17].

Comparison to the Literature

Our findings align with prior studies that underscore ChatGPT's potential in health care. Cascella et al [2] described ChatGPT's potential to reduce administrative burden and assist with clinical reasoning, which mirrors participants' reported use of ChatGPT for documentation and clinical queries. In medical education, Gilson et al [8] showed that ChatGPT achieved passing scores on all three components of the USMLE, highlighting its utility in medical education. Similarly, Kung et al [9] emphasized its role in creating standardized templates for patient education materials. These findings also align with our participants' views on its usefulness for both learners and patients alike. Sallam [18] highlighted ChatGPT's capacity to process and summarize complex medical data efficiently, which our participants also leveraged for research and evidence access.

However, our study adds unique insights by capturing global perspectives from diverse practice settings. Unlike prior reports focused on specific institutions or national populations, our results reflect a cross-disciplinary, international sample, offering a broader view of how generative AI is being perceived across diverse practice settings.

The main reasons behind the lack of use of ChatGPT in daily practice were mainly due to the nonapplicability to their practice, lack of information regarding its use, and concerns about the accuracy of ChatGPT's responses, and legal and ethical considerations. The reason behind not using ChatGPT due to lack of information may be partially attributed to insufficient training opportunities for health care professionals in the use of generative AI. Previous studies have also indicated similar concerns regarding its implementation [19]. For instance, the concern for the spread of wrong information is a major obstacle, and different languages may have inconsistent results [20,21]. Many studies have shown that up to 96.7% of users are concerned about ethical and legal obstacles [3,18], particularly plagiarism [21-23], and copyright issues [3,18]. In a study conducted by a university at Sweden, 62% of students considered the use of chatbots for assignments and exams as cheating [24]. Our study showed that 86 out of 101 participants defined copying from ChatGPT as plagiarism. These concerns show that the implementation of ChatGPT into clinical settings will require a transition period supported by extensive safety measures. Health care professional leaders need to work with technology experts to develop learning objectives, curricula, assessments and evaluations, and safety protocols for this emerging technology.

Regarding the accuracy of ChatGPT's responses, our study shows that health care professionals identified this as having a paramount importance. Similar studies have shown that ChatGPT should be used with caution due to potential biases of AI, which may lead to the generation of inaccurate information. When used in the health care system, this could potentially lead to harmful consequences [25].

Educational Implications

The educational relevance of our findings is especially important. Our study suggests several opportunities:

- Curriculum design: Educators can incorporate ChatGPT into simulation- and case-based learning modules to foster clinical reasoning and application of evidence-based medicine.

- Needs Assessment: Educators may use baseline familiarity and usage patterns to tailor AI training initiatives and address gaps in knowledge or ethical understanding.
- Institutional Strategies: ChatGPT may serve as a tool in flipped classrooms, interactive tutorials, and self-directed learning, offering real-time feedback and access to guideline-driven responses.
- Learner Outcomes: By providing immediate feedback and access to evidence-based guidelines, ChatGPT has the potential to improve learner performance on standardized assessments [16].

Additionally, ChatGPT's ability to generate accessible explanations for patients could enhance health literacy and improve communication between physicians and patients.

Strengths and Limitations

This study has several strengths. We examined ChatGPT adoption from a global perspective. By including participants from 21 countries and various clinical and academic backgrounds, the study provides a valuable overview of current usage patterns and attitudes toward generative AI tools in health care. The survey instrument was comprehensive, capturing a wide range of use cases and concerns across clinical, research, and educational domains.

However, several limitations must be acknowledged. Although participants were from diverse countries, they are unlikely to represent the full range of health care professionals within their regions. The sample was likely skewed toward individuals with greater access to technology and academic networks, especially in countries where access to ChatGPT or certain social media platforms may be restricted or limited. Therefore, findings should be interpreted with caution and may not be generalized to all health care professionals in low-resource or digitally restricted settings. The use of convenience and snowball sampling likely introduced self-selection bias, attracting participants with preexisting interest in technology or AI. Because of this sampling method, we could not calculate a response rate. Most respondents were from academic hospital settings in the United States, which may limit applicability to other regions or practice environments. Conducting the survey in English may have limited the global inclusivity. Given the swift pace of technological advancements, particularly in generative AI applications such as ChatGPT and the continuous process of learning and integration by health care professionals, the present survey may not accurately capture the current perceptions and attitudes of doctors and nurses toward these technologies [26], limiting the temporal relevance of our findings . Lastly, although our survey included open-ended questions, multiple-choice questions may have led participants to an available answer.

Future Directions

Further research is needed to address unanswered questions:

1. Long-term impact: Studies should evaluate how ChatGPT influences clinical outcomes, patient satisfaction, and educational performance over time.
2. Ethical frameworks: There is a pressing need for the development of institutional and regulatory guidelines governing AI use in health care [17].
3. Cross-language applications: Investigating how ChatGPT performs across different languages could help improve accessibility for non-English-speaking populations.
4. Training programs: Evidence-based strategies are needed to guide health care professionals in the ethical and effective use of generative AI technologies.

Conclusion

ChatGPT usage is expanding within health care settings due to its variety of capabilities, and the majority of health care professionals are likely aware of its availability. It can improve the caliber of writing papers, grants, and emails; help health care professionals in accessing the latest guidelines, diagnosis, and treatment suggestions; and possibly improve patient communication. There are several concerns related to the implementation of LLMs in clinical practice, including legal, ethical, and operational issues. Further research is necessary to clarify the role of ChatGPT and LLM-based generative AI tools in health care education, research, and clinical practice.

Acknowledgments

We thank Dr. Ognjen Gajic for critically reviewing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
ChatGPT Survey.
[PDF File, 59 KB - mededu_v11i1e58801_app1.pdf]

Multimedia Appendix 2
Diagram explaining survey flow.
[PNG File, 98 KB - mededu_v11i1e58801_app2.png]

Multimedia Appendix 3

Others within the demographic information table.
[DOCX File, 12 KB - mededu_v11i1e58801_app3.docx]

Multimedia Appendix 4
The distribution of answers to respondents with different levels of post-graduate experience.
[DOCX File, 24 KB - mededu_v11i1e58801_app4.docx]

References

1. Zhang K, Meng X, Yan X, et al. Revolutionizing health care: the transformative impact of large language models in medicine. J Med Internet Res 2025 Jan 7;27:e59069. [doi: 10.2196/59069] [Medline: 39773666]
2. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst 2023 Mar 4;47(1):33. [doi: 10.1007/s10916-023-01925-4] [Medline: 36869927]
3. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023;6:1169595. [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
4. Mu Y, He D. The potential applications and challenges of chatgpt in the medical field. Int J Gen Med 2024;17:817-826. [doi: 10.2147/IJGM.S456659] [Medline: 38476626]
5. Tangsrivimol JA, Darzidehkalani E, Virk HUH, et al. Benefits, limits, and risks of ChatGPT in medicine. Front Artif Intell 2025;8:1518049. [doi: 10.3389/frai.2025.1518049] [Medline: 39949509]
6. Sallam M, Salim NA, Barakat M, Al-Tammemi AB. ChatGPT applications in medical, dental, pharmacy, and public health education: a descriptive study highlighting the advantages and limitations. Narra J 2023 Apr;3(1):e103. [doi: 10.52225/narra.v3i1.103] [Medline: 38450035]
7. Thomae AV, Witt CM, Barth J. Integration of ChatGPT into a course for medical students: explorative study on teaching scenarios, students' perception, and applications. JMIR Med Educ 2024 Aug 22;10:e50545. [doi: 10.2196/50545] [Medline: 39177012]
8. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
9. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for ai-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
10. Chen SY, Kuo HY, Chang SH. Perceptions of ChatGPT in healthcare: usefulness, trust, and risk. Front Public Health 2024;12:1457131. [doi: 10.3389/fpubh.2024.1457131] [Medline: 39346584]
11. Kim JK, Chua M, Rickard M, Lorenzo A. ChatGPT and large language model (LLM) chatbots: the current state of acceptability and a proposal for guidelines on utilization in academic medicine. J Pediatr Urol 2023 Oct;19(5):598-604. [doi: 10.1016/j.jpurol.2023.05.018] [Medline: 37328321]
12. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023 Jul 6;6(1):120. [doi: 10.1038/s41746-023-00873-0] [Medline: 37414860]
13. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208. [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
14. Rossi R, Socci V, Pacitti F, et al. Mental health outcomes among frontline and second-line health care workers during the Coronavirus Disease 2019 (COVID-19) pandemic in Italy. JAMA Netw Open 2020 May 1;3(5):e2010185. [doi: 10.1001/jamanetworkopen.2020.10185] [Medline: 32463467]
15. Lu L, Zhu Y, Yang J, et al. Healthcare professionals and the public sentiment analysis of ChatGPT in clinical practice. Sci Rep 2025;15(1):1223. [doi: 10.1038/s41598-024-84512-y]
16. Khan AA, Khan AR, Munshi S, et al. Assessing the performance of ChatGPT in medical ethical decision-making: a comparative study with USMLE-based scenarios. J Med Ethics 2025 Jan 25:jme-2024-110240. [doi: 10.1136/jme-2024-110240] [Medline: 39863417]
17. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. J Med Internet Res 2023 Aug 11;25:e48009. [doi: 10.2196/48009] [Medline: 37566454]
18. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887. [doi: 10.3390/healthcare11060887] [Medline: 36981544]
19. De Angelis L, Baglivo F, Arzilli G, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front Public Health 2023;11:1166120. [doi: 10.3389/fpubh.2023.1166120] [Medline: 37181697]
20. Chatterjee J, Dethlefs N. This new conversational AI model can be your friend, philosopher, and guide... and even your worst enemy. Patterns (N Y) 2023 Jan 13;4(1):100676. [doi: 10.1016/j.patter.2022.100676] [Medline: 36699746]
21. Stokel-Walker C, Van Noorden R. What ChatGPT and generative AI mean for science. Nature New Biol 2023 Feb;614(7947):214-216. [doi: 10.1038/d41586-023-00340-6] [Medline: 36747115]

22. Gao CA, Howard FM, Markov NS, et al. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. NPJ Digit Med 2023 Apr 26;6(1):75. [doi: 10.1038/s41746-023-00819-6] [Medline: 37100871]
23. Rahimi F, Talebi Bezmin Abadi A. ChatGPT and publication ethics. Arch Med Res 2023 Apr;54(3):272-274. [doi: 10.1016/j.arcmed.2023.03.004] [Medline: 36990890]
24. Stöhr C, Ou AW, Malmström H. Perceptions and usage of AI chatbots among students in higher education across genders, academic levels and fields of study. Computers and Education: Artificial Intelligence 2024 Dec;7:100259. [doi: 10.1016/j.caeai.2024.100259] [Medline: 100259]
25. Sivarajah U, Wang Y, Olya H, Mathew S. Responsible artificial intelligence (AI) for digital health and medical analytics. Inf Syst Front 2023 Jun 5;2023:1-6. [doi: 10.1007/s10796-023-10412-7] [Medline: 37361886]
26. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. Acad Med 2024 Jan 1;99(1):22-27. [doi: 10.1097/ACM.0000000000005439] [Medline: 37651677]

Abbreviations

AI: artificial intelligence
LLM: large language model
USMLE: United States Medical Licensing Examination

Edited by B Lesselroth; submitted 25.03.24; peer-reviewed by F Chen, K Achara, S Mao; revised version received 11.04.25; accepted 19.04.25; published 12.05.25.

Please cite as:
Ozkan E, Tekin A, Ozkan MC, Cabrera D, Niven A, Dong Y
Global Health care Professionals' Perceptions of Large Language Model Use In Practice: Cross-Sectional Survey Study
JMIR Med Educ 2025;11:e58801
URL: https://mededu.jmir.org/2025/1/e58801
doi:10.2196/58801

Quantifying Emergency Medicine Residency Learning Curves Using Natural Language Processing: Retrospective Cohort Study

Carl Preiksaitis, MD, MEd; Joshua Hughes, MD; Rana Kabeer, MD, MPH; William Dixon, MD, MSED; Christian Rose, MD

Department of Emergency Medicine, Stanford University School of Medicine, 900 Welch Road, Suite 350, Palo Alto, CA, United States

Corresponding Author:
Carl Preiksaitis, MD, MEd
Department of Emergency Medicine, Stanford University School of Medicine, 900 Welch Road, Suite 350, Palo Alto, CA, United States

Abstract

Background: The optimal duration of emergency medicine (EM) residency training remains a subject of national debate, with the Accreditation Council for Graduate Medical Education considering standardizing all programs to 4 years. However, empirical data on how residents accumulate clinical exposure over time are limited. Traditional measures, such as case logs and diagnostic codes, often fail to capture the breadth and depth of diagnostic reasoning. Natural language processing (NLP) of clinical documentation offers a novel approach to quantifying clinical experiences more comprehensively.

Objective: This study aimed to (1) quantify how EM residents acquire clinical topic exposure over the course of training, (2) evaluate variation in exposure patterns across residents and classes, and (3) assess changes in workload and case complexity over time to inform the discussion on optimal program length.

Methods: We conducted a retrospective cohort study of EM residents at Stanford Hospital, analyzing 244,255 emergency department encounters from July 1, 2016, to November 30, 2023. The sample included 62 residents across 4 graduating classes (2020‐2023), representing all primary training site encounters where residents served as primary or supervisory providers. Using a retrieval-augmented generation NLP pipeline, we mapped resident clinical documentation to the 895 subcategories of the 2022 Model for Clinical Practice of Emergency Medicine (MCPEM) via intermediate mapping to the Systematized Nomenclature of Medicine, Clinical Terms, Clinical Observations, Recordings, and Encoding problem list subset. We generated cumulative topic exposure curves, quantified the diversity of topic coverage, assessed variability between residents, and analyzed the progression in clinical complexity using Emergency Severity Index (ESI) scores and admission rates.

Results: Residents encountered the largest increase in new topics during postgraduate year 1 (PGY1), averaging 376.7 (42.1%) unique topics among a total of 895 MCPEM subcategories. By PGY4, they averaged 565.9 (63.2%) topics, representing a 9.9% (51/515) increase over PGY3. Exposure plateaus generally occurred at 39 to 41 months, although substantial individual variation was observed, with some residents continuing to acquire new topics until graduation. Annual case volume more than tripled from PGY1 (mean 445.7, SD 112.7 encounters) to PGY4 (mean 1528.4, SD 112.7 encounters). Case complexity increased, as evidenced by a decrease in mean ESI score from 2.94 to 2.79, and a rise in high-acuity (ESI 1‐2) cases from 16% (4374/27,340) to 30.9% (9418/30,466).

Conclusions: NLP analysis of clinical documentation provides a scalable, detailed method for tracking EM residents' clinical exposure and progression. Many residents continue to gain new experiences into their fourth year, particularly in higher-acuity cases. These findings suggest that a 4-year training model may offer meaningful additional educational value, while also highlighting the importance of individualized assessment given the variability in learning trajectories.

(JMIR Med Educ 2025;11:e82326) doi:10.2196/82326

KEYWORDS

emergency medicine residency; clinical exposure; learning curves; natural language processing; electronic health records; graduate medical education

Introduction

The Challenge of Measuring Clinical Experience

The optimal duration of emergency medicine (EM) residency training remains a critical unresolved question in graduate medical education. As the Accreditation Council for Graduate Medical Education considers standardizing all programs to a 4-year model, this debate has highlighted a fundamental gap in our understanding. We lack reliable methods to measure how residents accumulate and master clinical experience across a vast spectrum of EM presentations [1]. While these clinical

encounters form the foundation of physician development and shape future practice patterns, programs have struggled to systematically track and optimize them, even within competency-based educational frameworks [2-5].

Current methods for measuring clinical exposure in EM suffer from both practical and conceptual limitations. Self-reported case logs, the traditional standard, demonstrate significant error rates due to recall bias [6,7]. Approaches using diagnostic coding systems such as the International Classification of Diseases, 10th revision, fundamentally misalign with EM's paradigm [8-11]. The core work of emergency physicians, evaluating and ruling out life-threatening conditions, often results in nonspecific final diagnoses (eg, "abdominal pain") that mask the complexity of care delivered [12]. Furthermore, we know that evaluating for life-threatening illnesses within the context of abdominal pain can be confounded by mimics such as acute coronary syndrome, which are nonspecific, have a high overlap with other conditions, and while they are considered, may not appear in the final diagnosis but nonetheless expose the resident to evaluating for that condition.

A Novel Natural Language Processing–Based Approach

Natural language processing (NLP) of clinical documentation offers a potential breakthrough. By analyzing the comprehensive narrative content of clinical notes that capture the diagnostic reasoning process, NLP can measure case exposure with greater granularity. When combined with the Systematized Nomenclature of Medicine–Clinical Terms, a clinical terminology system designed to represent complex medical concepts, this approach can systematically document both the breadth of conditions evaluated and the depth of clinical reasoning used. This methodological advance aligns with emerging calls for "precision education" in medical training, where the analysis of clinical data drives personalized learning optimization [5,13].

Study Objectives

In this study, we used NLP to provide a comprehensive, data-driven analysis of EM resident development. Using electronic health record data from a single academic medical center spanning multiple resident cohorts, we pursued three objectives: (1) quantify topic exposure curves and clinical progression, mapping how residents accumulate diagnostic topic exposure over time; (2) examine variation in clinical exposure patterns between individual residents and graduating classes; and (3) analyze the distribution of clinical experiences across presentation types and complexity levels to provide empirical evidence relevant to the debate on optimal training duration.

Methods

Study Design and Population

We conducted a retrospective analysis of emergency department (ED) encounters at Stanford Hospital between July 1, 2016, and November 30, 2023. Stanford Hospital serves as the primary training site for our residency and is a high-volume academic ED and level 1 trauma center. This period captured the primary training site experiences of 4 resident classes (2020 - 2023). The study included all EM residents who completed their full 4-year training during this period (n=62). The resident cohort had a mean age of 29.0 (SD 3.6) years, and 40 (64.5%) residents were male. The cohort was predominantly White (n=49, 79%) and non-Hispanic (n=60, 96.8%), with other racial identities including Asian (n=11, 17.7%), Black (n=1, 1.6%), and other (n=1, 1.6%). Encounters in which residents served as either the primary or supervisory resident were included. Resident-patient encounters were identified using the electronic health record's treatment team data. A rule-based algorithm was developed to attribute each encounter to the appropriate residents. Primary attribution was assigned to the first resident to document their involvement with a patient, and this was cross-referenced with shift schedule data to ensure that the encounter occurred during the resident's clinical duties. To account for both primary and supervisory roles, encounters were also coattributed to a senior resident if they were documented on the treatment team in close temporal proximity to a junior resident, reflecting an active supervisory role. When both junior and senior residents were assigned to an encounter, we credited topic exposure to both residents for resident-level analysis but counted each encounter once for patient and encounter-level summaries; per–postgraduate year (PGY) proportions used denominators based on unique, non-double-counted encounters. Encounters from nonclinical or administrative shifts were excluded. Our program is a 4-year PGY1 to 4 program where the PGY4 year includes 40 weeks of ED time with a specific emphasis on developing supervisory skills and practicing with graduated responsibility.

Residents also gain clinical experience at 2 high-volume, high-acuity affiliated sites (Kaiser Santa Clara and Santa Clara Valley Medical Center) that use a different electronic health record. These data were not included in our analysis. Depending on the PGY level, these external sites constitute approximately 30% to 35% of a resident's total ED training time. Encounters from off-service rotations (eg, intensive care unit, anesthesia, and obstetrics) were also excluded. Residents who did not complete the program were excluded from this analysis to ensure the integrity of longitudinal exposure curve construction.

Ethical Considerations

This study was approved as minimal-risk research by the Stanford University Institutional Review Board (IRB 69107). A waiver of informed consent was granted because the study involved secondary analysis of existing clinical documentation. Data access was authorized through the Stanford Research Repository. All data were deidentified and analyzed within a secure, Health Insurance Portability and Accountability Act–compliant environment, and no identifiable information left the repository. The study complied with institutional and national regulations for human participants research.

Data Sources and Variables

We extracted deidentified structured data, including patient demographics, Emergency Severity Index (ESI), and disposition status, as well as unstructured data in the form of clinical documentation from the Stanford Research Repository [14]. We focused on note sections capturing resident diagnostic

reasoning—history of present illness, medical decision-making, and ED course narratives.

NLP Overview

We developed a multistage NLP pipeline to map the narrative content of resident clinical documentation to the 895 clinical subcategories of the 2022 Model for Clinical Practice of Emergency Medicine (MCPEM) [15]. Our approach used a retrieval-augmented generation framework with a Health Insurance Portability and Accountability Act–compliant instance of Google's Gemini 1.5 Flash large language model, which was selected for its balance of cost-effectiveness and high performance within our institution's available tools [16-18]. This involved extracting key clinical concepts from resident notes and mapping them, as an intermediate step, to the Systematized Nomenclature of Medicine–Clinical Terms, Clinical Observations, Recordings, and Encoding subset before final classification into MCPEM topics [19]. This 2-stage process was chosen to preserve granular clinical detail while using a standardized clinical terminology.

Our retrieval-augmented pipeline can be conceptualized using standard NLP terminology as a 3-stage information retrieval (IR) to information extraction (IE) to classification process. The IR component selects relevant note sections and retrieves candidate concept matches from the Systematized Nomenclature of Medicine–Clinical Terms. The IE component, implemented implicitly through the language model, identifies and normalizes key medical entities and filters them for contextual relevance (eg, negated, historical, or uncertain findings). The classification stage maps these normalized concepts to the MCPEM topics.

Validation Methodology

We validated this pipeline through a manual review of 500 randomly selected encounters by 4 board-certified emergency physicians (CP, WD, JH, and RK). This process confirmed the high accuracy of our automated approach, with the model's classifications agreeing with the expert consensus 89.76% (377/420) of the time. The interrater reliability among the physician reviewers was substantial (κ=0.71). A comprehensive description of the NLP architecture, model configuration, and validation methodology is provided in Multimedia Appendix 1.

Analysis

Our analysis focused on 3 key aspects of resident development. To appropriately account for the resident as the primary unit of analysis and the clustered nature of the data (ie, multiple encounters nested within each resident), all encounter-level data were first aggregated to the individual resident level. Statistical comparisons were performed on this resident-level dataset (N=62).

First, we constructed topic exposure curves by tracking cumulative unique topics over time. Topic exposure rates were calculated using 30-day sliding windows. We defined exposure plateaus as periods where residents encountered fewer than 1 new topic per 100 patients over 3 consecutive measurement windows. Second, we examined variation in exposure by analyzing differences in case volumes, topic coverage, and patient acuity. To quantify the equity of exposure distribution among residents within the same PGY level, we calculated the Gini coefficients. Originally developed to measure income inequality, the Gini coefficient quantifies the inequality of a frequency distribution, with values ranging from 0 (representing perfect equality) to 1 (representing perfect inequality) [20]. In this context, a Gini coefficient of 0 would indicate that all residents in a cohort had the exact same volume of exposure to a given measure (eg, high-acuity cases), while a value of 1 would indicate that a single resident received all of the exposure and all others received none. This metric provides a standardized way to compare the degree of interresident variability across different clinical domains and training years. Third, we tracked clinical complexity progression using ESI scores and admission rates as proxies.

Analyses were performed using R (version 4.3.1; R Foundation for Statistical Computing) and Python (version 3.11). Statistical comparisons between classes were performed using Kruskal-Wallis tests with eta-squared effect sizes. A *P* value of <.05 was considered to indicate statistical significance.

Results

Resident Cohort

Our analysis included the primary-site training experiences of 62 EM residents from 4 graduating classes (2020 - 2023). Over the course of their training, this cohort managed 244,255 patient encounters, representing 133,748 (54.8%) unique patients. Detailed demographic and clinical characteristics of the patient encounters are provided in Table 1.

Table . Demographic and clinical characteristics of emergency department patient encounters managed by emergency medicine residents at Stanford Hospital, 2016 to 2023 (N=133,748).

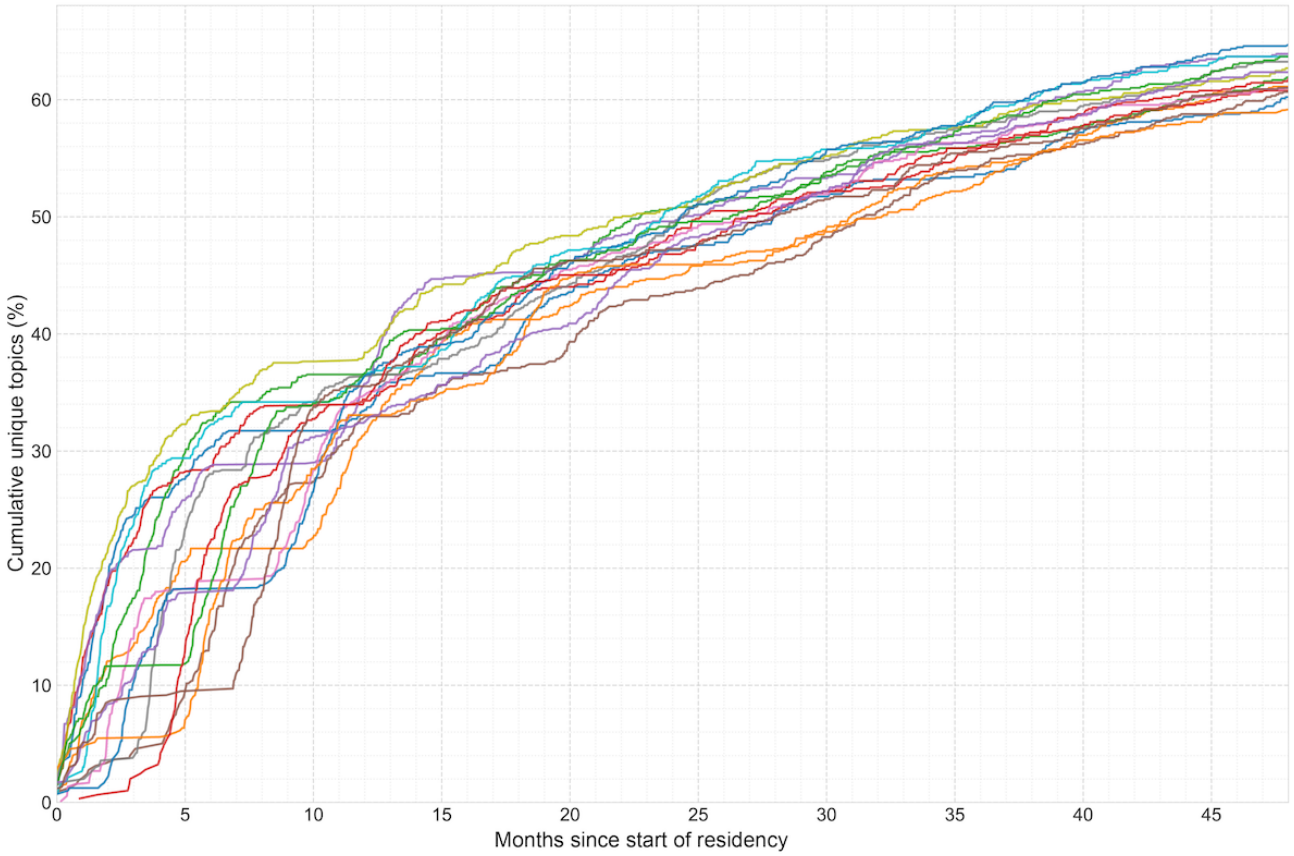
Characteristic	Values
Patient demographics	
Age (y), mean (SD)	47.2 (25.7)
Sex, n (%)	
Female	69,015 (51.6)
Male	64,666 (48.4)
Unknown	67 (0.0)
Race, n (%)	
Asian	22,994 (17.2)
Black	6418 (4.8)
Native American	414 (0.3)
Pacific Islander	2469 (1.9)
Unknown	1564 (1.2)
White	55,948 (41.8)
Other or multiple	43,939 (32.9)
Ethnicity, n (%)	
Hispanic or Latino	35,132 (26.3)
Non-Hispanic	96,965 (72.5)
Unknown	1649 (1.2)
Insurance, n (%)	
Commercial	48,894 (36.6)
Medicare	30,357 (22.7)
Medicaid or Medi-Cal	33,632 (25.2)
Unknown	11,617 (8.7)
Other	9242 (6.9)
Primary language, n (%)	
English	110,165 (82.4)
Spanish	16,371 (12.2)
Chinese languages	2257 (1.7)
Southeast Asian languages	1513 (1.1)
Other	3339 (2.5)
Interpreter services, n (%)	
Interpreter needed	20,483 (15.3)
No interpreter needed	113,162 (84.6)
Unknown	103 (0.1)
Encounter characteristics	
Total patient encounters, n	244,255
Encounters per resident, mean (95% CI)	3,940 (3,794-4,086)
Admission rate, %	35.7
Length of stay (h), median (IQR)	4.8 (3.1 - 7.2)
Emergency Severity Index, n (%)	
1	2759 (1.2)

Characteristic	Values
2	57,870 (24.1)
3	157,556 (65.7)
4	19,766 (8.2)
5	1894 (0.8)
Missing	4410 (1.8)

Topic Exposure Progression and Interresident Variation

EM residents demonstrated a clear, progressive acquisition of clinical topic exposure throughout training (Figures 1-4). The most rapid new topic exposure occurred during postgraduate year 1 (PGY1), with residents encountering a mean of 376.7 (42.1%) unique topics of the 895 MCPEM subcategories. However, PGY1 was also the year with the greatest interresident variability in topic coverage (coefficient of variation [CV]=9.1%). As residents progressed through training, the variation in total topic exposure decreased (PGY4 CV=2.8%).

Figure 1. Cumulative clinical topic exposure curve for the emergency medicine graduating class of 2020 at Stanford Hospital, 2016 - 2020 (n=15). The x-axis represents months of training, and the y-axis represents the mean cumulative number of unique clinical topics from the Model for Clinical Practice of Emergency Medicine encountered by residents. The curve does not reach the total of 895 Model for Clinical Practice of Emergency Medicine topics, indicating that no resident achieved 100% topic exposure during their training at the primary academic site.



Exposure coverage continued to increase in subsequent years, reaching a mean of 447.6 (50%) of MCPEM topics in PGY2, 515.0 (57.5%) in PGY3, and 565.9 (63.2%) in PGY4. Exposure plateaus, defined as periods with minimal new topic exposure, typically occurred in the fourth year of training; for example, the Class of 2023 reached plateaus at a mean of 39.8 (SD 3.0) months (mean 3268 encounters). However, individual variation was substantial: some residents plateaued as early as 38.2 months (2742 encounters), while others continued to encounter new topics through their final month, with final topic coverage ranging from 59.44% (532/895) to 67.26% (602/895) of all possible topics among PGY4 residents.

Figure 2. Cumulative clinical topic exposure curve for the emergency medicine graduating class of 2021 at Stanford Hospital, 2017 - 2021 (n=16). The x-axis represents months of training, and the y-axis represents the mean cumulative number of unique clinical topics from the Model for Clinical Practice of Emergency Medicine encountered by residents. The curve does not reach the total of 895 Model for Clinical Practice of Emergency Medicine topics, indicating that no resident achieved 100% topic exposure during their training at the primary academic site.

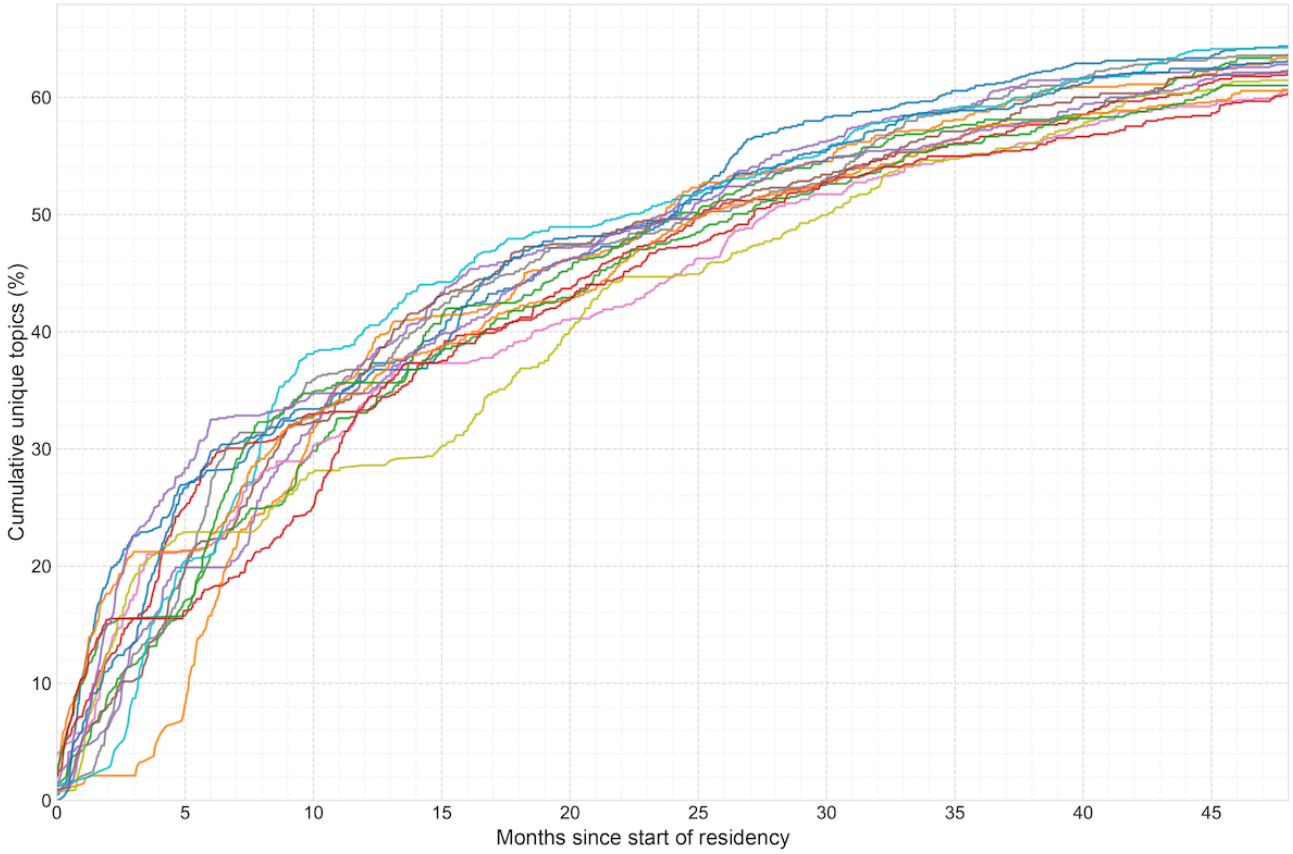


Figure 3. Cumulative clinical topic exposure curve for the emergency medicine graduating class of 2022 at Stanford Hospital, 2018 - 2022 (n=15). The x-axis represents months of training, and the y-axis represents the mean cumulative number of unique clinical topics from the Model for Clinical Practice of Emergency Medicine encountered by residents. The curve does not reach the total of 895 Model for Clinical Practice of Emergency Medicine topics, indicating that no resident achieved 100% topic exposure during their training at the primary academic site.

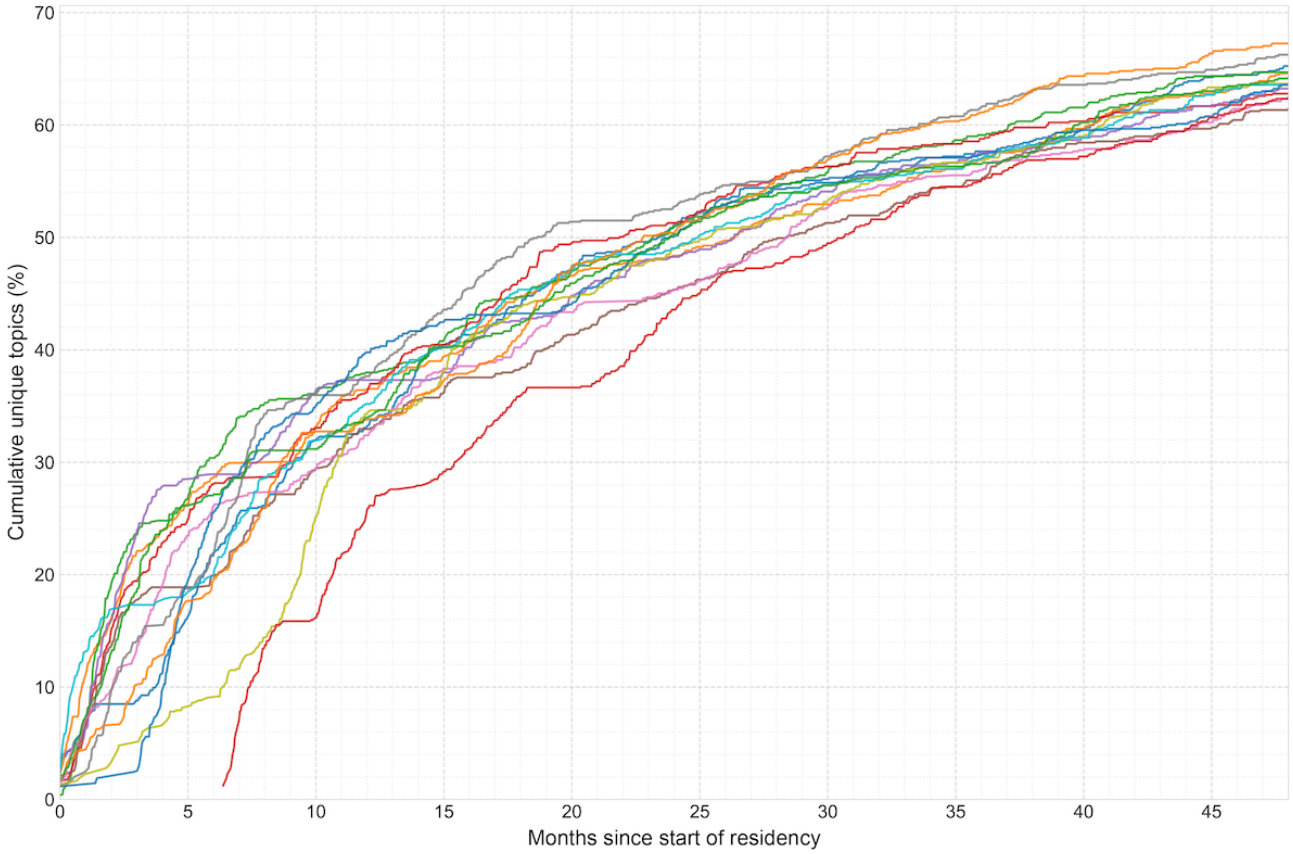
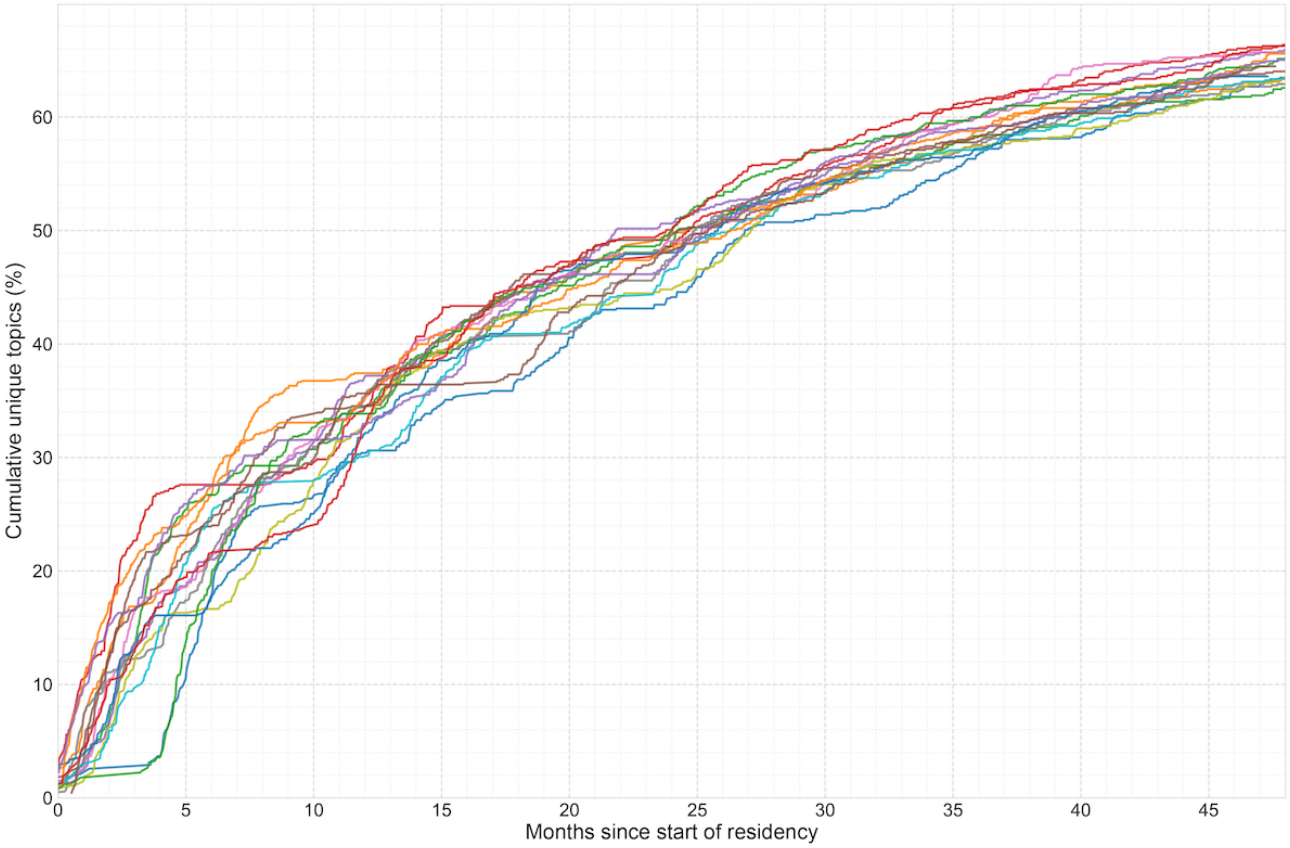


Figure 4. Cumulative clinical topic exposure curve for the emergency medicine graduating class of 2023 at Stanford Hospital, 2019 - 2023 (n=16). The x-axis represents months of training, and the y-axis represents the mean cumulative number of unique clinical topics from the Model for Clinical Practice of Emergency Medicine encountered by residents. The curve does not reach the total of 895 Model for Clinical Practice of Emergency Medicine topics, indicating that no resident achieved 100% topic exposure during their training at the primary academic site.



Progression of Clinical Workload and Complexity

In parallel with increasing topic exposure, residents demonstrated a progressive increase in their clinical workload and the complexity of cases managed (Table 2). Annual case volumes more than tripled from PGY1 (mean 445.7 encounters) to PGY4 (mean 1528.4 encounters). This growth in volume was accompanied by increasing case acuity, as mean ESI scores progressively decreased from 2.94 in PGY1 to 2.79 in PGY4. Correspondingly, the proportion of high-acuity patients (ESI 1 - 2) managed by residents increased from 16% (4374/27,340) in PGY1 to 30.91% (9418/30,466) in PGY4, and admission rates rose from 31% (8475/27,340) to 37.50% (11425/30,466) over the same period. Notably, the CV for both clinical volume and admission rates followed a U-shaped pattern, decreasing from PGY1 to PGY3 before increasing again in PGY4.

Table . Progression of clinical experience by postgraduate year (PGY) for emergency medicine residents (N=62), Stanford Hospital, 2016 to 2023.

	PGY1	PGY2	PGY3	PGY4
Annual clinical volume and complexity				
Number of encounters, mean (95% CI)	445.7 (417.6 - 473.7)	772.1 (738.5 - 805.8)	1193.4 (1138.5 - 1248.2)	1528.4 (1429.1 - 1627.8)
ESI[a] score, mean (95% CI)	2.94 (2.93 - 2.96)	2.87 (2.85 - 2.88)	2.85 (2.84 - 2.87)	2.79 (2.76 - 2.81)
High-acuity cases (ESI 1 - 2), n/N (%)	4374/27,340 (16)	5575/27,479 (20.2)	7398/30,101 (24.5)	9418/30,466 (30.9)
Admission rate, % (95% CI)	31.0 (29.7 - 32.4)	38.3 (37.3 - 39.4)	35.5 (34.6 - 36.3)	37.5 (36.3 - 38.8)
Cumulative topic exposure				
Cumulative unique MCPEM[b] topics covered, mean (95% CI)	376.7 (367.9 - 385.5)	447.6 (442.2 - 453.0)	515.0 (510.9 - 519.0)	565.9 (561.9 - 569.9)
Cumulative total percentage of MCPEM covered, %	41.1	50	57.5	63.2
New topics encountered each year, mean (SD)	376.7 (28.4)	71.0 (18.1)	67.3 (14.1)	50.9 (13.3)
Interresident variability (CV[c]), %				
CV for clinical volume	24.6	17	18	25.4
CV for admission rate	16.4	11.1	9.5	13.4
CV for topic coverage	9.1	4.7	3.1	2.8

[a]ESI: Emergency Severity Index.

[b]MCPEM: Model for Clinical Practice of Emergency Medicine.

[c]CV: coefficient of variation.

Distribution of Clinical Experiences

Residents' exposure to the 895 distinct clinical topics followed a consistent pattern—they encountered a small core of presentations (n=49, 5.5%) more than 100 times each, a larger set (n=284, 31.7%) between 10 and 100 times, and the majority (n=562, 62.9%) fewer than 10 times. Topic distribution showed moderate inequality (mean Gini coefficient=0.611), with consistency between the graduating classes (P=.61). High-acuity case exposure was more unequally distributed (Gini=0.292) than the overall case volumes (Gini=0.117).

Discussion

Principal Findings

Our analysis of 62 EM residents across 4 years of training revealed distinct and progressive patterns in the arc of their clinical experience. Using a novel NLP methodology on over 244,000 clinical encounters, we found that residents demonstrated a rapid acquisition of topic exposure in their first year, which continued, albeit at a slower rate, deep into their fourth year. Importantly, this continued exposure occurs in the context of increasing clinical complexity and, based on our program's structure, escalating supervisory responsibility. These findings provide empirical evidence that can inform the national debate on the optimal length of EM training and highlight the potential for data-driven, precision education.

Implications for Competency-Based Medical Education

The observed pattern of topic exposure, rapid initial acquisition followed by a plateau, aligns with the power-law "experience curves" documented in medical education by Pusic et al [21] However, it is critical to note that case exposures are merely the substrate for learning and are not all equal in value toward developing competence. The conversion of these experiences into durable competence is a complex process mediated by the *deliberate* components of the theory of deliberate practice by Ericsson [22], such as feedback, reflection, and coaching [22-24]. Indeed, recent research has shown that even established competency measures, such as milestones, may not directly correlate with early-career patient outcomes, cautioning against a simple equation of more exposure with more competence [25]. Therefore, the primary value of the exposure data we present lies in its ability to serve as a powerful objective input for these established educational frameworks. At our institution, for example, these components are formalized through a resident coaching program and a quarterly Clinical Competency Committee, systems for which this objective exposure data can provide a more precise foundation for assessment and goal-setting, and tracking progress toward the Accreditation Council for Graduate Medical Education milestones [2,26].

Informing the Debate on Training Duration

A central question for the EM as a specialty is whether a 3- or 4-year training model is optimal. Our data provide empirical

evidence relevant to this debate. The finding that residents continue to acquire a mean of 50.9 new core topics in their PGY4 year, representing a 9.9% (51/515) increase over PGY3, suggests that the fourth year offers more than just redundant experience. This quantitative increase is accompanied by a significant qualitative shift—PGY4 residents manage a higher proportion of high-acuity patients (ESI 1 - 2) and cases requiring hospitalization. This exposure to a more complex and challenging case mix, as supported by work from Lam et al [10] and Zhou et al [27], is critical for developing the advanced diagnostic reasoning necessary for independent practice.

Understanding Interresident Variation

Our analysis also revealed significant interresident variation, particularly in the timing of exposure plateaus and in the experience with high-acuity cases. This aligns with findings from other specialties and supports a more personalized view of competency development [28]. This variability is likely to be driven by a combination of systemic factors and resident choices. We propose framing this variability within the concept of "warranted versus unwarranted variation" as described by Holmboe and Kogan [29]. While some variation is an expected and even desirable feature of individualized learning, our NLP-based tool provides a mechanism for programs to identify potentially unwarranted gaps in exposure to core experiences. Notably, the decreasing CV in total topic coverage from PGY1 (9.1%) to PGY4 (2.8%) suggests that a longer training duration may lead to a more standardized and equitable clinical experience among graduates.

The U-shaped pattern observed in the variability of clinical volume and admission rates is an intriguing finding. We hypothesize that this reflects the evolving roles within our program—PGY1 residents have variable schedules due to numerous off-service rotations; PGY2 and PGY3 residents assume more structured supervisory roles within a pod system, which may standardize their workflow and decrease variability; finally, in PGY4, residents are granted greater autonomy to run a pod independently while also supervising junior residents, potentially allowing individual practice styles to reemerge at increasing variability.

Bridging the Curriculum-Practice Gap and Methodological Advantages

A key finding of this study is that even after 4 years, residents were not exposed to over a third of the topics listed in the MCPEM (329/895, 36.76%). This highlights the fundamental tension between the prescribed curriculum and the reality of clinical practice. Achieving 100% topic coverage is likely an unachievable goal for any single residency program. The true aim of training is not encyclopedic exposure but rather developing the core competencies and adaptive expertise required for safe, independent practice and effective lifelong learning. Our NLP-based methodology offers a powerful, dual-pronged approach to address this gap. At the local program level, it serves as a diagnostic tool, enabling educators to identify and amend exposure gaps through targeted interventions, such as simulation. More broadly, the scalability of this approach presents an opportunity to transform the specialty's understanding of its own work. If applied across multiple institutions, this method could create a dynamic, data-driven map of the actual clinical practice of EM, providing an evidence base for future revisions of the MCPEM. This would allow the model to better reflect the true prevalence and complexity of conditions encountered in contemporary practice, ensuring a more authentic alignment between what is taught, what is tested, and what is practiced. Nonetheless, there will likely always remain a set of high-acuity, low-frequency conditions for which training programs must prescribe exposure through didactics, as real-world encounters will be too rare to ensure universal competence.

A key advantage of this methodology is its practical implementation. Traditional NLP pipelines often require a substantial upfront investment in manual data annotation for model training and specialized expertise. In contrast, our retrieval-augmented generation approach takes advantage of a pretrained large language model that requires no model-specific training, dramatically reducing the implementation complexity. To offer a concrete sense of financial feasibility, the total cost for the application programming interface calls to process all 244,255 clinical encounters for this study was approximately US $180, demonstrating the financial accessibility of this approach for programs seeking to adopt data-driven educational strategies. Framing our pipeline in IR and IE terms also clarifies where most misclassifications arise—either from retrieval scope or from information-extraction phenomena such as negation or temporality—providing a useful structure for future error analysis and comparison to traditional rule-based clinical NLP systems.

Limitations

This study has several limitations. As a single-institution study, our findings may not be generalizable to programs with a 3-year training format, or to institutions operating in different clinical settings. A primary limitation is that our analysis excludes data from additional training sites, which constitute a significant proportion of resident training (approximately 30% - 35% of ED time, depending on the PGY level), and from off-service rotation where key procedural and critical care exposures occur (eg, anesthesia, intensive care unit, etc). The training period for our cohorts also overlapped with the COVID-19 pandemic, which may have influenced case volumes and mix, although the remarkable consistency of exposure patterns we observed across the 4 classes mitigates this concern. Finally, although our NLP approach achieved high accuracy in classifying clinical encounters (89.7% agreement with physician review), this methodology relies on the comprehensiveness of resident documentation, which may vary between individuals and over time.

In conclusion, our analysis reveals both consistent patterns in resident clinical exposure and substantial individual variation in topic exposure trajectories. The finding that many residents continue to encounter new clinical topics into their fourth year provides empirical evidence for the potential educational value of a 4-year training model. NLP of clinical documentation offers EM programs a powerful and accessible tool to objectively measure and optimize resident clinical experiences based on actual exposure patterns, moving beyond traditional metrics to

foster a more precise, data-informed, and equitable approach to graduate medical education.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

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

Authors' Contributions

Conceptualization (lead): CP, CR
Methodology (lead): CP, JH; (supporting): RK, CR
Software (lead): CP
Data curation (lead): CP
Formal analysis (lead): CP
Writing–original draft (lead): CP, JH, CR; (supporting): RK, WD
Writing–review & editing (equal): CP, JH, RK, WD, CR

Conflicts of Interest

None declared.

Multimedia Appendix 1
Technical description of the natural language processing pipeline, validation procedures, and statistical analysis methods.
[DOCX File, 36 KB - mededu_v11i1e82326_app1.docx]

References

1. ACGME program requirements for graduate medical education in emergency medicine. Accreditation Council for Graduate Medical Education. 2025. URL: https://www.acgme.org/globalassets/pfassets/reviewandcomment/2025/110_emergencymedicine_rc_02122025.pdf [accessed 2025-03-17]
2. Sebok-Syer SS, Shepherd L, McConnell A, Dukelow AM, Sedran R, Lingard L. “EMERGing” electronic health record data metrics: insights and implications for assessing residents’ clinical performance in emergency medicine. AEM Educ Train 2020 Aug 9;5(2):e10501. [doi: 10.1002/aet2.10501] [Medline: 33898906]
3. Sirovich BE, Lipner RS, Johnston M, Holmboe ES. The association between residency training and internists’ ability to practice conservatively. JAMA Intern Med 2014 Oct;174(10):1640-1648. [doi: 10.1001/jamainternmed.2014.3337] [Medline: 25179515]
4. Chen C, Petterson S, Phillips R, Bazemore A, Mullan F. Spending patterns in region of residency training and subsequent expenditures for care provided by practicing physicians for Medicare beneficiaries. JAMA 2014 Dec 10;312(22):2385-2393. [doi: 10.1001/jama.2014.15973] [Medline: 25490329]
5. Burk-Rafel J, Drake CB, Sartori DJ. Characterizing residents’ clinical experiences-a step toward precision education. JAMA Netw Open 2024 Dec 2;7(12):e2450774. [doi: 10.1001/jamanetworkopen.2024.50774] [Medline: 39693075]
6. Langdorf MI, Strange G, Macneil P. Computerized tracking of emergency medicine resident clinical experience. Ann Emerg Med 1990 Jul;19(7):764-773. [doi: 10.1016/s0196-0644(05)81700-7] [Medline: 2389860]
7. Nagler J, Harper MB, Bachur RG. An automated electronic case log: using electronic information systems to assess training in emergency medicine. Acad Emerg Med 2006 Jul;13(7):733-739. [doi: 10.1197/j.aem.2006.02.010] [Medline: 16723724]
8. Rhee DW, Reinstein I, Jrada M, et al. Mapping hospital data to characterize residents’ educational experiences. BMC Med Educ 2022 Jun 25;22(1):496. [doi: 10.1186/s12909-022-03561-x] [Medline: 35752814]
9. Rhee DW, Chun JW, Stern DT, Sartori DJ. Experience and education in residency training: capturing the resident experience by mapping clinical data. Acad Med 2022 Feb 1;97(2):228-232. [doi: 10.1097/ACM.0000000000004162] [Medline: 33983144]
10. Lam AC, Tang B, Liu C, et al. Variation in case exposure during internal medicine residency. JAMA Netw Open 2024 Dec 2;7(12):e2450768. [doi: 10.1001/jamanetworkopen.2024.50768] [Medline: 39693070]
11. Lam AC, Tang B, Lalwani A, et al. Methodology paper for the General Medicine Inpatient Initiative Medical Education Database (GEMINI MedED): a retrospective cohort study of internal medicine resident case-mix, clinical care and patient outcomes. BMJ Open 2022 Sep 23;12(9):e062264. [doi: 10.1136/bmjopen-2022-062264] [Medline: 36153026]

12. Bischof JJ, Emerson G, Mitzman J, Khandelwal S, Way DP, Southerland LT. Does the emergency medicine in-training examination accurately reflect residents' clinical experiences? AEM Educ Train 2019 Sep 19;3(4):317-322. [doi: 10.1002/aet2.10381] [Medline: 31637348]
13. Triola MM, Burk-Rafel J. Precision medical education. Acad Med 2023 Jul 1;98(7):775-781. [doi: 10.1097/ACM.0000000000005227] [Medline: 37027222]
14. Weber SC, Pallas J, Olson G, et al. Compliant self service access to secondary use clinical data at stanford medicine. arXiv. Preprint posted online on Dec 5, 2024. [doi: 10.48550/arxiv.2412.04248]
15. Beeson MS, Bhat R, Broder JS, et al. The 2022 model of the clinical practice of emergency medicine. J Emerg Med 2023 Jun;64(6):659-695. [doi: 10.1016/j.jemermed.2023.02.016] [Medline: 37244783]
16. Gemini models. Google AI for Developers. URL: https://ai.google.dev/gemini-api/docs/models [accessed 2025-03-17]
17. Klang E, Tessler I, Apakama DU, et al. Assessing retrieval-augmented large language model performance in emergency department ICD-10-CM coding compared to human coders. medRxiv. Preprint posted online on Oct 17, 2024. [doi: 10.1101/2024.10.15.24315526] [Medline: 39484238]
18. D'Oosterlinck K, Khattab O, Remy F, Demeester T, Develder C, Potts C. In-context learning for extreme multi-label classification. arXiv. Preprint posted online on Jan 22, 2024. [doi: 10.48550/arXiv.2401.12178]
19. The CORE problem list subset of SNOMED CT®. US National Library of Medicine. URL: https://www.nlm.nih.gov/research/umls/Snomed/core_subset.html [accessed 2025-03-17]
20. Thomas V, Wang Y, Fan X. Measuring education inequality: Gini coefficients of education. Social Science Research Network. 2001. URL: https://papers.ssrn.com/abstract=258182 [accessed 2025-06-12]
21. Pusic MV, Kessler D, Szyld D, Kalet A, Pecaric M, Boutis K. Experience curves as an organizing framework for deliberate practice in emergency medicine learning. Acad Emerg Med 2012 Dec;19(12):1476-1480. [doi: 10.1111/acem.12043] [Medline: 23230958]
22. Ericsson KA. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. Acad Med 2004 Oct;79(10 Suppl):S70-S81. [doi: 10.1097/00001888-200410001-00022] [Medline: 15383395]
23. Natesan S, Jordan J, Sheng A, et al. Feedback in medical education: an evidence-based guide to best practices from the Council of Residency Directors in Emergency Medicine. West J Emerg Med 2023 May 5;24(3):479-494. [doi: 10.5811/westjem.56544] [Medline: 37278777]
24. Fredette J, Michalec B, Billet A, et al. A qualitative assessment of emergency medicine residents' receptivity to feedback. AEM Educ Train 2021 Aug 1;5(4):e10658. [doi: 10.1002/aet2.10658] [Medline: 34527849]
25. Kendrick DE, Thelen AE, Chen X, et al. Association of surgical resident competency ratings with patient outcomes. Acad Med 2023 Jul 1;98(7):813-820. [doi: 10.1097/ACM.0000000000005157] [Medline: 36724304]
26. Sebok-Syer SS, Shaw JM, Sedran R, et al. Facilitating residents' understanding of electronic health record report card data using faculty feedback and coaching. Acad Med 2022 Nov 1;97(11S):S22-S28. [doi: 10.1097/ACM.0000000000004900] [Medline: 35947480]
27. Zhou L, Cai C, Wu R, Qi Y. Effectiveness of scaffolded case-based learning in anesthesiology residency training: a randomized controlled trial. BMC Med Educ 2025 May 8;25(1):672. [doi: 10.1186/s12909-025-07236-1] [Medline: 40340724]
28. Frank JR, Snell LS, Cate OT, et al. Competency-based medical education: theory to practice. Med Teach 2010;32(8):638-645. [doi: 10.3109/0142159X.2010.501190] [Medline: 20662574]
29. Holmboe ES, Kogan JR. Will any road get you there? Examining warranted and unwarranted variation in medical education. Acad Med 2022 Aug 1;97(8):1128-1136. [doi: 10.1097/ACM.0000000000004667] [Medline: 35294414]

Abbreviations

CV: coefficient of variation
ED: emergency department
EM: emergency medicine
ESI: Emergency Severity Index
IE: information extraction
IR: information retrieval
MCPEM: Model for Clinical Practice of Emergency Medicine
NLP: natural language processing
PGY: postgraduate year

Edited by R Pellegrino; submitted 13.08.25; peer-reviewed by D Chartash, LH Yao; revised version received 10.10.25; accepted 09.11.25; published 09.12.25.

Please cite as:
Preiksaitis C, Hughes J, Kabeer R, Dixon W, Rose C
Quantifying Emergency Medicine Residency Learning Curves Using Natural Language Processing: Retrospective Cohort Study
JMIR Med Educ 2025;11:e82326
URL: https://mededu.jmir.org/2025/1/e82326
doi:10.2196/82326

AI-Generated "Slop" in Online Biomedical Science Educational Videos: Mixed Methods Study of Prevalence, Characteristics, and Hazards to Learners and Teachers

Eric M Jones[1], PhD; Jane D Newman[1], PhD; Boyun Kim[2], MA; Emily J Fogle[3], PhD

[1]Department of Foundational Medical Studies, Oakland University William Beaumont School of Medicine, 586 Pioneer Drive, Rochester, MI, United States

[2]Department of Human Development and Child Studies, Oakland University, Rochester, MI, United States

[3]Department of Chemistry and Biochemistry, California Polytechnic State University, San Luis Obispo, CA, United States

Corresponding Author:
Eric M Jones, PhD
Department of Foundational Medical Studies, Oakland University William Beaumont School of Medicine, 586 Pioneer Drive, Rochester, MI, United States

Abstract

Background: Video-sharing sites such as YouTube (Google) and TikTok (ByteDance) have become indispensable resources for learners and educators. The recent growth in generative artificial intelligence (AI) tools, however, has resulted in low-quality, AI-generated material (commonly called "slop") cluttering these platforms and competing with authoritative educational materials. The extent to which slop has polluted science education video content is unknown, as are the specific hazards to learning from purportedly educational videos made by AI without the use of human discretion.

Objective: This study aimed to advance a formal definition of slop (based on the recent theoretical construct of "careless speech"), to identify its qualitative characteristics that may be problematic for learners, and to gauge its prevalence among preclinical biomedical science (medical biochemistry and cell biology) videos on YouTube and TikTok. We also examined whether any quantitative features of video metadata correlate with the presence of slop.

Methods: An automated search of publicly available YouTube and TikTok videos related to 10 search terms was conducted in February and March 2025. After exclusion of duplicates, off-topic, and non-English results, videos were screened, and those suggestive of AI were flagged. The flagged videos were subject to a 2-stage qualitative content analysis to identify and code problematic features before an assignment of "slop" was made. Quantitative viewership data on all videos in the study were scraped using automated tools and compared between slop videos and the overall population.

Results: We define "slop" according to the degree of human care in production. Of 1082 videos screened (814 YouTube, 268 TikTok), 57 (5.3%) were deemed probably AI-generated and low-quality. From qualitative analysis of these and 6 additional AI-generated videos, we identified 16 codes for problematic aspects of the videos as related to their format or contents. These codes were then mapped to the 7 characteristics of careless speech identified earlier. Analysis of view, like, and comment rates revealed no significant difference between slop videos and the overall population.

Conclusions: We find slop to be not especially prevalent on YouTube and TikTok at this time. These videos have comparable viewership statistics to the overall population, although the small dataset suggests this finding should be interpreted with caution. From the slop videos that were identified, several features inconsistent with best practices in multimedia instruction were defined. Our findings should inform learners seeking to avoid low-quality material on video-sharing sites and suggest pitfalls for instructors to avoid when making high-quality educational materials with generative AI.

(JMIR Med Educ 2025;11:e80084) doi:10.2196/80084

KEYWORDS

generative AI; artificial intelligence; YouTube; TikTok; biochemistry education; medical biochemistry; cell biology education; basic medical sciences education; medical education; slop; careless speech

Introduction

Background

Video-sharing platforms such as YouTube (Google) and TikTok (ByteDance) have become entrenched features of the educational landscape. Both instructors and students rely on these resources for a variety of purposes relating to teaching and learning [1-3], and the inherent benefits of video instruction in science education, specifically, have been well-documented [4-6]. Nonetheless, these video-sharing platforms' greatest advantages—accessibility and low barriers to creating and sharing—are also arguably their greatest weaknesses, as the lack of barriers leads naturally to large amounts of low-quality material appearing alongside authoritative, high-value material. Because these platforms rely mostly on advertising for revenue, algorithms that recommend videos based on past views [7] and prioritize engaging over reliable videos [8] make it difficult to find the most credible and useful videos, not simply those most likely to maximize time on the site [8]. Furthermore, not all audiences may have the motivation or ability to assess the reliability of online videos [9,10], and in the absence of a shared standard of quality, one audience might find a video informative while another finds it inappropriate. Frameworks for assessing the quality of multimedia instruction have been advanced over the past few decades [6,11-13], but these are directed toward educators and instructional designers. How learners judge the quality of educational content remains poorly studied, and in the absence of guidance, students may simply defer to intuition [14].

The accessibility of social media platforms, and the lack of any uniform standard for judging quality, inherently present a challenge to any learner searching for educational content. The recent explosion in generative artificial intelligence (genAI) technologies, including large language models (LLMs) like ChatGPT (OpenAI) and Claude (Anthropic), and image generators like Stable Diffusion (Stable Diffusion AI) and Midjourney [15], has added further complication to this situation. To genAI users, time and effort are no longer barriers to generating shareable content. As a result, writing websites [16], social networks [17,18], and online markets [19] are increasingly cluttered with fake artificial intelligence (AI)–generated essays, posts, artwork, and merchandise. More troublingly, the scientific literature is now polluted with false machine-generated studies and data [20,21], often to push contrarian agendas [22]. The low-quality, high-volume, AI-generated content behind these examples has been called "slop" [19]; its proliferation led *MIT Technology Review* to deem slop the "biggest AI flop" of 2024 [23]. Slop lacks a single widely accepted definition, but journalists and industry commentators generally agree that slop is of low quality, ubiquitous, lacking in artistic or scientific value, and generated to maximize exposure or engagement, or simply to fill space on sharing platforms [16,19,24]. At best, slop is a distraction requiring time and effort to sift through in search of good material; at worst, it presents specific dangers to learners, educators, creative professions, and the overall atmosphere of public information [24].

The degree to which slop has crept into educational materials is largely unknown. Exploratory studies in medical and undergraduate science education suggest that carelessly produced genAI content can pose a significant risk of misunderstanding, spread outright misinformation [25,26], or promote deskilling or "metacognitive laziness" [25,27] in learners. Learners turn to AI content when they are most easily influenced—while uncertain or confused—and thus are most likely to be persuaded by errors or biases in the genAI output [28]. Yet errors abound in such content; in medical teaching, genAI output has been found to misrepresent rare diseases or those with variable presentations [29,30], and AI-generated anatomical diagrams often contain gross inaccuracies [31,32]. In other fields, genAI has been shown to create garbled chemical models and biased depictions of researchers [33,34], inaccurate but plausible-sounding descriptions of metabolic processes [35], and realistic images of nonexistent animal species, leading to confusion in biodiversity conservation efforts [36].

As for video, it is possible to create entirely AI-generated video clips using tools such as Synthesia and Sora (OpenAI); however, at the time of writing, these tools (unlike those above) are not yet available for free use, or free users are limited to very short clips. Consequently, fully AI-generated videos are not yet as widespread as images and text. This does not, however, mean video-sharing sites are free of slop. Free users can, for example, use AI tools to animate a photo of a "narrator," or stitch together stock or AI-generated images to create a longer video. Tutorials on making videos in this fashion are widely available [19,37].

The purpose of this study is to examine the reach and characteristics of lazily-made genAI content in online videos on preclinical biomedical sciences (medical biochemistry and cell biology; eg, Biochemistry & Nutrition and Cell Biology & Histology topics in the USMLE Foundational Sciences area, as these are the authors' disciplines of expertise). Although slop has been widely discussed in popular media, it has not yet received much scholarly attention, so our first priority is to establish a useful definition of slop in educational media. To this end, we use the theoretical framework of "careless speech" recently advanced [38] to propose legal-ethical responsibilities of genAI.

Theoretical Framework

The characteristics of slop are ultimately a consequence of genAI's intrinsic structure. All present genAI tools work by predicting associations between linguistic elements, based on human-reinforced training with real (human-generated) data. They cannot directly access external reality. They may therefore state falsehoods as fact or realistically depict impossible scenes (so-called hallucinations or confabulations), so long as the output correlates with training data [38]. GenAI is also prone to subtle errors or omissions in addition to outright falsehoods; it has difficulty grasping humor, nuance, or insinuation [26]; it does not exhibit a clear concept of uncertainty and tends to make confident assertions even where there is no clear answer [38,39]. Most genAI tools are programmed to sound authoritative and to give responses deferentially and in accord with a user's desires (sycophancy) [40]. For any AI-generated task, biased or incomplete information in the training data (eg, absence of

an important but uncommon viewpoint) will result in biased output [41,42]. The "speech" generated by AI (which we mean here to encompass not only language but also images, video, sounds, and other output), therefore, appears authoritative and competent but is unmoored from physical reality, lacks any motivations or principles, and carries whatever biases are present in the training data.

Can AI-generated speech be trusted, then? A recent paper by Wachter and colleagues [38] suggests the answer is no, unless there is cross-validation with the outside world, for example, using "human in the loop" [43,44] or "zero-shot translation" [45] approaches to verify accuracy, forestall errors, and account for uncertainty or caveats in the output. Wachter and colleagues use the term "careless speech" to describe the type of output that unsupervised AI produces. Careless speech is quasi-factual output that correlates with what humans say is reality (the training data), but without direct access to that reality; it is a coherent statement approximating a factual statement. Careless speech is not necessarily misinformation, as it is not always false. Rather, it is independent of reality; it may resemble truth, but exists separately.

We believe careless speech is a useful framework for establishing a definition of slop relevant to educational material. From this point onward, we use the term "material" to refer to a specific created object (eg, a video), and reserve the word "content" for the subject matter (contents) of these materials.

We define slop as follows: slop is any material, created mostly or entirely by generative AI, with little or no apparent human care toward the accuracy, fluency, or helpfulness of the material or of its most likely use or interpretation.

The operative word in this definition is *care*, to which we assign two meanings [46]: (1) deliberate attention and effort (eg, prompt design, editing, and fact-checking) toward ensuring the material has desirable characteristics, that is, to "care for" (taking on "responsibility to meet a need that has been identified" [46]); and (2) some professional or personal stake in the outcome, implying an ownership of and accountability for the product and its likely uses, that is, to "care about." AI-generated material that does not discernibly exhibit care in *both* senses of the term is slop, regardless of its accuracy. Our definition implies that any material made entirely by genAI is slop; material made by genAI with human intervention may or may not be slop, depending on the degree of care in the intervention.

Our definition makes no direct reference to the quality of the material. This is intentional. If one defines "quality" in terms of accuracy and realism, genAI is making tremendous strides in improving its quality. Yet just as the content (meaning here the messaging and subject matter) of AI output is, at best, incidentally accurate—in Wachter's words, "True responses are an accident of probability and reinforcement via human feedback, not agency or a conception of truth or intent to tell the truth" [38]—the quality of AI output is, at best, incidentally good. Without a caring human in the loop, AI output can only approximate, by correlation, characteristics associated with quality. Thus, slop is independent of quality in the same way that careless speech is independent of reality.

Likewise, our definition is agnostic regarding the purpose for which the AI-generated material is made. Slop is often made and disseminated to game engagement metrics (eg, clickthroughs, likes, and views), ultimately for the creator's financial or political gain [19,24]. However, it is possible that some slop is made with the genuine intent of informing or entertaining—only without adequate care to ensure this intent is fulfilled. We therefore believe intent and purpose are irrelevant to the definition of slop, particularly since the intentions of those generating the slop are generally unknown.

Educational Implications of Slop and Careless Speech

Careless speech was proposed as a framework for understanding the dangers of genAI output in legal settings. In educational settings, genAI is likely to present a distinct set of risks. Much of the research on educational hazards of genAI focuses narrowly on non-factual or hallucinated output [30-32,35,36] or on bias and ethical risks [34,47-50]. A more holistic understanding of the impact of careless speech on learning requires a broad framework for what makes educational materials effective at all.

With respect to video, the theory of multimedia learning [51] is one such framework and is supported by considerable empirical evidence [4]. Drawing from cognitive load theory, this model considers structural and design features that influence the effectiveness of multimedia materials. Complementary to this theory, other models emphasize the content of multimedia materials, specifically themes of active learning and features that promote engagement with the video [6,12,13]. Thus, the effectiveness of multimedia educational tools may be seen as having both design or structural components and content components that support learning.

Because the concern of this study is the educational effectiveness of video, some might question our choice to focus on AI slop rather than low-quality video more generally. There are at least 2 reasons why slop deserves particular attention. First, the sheer volume of slop is already overwhelming online platforms [16,18,19]. Students are thus extremely likely to encounter slop, which may soon comprise the majority of low-quality video. Second, instructors wishing to use genAI responsibly need to know the likely failure modes of the technology, so that proper attention can be paid to avoiding these pitfalls, maximizing educational impact, and reinforcing the necessity of human judgment in human-centered professions [52].

This study examines the current prevalence and problematic characteristics of slop educational videos on popular platforms, according to the following three research questions (RQs):

- RQ1: What are the qualitative characteristics of slop videos that might imperil learning or trust in educational systems?
- RQ2: What is the prevalence and reach of slop, according to our definition, in medical biochemistry or cell biology material on YouTube and TikTok, as discerned from viewership data?
- RQ3: Are there any quantitative metrics, such as view or like rates, that can reliably identify slop videos?

To address RQ1, we rely on a 2-stage qualitative content analysis in which we identify the educationally hazardous traits

of likely slop videos and map them to the 7 characteristics of careless speech identified by Wachter et al [38]. We approach RQ2 and RQ3 using basic data mining methods. We hypothesize that slop videos display features contravening both the structural and content-thematic elements of effective multimedia instruction and present a significant (and growing) share of the educational space on these platforms.

Methods

Study Design and Approach

A complete description of the search and screening procedure is given in Multimedia Appendix 1. The overall strategy was to search YouTube and TikTok for videos on biochemical topics that first-year medical students often find challenging, to examine these videos for signs of careless AI use, and to compile a list of problematic features of the suspected genAI videos (RQ1). Data on viewership, video age, and duration were also collected to infer the reach and popularity of these videos, compared with the entire dataset (RQ2), and to see if these correlated with slop (RQ3).

Searching and Screening of Videos

YouTube and TikTok were searched using third-party application programming interfaces (APIs; SerpAPI YouTube Search Engine and Apify TikTok Search API) over a 2-week period in late February and early March 2025. In total, 10 queries were used for each platform (Table 1), incorporating both single-word and sentence-like queries to obtain a variety of results.

Table . Summary of search results for each of the 10 queries. Unique URLs include off-topic and non-English videos that were excluded at later stages of screening.

Search query	YouTube URLs (raw), n	YouTube URLs (unique), n	TikTok URLs (raw), n	TikTok URLs (unique), n
How do enzymes work	209	114	60	—[a]
Protein secondary versus tertiary structure	210	96	60	—
Ion channel function	154	66	89	—
Cell cycle regulation	198	105	60	—
Carbohydrate metabolism	59	51	60	—
Electron transport chain	208	101	84	—
Urea cycle	68	67	42	—
Pentose phosphate pathway	195	109	60	—
Cytoskeleton	185	114	60	—
What are eicosanoids	110	85	88	—
Total	1596	908	663	617

[a]Not available, as the TikTok URLs were deduplicated as a single group.

Residential proxy servers were used, and stored browser data were cleared before each search. The search results were exported in JSON format, from which bare URLs were extracted and deduplicated, resulting in 908 and 617 unique video links for YouTube and TikTok, respectively. These videos were viewed for a few seconds each to identify off-topic and non-English videos, which were excluded. The 1082 on-topic videos remaining (814 YouTube and 268 TikTok) were then screened briefly for common "tells" of AI-generated material (refer to Multimedia Appendix 1 for rubric); 1 screener (EMJ) viewed each video for a minimum of 30 seconds (or the whole video, if less than 30 s) and flagged any showing signs suggestive of genAI. A list of these videos was then distributed among 3 reviewers, who viewed them in their entirety to verify that the "tells" were present and to provide a score (0=not AI-generated, 1=partially AI-generated, and 2=mostly AI-generated), and to note any factual errors. These videos were labeled as "likely AI-generated" (but not necessarily slop) if the average score was 1.0 or higher (only 1 video did not meet this criterion). AI detection tools were not used, due to their known unreliability [53].

To determine viewership, video metadata (including days online, duration, number of views, likes, and comments) was scraped for the entire 1082-video dataset using commercial data-scraping agents (Apify YouTube and TikTok Scrapers). Metrics were compared between the AI-generated and overall population datasets using a permutation test [54], a nonparametric test deemed appropriate because of the highly nonnormal distributions of data and the fact that the AI dataset was contained within the overall dataset.

Qualitative Analysis

A detailed description of the qualitative analysis method is given in the Multimedia Appendix 1. A 2-stage procedure was used. The first stage consisted of an inductive content analysis [55] to categorize features that we deemed educationally problematic. We define "educationally problematic" as violating one or more tenets of effective multimedia instruction according to Mayer's [51] cognitive theory of multimedia instruction, or principles of quality explanatory video design, including precise and descriptive language, clear learning objectives, and opportunities for engagement and reflection, as outlined by Brame [6], Kulgemeyer [12], and Ring and Brahm [13]. The objects of

analysis were decided to be any audiovisual feature (such as graphics, narration, linguistic features, sounds, or combinations of these) that were potentially inaccurate, misleading, distracting, irrelevant, or clearly biased. The videos deemed "likely AI-generated" (plus 6 additional videos found independently; not included in above statistics) were viewed separately, in their entirety, by 2 faculty (EMJ and JDN), one of whom was not involved in the screening steps (the additional videos were found in a YouTube video search for an unrelated project, using queries "enzyme catalysis," "metabolic pathways," "what is an enzyme mechanism," or "lipid bilayer structure"). Each viewer independently compiled a list of such features in all videos and then compared observations. Features tended to relate to either the arrangement and layout of audiovisual and linguistic elements (structural or design features) or the informational content of the videos (content features). An effort was thus made to divide all problematic features between these categories. This was deemed insufficient because certain features involved an inappropriate pairing of structural with content elements. A third category of "content – structure/language" features was thus created for audiovisual features that were "conditionally" problematic based on content, or vice-versa. These categories formed the core of the coding frame. The viewers independently assigned preliminary codes to features in each category, rewatched videos, and modified as appropriate. The viewers then met again to compare codes, re-view videos, and revise until agreement was reached on all codes.

Following the inductive content analysis, a deductive coding stage was performed [56], in which the viewers separately mapped the consensus codes onto the 7 characteristics of careless speech [38]. Reviewers independently assigned codes from the first stage to these 7 characteristics (except "lack of references to source material," as references are not typically provided in teaching videos), then met and discussed to obtain internal consistency. Some of the codes could not be mapped to the characteristics of careless speech, so 2 new characteristics of slop were proposed, and the reviewers again separately assigned codes to these characteristics, met, and revised until agreement was reached.

After completion of the qualitative analysis, an assessment of each video in the AI dataset as "slop" or "not slop" was made. We considered a video "slop" if it contained at least 2 of our codes of problematic content and exhibited at least 1 of the 7 characteristics of careless speech, except "lack of references to source material." Agreement by both reviewers was required for a "slop" assignment. All videos in the likely-AI dataset were judged to be slop by these criteria.

Ethical Considerations

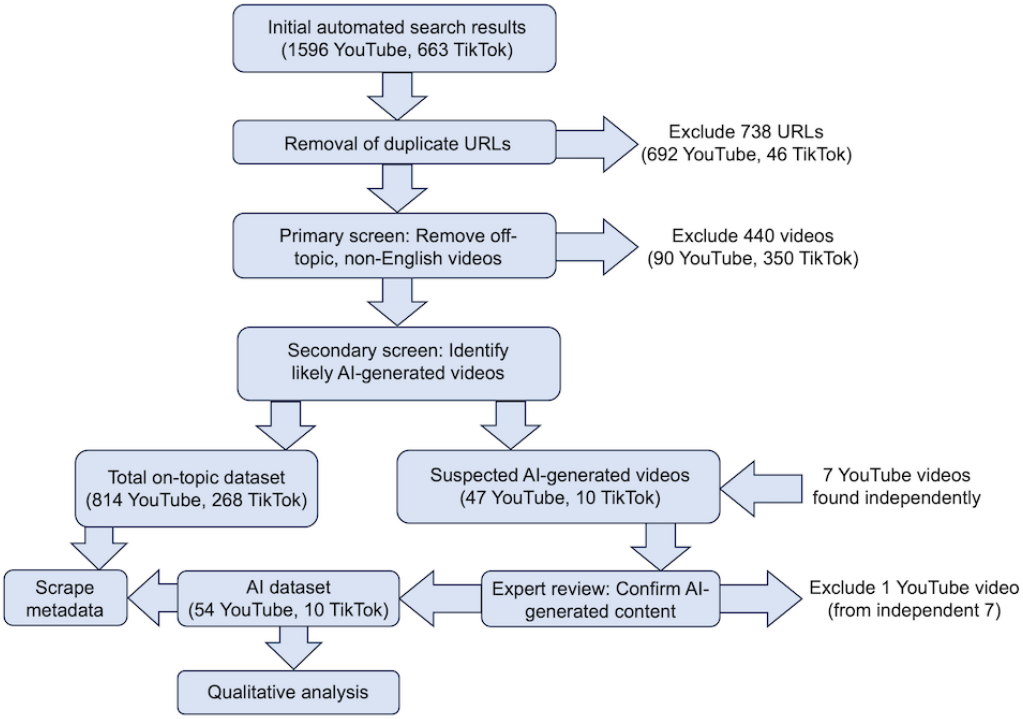
Because this study uses only publicly available data and videos shared with the public are accessible via general search, it is not a research involving human participants, and no ethical review was sought.

Results

Summary of Dataset and Prevalence of Slop

The study design is summarized in Figure 1. Summary statistics for the 814 YouTube and 268 TikTok videos examined, and a complete numbered listing of all videos, are available in the Multimedia Appendix 2. Regarding RQ2, 47 of 814 YouTube videos (5.8%) were judged to be slop according to our definition. We found that slop on YouTube was concentrated among YouTube Shorts, short-format videos that play in a loop: although only 279 of the YouTube videos examined (34.3%) were Shorts, 37 of 47 videos identified as slop (78.7%) were Shorts, with only 10 standard YouTube videos being slop. This finding is unsurprising, given YouTube's recent integration of genAI tools with Shorts [57], although most of the videos on the list predate this development. On TikTok, 10 of 268 on-topic videos (3.7%) were judged to be slop; across both platforms, the proportion was 57 of 1082 videos (5.3%) slop. We caution that these numbers likely underestimate the true prevalence of slop, as our method was designed to only identify obvious, low-quality AI-generated videos, and many better-quality videos may have been missed. Furthermore, since the platforms were searched using an automated tool, links to suggested videos, possibly containing more slop, were not retrieved.

Figure 1. Study design. The final AI dataset for qualitative analysis is a subset of the total on-topic dataset, plus 6 additional YouTube videos found after the initial search (these 6 not included in summary statistics). AI: artificial intelligence.

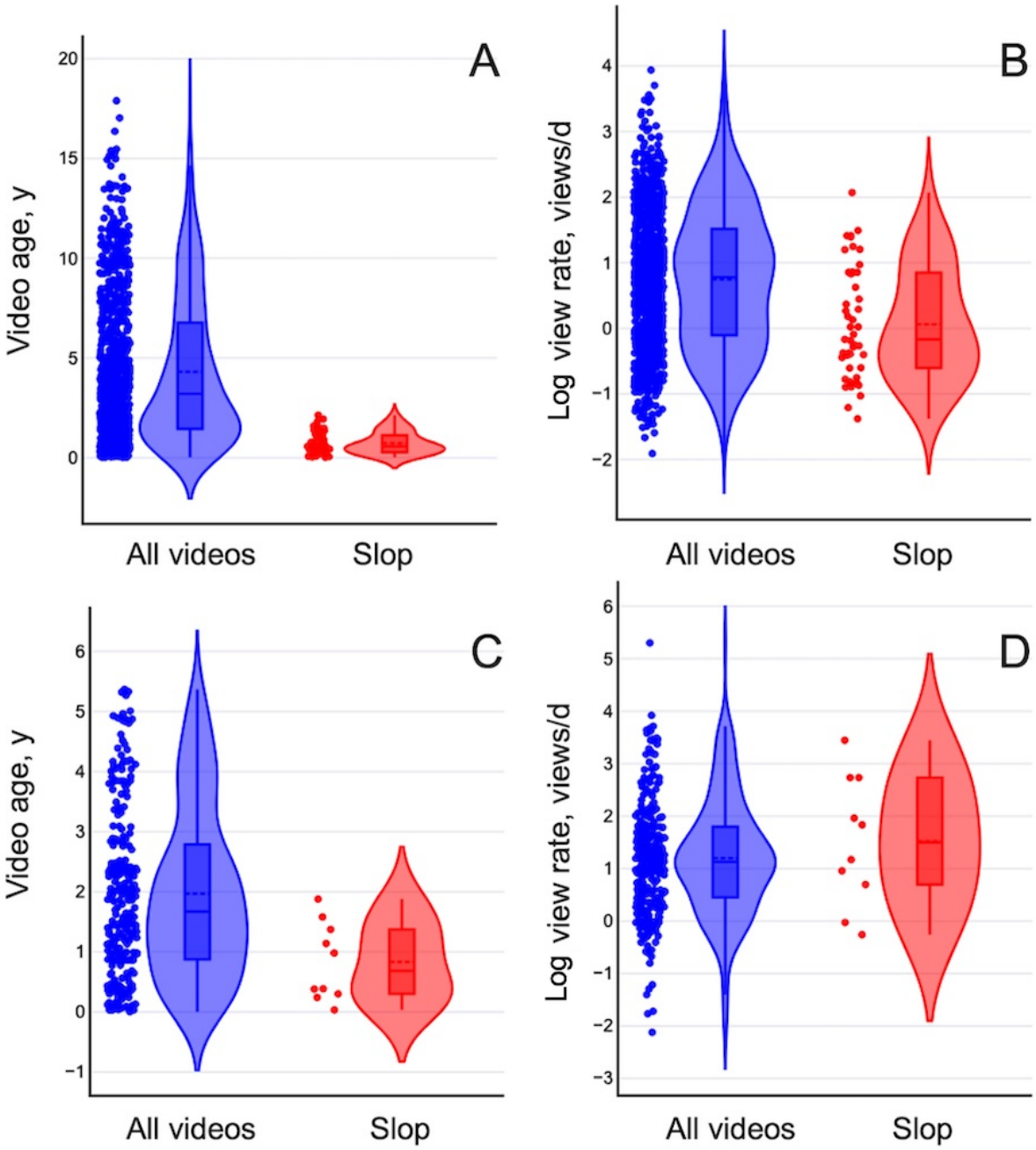


Regarding RQ3, video metadata revealed the videos varied widely in terms of age (number of days online at time of data collection), duration, number of views ("plays" on TikTok), and number of likes and comments (Tables S1 and S2 in Multimedia Appendix 1). Slop video durations on YouTube were, on average, shorter than the population at large, due to the overrepresentation of YouTube Shorts. On TikTok, the opposite was true; however, the TikTok average is skewed by 1 very long (24 min) video. View and like rates were calculated by dividing the total number of likes, views, and so on for each video by the age of the video in days to obtain average views, likes, and so on per day. This step was essential due to the widely varying ages of the videos, making raw counts of these figures misleading. The distributions of video age and view rate (log scale) are presented graphically in Figure 2; rates of likes and comments are not shown because many videos had no likes or comments and would thus not appear on a log-scale plot.

On both platforms, slop videos tended to have lower rates of engagement (views, likes, shares, and comments) than the population at large, although the difference was not statistically significant (eg, P=.87 for TikTok collect rate) according to permutation tests. The difference in engagement was more pronounced on YouTube. YouTube slop videos were engaged with about an order of magnitude less frequently, on average, than the population (Table S1 in Multimedia Appendix 1, last 3 columns), although the extremely broad and asymmetric distributions made the difference insignificant (P=.11 for view rate; data not shown). Notably, 21.3% (10/47) of the slop YouTube videos had no likes, and 78.7% (37/47) had no comments at the time of scraping.

Engagement with TikTok videos was generally higher than with YouTube videos (last 4 columns of Table S2 in Multimedia Appendix 1). Rates of views (plays), collects (which we regard as analogous to YouTube likes), and comments are all higher than the corresponding YouTube metrics, which may reflect broader differences in the manner of use of the 2 platforms, or simply a greater number of videos to choose from on the much larger YouTube. TikTok also has a "share" feature, which lacks a direct YouTube analog. All of these metrics were lower in the slop group than in the population (Figure 2D and Table S2 in Multimedia Appendix 1), although the differences were less pronounced than on YouTube (and, again, statistically insignificant; eg, P=.38 for view rate; data not shown). All numbers should be taken with caution due to the small sample size (n=10) of slop TikTok videos. Slop thus appears less popular than general materials on TikTok, although proportionately more so than on YouTube. The reason for the difference in relative visibility or popularity of slop between the 2 platforms is not clear from our data. Specifically addressing RQ3, it appears none of the metrics we collected correlates significantly with the presence of slop.

Figure 2. Distributions of the age of videos at time of collection (A and C) and log of view rate in views/day (B and D) for YouTube (A and B) and TikTok (C and D). Dashed and solid lines are mean and median, respectively. Violin plots generated using StatsKingdom Violin Plot Maker.



Qualitative Characteristics of Slop

Overview

The qualitative analysis resulted in 16 codes for problematic features. Code categories encompassed features of either the video content, the video structure or format, or features with both structural and content components (ie, structural features not suited to the content). We assigned these categories as groups A, B, and C, respectively, organized hierarchically in Table 2 (the distributions of these codes among our video dataset are given in Multimedia Appendix 2, along with a precise definition of each code, including inclusion and exclusion criteria). Following is a brief description of each code, with examples where appropriate.

Table . Codes for problematic features of AI-generated videos in our dataset. The “common variants” list is not exhaustive.

Category and codes	Common variants
Content codes	
A1. Factual inaccuracies	• Hallucinations or inventions presented as truthful
A2. Omissions of facts or context	• Missing details or definitions • Missing facts • Lack of examples • Lack of adherence to best practices of discipline
A3. Overgeneralization or oversimplification	• Superficiality • Lack of qualifications or contexts for facts • Inappropriate conflation of distinct items
A4. Inappropriate or inconsistent level of depth or inattention to audience needs	• Uncertain target audience or purpose • Missing, unmet, or unclear learning objectives • Mixing of beginning and advanced topics
A5. Sloppy analogies	• Meaningless analogies (lack of correspondence) • Misleading or misemphasized analogies (distracting correspondence) • Overextended analogies
Structure and language codes	
B1. Poor graphic or animation quality	• Poor clarity, resolution, or size • Animation artifacts
B2. Poor audio quality	• Inappropriate volume or speed • Compression or conversion artifacts
B3. Poor grammar and vocabulary	• Nongrammatical speech or text • Limited or repetitive vocabulary
B4. Speech or narration irregularities	• Inconsistent or unnatural tone, pitch, or emphasis • Unnatural pace, cadence, or stress • Mispronunciations • “Script read aloud” narration
B5. Poor editing or sequencing	• Excessive transitions • Video and audio transition asynchronously • Abrupt beginnings or ends • Inappropriate speed
Content – structure and language codes	
C1: Problematic descriptiveness	• Overuse of descriptive words and clichés • Verbose scripts • Indirect, repetitive language • Needless or inappropriate emotion • Vague, empty descriptions
C2: Mismatching audio-visual elements	• On-topic but irrelevant graphic elements • Graphics and narration covering different aspects of subject • Narrator – narration mismatch • Text unrelated to content or speech
C3: Distracting or off-topic material	• Music • Distracting visual items (eg, watermarks) • Needless text overlays
C4: Meaningless graphics	• Nonsense graphics or diagrams • Nonphysical depictions of physical objects
C5: Text irregularities	• Garbled text • Illegible text

Category and codes	Common variants
C6: Disorganization	• Illogical sequence of material • Poor flow or fluency • Lack of linkages between topics or sections

A1: Factual Inaccuracies

This code refers to direct errors of fact, that is, hallucinations of the genAI. Factual errors were fairly common in the slop dataset. Some are glaring (eg, Video 1088 claims that biochemistry "allows the sun to rise and set"), but most are subtle and plausible-sounding. For example, Video 546 (nominally about protein structure) discussed "primary," "secondary," "tertiary," and "quaternary proteins," as if these labels refer to types of protein rather than organizational levels of protein conformation. Video 1083 states that the rate of an enzyme-catalyzed reaction increases, but only up to a limiting value, as the enzyme concentration increases (this is only true if the enzyme concentration exceeds the substrate concentration, which is almost never the case; typically, the rate increases up to a limiting value as the substrate concentration increases, ie, the Michaelis-Menten model). As this latter example illustrates, errors of fact often coincided with misframing of facts (next code).

A2: Omissions of Facts or Context

This code reflects content that is narrowly or technically correct, but which does not present needed additional information, that is, misleadingly framed content. This may take the form of missing facts, details, or categorizations, a lack of examples, a lack of nuance or uncertainty, bias, or a lack of adherence to best practices in presenting the subject matter. For example, Video 15 discusses the cytoskeleton keeping cells from collapsing without mentioning that this only applies to eukaryotes, as prokaryotes use the cell wall for this purpose; Video 510 states enzyme active sites fit substrates "perfectly," which is true only for a small subset of enzymes.

A3: Overgeneralization or Oversimplification

Several videos made generalizations about phenomena with important exceptions or simplified topics to a misleading degree. A tendency to generate oversimplified summaries is a known feature of LLMs [58]. Accordingly, videos made oversimplified claims such as quaternary structure being defined as "how multiple protein molecules interact" (Video 691), or treated multiple related topics as a single subject (Video 493, which referred to "urea cycle disorder" as a single disease).

A4: Inconsistent Level of Depth or Inattention to Audience Needs

The videos in the AI-generated dataset were extremely diverse in terms of detail, professionalism, and style, and it was often not clear who the intended audience was. Some videos were simply inexplicable, for example, Video 643, a description of the electron transport chain (ETC) atop a clip from the children's television series "Barney & Friends," set to a synthesized version of "Yankee Doodle." Video 875 was apparently intended as a meme. Even the most professional videos, however, often had no apparent audience in mind and covered material to inconsistent levels of detail and depth. Some gave entry-level overviews of a topic, but in a manner that assumed knowledge of more advanced topics (eg, Video 866 ostensibly gives an introduction to the cell cycle, but mentions the functions of maturation-promoting factor and platelet-derived growth factor). Learners would likely find these videos confusing in terms of how much content they should know, or what aspects of the content were most important.

A5: Sloppy Analogies

One of the most insidious features of AI-generated videos is the frequency of almost-accurate, yet misleading analogies, which we have termed "sloppy analogies." In an effective analogy, the items being compared have similar meanings (semantic correspondence) and similar positions or relationships toward other items (structural correspondence), and ideally do not have coincidental, misleading similarities (for detailed discussion, refer to the study by Thagard [59] and references therein). Sloppy analogies violate one of these correspondences (typically the structural correspondence), or make use of distracting, irrelevant similarities between analogs, or extend the analogy to situations where it is not helpful. A learner may thus gain a misrepresentation of the subject, or an improper sense of importance of an irrelevant feature of the subject.

Some sloppy analogies were particularly terrible. For example, Video 679 compares nicotinic acetylcholine receptors to a gateway leading into a beehive, with acetylcholine molecules as "worker bees" that open the gate to allow the "queen bee" (a Na^+ cation) into the hive (the cell). While the semantic correspondence (the ion channel and the gate) is sound, the structural correspondence is nonexistent: beehives do not have gates, a queen bee does not need to be "let in" to the hive (a queen seldom leaves the hive), and worker bees, unlike acetylcholine, have numerous roles both inside and outside the hive (cell), which they can enter and leave freely. This analogy is baffling, not illuminating.

However, not all sloppy analogies are this obviously bad. Video 1085, for instance, compares metabolic regulation to the coordination of an orchestra by a conductor ("just as a conductor ensures each instrument plays in harmony, enzymes coordinate the complex symphony of biochemical reactions in our bodies"). Both structural and semantic correspondences exist between a conductor and a regulatory enzyme, but the analogy suggests enzymes somehow "choose" which pathways to accelerate or inhibit. It ignores the distributed nature of most metabolic regulation and implies a single, central locus of metabolic control. This analogy captures only one similarity between analogs, ignoring several major dissimilarities, and thus creates a misleading view of how metabolic regulation works in reality.

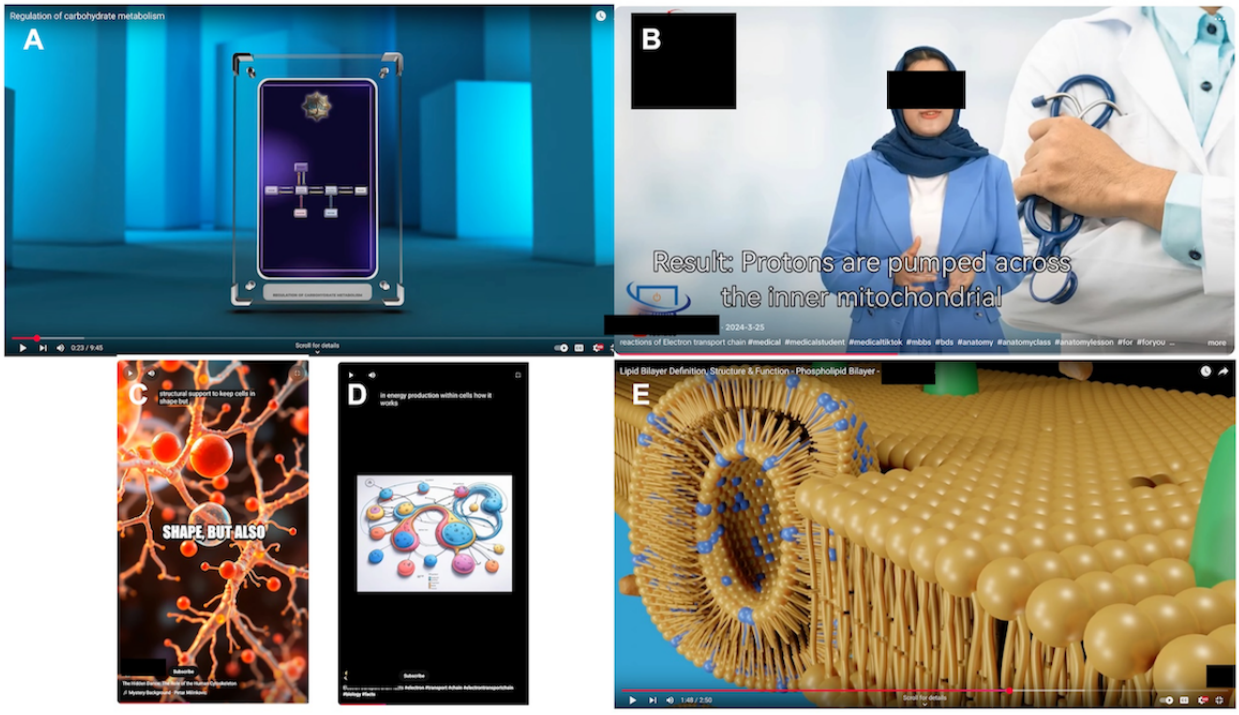
Other sloppy analogies in the dataset included the ETC being likened to a "relay race" (several videos), protein quaternary

structure as a “protein party” (Video 864), and enzymes being akin to a “set of instructions” for making molecules (Video 752).

B1: Poor Graphic or Animation Quality

Most videos in the dataset made use of cartoon graphics or animations, or joined still images with transitions. In making the videos, insufficient attention to detail or editing led to poor-quality graphics, such as low-resolution, pixelated, or blurry images, pictures too small to be clearly seen, and jerky or flickering movement (Figure 3).

Figure 3. Gallery of video stills illustrating problematic video features. (A) Video 388, diagram too small to be read, unrelated graphic background (codes B1, C2). (B) Video 1045: Monotonous narration, unrelated graphic background, narrator moves and gestures unnaturally (codes B4, C2). (C) Video 559: Distracting text overlay, nonphysical depiction of cytoskeletal fibers (codes C3, C4). (D) Video 712: Meaningless metabolic pathway diagram with garbled text (codes C4, C5). (E) Video 1086: Nonphysical depiction of physical objects (vesicle embedded in bilayer; code C4). Names and logos of content creators have been redacted.



B2: Poor Audio Quality

Similarly, inattention to editing caused many videos to have audio that was too fast or slow, varied wildly in volume, or contained sudden cuts. This code applies to general audio; spoken language is accounted for by codes B3 and B4.

B3: Poor Grammar and Vocabulary

Several videos contain grammatical errors in either spoken words or on-screen text (eg, “Do you know what is cytoskeleton?” [Video 222]) or use inappropriate vocabulary. Most of these errors are minor and unlikely to affect understanding of the subject, but are distracting to native English speakers and may be confusing to learners whose native language is not English.

B4: Speech or Narration Irregularities

Many videos with spoken narration exhibit the flaws commonly seen in text-to-voice translation: Unnatural tone or stress, mispronunciations of certain words, awkward pace or cadence, emotionless (or overly emotional) tone, and narration that sounds like text being read aloud. Abbreviations are often read out like words, for example, K_A, association constant, was pronounced “ka” in Video 864; in Video 712, ETC (electron transport chain) was read as “et cetera.” These flaws are generally not sufficient to affect understanding of the material, but they are distracting and require extraneous processing to ignore [6].

B5: Poor Editing or Sequencing

Videos in the list often exhibited excessive or poorly executed transitions between images or sections, started or stopped abruptly (often mid-sentence), shifted rapidly between different topics, or featured visual and auditory elements that did not transition together, leading to an audio-visual mismatch (also refer to code C2). In most cases, this lack of attention to fluent editing created only a distraction; poor sequencing impacting understanding of the content is captured by code C6.

C1: Problematic Descriptiveness

Several studies have found LLM-generated writing tends to overuse adjectives [60] or create prose with an effusive, grandiose style [61,62]. We observed this tendency in nearly all of the videos in our AI-generated dataset, most of which were apparently based on LLM-written scripts. Words frequently overused by AI (“amazing,” “crucial,” and “delve”) were superabundant. In addition to descriptive words, the videos in our list frequently overused certain cliches [63]: “deep dive,” “break it down,” “unsung hero” (Video 1085 used this term

four times in nine minutes), and constructions like "from … to …" (or "whether it's … or …") all appeared far more frequently than would be expected in ordinary narration. More generally, the scripts of our videos tended toward repetitive and indirect language, often incorporating needless emotion or forced casualness ("pretty cool, huh?"). In extreme cases (Video 1088), the scripts gave elaborate, emphatic declarations of a topic's importance without ever incorporating actual facts. Several videos also made analogies for simple concepts not needing an analogy, such as the cytoplasm filling the inside of a cell "like the water in a water balloon" (Video 638; also refer to code A5).

C2: Mismatch of Audio and Visual Elements

In multimedia educational materials, visual elements should be paired with relevant auditory elements so that inputs to the 2 cognitive channels can reinforce one another; off-topic and unnecessary visuals should be minimized [6]. This principle was frequently violated in the AI-generated videos in our dataset. Off-topic graphic backdrops were present in many videos (Figure 3A–B), and in others, the graphics and narration described different aspects of the subject, or text not matching or reinforcing the narration was displayed. Video 1087, for example, showed a model of DNA while discussing proteins. Some videos (eg, Videos 940 and 1087) displayed animated or avatar narrators whose hand gestures did not match the points of emphasis in the script (Figure 3B), a common artifact of photo-animation software like HeyGen. These distracting visual elements require cognitive processing to ignore, diluting the educational effectiveness of the videos. They may also, without careful structuring, suggest misleading connections between audio and visual content, resulting in a misconception of the topic being presented [64].

C3: Distracting or Off-Topic Material

Beyond mismatched visuals and sound, many videos displayed miscellaneous off-topic and distracting features, such as music, watermarks, animations, or unnecessary text overlays (which often obscured relevant imagery). Some videos included unrelated or loosely related stock footage (eg, Video 541, about fatty acids, showed supplement pills on a tray). These so-called "seductive details" contribute to cognitive load without imparting real information [65].

C4: Meaningless Graphics

Some of the visual elements in the videos were nonphysical representations of real objects, or completely meaningless diagrams (Figure 3C–E). Many of these graphics could be scientifically misleading, such as an inaccurate rendering of a protein structure (Video 691) or a phospholipid vesicle embedded in a bilayer membrane (Video 1086, Figure 3E).

C5: Text Irregularities

AI image generators struggle to create realistic text, and accordingly, several of the videos featured nonsense words resembling real words, such as "eectron." Properly rendered text was also often illegible, either due to poor resolution, inadequate size, or cropping.

C6: Disorganization

Videos in the dataset frequently suffered from general disorganization. Topics were presented in a nonintuitive sequence, concepts did not flow naturally from one to the next, and linkages between subjects were often not explained or insinuated. Disorganization occurred at the level of structure and editing (eg, a rapid series of images being flashed in the background while a single topic was explained, as in Video 638) or content (eg, Video 1083 described the effects of enzyme inhibitors on an enzyme's kinetic parameters before discussing enzyme kinetics). Disorganization contributes to cognitive load by requiring the learner to "hold" relevant information in working memory while waiting for complementary information [65].

Many of the 16 codes of problematic slop content could be directly matched with the seven characteristics of careless speech: (1) factual inaccuracies or inventions (hallucinations or obsolete ideas); (2) nonrepresentativeness of sources (bias; statements not proportionally representing the totality of views); (3) incompleteness (statements that are narrowly correct but omit needed context); (4) lacking signifiers of uncertainty (unwarranted confidence or failure to account for variability in responses); (5) lacking references to source material (failure to cite relevant sources, where appropriate); (6) references not based on referred text (hallucinated or off-topic references); and (7) inaccurate summaries of referenced text (incorrect or incomplete summary of a real reference) [38]. These mappings of codes are presented in Table 3.

Table . Alignment of qualitative codes from Table 2 with characteristics of careless speech. Two additional code groupings, which we designate "communicative nonfluency" and "message – delivery incoherence," were identified in addition to the 7 features of careless speech published previously.

Careless speech characteristic	Codes from Table 2
Factual inaccuracies	A1, A2, C4, C5
Nonrepresentativeness of sources	A2, A3
Incompleteness	A2, A3, A4, A5
Lacking signifiers of uncertainty	A3, A5
Lacking references to sources	A2
References not based on referred text	A1, A2
Inaccurate summaries of referred text	A1, A2, A3, C1, C4
Additional characteristics of slop	
Communicative nonfluency	B1, B2, B3, B4, B5, C6
Message – delivery incoherence	A4, C2, C3, C6

Because the features of careless speech describe content, they align most closely with groups A and C of our codes for characteristics of slop. For instance, A1 is virtually identical to "factual inaccuracies or inventions." We note that the careless speech codes involving references are less relevant to the case of educational videos or lessons, which often do not cite references (references are presumed to be the latest discipline-standard textbooks or review articles). Only 2 of the videos in our slop dataset (681 and 697) cited a reference.

The group B codes, and codes C2, C3, and C6, did not specifically align with any of the characteristics of careless speech, as these codes are primarily concerned with the form and structure of the speech. We consider these codes to embody 2 additional common characteristics of slop (if not of careless speech more generally): "communicative nonfluency" (C6 and all group B) and "incoherence of message and delivery" (A4, C2, C3, and C6). Communicative nonfluency often makes low-end genAI materials recognizable (telltale linguistic features, robotic narration, unnatural animations, and so forth), while the incoherence of message and delivery (features such as excessive or unnecessary transitions, mismatching visual and auditory output, confusing or inappropriate diagrams, and illogical sequencing) limits its usefulness as a teaching tool, even when factually accurate. Since many videos incorporated content or styles that were incompatible with a particular audience or set of learning goals, we included code A4 in this group as well.

Discussion

Prevalence and Reach of Slop

Our results show that slop accounts for a small but nonnegligible portion of medical biochemistry and cell biology videos, seems to be comparably popular to nonslop videos, and cannot be reliably distinguished from nonslop on the basis of that quantitative features. These findings accord with previous studies of educational YouTube videos, which found quantitative metrics do not correlate with video quality [11,66]. Some studies have suggested that the number or text of comments may correlate with quality [11], but we did not observe any strong correlation of comment rates with slop (Tables S1 and S2 in Multimedia Appendix 1). The text content of comments was not examined in this study. We may thus conclude that slop cannot easily be identified without viewing the video. We also emphasize that our method is only able to detect obvious and low-quality genAI output, so the true prevalence of slop is certainly higher than our numbers suggest, and as the apparent quality of genAI content improves, even viewing a video may soon be insufficient to identify it as slop.

In the course of this research, we observed that many of the slop videos were posted by channels that consisted mostly or entirely of slop material, suggesting that characteristics of the channel or creator, and not the individual video, may provide evidence that a video is slop or otherwise questionable. Studies of channel characteristics, rather than video characteristics, should thus be a productive line of future slop research.

Problematic Features of Slop

At present, research on the educational effects of genAI video is scant. At least 3 recent small-scale studies have examined the effectiveness of AI-constructed video on learning, and all found little difference in learning outcomes between genAI and traditional materials [67-69]. Critically, however, all of these studies used extensively edited and fact-checked videos, designed by disciplinary experts. In other words, the videos in these studies were not slop. A full understanding of the hazards of slop must be drawn not from well-designed genAI materials but from slop typical of online video platforms.

When addressing RQ1, we identified 16 problematic features of slop videos that may impact educational value (Table 2). These features encompass both subject matter and structure-based aspects, and thus imperfectly align with the features of careless speech (Table 3), which is a subject matter–based construct. We feel it is relevant to include structure-based features in consideration of slop, since proper formatting and editing of multimedia educational materials contribute to educational effectiveness [6,12,65,66]. Thus, some consideration of video format and structure is appropriate in assessing the impact of slop.

While all 16 features can dilute the educational effectiveness of videos, we think 2 deserve additional discussion. The first

of these is sloppy analogies (code A5). The educational perils of imperfect analogies and metaphors have been described for disciplines including the biological sciences [70,71], chemistry [72,73], and physics [74,75]; genAI does not add any new hazards. Rather, genAI removes the effort barrier to creating a weak analogy, and along with it, the mental check of whether the analogy makes sense (unless further prompting or editing, ie, care, is performed by the content creator). What is particularly damaging about sloppy analogies is the illusion of understanding [76] generated by an intuitive but inaccurate or unnuanced analogy. Previous studies have shown that analogies can have negative effects on metacomprehension if not reinforced by experiential inputs, such as experimentation [77]. Video, being an inherently passive medium, is thus especially inclined to mislead by analogies presented without real-world reinforcement, and experimental data confirm that video instruction is prone to an illusion of understanding effect when misconceptions are present [76,78]. While this is equally true of all videos (not just slop), the ease with which genAI conjures plausible-sounding analogies makes slop videos especially likely to contain poor analogies and metaphors, as seen in our dataset. Incidentally, at least 1 popular book on genAI in teaching [79] specifically recommends asking an LLM to create analogies for unfamiliar topics. Based on our results, we strongly feel unsupervised beginning learners should not follow this suggestion (to be fair, this book emphasizes the importance of careful prompting when asking an LLM for an analogy, but since beginners are typically not capable of evaluating the analogy, there would be no way for a beginner to know if a prompt was effective). However, experienced instructors might find this suggestion useful as part of a carefully curated activity, for example, asking students to identify problems with the analogy.

The second educational hazard worth further discussion is problematic descriptiveness, code C1. Numerous studies have commented on the tendency of LLMs to overdescribe [60,62] as in our slop videos. For example, the passage “a toolbox of folds and domains that evolution has mixed and matched over billions of years to create the incredible diversity of proteins we see today” (Video 864) has to be read or heard several times to extract the main point: that proteins contain modular folds and domains that perform discrete functions. Apart from being distracting (and annoying), this overly descriptive style creates a problem of misplaced emphasis in many of the videos. Unable to know which terms or ideas are really important and which are incidental, the LLM attaches descriptive words or phrases to everything it can. A human teacher, however, would preferentially reserve description for the most salient words and topics, thus cueing the learner toward the important material. Goodwin [80] described this practice as “highlighting,” and identified it as one component of the so-called professional vision that frames expertise in any discipline. An LLM lacks the professional vision of an educator (or any other profession) and thus cannot model a professional’s practice of sense-making, except insofar as a human professional shapes the genAI output. Instead, all aspects are treated as potentially equal by the LLM, resulting in an unfocused, directionless treatment of the subject.

Other features of slop identified in our qualitative analysis generally align with the 7 features of careless speech [38] or violate good practices of multimedia teaching [6,51]. We identify 2 clusters of features that do not neatly map onto the careless speech framework, “communicative nonfluency” and “message – delivery incoherence” (Table 3). Communicative nonfluency may be loosely defined as the property of requiring undue cognitive effort to understand; it aligns closely with so-called perceptual fluency, or the ease of making sense of inputs based on sensory features [81]. Perceptual fluency is strongly associated with metacognition, specifically judgment-of-learning [82] and perceived accuracy or truth [83,84]. Accordingly, students have perceived fluent delivery (in video or live lecture) as more instructive even though fluency did not affect learning gains [85,86]. These findings suggest students prefer fluent over nonfluent learning materials, and thus would be less likely to perceive slop videos as reliable, even if the slop video lacked any factual errors. However, as the realism and fluency of genAI output increase, this effect would be expected to diminish as technology improves, and thus communicative nonfluency may not be a characteristic marker of slop in the future.

Message – delivery incoherence refers to a mismatch between the concept being communicated (the message) and the object or language ostensibly used to communicate it (the delivery). In our video dataset, this most frequently took the form of mismatching audio and visual elements (code C2) or superfluous content (code C3). Either of these will increase the amount of cognitive processing needed to encode the underlying message, and thus hinder learning [65]. For instance, extraneous visual content (such as watermarks and text overlays duplicating the narration) conveys no relevant information and competes for working memory with the educational content [6]. Likewise, mismatched content between channels (such as a description of cell cycle regulation over a schematic of ligand-receptor binding; Video 1051) requires processing to identify which channel contains the relevant information, so-called extraneous overload [65], or may create a misimpression that the two channels are, in fact, related. It is thus considered best practice in video instruction to remove superfluous material (“weeding”) and to ensure information in audio and visual channels complement each other [6,65]. Message – delivery incoherence also sometimes took the form of presentations that were inappropriate for the apparent learning goal of the video, or were so muddled that the learning goal was indiscernible (codes A4 and C6). Relevance to learners and links to previous knowledge are considered important elements of effective instructional video design [13], and were conspicuously weak in most of the slop videos. These deficiencies could impact engagement with the videos [6], even if factual accuracy were not a concern.

Of course, none of our problematic features of slop is entirely unique to AI-generated material, and most are not unique to video. Classroom lectures may use flat, repetitive language or incorporate unnecessary content; human teachers may make bad analogies or fail to highlight key points. Living instructors, however, generally bear some risk of consequence for ineffective teaching (care in the second sense of the term), such as poor evaluations, career stagnation, or a personal sense of failure.

GenAI is totally unencumbered by such consequences, and the human who uses genAI to make educational slop for free public platforms (particularly when posting anonymously) is, to some extent, insulated from these consequences—although not from gain in the form of monetization or self-satisfaction. Additionally, living instructors often have the chance to immediately correct student misconceptions that may result from a bad analogy or other errors through just-in-time teaching techniques; for asynchronous video, this is not a possibility, requiring videos to be high-quality from the beginning. Slop is likely to remain a significant problem on video-sharing platforms as long as there is an asymmetry between the risks and rewards of sharing it. Students lacking the expertise to evaluate content on unfamiliar subjects will be especially vulnerable.

Implications for Content Creators and Learners

Fortunately, our inventory of slop characteristics provides a ready checklist of pitfalls that caring creators of genAI material can work to avoid. The central pillar of making good genAI material is maintaining a human-in-the-loop [44] at every point in the creation and dissemination process. Prompting (and reprompting or iterative prompting) needs to be done in a planned and deliberate fashion; tools such as prompt-design frameworks [87] are helpful at this stage. The output must then be evaluated not only for accuracy and appropriate context, but also for alignment with intended learning goals and audience characteristics, which should be explicitly stated. GenAI descriptions, explanations, and metaphors should never be accepted at face value, but rather examined closely (and amended, if necessary) to avoid misleading, oversimplified, or misemphasized statements. If possible, the output should align with the viewer's experience of reality and use meaningful mental models (eg, the human experience, expertise, accuracy, trust, or HEAT heuristic [44]) and be coherent with respect to subject matter and presentation. The output should give a balanced overview of the field without unduly favoring a particular viewpoint [41]. Attention should be paid to the fluency of the output (video quality and realism, natural-sounding speech and audio, and appropriate transitions). Once posted to an online forum, material should be monitored for signs of misinterpretation (eg, comments reflecting confusion or dislike) and edited or removed as necessary. Finally, clear disclosure of the use of genAI will help establish trust with the audience. While this will not impact the quality or effectiveness of the material, these efforts toward building trust may be seen as a sign of care. Slop-free genAI material requires more than just good prompting; human involvement throughout the lifecycle of the material is essential to make sure good AI creations continue to fulfill their intended functions. We are currently in the process of developing lifecycle guidelines for genAI educational materials.

For learners, the presence of slop on video-sharing sites presents a challenge. The videos encountered in the course of this study are at the lowest end of quality and usefulness; future slop may be far more realistic and seemingly helpful than today's slop, but as long as genAI works in an associative fashion, it will always be careless speech (and thus unreliable). Learners should thus focus not on detecting and excluding suspected genAI material, but on cross-checking claims made in online videos with other sources, verifying the credentials of the creators, and discussing points of concern with subject matter experts. Sadly, vetting and comparing sources of information is a skill that takes time and effort to develop. Furthermore, genAI is often presented to learners as a shortcut to learning (eg, as a way to quickly digest and summarize dense information), or as a source of information [39] while its ability to introduce misconceptions and misframings of reality [28,38] is far less recognized. We therefore recommend that learners approach known genAI output with great caution, always validate sources, and consult with experts when possible, and generally exercise judgment when using video-sharing platforms for educational purposes.

Limitations

Our definition of slop is intended only for educational materials and may not be appropriate in other contexts, such as art, advertising, or political speech. Since we restricted our attention to videos, not all characteristics of slop may be relevant to other output types (such as still images or text). Our numerical data should be considered semiquantitative at best, since we believe our methods underestimate the prevalence of slop for reasons given above. Our estimates of slop's prevalence for RQ2 may not apply to other major platforms where slop is endemic (eg, Facebook [Meta] or Instagram [Meta]) due to the different user bases and content moderation practices on these platforms. We should also emphasize that our aim in RQ1 was only to identify problematic features of genAI material—any potential educational benefits of AI-generated learning materials were not considered for the purposes of this study.

A well-known problem with genAI output is bias, which in this study was encompassed by code A2, incompleteness. We did not examine bias in our dataset in any further detail (eg, biased overviews of a discipline and bias in examples), but this aspect of slop clearly needs deeper attention in future work. We also note, as mentioned previously, that our own dataset is biased toward obviously bad genAI content, due to our screening method, so some of our findings may not be fully generalizable to higher-quality slop.

Our methodology made no effort to identify the purposes for which the slop videos were made. It is likely that most of the AI videos in our dataset were made not to educate, but to accumulate views and other markers of engagement. While irrelevant to the educational impacts of slop, further study of the motivating factors behind slop's rapid proliferation on the internet may help to curb its influence.

Perhaps most significantly, this study does not address the enormous ethical concerns raised by slop, such as unmitigated biases, the unpaid and uncredited use of intellectual property for training, and the environmental impacts of needless AI use [41,50]. Slop disregards ethical considerations, as care forms a basis for ethics [46], and a definition of slop that is based in an ethics framework (rather than on features of its content, as in this study) would be needed for a proper accounting of the ethical dimension of slop. Regardless, ethical matters feature prominently in UNESCO's (United Nations Educational, Scientific and Cultural Organization) recent statement of guidelines for responsible AI use in education [88], and while

ethics is only one of many risks of improper educational use of genAI [50], it deserves attention in future research on slop.

Conclusions

We have suggested a definition of slop that, to our knowledge, is the first in the scholarly literature. We define slop as AI-generated material that is produced with little or no care toward its educational usefulness or quality, situated in the broader conceptualization of careless speech [38]. Using this definition, we find that slop composes a small percentage of preclinical biochemistry and cell biology videos on YouTube and TikTok, but these videos could be found without considerable effort and present specific educational risks to learners. Among these risks are misleading content (eg, inaccuracies and improper comparisons), nonfluency endangering effective cognitive processing, and incoherent presentation (eg, problematic descriptiveness and disorganization). Our findings can hopefully inform better practices among responsible creators of genAI material for education. We also hope our findings will be informative to other disciplines, such as journalism, in which vetting of information sources is critical and under threat by genAI materials.

Some commentators [16,24] have suggested that slop is a temporary problem that will be solved by technological means, much as spam email has been curtailed by email filtering. Even if this proves to be true, existing slop videos will persist on the internet for some time, potentially training future genAI applications, and thus propagating subtle misconceptions or misemphases throughout future, higher-quality genAI output. As genAI becomes more central to education and training, these propagated errors may be used to train human beings, leading to a generational erosion of understanding and expertise [38]. Slop is therefore a problem that needs to be recognized and fought right now, and we hope this study provides a useful starting point for this fight.

Acknowledgments

We thank Prof Deidre Hurse for assistance with preparing the Qualitative Analysis section of the Methods and for identifying helpful sources; Profs Dwayne Baxa and Serena Kuang for helpful discussions on organization of the manuscript; and Profs Tammy Campbell and Anya Goodman for assistance with video screening and suggestions on study design (respectively). The work in this paper was self-funded by the lead author.

Disclaimer

No generative artificial intelligence tools were used at any point during the collection and analysis of data for this project, or in the preparation of the manuscript.

Data Availability

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

Authors' Contributions

Conceptualization: EMJ
Data curation: EMJ
Formal analysis: BK
Investigation: EMJ, JDN, EJF
Methodology: EMJ, JDN, EJF
Project administration: EMJ
Resources: EMJ
Supervision: EMJ
Validation: JDN, BK
Visualization: EMJ
Writing – original draft: EMJ, JDN
Writing – review and editing: EMJ, JDN, BK, EJF

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed methods and descriptions of qualitative codes.
[DOCX File, 34 KB - mededu_v11i1e80084_app1.docx]

Multimedia Appendix 2
Complete listing of videos in datasets with URLs and content codes.

[XLSX File, 169 KB - mededu_v11i1e80084_app2.xlsx]

References

1. Curran V, Simmons K, Matthews L, et al. YouTube as an educational resource in medical education: a scoping review. Med Sci Educ 2020 Dec;30(4):1775-1782. [doi: 10.1007/s40670-020-01016-w] [Medline: 34457845]
2. Conde-Caballero D, Castillo-Sarmiento CA, Ballesteros-Yánez I, Rivero-Jiménez B, Mariano-Juárez L. Microlearning through TikTok in higher education. An evaluation of uses and potentials. Educ Inf Technol (Dordr) 2023 Jun 2;29(2):1-21. [doi: 10.1007/s10639-023-11904-4] [Medline: 37361804]
3. Greenhow C, Lewin C. Social media and education: reconceptualizing the boundaries of formal and informal learning. Learn Media Technol 2016 Jan 2;41(1):6-30. [doi: 10.1080/17439884.2015.1064954]
4. Noetel M, Griffith S, Delaney O, et al. Multimedia design for learning: an overview of reviews with meta-meta-analysis. Rev Educ Res 2022 Jun;92(3):413-454. [doi: 10.3102/00346543211052329]
5. Noetel M, Griffith S, Delaney O, et al. Video improves learning in higher education: a systematic review. Rev Educ Res 2021 Apr;91(2):204-236. [doi: 10.3102/0034654321990713]
6. Brame CJ. Effective educational videos: principles and guidelines for maximizing student learning from video content. CBE Life Sci Educ 2016;15(4):1-6. [doi: 10.1187/cbe.16-03-0125] [Medline: 27789532]
7. Singh S. Why am I seeing this? How video and e-commerce platforms use recommendation systems to shape user experiences. New America. 2020. URL: https://www.newamerica.org/oti/reports/why-am-i-seeing-this/ [accessed 2025-06-23]
8. Vidal Bustamante CM, Candela JQ, Wright L, et al. Technology primer: social media recommendation algorithms. : Belfer Center for Science and International Affairs, Harvard Kennedy School; 2022 URL: https://www.belfercenter.org/publication/technology-primer-social-media-recommendation-algorithms [accessed 2025-11-11]
9. Metzger MJ. Making sense of credibility on the web: models for evaluating online information and recommendations for future research. J Am Soc Inf Sci 2007 Nov;58(13):2078-2091. [doi: 10.1002/asi.20672]
10. Bitzenbauer P, Teußner T, Veith JM, Kulgemeyer C. (How) do pre-service teachers use YouTube features in the selection of instructional videos for physics teaching? Res Sci Educ 2024 Jun;54(3):413-438. [doi: 10.1007/s11165-023-10148-z]
11. Bitzenbauer P, Höfler S, Veith JM, Winkler B, Zenger T, Kulgemeyer C. Exploring the relationship between surface features and explaining quality of YouTube explanatory videos. Int J of Sci and Math Educ 2024 Jan;22(1):25-48. [doi: 10.1007/s10763-022-10351-w]
12. Kulgemeyer C. A framework of effective science explanation videos informed by criteria for instructional explanations. Res Sci Educ 2020 Dec;50(6):2441-2462. [doi: 10.1007/s11165-018-9787-7]
13. Ring M, Brahm T. A rating framework for the quality of video explanations. Tech Know Learn 2024 Dec;29(4):2117-2151. [doi: 10.1007/s10758-022-09635-5]
14. Gyamfi G, Hanna B, Khosravi H. Supporting peer evaluation of student-generated content: a study of three approaches. Assessment & Evaluation in Higher Education 2022 Oct 3;47(7):1129-1147. [doi: 10.1080/02602938.2021.2006140]
15. Chiu TKF. The impact of generative AI (GenAI) on practices, policies and research direction in education: a case of ChatGPT and Midjourney. Interactive Learning Environments 2024 Nov 25;32(10):6187-6203. [doi: 10.1080/10494820.2023.2253861]
16. Knibbs K. AI slop is flooding medium. WIRED. 2024. URL: https://www.wired.com/story/ai-generated-medium-posts-content-moderation/ [accessed 2025-07-02]
17. DiResta R, Goldstein JA. How spammers and scammers leverage AI-generated images on Facebook for audience growth. HKS Misinfo Review 2024 Aug 15;5(4). [doi: 10.37016/mr-2020-151]
18. Knibbs K. Yes, that viral LinkedIn post you read was probably AI-generated. WIRED. 2024. URL: https://www.wired.com/story/linkedin-ai-generated-influencers/ [accessed 2025-07-02]
19. Read M. Drowning in slop: a thriving underground economy is clogging the internet with AI garbage--and it's only going to get worse. New York Magazine. 2024. URL: https://nymag.com/intelligencer/article/ai-generated-content-internet-online-slop-spam.html [accessed 2025-06-23]
20. Strzelecki A. 'As of my last knowledge update': how is content generated by ChatGPT infiltrating scientific papers published in premier journals? Learn Publ 2025 Jan;38(1):e1650 [FREE Full text] [doi: 10.1002/leap.1650]
21. Lei F, Du L, Dong M, Liu X. Global retractions due to randomly generated content: Characterization and trends. Scientometrics 2024 Dec;129(12):7943-7958. [doi: 10.1007/s11192-024-05172-3]
22. Jacob M. Experts warn "AI-written" paper is latest spin on climate change denial. Tech Xplore. 2025. URL: https://techxplore.com/news/2025-04-experts-ai-written-paper-latest.html [accessed 2025-06-25]
23. Williams R. The biggest AI flops of 2024. MIT Technology Review. 2024. URL: https://www.technologyreview.com/2024/12/31/1109612/biggest-worst-ai-artificial-intelligence-flops-fails-2024/ [accessed 2025-06-25]
24. Adami M. AI-generated slop is quietly conquering the internet. is it a threat to journalism or a problem that will fix itself? Reuters Institute. 2024. URL: https://reutersinstitute.politics.ox.ac.uk/news/ai-generated-slop-quietly-conquering-internet-it-threat-journalism-or-problem-will-fix-itself [accessed 2025-06-25]
25. Fan Y, Tang L, Le H, et al. Beware of metacognitive laziness: effects of generative artificial intelligence on learning motivation, processes, and performance. Brit J Educational Tech 2025 Mar;56(2):489-530. [doi: 10.1111/bjet.13544]

26. Sun Y, Sheng D, Zhou Z, Wu Y. AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content. Humanit Soc Sci Commun 2024;11(1):1278. [doi: 10.1057/s41599-024-03811-x]
27. Bastani H, Bastani O, Sungu A, Ge H, Kabakcı Ö, Mariman R. Generative AI can harm learning. SSRN. Preprint posted online on Jul 18, 2024. [doi: 10.2139/ssrn.4895486]
28. Kidd C, Birhane A. How AI can distort human beliefs. Science 2023 Jun 23;380(6651):1222-1223. [doi: 10.1126/science.adi0248] [Medline: 37347992]
29. Shikino K, Shimizu T, Otsuka Y, et al. Evaluation of ChatGPT-generated differential diagnosis for common diseases with atypical presentation: descriptive research. JMIR Med Educ 2024 Jun 21;10:e58758. [doi: 10.2196/58758] [Medline: 38915174]
30. Tabuchi H, Nakajima I, Day M, et al. Comparative educational effectiveness of AI generated images and traditional lectures for diagnosing chalazion and sebaceous carcinoma. Sci Rep 2024 Nov 25;14(1):29200. [doi: 10.1038/s41598-024-80732-4] [Medline: 39587233]
31. Buzzaccarini G, Degliuomini RS, Borin M, et al. The promise and pitfalls of AI-generated anatomical images: evaluating midjourney for aesthetic surgery applications. Aesthetic Plast Surg 2024 May;48(9):1874-1883. [doi: 10.1007/s00266-023-03826-w] [Medline: 38238569]
32. Temsah MH, Alhuzaimi AN, Almansour M, et al. Art or artifact: evaluating the accuracy, appeal, and educational value of AI-generated imagery in DALL·E 3 for illustrating congenital heart diseases. J Med Syst 2024 May 23;48(1):54. [doi: 10.1007/s10916-024-02072-0] [Medline: 38780839]
33. Kaufenberg-Lashua MM, West JK, Kelly JJ, Stepanova VA. What does AI think a chemist looks like? An analysis of diversity in generative AI. J Chem Educ 2024 Nov 12;101(11):4704-4713. [doi: 10.1021/acs.jchemed.4c00249]
34. Blonder R, Feldman-Maggor Y. AI for chemistry teaching: responsible AI and ethical considerations. Chemistry Teacher International 2024 Dec 27;6(4):385-395. [doi: 10.1515/cti-2024-0014]
35. Elmas R, Adiguzel-Ulutas M, Yılmaz M. Examining ChatGPT's validity as a source for scientific inquiry and its misconceptions regarding cell energy metabolism. Educ Inf Technol 2024 Dec;29(18):25427-25456. [doi: 10.1007/s10639-024-12749-1]
36. Campos DS, Oliveira RD, Vieira LDO, et al. Revisiting the debate: documenting biodiversity in the age of digital and artificially generated images. Web Ecol 2023;23(2):135-144. [doi: 10.5194/we-23-135-2023]
37. Koebler J. Inside the world of TikTok spammers and the AI tools that enable them. 404 Media. URL: https://www.404media.co/inside-the-world-of-tiktok-spammers-and-the-ai-tools-that-enable-them/ [accessed 2025-07-02]
38. Wachter S, Mittelstadt B, Russell C. Do large language models have a legal duty to tell the truth? R Soc Open Sci 2024 Aug;11(8):240197. [doi: 10.1098/rsos.240197] [Medline: 39113763]
39. Sundar SS, Liao M. Calling BS on ChatGPT: reflections on AI as a communication source. Journal Commun Monogr 2023 Jun;25(2):165-180. [doi: 10.1177/15226379231167135]
40. Sharma SS, Tong M, Korbak T, et al. Towards understanding sycophancy in language models. Presented at: 12th International Conference on Learning Representations; May 7-11, 2024 URL: https://openreview.net/forum?id=tvhaxkMKAn [accessed 2025-07-02]
41. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? Presented at: FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency; Mar 3-10, 2021. [doi: 10.1145/3442188.3445922]
42. Weidinger L, Uesato J, Rauh M, et al. Taxonomy of risks posed by language models. Presented at: FAccT '22; Jun 21-24, 2022. [doi: 10.1145/3531146.3533088]
43. Brundage M, Avin S, Wang J, et al. Toward trustworthy AI development: mechanisms for supporting verifiable claims. arXiv. Preprint posted online on Apr 20, 2020. [doi: 10.48550/arXiv.2004.07213]
44. Verhulsdonck G, Weible J, Stambler DM, Howard T, Tham J. Incorporating human judgment in AI-assisted content development: the HEAT heuristic. tech comm 2024 Aug 1;71(3):60-72. [doi: 10.55177/tc286621]
45. Mittelstadt B, Wachter S, Russell C. To protect science, we must use LLMs as zero-shot translators. Nat Hum Behav 2023 Nov;7(11):1830-1832. [doi: 10.1038/s41562-023-01744-0] [Medline: 37985912]
46. Tronto JC. An ethic of care. Generations 1998;22(3):15-20. [Medline: 12785337]
47. Gisselbaek M, Minsart L, Köselerli E, et al. Beyond the stereotypes: artificial Intelligence image generation and diversity in anesthesiology. Front Artif Intell 2024;7:1462819. [doi: 10.3389/frai.2024.1462819] [Medline: 39444664]
48. Franco D'Souza R, Mathew M, Mishra V, Surapaneni KM. Twelve tips for addressing ethical concerns in the implementation of artificial intelligence in medical education. Med Educ Online 2024 Dec 31;29(1):2330250. [doi: 10.1080/10872981.2024.2330250] [Medline: 38566608]
49. Aksoy DA, KurSun E. Behind the scenes: a critical perspective on GenAI and open educational practices. Open Praxis 2024 Aug 29;16(3):457-470. [doi: 10.55982/openpraxis.16.3.674]
50. Al-Zahrani AM. Unveiling the shadows: beyond the hype of AI in education. Heliyon 2024 May 15;10(9):e30696. [doi: 10.1016/j.heliyon.2024.e30696] [Medline: 38737255]
51. Mayer RE. Cognitive theory of multimedia learning. In: The Cambridge Handbook of Multimedia Learning, 2nd edition: Cambridge University Press; 2014.

52. Brown P. Education, opportunity and the future of work in the fourth industrial revolution. Br J Sociol Educ 2024 May 18;45(4):475-493. [doi: 10.1080/01425692.2023.2299970]
53. Dugan L, Hwang A, Trhlík F, et al. RAID: a shared benchmark for robust evaluation of machine-generated text detectors. Presented at: 62nd Annual Meeting of the Association for Computational Linguistics; Aug 11-16, 2024. [doi: 10.18653/v1/2024.acl-long.674]
54. Pitman EJG. Significance tests which may be applied to samples from any populations. Journal of the Royal Statistical Society Series B 1937 Jan 1;4(1):119-130. [doi: 10.2307/2984124]
55. Vears DF, Gillam L. Inductive content analysis: a guide for beginning qualitative researchers. FoHPE 2022;23(1):111-127. [doi: 10.11157/fohpe.v23i1.544]
56. Hamad EO, Savundranayagam MY, Holmes JD, Kinsella EA, Johnson AM. Toward a mixed-methods research approach to content analysis in the digital age: the combined content-analysis model and its applications to health care Twitter feeds. J Med Internet Res 2016 Mar 8;18(3):e60. [doi: 10.2196/jmir.5391] [Medline: 26957477]
57. Silberling A. YouTube Shorts adds Veo 2 so creators can make GenAI videos. TechCrunch. 2025. URL: https://techcrunch.com/2025/02/13/youtube-shorts-adds-veo-2-so-creators-can-make-gen-ai-videos/ [accessed 2025-07-01]
58. Peters U, Chin-Yee B. Generalization bias in large language model summarization of scientific research. R Soc Open Sci 2025 Apr;12(4):241776. [doi: 10.1098/rsos.241776] [Medline: 40309181]
59. Thagard P. Analogy, explanation, and education. J Res Sci Teach 1992 Aug;29(6):537-544. [doi: 10.1002/tea.3660290603]
60. Markowitz DM, Hancock JT, Bailenson JN. Linguistic markers of inherently false AI communication and intentionally false human communication: evidence from hotel reviews. J Lang Soc Psychol 2024 Jan;43(1):63-82. [doi: 10.1177/0261927X231200201]
61. Kobak D, González-Márquez R, Horvát E, Lause J. Delving into LLM-assisted writing in biomedical publications through excess vocabulary. Sci Adv 2025 Jul 4;11(27):eadt3813. [doi: 10.1126/sciadv.adt3813] [Medline: 40601754]
62. Ghiurău D, Popescu DE. Distinguishing reality from AI: approaches for detecting synthetic content. Computers 2024;14(1):1. [doi: 10.3390/computers14010001]
63. Tiffany K. Welcome to the golden age of clichés. The Atlantic. 2023. URL: https://www.theatlantic.com/technology/archive/2023/02/ai-chatbots-cliche-writing/673143/ [accessed 2025-07-02]
64. Karlsson G. Animation and grammar in science education: learners' construal of animated educational software. Computer Supported Learning 2010 Jun;5(2):167-189. [doi: 10.1007/s11412-010-9085-5]
65. Mayer RE, Fiorella L. Principles for reducing extraneous processing in multimedia learning: coherence, signaling, redundancy, spatial contiguity, and temporal contiguity principles. In: The Cambridge Handbook of Multimedia Learning, 2nd edition: Cambridge University Press; 2014. [doi: 10.1017/CBO9781139547369.015]
66. Kulgemeyer C, Peters CH. Exploring the explaining quality of physics online explanatory videos. Eur J Phys 2016 Nov 1;37(6):065705. [doi: 10.1088/0143-0807/37/6/065705]
67. Worthley B, Guo M, Sheneman L, Bland T. Antiparasitic pharmacology goes to the movies: leveraging generative AI to create educational short films. AI 2025;6(3):60. [doi: 10.3390/ai6030060]
68. Arkün-Kocadere S, Çağlar Özhan Ş. Video lectures with AI-generated instructors: low video engagement, same performance as human instructors. IRRODL 2024;25(3):350-369. [doi: 10.19173/irrodl.v25i3.7815]
69. Netland T, von Dzengelevski O, Tesch K, Kwasnitschka D. Comparing human-made and AI-generated teaching videos: an experimental study on learning effects. Comput Educ 2025 Jan;224:105164. [doi: 10.1016/j.compedu.2024.105164]
70. Wahlberg SJ, Haglund J, Gericke NM. Metaphors on protein synthesis in Swedish upper secondary chemistry and biology textbooks – a double-edged sword. Res Sci Educ 2025 Apr;55(2):425-444. [doi: 10.1007/s11165-024-10197-y]
71. Wernecke U, Schwanewedel J, Harms U. Metaphors describing energy transfer through ecosystems: helpful or misleading? Sci Educ 2018 Jan;102(1):178-194. [doi: 10.1002/sce.21316]
72. Orgill M, Bussey TJ, Bodner GM. Biochemistry instructors' perceptions of analogies and their classroom use. Chem Educ Res Pract 2015;16(4):731-746. [doi: 10.1039/C4RP00256C]
73. Raviolo A, Garritz A. Analogies in the teaching of chemical equilibrium: a synthesis/analysis of the literature. Chem Educ Res Pract 2009;10(1):5-13. [doi: 10.1039/B901455C]
74. Didiş Körhasan N, Hıdır M. How should textbook analogies be used in teaching physics? Phys Rev Phys Educ Res 2019 Feb;15(1):010109. [doi: 10.1103/PhysRevPhysEducRes.15.010109]
75. Haglund J, Jeppsson F. Using self‐generated analogies in teaching of thermodynamics. J Res Sci Teach 2012 Sep;49(7):898-921. [doi: 10.1002/tea.21025]
76. Kulgemeyer C, Wittwer J. Misconceptions in physics explainer videos and the illusion of understanding: an experimental study. Int J Sci Math Educ 2023;21(2):417-437. [doi: 10.1007/s10763-022-10265-7] [Medline: 35342380]
77. Wiley J, Jaeger AJ, Taylor AR, Griffin TD. When analogies harm: the effects of analogies on metacomprehension. Learn Instr 2018 Jun;55:113-123. [doi: 10.1016/j.learninstruc.2017.10.001]
78. Paik ES, Schraw G. Learning with animation and illusions of understanding. J Educ Psychol 2013;105(2):278-289. [doi: 10.1037/a0030281]
79. Bowen JA, Watson CE. Teaching With AI: A Practical Guide to A New Era of Human Learning: Johns Hopkins University Press; 2024. URL: https://muse.jhu.edu/book/123216 [accessed 2025-07-01] [doi: 10.56021/9781421449227]

80. Goodwin C. Professional vision. Am Anthropol 1994 Sep;96(3):606-633. [doi: 10.1525/aa.1994.96.3.02a00100]
81. Alter AL, Oppenheimer DM. Uniting the tribes of fluency to form a metacognitive nation. Pers Soc Psychol Rev 2009 Aug;13(3):219-235. [doi: 10.1177/1088868309341564] [Medline: 19638628]
82. Finn B, Tauber SK. When confidence is not a signal of knowing: how students' experiences and beliefs about processing fluency can lead to miscalibrated confidence. Educ Psychol Rev 2015 Dec;27(4):567-586. [doi: 10.1007/s10648-015-9313-7]
83. King D, Auschaitrakul S. Symbolic sequence effects on consumers' judgments of truth for brand claims. J Consum Psychol 2020 Apr;30(2):304-313. [doi: 10.1002/jcpy.1132]
84. Unkelbach C. Reversing the truth effect: learning the interpretation of processing fluency in judgments of truth. J Exp Psychol Learn Mem Cogn 2007 Jan;33(1):219-230. [doi: 10.1037/0278-7393.33.1.219] [Medline: 17201563]
85. Silaj KM, Frangiyyeh A, Paquette‐Smith M. The impact of multimedia design and the accent of the instructor on student learning and evaluations of teaching. Appl Cogn Psychol 2024 Jan;38(1):e4143. [doi: 10.1002/acp.4143]
86. Carpenter SK, Northern PE, Tauber SU, Toftness AR. Effects of lecture fluency and instructor experience on students' judgments of learning, test scores, and evaluations of instructors. J Exp Psychol Appl 2020 Mar;26(1):26-39. [doi: 10.1037/xap0000234] [Medline: 31169395]
87. Brand S. Meet TRACI: User's Guide to the TRACI Prompt Framework for ChatGPT: Structured Prompt; 2023. URL: https://structuredprompt.com/free-traci-users-guide-white-paper/ [accessed 2025-07-02]
88. Miao F, Holmes W, Huang R, Zhang H. AI and Education: Guidance for Policy-Makers: United Nations Educational, Scientific and Cultural Organization; 2021. URL: https://tinyurl.com/mr4dzxtv [accessed 2025-07-02]

Abbreviations

AI: artificial intelligence
API: application programming interface
ETC: electron transport chain
genAI: generative artificial intelligence
LLM: large language model
RQ: research question

Edited by T Leung; submitted 03.07.25; peer-reviewed by C Wardle, I Park; revised version received 15.09.25; accepted 30.09.25; published 20.11.25.

Please cite as:
Jones EM, Newman JD, Kim B, Fogle EJ
AI-Generated "Slop" in Online Biomedical Science Educational Videos: Mixed Methods Study of Prevalence, Characteristics, and Hazards to Learners and Teachers
JMIR Med Educ 2025;11:e80084
URL: https://mededu.jmir.org/2025/1/e80084
doi:10.2196/80084

Application of AI Communication Training Tools in Medical Undergraduate Education: Mixed Methods Feasibility Study Within a Primary Care Context

Chris Jacobs[1], MB BChir, MRes, MD; Hans Johnson[2], MBChB, MRES; Nina Tan[1], BSc; Kirsty Brownlie[2], MBChB, BA; Richard Joiner[1], MSc, PhD; Trevor Thompson[2], MBBS, MSc, PhD

[1]Department of Psychology, University of Bath, Claverton Down, Bath, United Kingdom

[2]Bristol Medical School, University of Bristol, Bristol, United Kingdom

Corresponding Author:
Chris Jacobs, MB BChir, MRes, MD
Department of Psychology, University of Bath, Claverton Down, Bath, United Kingdom

Abstract

Background: Effective communication is fundamental to high-quality health care delivery, influencing patient satisfaction, adherence to treatment plans, and clinical outcomes. However, communication skills training for medical undergraduates often faces challenges in scalability, resource allocation, and personalization. Traditional methods, such as role-playing with standardized patients, are resource intensive and may not provide consistent feedback tailored to individual learners' needs. Artificial intelligence (AI) offers realistic patient interactions for education.

Objective: This study aims to investigate the application of AI communication training tools in medical undergraduate education within a primary care context. The study evaluates the effectiveness, usability, and impact of AI virtual patients (VPs) on medical students' experience in communication skills practice.

Methods: The study used a mixed methods sequential explanatory design, comprising a quantitative survey followed by qualitative focus group discussions. Eighteen participants, including 15 medical students and 3 practicing doctors, engaged with an AI VP simulating a primary care consultation for prostate cancer risk assessment. The AI VP was designed using a large language model and natural voice synthesis to create realistic patient interactions. The survey assessed 5 domains: fidelity, immersion, intrinsic motivation, debriefing, and system usability. Focus groups were used to explore participants' experiences, challenges, and perceived educational value of the AI tool.

Results: Significant positive responses emerged against a neutral baseline, with the following median scores: intrinsic motivation 16.5 of 20.0 (IQR 15.0 - 18.0; d=2.09, P<.001), system usability 12.0 of 15.0 (IQR 11.5 - 12.5; d=2.18, P<.001), and psychological safety 5.0 of 5.0 (IQR 5.0 - 5.0; d=4.78, P<.001). Fidelity (median score 6.0/10.0, IQR 5.2 - 7.0; d=–0.08, P=.02) and immersion (median score 8.5/15.0, IQR 7.0 - 9.8; d=0.25 P=.08) were moderately rated. The overall Immersive Technology Evaluation Measure scores showed a high positive learning experience: median 47.5 of 65.0 (IQR 43.0 - 51.2; d=2.00, P<.001). Qualitative analysis identified 3 major themes across 11 subthemes, with participants highlighting both technical limitations and educational value. Participants valued the safe practice environment and the ability to receive immediate feedback.

Conclusions: AI VP technology shows promising potential for communication skills training despite the current realism limitations. While it does not yet match human standardized patient authenticity, the technology has achieved sufficient fidelity to support meaningful educational interactions, and this study identified clear areas for improvement. The integration of AI into medical curricula represents a promising avenue for innovation in medical education, with the potential to improve the quality and effectiveness of training programs.

(JMIR Med Educ 2025;11:e70766) doi:10.2196/70766

KEYWORDS

artificial intelligence; technology-enhanced learning; virtual patient; communication skills; simulation

Introduction

Effective communication is fundamental to high-quality health care delivery, influencing patient satisfaction, adherence to treatment plans, and clinical outcomes [1]. Despite its significance, communication skills training for medical undergraduates often faces challenges in scalability, resource allocation, and personalization. Although traditional methods such as role-playing with standardized patients or observational feedback sessions are effective, they are resource intensive and

may not provide consistent feedback tailored to individual learners' needs in all circumstances [2,3].

Artificial intelligence (AI) offers an innovative solution to address these challenges by simulating realistic patient-provider interactions in controlled, scalable environments. Given their ability to show proficiency in medical knowledge, large language models (LLMs) have been used by medical students in the preclinical phase, as the models aid them in differential diagnosis and provide interactive practice cases to support learning [4]. Recent applications of LLMs in medical education demonstrate expanding capabilities across diverse educational contexts. Advanced LLMs (ChatGPT, Copilot, PaLM, Bard, and Gemini) show comparable performance to students in gross anatomy assessments, indicating potential for educational support in foundational medical sciences [5]. Furthermore, LLMs have shown competitive performance against medical students in specialized topic areas, suggesting potential for supplementing traditional learning approaches [6].

A narrative review of AI in health care communication indicated that these systems have the potential to replicate the complexities of clinical encounters, including interpreting verbal and nonverbal cues, ensuring empathy in responses, and managing dynamic conversational scenarios [7]. Furthermore, communication, as an emotional exchange of information, can be considered a complex system characterized by diversity, nonlinearity of response, interconnectedness of individuals (akin to a neural network), dynamism, feedback orientation, and constant information flow. AI encompasses these facets and can help predict emergent behaviors with machine learning.

AI tools, such as conversational agents and LLMs, can provide health care learners with opportunities to practice communication skills repeatedly and independently while receiving immediate and detailed feedback on performance [8,9]. Simulated virtual patients (VPs) have shown promise in prenatal counseling education, where ChatGPT-generated dialogues increased the realism and variety of patient interactions while maintaining clinical accuracy [10]. In emergency medical services, generative pretrained transformer–based VPs integrated with mixed reality simulation improved the effectiveness of communication training for medical first responders, although technical delays remained a limiting factor [9]. Large-scale implementations have achieved substantial reach, with platforms such as the Geeky Medics Virtual Patient Simulator conducting over 45,000 AI-powered clinical consultations using GPT-3.5 and GPT-4 technologies [11]. However, training needs are substantial yet largely unmet: a previous study reported that 74.8% of medical students wanted structured AI training in medical curricula, but only 26.8% felt competent to inform patients about AI applications [12].

Primary care settings are ideal for communication skills training [13], given the frequent need for effective patient-provider communication to address complex, multidisciplinary health issues. Training future medical professionals in this context requires tools that mirror the intricacies of real-world interactions, including handling diverse patient populations, managing sensitive topics, and navigating the nuances of shared decision-making. AI platforms offer an opportunity to simulate such scenarios with high fidelity while ensuring accessibility and scalability, thus addressing key gaps in current educational approaches [14-16]. Previous studies have used AI in primary care education [17,18]; however, publications using advanced language models to simulate VP encounters remain limited, as the vast majority are discussion papers and viewpoints [19].

To systematically evaluate the educational effectiveness of AI-powered learning tools in health care education, robust measurement frameworks are essential. The rapid proliferation of immersive technologies in medical training, from virtual reality simulations to AI VPs, has created a need for validated assessment instruments that can capture the multidimensional nature of technology-enhanced learning experiences. Recent frameworks, such as the Immersive Technology Evaluation Measure (ITEM) developed by Jacobs et al [20], provide a structured approach to evaluating the educational efficacy of technology-enhanced learning tools and have demonstrated use in conversational AI in medical student evaluation [21]. The ITEM assesses the following domains: engagement, intrinsic motivation, cognitive load, system usability, and postexperience debriefing. These domains are particularly relevant in simulation training with technology, where the fidelity of the simulated interaction as well as the learners' motivation to engage and ability to reflect on performance are pivotal to achieving meaningful educational outcomes [22]. Applying such frameworks, this study aims to investigate the use of AI technologies in communication training within the primary care context for medical undergraduates.

This study aims to evaluate AI VP technology as a tool to support communication-based learning and to compare this technology with standardized patients used in primary care educational sessions, using quantitative and qualitative analysis. Through quantitative inquiry, the study seeks to evaluate how medical undergraduates perceive the effectiveness and usability of conversational AI, as well as its impact on their confidence and self-reported competence in communication skills. Complementing this, a qualitative analysis explores the experiences of learners using these AI-powered tools, focusing on the challenges and benefits they encounter while practicing communication skills in simulated primary care scenarios. Together, the purpose of these analyses is to provide a comprehensive understanding of the tool's utility and effectiveness and areas for potential improvement in enhancing communication training for future health care professionals.

Methods

Ethical Considerations

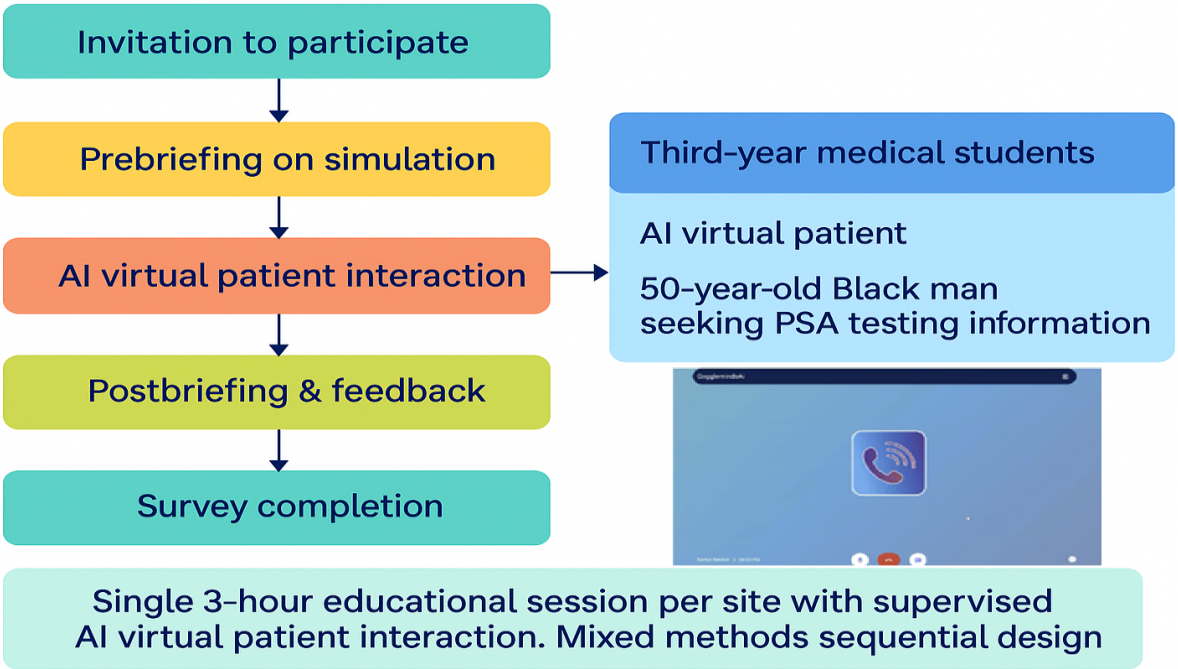
Ethical approval was obtained from a regional Independent Research Board, Swindon, United Kingdom (CJ062023). Furthermore, the senior academic staff at the University of Bristol approved the study. Informed consent was obtained from all participants, who were free to opt out at any time. All data were anonymized prior to analysis to ensure privacy and confidentiality. No financial or other compensation was provided for participation.

Study Design

The study used a mixed methods sequential explanatory design comprising 2 distinct phases: a quantitative phase followed by a qualitative phase. The mixed methods design strengthens the research findings and provides a holistic approach to answer the research aims [23]. The qualitative focus group phase enables a more detailed understanding of the quantitative phase.

The study sought to provide a nuanced understanding of how AI can enhance communication skills training for medical undergraduates. Primary care training sites were invited to participate via email. Three sites were randomly selected from email responses, with the number limited to 3 for technology oversight purposes. Third-year medical students completing primary care clerkships and general practitioner (GP) facilitators were recruited from 3 selected sites, with 5 students and 1 GP facilitator per site. All approached participants consented to participate (18/18, 100% response rate). The participating students (n=15) represented 6% of the total third-year medical student population (N=240), with GP facilitators totaling 3 participants. Students in their third year of study were selected, as the third-year curriculum focuses on primary care communication skills. Practicing doctors were included because they represent the educator demographic who would implement AI VP technology in clinical teaching. Their participation provided stakeholder perspectives on feasibility, usability, and educational value from the instructor viewpoint. Educational sessions at the primary care sites lasted 3 hours and were conducted as single-exposure experiences at each site. The theme of the day was evaluating urological cancer, and students interacted with the AI VP (20 min per consultation) only during these supervised sessions, with no access provided outside the research period. Standardized materials were provided to the facilitators to conduct a prebriefing on simulation and important aspects of communication. Learning objectives included developing shared decision-making communication skills for prostate cancer screening discussions; explaining the benefits and limitations of prostate-specific antigen (PSA) testing, including the concepts of specificity and sensitivity; exploring patient concerns and health anxieties; applying evidence-based counseling techniques for screening decisions; and addressing patient questions about ethnicity-related cancer risks. Furthermore, the facilitator at each site conducted a postbriefing on student performance on consultations. The students and facilitators were invited to attend 3 focus group discussions following their participation in AI simulation and completion of the initial survey. Figure 1 outlines the procedure and shows a screenshot of the display seen by participants. Survey methods are reported using the Consensus-Based Checklist for Reporting of Survey Studies (CROSS; Checklist 1).

Figure 1. Participant recruitment and study protocol. AI: artificial intelligence; PSA: prostate-specific antigen.



Technical Implementation

The VP was constructed using OpenAI's GPT-3.5-turbo LLM accessed via RESTful application programming interface (API) end points. The system was hosted on the Firebase cloud infrastructure (Google LLC), using Firebase Authentication for secure user access, Firestore database for session management, and Firebase Hosting for web application deployment. All communications with the OpenAI API were encrypted using HTTPS protocols. The language model was configured with specific parameters to optimize conversational flow and educational authenticity. API requests used a temperature setting of 0.7 to balance response creativity with consistency.

The AI VP was programmed with a comprehensive system prompt defining the patient persona as a 50-year-old Black man seeking information about PSA testing. The prompt included specific instructions to maintain character consistency, use an appropriate level of medical knowledge for a lay person, express realistic health concerns and anxieties, and avoid breaking the illusion of being a real patient (see Multimedia Appendix 1 for more information on the prompt). Conversational context was

maintained through session-based memory management. Students accessed the AI VP through a secure web-based interface featuring a conversational chat window with dual input modalities. The interface supported both typed text responses and voice input captured via Web Speech API (Google LLC) for speech-to-text transcription. User authentication was implemented through Firebase Authentication to ensure secure access and session isolation between participants. Student speech was captured using browser-based microphone interfaces and converted to text using the Web Speech API configured for British English language recognition. AI responses were converted to natural speech using the ElevenLabs text-to-speech synthesis service, configured with a professional male voice model emulating a London accent. Prior to deployment, the research team extensively tested the AI VP by interacting with the VP as a doctor to ensure consistent character portrayal, appropriate medical knowledge representation, and reliable technical performance. Content validity was achieved by testing, and ecological validity was sought by real-world deployment in primary care settings. A video example of this system is available in Multimedia Appendix 2.

Survey Instrument and Focus Group Development

The survey used an abridged version of the ITEM, which assesses 5 domains of experience in a health care education context using technology-enhanced learning in simulation [24,25]. The engagement subdomain included 2 questions related to the realism (fidelity) of the experience. The measure was developed on the basis of the Model for Immersive Technology in Healthcare Education (MITHE), which proposes that the learners' perception of the usability of technology is integral to experiential learning [20]. System usability was measured as learners' perceptions of technology ease of use (using adapted System Usability Scale items), recognizing that these perceptions are influenced by individual learners' characteristics such as prior technology experience and adaptability to novel interfaces. The study team reviewed the full ITEM questionnaire (40 questions) and reduced the number of questions to 12, which is supported by Jacobs and Rigby's [24] work on measure development, minimizing the effect on internal consistency and maintaining construct validity. Domain scores were calculated by summing the individual item responses within each ITEM domain: fidelity (2 items, range 2 - 10), immersion (3 items, range 3 - 15), intrinsic motivation (4 items, range 4 - 20), debriefing psychological safety (1 item, range 1 - 5), and system usability (3 items, range 3 - 15). All items were rated on a 5-point Likert scale with a consistent scoring direction. The Intrinsic Motivation Inventory was used to assess the perceived interest, learning value, and user competence, as proposed by the self-determination theory [26]. The survey was designed by medical tutors, and the construct was derived from the work of a consortium of simulation specialists on using simulation educational technology using the Delphi methodology [25].

In the subsequent focus group sessions (n=3), the interview data were enhanced by 4 open-ended questions in the initial survey to explore the authenticity, realism, and potential learning value of the interaction and to compare AI to human patients. Each focus group was facilitated and led by a medical tutor (KB), and 5 areas of questioning were developed by the study team. An interview guide with topics on educational value, realism, technical aspects, further development, and future applications was created, allowing for deeper probing through conversation (see Multimedia Appendix 3 for topic guide). Focus group interviews were conducted after the AI interaction, and the end data produced were obtained from participants from AI VP groups. Additionally, each focus group included a medical tutor to enrich the perspectives, with a facilitator encouraging student input to minimize the power differential this created.

Data Analysis

Quantitative Component

Quantitative data obtained through the survey instrument ITEM were analyzed using both descriptive and inferential statistical methodologies, using nonparametric methods suitable for Likert scale data. Statistical analysis was performed using R software (version 4.3.2; R Foundation for Statistical Computing) [27]. The primary metrics used to summarize the responses for each domain (fidelity, immersion, intrinsic motivation, debriefing, and usability scores) were the median and IQR, as these are the most appropriate measures of central tendency and variability for ordinal data. This inference provides an accurate representation of participants' response without assuming a normal distribution. The median rating for each domain was standardized as a percentage of the maximum possible score to allow for fair comparisons of ratings. Given the small sample size, inferential statistical analysis was limited to exploratory purposes. Single group nonparametric statistical analysis was conducted using 1-sample Wilcoxon signed rank tests comparing participant domain scores against neutral response values (domain midpoints) [28]. Neutral values represent a theoretical reference point where there is a lack of prior experience and represent a conceptual no-effect condition [29]. This approach was selected due to the ordinal nature of Likert scale data and small sample size precluding parametric assumptions. Effect sizes were calculated using Cohen *d* conventions (0.2=small, 0.5=medium, 0.8=large effect) to assess practical significance alongside statistical significance. Statistical significance was set at $P<.05$ for this exploratory analysis, with no corrections applied for multiple comparisons given the pilot study design.

Qualitative Component

Open-ended survey responses were uploaded to NVivo software (version 12; QSR International Pty Ltd) for an initial content analysis. A thematic content analysis was performed to investigate the participants' experience of using an AI VP to simulate a primary care consultation. The thematic analysis was performed using a hybrid inductive-deductive approach guided by the MITHE framework and ITEM domains while remaining open to emergent themes. The focus group data were analyzed by 1 author (KB), and the open-ended response codes and themes were explored by 2 other authors (CJ and HJ). The thematic analysis was conducted following the 6-step framework outlined by Braun and Clarke [30], encompassing getting familiarized with the data, generating initial codes, searching for themes, reviewing themes, defining and naming themes, and producing the final report. A provisional codebook was developed from theoretical constructs and then iteratively refined through independent dual coding by 2 authors (CJ and HJ).

Codebook development used a collaborative shared document approach with interrater agreement, and discrepancies were resolved through a subsequent consensus discussion. Finally, 2 authors (CJ and HJ) refined and revised the datasets to capture the overall narrative and major themes. Furthermore, the open-ended question dataset was analyzed on the basis of a study by Braun et al [31], who suggest that a reduced transcript length can be both concise and informative, retaining the unguarded perspectives of participants.

Results

Demographics

The survey respondents included third-year clinical medicine undergraduate students from a UK medical school. Of the 18 participants, 15 were medical students (83%), and 3 (17%) were practicing doctors in primary care and group tutors. No participants had prior experience interacting with an AI-simulated patient.

Quantitative Results

There were no missing data from the 18 participants.

A quantitative survey assessed the perceived effectiveness and usability of the AI tool and the medical undergraduates' engagement with the tool. Results are summarized in Table 1, and the results of the 5 domains are standardized in Figure 2.

Table . Summary of the AI[a] educational domain ratings.

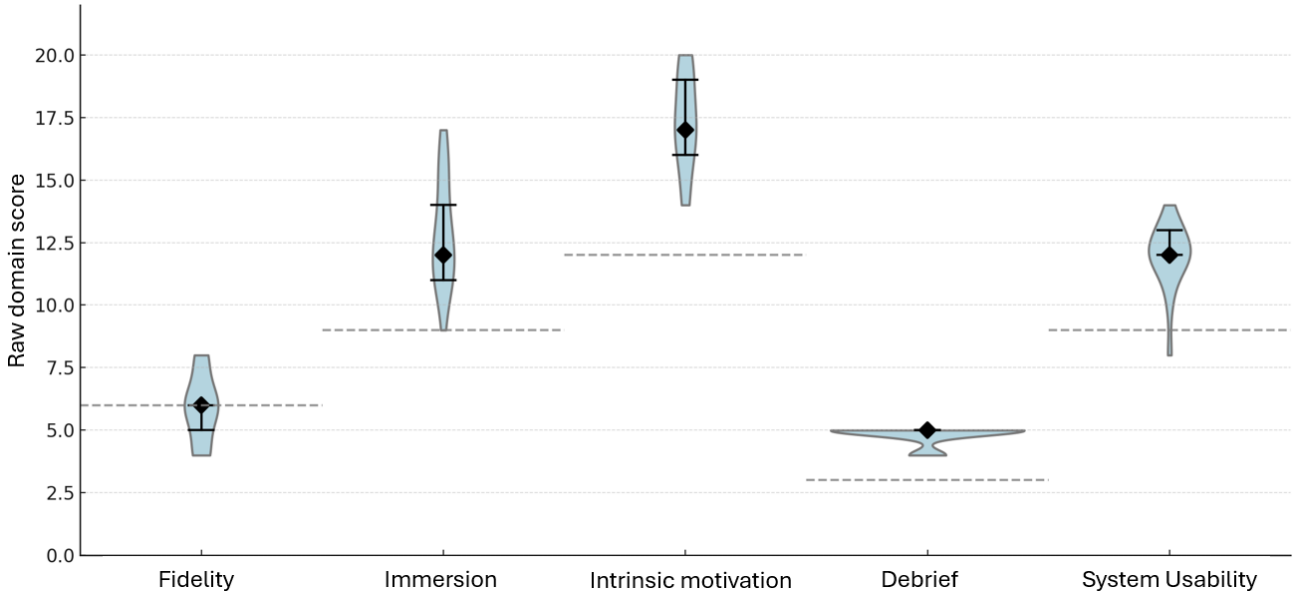
Domain	Raw median score/maximum possible score (IQR)	Difference between the median and neutral scores	Standardized median score/5.0 (IQR)	Median score as percentage of maximum possible score (%)	Effect size (*d*)	*P* value[b]	Context[c]
Fidelity	6.0/10.0 (5.2-7.0)	0.0	3.0 (2.6-3.5)	60.0	–0.08	.02	Realism of AI virtual patient communication
Immersion	8.5/15.0 (7.0-9.8)	0.5	2.8 (2.3-3.3)	56.7	0.25	.08	Participant engagement and immersion
Intrinsic motivation	15.5/20.0 (15.0-18.0)	4.5	4.1 (3.8-4.5)	82.5	2.09	<.001	Internal motivation and learning potential
Debriefing	5.0/5.0 (5.0-5.0)	2.0	5.0 (5.0-5.0)	100.0	4.78	<.001	Quality and safety of reflective discussion
System usability	12.0/15.0 (11.5-12.5)	3.0	4.0 (3.8-4.2)	80.0	2.18	<.001	Ease of use and accessibility of the platform
Total	47.5/65.0 (43.0-51.2)	9.5	3.6 (3.3-3.9)	73.1	2.00	<.001	Combined score across all domains

[a]AI: artificial intelligence.

[b]*P* values derived from the Wilcoxon signed rank test comparing each domain's score to a neutral reference point.

[c]Context summarizes what each domain aims to measure in relation to the AI-enhanced learning activity.

Figure 2. Violin plot of raw domain scores versus a neutral benchmark.



The fidelity or realism of the AI communication had a moderate median score of 6.0 (IQR 5.2 - 7.0), of a maximum possible score of 10.0. Furthermore, the median score for immersion was 8.5 (IQR 7.0 - 9.8), with a maximum possible score of 15.0. The difference between the median score and a neutral response comparator (ie, median difference) was equivalent in both domains (fidelity: median difference=0.0; *P*=.02; immersion: median difference=0.5; *P*=.08).

Intrinsic motivation had a high median score of 15.5 (IQR 15.0 - 18.0), with a maximum possible score of 20.0. There was a significant difference compared to the neutral response, with a large effect size (median difference=4.5; *d*=2.09, *P*<.001).

A universally high score was reported for the psychological safety of interacting with an AI VP during simulation debriefing (median 5.0, IQR 5.0 - 5.0). There was a significant difference compared to the neutral response, with a large effect size (median difference=2.0; *d*=4.78, *P*<.001).

System usability was highly rated, with a median score of 12.0 (IQR 11.5 - 12.5) on a maximum possible score of 15.0, which showed a significant positive difference from the neutral response, with a large effect size (median difference=3.0; *d*=2.18, *P*<.001).

The total measure score was high, with a median of 47.5 (IQR 43.0 - 51.2), and showed a significant positive median difference compared to the neutral response, with a large effect size (*d*=2.00, *P*<.001).

Qualitative Results

Following a collective review of the responses of 18 participants to the focus group questions and open-ended questions, 3 overarching themes were identified, supported by 11 subthemes (Table 2 and Figure 3). Convergence of these themes and integration with quantitative results are described later under Key Findings and Implications in the Discussion section.

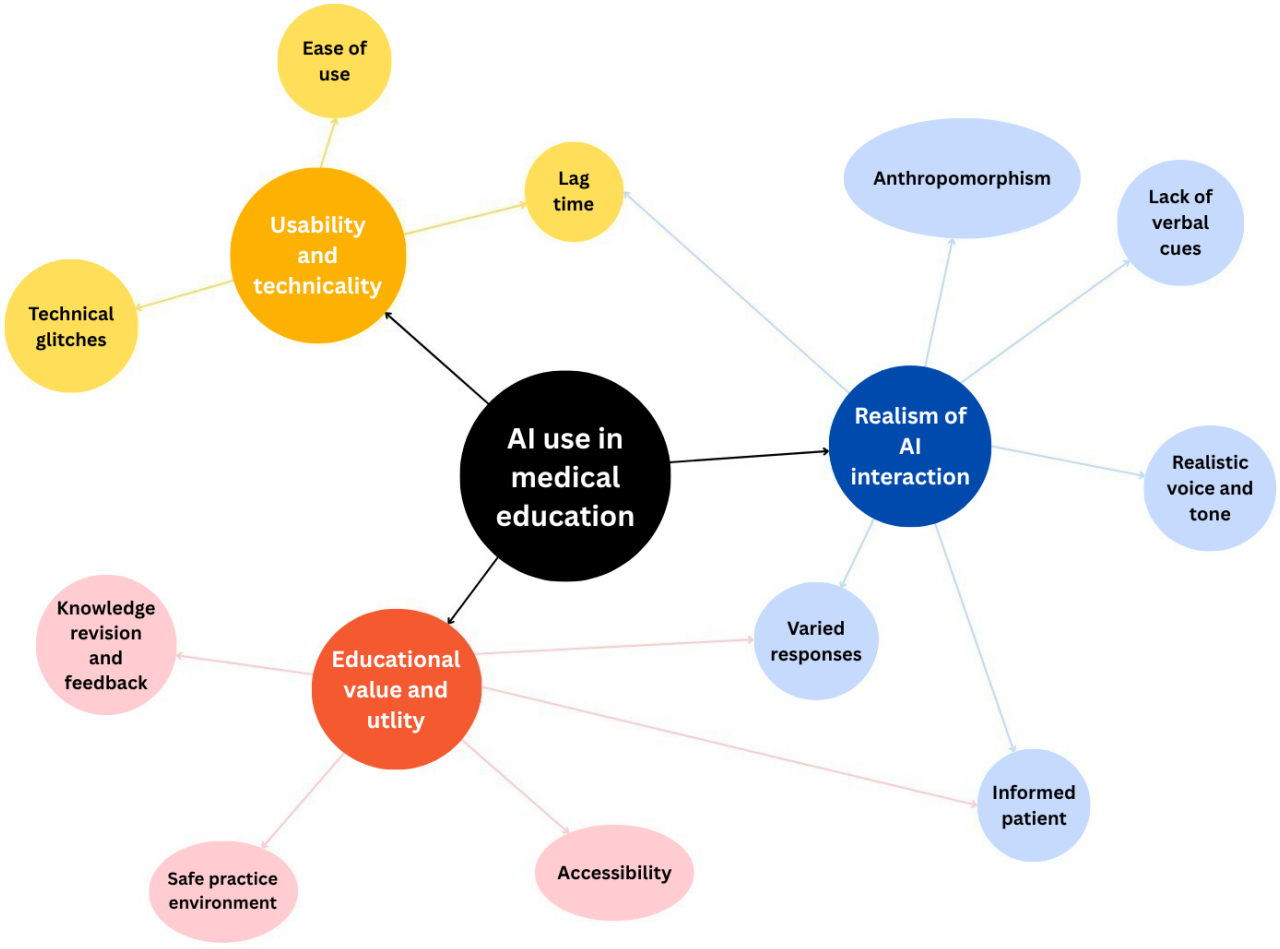
Table . Qualitative summary of the major themes and subthemes identified from responses to open-ended survey questions and focus group interviews.

Theme and subtheme	Description	Illustrative quote (participant number; quote ID)
Usability and practicality		
Time lag	Significant pauses disrupted the natural flow of conversation and reduced realism.	"Long pauses between questions. Hard [*sic*] felt like emotionless answers." (P3; Q1)
		"The pause between question and answering made it difficult for the consultation to flow and didn't feel it replicated a 'real' consultation." (P6; Q2)
Technical glitches	System crashes or repeated questions caused frustration and disrupted learning.	"There were a lot of faults and hiccups as we did it but wasn't too bad overall." (P12; Q3)
		"There was also some glitches that required questions to be asked multiple times or for the consulting student to begin again." (P17; Q4)
Ease of use	The technology was generally easy to access and use.	"The technology [is] straightforward to access and use." (Focus group 1; Q5)
Realism of AI[a] interaction		
Anthropomorphism	Some interactions felt mechanical, with reduced empathy and natural conversational flow.	"It felt very robotic with a lot of miscommunication." (P12; Q6)
		"Robotic answers and there was a delay in responding." (P16; Q7)
		"It felt like a realistic conversation." (Focus group 2; Q8)
Informed patient	AI responses were either too short or overwhelming with unnecessary details.	"The AI gave more information unprovoked/not asked for than a patient normally would." (P17; Q9)
		"The patient was very well informed." (P13; Q10)
		"Gave very scientific responses." (P10; Q11)
	AI used technical language inappropriate for a simulated patient role.	"The AI was using medical jargon that you wouldn't normally expect a patient to use." (P17; Q12)
Lack of nonverbal cues	The absence of body language and facial expressions limited realism and engagement.	"Also not being able to see the AI's body language." (P16; Q13)
Educational value and utility		
Realistic voice and tone	AI's voice and tone were seen as realistic and contributed to a sense of authenticity.	"AI voice was human-like, but the answers were a bit too succinct." (P5; Q14)
		"Realistic voice." (P13; Q15)
		"I think the voice was very realistic, the time delay wasn't too much of an issue." (P18; Q16)
Varied responses	AI responses were sometimes unexpected, adding a layer of realism.	"The AI patient gives unpredictable answers and asks questions like a patient would." (P16; Q17)
		"Was really impressed by how the AI bot could interpret conversational language and respond appropriately." (P9; Q18)
	AI responses were repetitive, leading to a lack of conversational depth.	"There were limited responses however and we got a sense of déjà vu with some of the responses." (P9; Q19)
		"It answered some questions that we didn't ask." (P17; Q20)
Knowledge revision and feedback	AI reinforced clinical knowledge effectively and provided a platform for self-assessment.	"Very useful knowledge checking in a clinical setting." (P1; Q21)
		"The AI can ask questions which helps with the learning." (P7; Q22)
		"Good to learn clinical knowledge." (P4; Q23)

Theme and subtheme	Description	Illustrative quote (participant number; quote ID)
Safe practice environment	AI provided a low-risk space for students to practice and make mistakes without judgment.	"It would be very useful as a revision tool to check how we as students present knowledge/consult with patients." (P5; Q24) "This was a good tool for practicing giving patients information in a safe space." (P9; Q25)
Accessibility	The ability to use AI for practice outside formal clinical settings was highly valued.	"Allows us to practice at home." (P10; Q26) "It would be really useful to have access to it at home to practice with." (P17; Q27)

[a]AI: artificial intelligence.

Figure 3. Thematic map displaying the interlinking themes and subthemes. AI: artificial intelligence.



Theme 1: Usability and Practicality

The AI VP interface demonstrated mixed usability outcomes across 3 key domains. Response delays of approximately 2 - 3 seconds consistently disrupted the natural flow of conversation and reduced perceived realism (Q1 and Q2), impacting participants' ability to maintain authentic consultation dynamics. System instability presented additional challenges, with technical faults requiring question repetition or consultation restarts (Q3 and Q4), interfering with the intended focus on communication skills development. Despite these technical limitations, participants found the interface straightforward to access and operate without requiring technical support or extensive instruction (Q5), facilitating smooth adoption across focus groups.

Theme 2: Realism of AI Interaction

Realism emerged as the most debated aspect among participants, with varied perceptions across multiple dimensions of the VP experience. Participant perspectives on the AI's humanlike qualities were mixed: while some participants found the interactions mechanical and emotionless (Q6 and Q7), focus groups generally perceived the overall experience as reasonably authentic (Q8). The AI's knowledge level created divergent opinions, with some viewing the detailed, scientific responses as unrealistic for typical patients (Q9, Q10, and Q11), particularly when medical jargon was inappropriately used (Q12). The audio-only format limited realism owing to the absence of visual and nonverbal communication cues (Q13), reducing engagement and authenticity compared to face-to-face consultations. However, participants praised the AI's voice

quality and tone as contributing to authenticity (Q14, Q15, and Q16), supporting the simulation's credibility as a telephone consultation format. The AI's ability to provide unpredictable responses and interpret conversational language enhanced perceived authenticity (Q17 and Q18), although repetitive responses occasionally created a sense of déjà vu, diminishing the conversational depth (Q19 and Q20).

Theme 3: Educational Value and Utility

Participants recognized significant educational potential across 3 key areas that could enhance medical training. The AI effectively reinforced clinical knowledge and provided opportunities for self-assessment (Q21, Q22, and Q23), with participants valuing the AI's ability to prompt a discussion of clinical concepts they might not have considered. The virtual environment offered a psychologically safe practice space without the risk of patient harm or judgment (Q24 and Q25), enabling students to experiment with consultation approaches and make mistakes without consequences. Participants particularly valued the potential for accessible, asynchronous learning outside formal clinical settings (Q26 and Q27), representing aspirational use for home-based practice and revision, although access was limited to supervised study sessions during the research period.

Discussion

This mixed methods study explored the application of AI communication training tools in medical undergraduate education within a primary care context. The findings indicate that AI VP encounters can enhance communication skills training by providing a scalable, accessible, and realistic simulation environment.

Key Findings and Implications

Participants reported moderate levels of perceived realism (fidelity score 6/10), with the voice quality achieving authenticity but technical limitations, such as response delays and overly clinical language, reducing overall fidelity. Despite these limitations, the AI VP demonstrated sufficient realism for educational engagement, as evidenced by the high intrinsic motivation score (16.5/20.0) and positive learning value feedback.

In particular, the results demonstrated high levels of perceived effectiveness, usability, and psychological safety among participants. High median scores for intrinsic motivation and system usability were particularly notable, indicating strong engagement and ease of use. These findings align with previous research suggesting that AI tools can effectively support medical education by providing immediate feedback and opportunities for repeated practice [32,33]. The high intrinsic motivation score suggests that students found the AI tool engaging and beneficial for their learning, which is crucial for the adoption of new educational technologies [34]. Learning motivation can be seen as a mediator between technological acceptance and self-efficacy on the task [35,36]. Importantly, AI technologies create a dynamic and adaptive learning ecosystem that tailors educational experiences to individual preferences, enhancing engagement and effectiveness.

Immersion scores (median 8.5/15) aligned with prior ITEM validation research demonstrating that computer screen–based educational experiences typically achieve moderate immersion levels [37]. These scores may reflect the inherent limitations of a 2D interface rather than deficiencies of AI VPs. Importantly, prior studies have found equivalent learning outcomes between high-immersion virtual reality and 2D screens, suggesting moderate immersion may be optimal because it avoids excessive cognitive load while maintaining educational presence.

Technological acceptance on the basis of reasoned action was first proposed as the technology acceptance model [38], which is used to explain the association between acceptance of a computer system and the behavioral intention thereof. System usability within human-computer interaction is critical in the use of technology [39]. The MITHE framework provided an appropriate theoretical basis for our AI VP evaluation, as conversational AI creates immersive learning experiences through cognitive and social engagement rather than visuospatial immersion. System usability within MITHE encompasses both learner perceptions of technology ease of use and individual learner characteristics that influence technology adoption, both of which were evident in our participants' varied responses to the AI VP interface. Furthermore, in the context of technological learning, this means that students who find technology-based learning activities inherently interesting are more likely to engage deeply and persistently. Thus, there is a paradigm of interrelating concepts between engagement and usability. The AI tool enabled both to be established and demonstrated an educational value to students.

Realism in simulation refers to the degree to which a simulation accurately represents real-life scenarios. It involves creating an environment that closely mimics actual clinical situations, allowing learners to engage in tasks and decision-making processes as they would in real clinical settings. Realism is necessary for ensuring that the skills and knowledge acquired during simulation training are transferable to real-world practice [14,40]. Fidelity in simulation is the extent to which the simulation replicates the real-world environment and experiences. It encompasses various dimensions, such as physical, conceptual, and psychological dimensions [41]. The degree of fidelity in this pilot study was reported as moderate by participants. Participants highlighted issues such as time lag, technical glitches, and the lack of nonverbal cues, which impacted the perceived fidelity of the AI interactions. The AI VP provided a naturalistic voice, which was the foundation for a humanlike experience. However, the psychological fidelity was impacted by the extent to which learners could emotionally and cognitively engage with the AI as if they were in a real situation. High psychological fidelity scenarios have the potential for students to develop not only their communication skills but also their empathy and bedside manner.

While AI VPs may not have a physical presence, their interactions can be designed to closely replicate real patient encounters. This includes realistic voice modulation, appropriate use of medical terminology, and the ability to simulate complex medical scenarios. In this study, the scenario required the AI VP to simulate a patient asking a doctor questions, leading to a balanced discussion on PSA testing. Participants appreciated

the AI's ability to simulate informed patient interactions, which enhanced their clinical knowledge and decision-making skills. However, some noted that the AI's responses could sometimes be overly detailed or robotic, detracting from the realism of the simulation. This feedback suggests that future iterations of AI VP should aim to balance informativeness with a natural conversational flow. The ability of the AI to provide detailed responses was seen as both a strength and a limitation, indicating the need for a more nuanced approach to AI response generation.

AI models, especially advanced ones such as LLMs, can understand the context of a prompt. This means they can interpret the nuances of a question or statement and generate responses that are contextually appropriate. For example, if a medical student asks about symptoms of a specific condition, the AI model can provide detailed information relevant to that condition [42]. This is somewhat incompatible with the patient perspective of lacking the understanding of a problem because of which they consult with a physician. A dynamic interaction improves realism, and the varied responses by the AI VP provided users with adaptive and personalized responses. Advanced AI models can simulate emotions and empathy in their responses [7,43]. For example, if a student is practicing delivering bad news to a patient, the AI can respond with appropriate emotional cues, such as expressing concern or asking for clarification, to mimic a real patient's reaction. A narrative synthesis of learning empathy through simulation suggests that simulation is an appropriate method to teach empathy to preservice health professional students [44].

In a pilot study among medical students, earlier LLM models, for example, GPT-3.5, that generated text-only responses were reported to create a plausible experience for students. The analysis revealed that most answers provided by the LLM were medically plausible and in line with the illness script [45]. Vaughan et al [46] recently reported that LLMs can create realistic simulation examples following text-based review. Furthermore, the accuracy, relevance, and structure of the AI programs benefit review prior to adoption in medical education settings [47].

The thematic analysis also revealed that participants valued the safe practice environment provided by the AI VPs. The ability to practice communication skills without the risk of harm to real patients was seen as a significant advantage, supporting the use of AI in medical training [48]. Additionally, the accessibility of the AI tool, allowing for asynchronous learning, was highlighted as a key benefit, particularly for students needing flexible study options [49]. Cognitive presence and social presence are essential for a comprehensive learning experience. While asynchronous learning environments offer flexibility to accommodate individual schedules, they also present challenges in enhancing cognitive and social presence. Despite an informative and realistic interaction, this aspect of learning may have been impacted, as participants assigned immersion a moderate median score. Social presence is the ability of learners to project themselves socially and emotionally in a learning environment. It involves the sense of being "there" and being able to interact meaningfully with others. The illusion of presence—that is, a student wholly experiencing the interaction as if they are a doctor talking to a real patient—may be broken by a simple time lag in response [50].

Educational Value and Future Applications

The study's findings suggest that AI-powered VPs have substantial educational value, particularly in enhancing communication skills in primary care settings. The positive feedback on knowledge reinforcement and the potential for personalized feedback indicates that AI tools can complement traditional training methods, providing a more comprehensive learning experience. The ability to receive immediate, detailed feedback from the AI was particularly valued by participants, highlighting the potential of AI to support continuous improvement in communication skills, contributing to the broader argument of adoption of AI in medical education curricula. Further advances in AI computational complexity with a reduced time lag in response and authenticity of character will further improve learner experience. There are numerous use cases beyond this pilot study, including patient standardization for objective structured clinical examinations, simulated consulting, advanced manikin simulation with realistic patient response, and procedural skills communication feedback. Integrating machine learning into the review of a participant's behavior and response to AI VP could add a degree of automation to medical and allied medical licensing examinations. Model variations allow for changes in VP characteristics; however, model stability and sensitivity need further investigation. Future development should prioritize technical infrastructure to minimize response delays using API optimization, advanced prompt engineering to achieve more natural conversational patterns, and adaptive information delivery systems that balance patient authenticity with educational value. Conversational analysis would be an important step to evaluate the clinical narrative and naturalness. Eventual integration of visual avatars and enhanced voice prosody could address nonverbal communication limitations while maintaining the accessibility advantages of current screen-based implementation.

Strengths and Limitations

To our knowledge, this is the first study to use an advanced, realistic voice on a recent LLM model, with the aim to assess simulation realism for students for use in practicing consultation skills. The context of primary care consulting that focused on discussing a case of PSA testing additionally tested the AI in a conversation that helped students develop their cognitive reasoning skills. The use case provides evidence for future work in assessing the accuracy of responses and AI feedback on the consultation via transcript analysis.

Being a pilot study, this study was underpowered, and the results demonstrating statistical significance ($P<.05$) with large effect sizes need to be interpreted with caution. This study was conducted at 3 randomly selected sites with a small sample of students (n=15) representing 6% of the third-year cohort (N=240), which may limit the generalizability of findings to the broader medical student population despite the random site selection. Furthermore, the measures were abridged versions of validated questionnaires. This study did not aim to analyze the content validity; however, construct validity was assumed

from prior work in this field. Additionally, this was an observational study, and a comparison of human consultation was not conducted, which is important for future work. The assumption of high intrinsic motivation and other domains requires further comparison to standard learning methods such as using actors.

Learning objectives focused on shared decision-making communication skills for prostate cancer screening discussions. This feasibility study assessed the perceived educational value of an AI VP through ITEM scores and qualitative feedback rather than a formal pre- and postcompetency assessment, representing a limitation. Future research should incorporate validated communication skills assessments to quantify learning outcomes objectively.

Focus groups enable the inclusion of a social element into the development of ideas on a topic; however, the presence of a tutor during the data collection may have influenced the responses by students, even when the facilitator is attuned to this and adapts questions to promote collective contribution.

Conclusion

The research aimed to explore the practical application of an AI VP tool in preparing future health care professionals for real-world patient interactions. AI VP technology shows promising potential for communication skills training despite the current limitations in realism. While it does not yet match human standardized patient authenticity, the technology achieved sufficient fidelity to support meaningful educational interactions. Furthermore, the study identified clear areas for improvement. The integration of AI into medical curricula represents a promising avenue for innovation in medical education, with the potential to improve the quality and effectiveness of training programs. Future research focusing on comparison with medical actors in assessment and AI-generated participant score will assist in strengthening the argument for AI integration into health care education.

Acknowledgments

The authors thank Goggleminds Ltd (London, United Kingdom) for providing access to the software platform used in this study at no charge. The company had no role in study design, data collection or analysis, manuscript preparation, or the decision to publish.

Data Availability

The anonymized datasets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: CJ (lead), KB (supporting)
Data curation: CJ (lead), NT (supporting)
Formal analysis: CJ (lead), HJ (supporting), NT (supporting)
Investigation: KB (lead), CJ (supporting), NT (supporting)
Methodology: CJ (lead), HJ (supporting), RJ (supporting), TT (supporting)
Project administration: CJ (lead), HJ (supporting)
Resources: KB (lead), CJ (supporting), TT (supporting)
Supervision: RJ (equal), TT (equal)
Validation: HJ (lead)
Visualization: CJ (lead), HJ (supporting), NT (supporting)
Writing – original draft: CJ (lead), HJ (supporting), KB (supporting), RJ (supporting), TT (supporting), NT (supporting)
Writing – review & editing: CJ (equal), HJ (equal), KB (equal), RJ (supporting), TT (supporting), NT (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
System prompt for artificial intelligence (AI) virtual patient.
[DOCX File, 297 KB - mededu_v11i1e70766_app1.docx]

Multimedia Appendix 2
Video example of the artificial intelligence (AI) system.
[MP4 File, 26817 KB - mededu_v11i1e70766_app2.mp4]

Multimedia Appendix 3

Artificial intelligence (AI) virtual patient focus group facilitator topic guide.
[PDF File, 107 KB - mededu_v11i1e70766_app3.pdf]

Checklist 1
Consensus-Based Checklist for Reporting of Survey Studies (CROSS).
[PDF File, 75 KB - mededu_v11i1e70766_app4.pdf]

References

1. Mohammed K, Nolan MB, Rajjo T, et al. Creating a patient-centered health care delivery system: a systematic review of health care quality from the patient perspective. Am J Med Qual 2016;31(1):12-21. [doi: 10.1177/1062860614545124] [Medline: 25082873]
2. Gelis A, Cervello S, Rey R, et al. Peer role-play for training communication skills in medical students: a systematic review. Simul Healthc 2020 Apr;15(2):106-111. [doi: 10.1097/SIH.0000000000000412] [Medline: 32168292]
3. D'Agostino TA, Atkinson TM, Latella LE, et al. Promoting patient participation in healthcare interactions through communication skills training: a systematic review. Patient Educ Couns 2017 Jul;100(7):1247-1257. [doi: 10.1016/j.pec.2017.02.016] [Medline: 28238421]
4. Safranek CW, Sidamon-Eristoff AE, Gilson A, Chartash D. The role of large language models in medical education: applications and implications. JMIR Med Educ 2023 Aug 14;9:e50945. [doi: 10.2196/50945] [Medline: 37578830]
5. Mavrych V, Ganguly P, Bolgova O. Using large language models (ChatGPT, Copilot, PaLM, Bard, and Gemini) in Gross Anatomy course: comparative analysis. Clin Anat 2025 Mar;38(2):200-210. [doi: 10.1002/ca.24244] [Medline: 39573871]
6. Mavrych V, Yaqinuddin A, Bolgova O. Claude, ChatGPT, Copilot, and Gemini performance versus students in different topics of neuroscience. Adv Physiol Educ 2025 Jun 1;49(2):430-437. [doi: 10.1152/advan.00093.2024] [Medline: 39824512]
7. Clay TJ, Da Custodia Steel ZJ, Jacobs C. Human-computer interaction: a literature review of artificial intelligence and communication in healthcare. Cureus 2024 Nov;16(11):e73763. [doi: 10.7759/cureus.73763] [Medline: 39677224]
8. Stamer T, Steinhäuser J, Flägel K. Artificial intelligence supporting the training of communication skills in the education of health care professions: scoping review. J Med Internet Res 2023 Jun 19;25:e43311. [doi: 10.2196/43311] [Medline: 37335593]
9. Gutiérrez Maquilón R, Uhl J, Schrom-Feiertag H, Tscheligi M. Integrating GPT-based AI into virtual patients to facilitate communication training among medical first responders: usability study of mixed reality simulation. JMIR Form Res 2024 Dec 11;8:e58623. [doi: 10.2196/58623] [Medline: 39661979]
10. Gray M, Baird A, Sawyer T, et al. Increasing realism and variety of virtual patient dialogues for prenatal counseling education through a novel application of ChatGPT: exploratory observational study. JMIR Med Educ 2024 Feb 1;10:e50705. [doi: 10.2196/50705] [Medline: 38300696]
11. Bowers P, Graydon K, Ryan T, Lau JH, Tomlin D. Artificial intelligence-driven virtual patients for communication skill development in healthcare students. Australas J Educ Technol 2024;40(3):39-57. [doi: 10.14742/ajet.9307]
12. Jackson P, Ponath Sukumaran G, Babu C, et al. Artificial intelligence in medical education - perception among medical students. BMC Med Educ 2024 Jul 27;24(1):804. [doi: 10.1186/s12909-024-05760-0] [Medline: 39068482]
13. Thuraisingham C, Abd Razak SS, Nadarajah VD, Mamat NH. Communication skills in primary care settings: aligning student and patient voices. Educ Prim Care 2023 May;34(3):123-130. [doi: 10.1080/14739879.2023.2210097] [Medline: 37194600]
14. Elendu C, Amaechi DC, Okatta AU, et al. The impact of simulation-based training in medical education: a review. Medicine (Baltimore) 2024 Jul 5;103(27):e38813. [doi: 10.1097/MD.0000000000038813] [Medline: 38968472]
15. Yamamoto A, Koda M, Ogawa H, et al. Enhancing medical interview skills through AI-simulated patient interactions: nonrandomized controlled trial. JMIR Med Educ 2024 Sep 23;10:e58753. [doi: 10.2196/58753] [Medline: 39312284]
16. Cook DA, Overgaard J, Pankratz VS, Del Fiol G, Aakre CA. Virtual patients using large language models: scalable, contextualized simulation of clinician-patient dialogue with feedback. J Med Internet Res 2025 Apr 4;27:e68486. [doi: 10.2196/68486] [Medline: 39854611]
17. Sarkar U, Bates DW. Using artificial intelligence to improve primary care for patients and clinicians. JAMA Intern Med 2024 Apr 1;184(4):343-344. [doi: 10.1001/jamainternmed.2023.7965] [Medline: 38345801]
18. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. JMIR Med Educ 2023 Oct 20;9:e48785. [doi: 10.2196/48785] [Medline: 37862079]
19. Kavadella A, Dias da Silva MA, Kaklamanos EG, Stamatopoulos V, Giannakopoulos K. Evaluation of ChatGPT's real-life implementation in undergraduate dental education: mixed methods study. JMIR Med Educ 2024 Jan 31;10(1):e51344. [doi: 10.2196/51344] [Medline: 38111256]
20. Jacobs C, Wheeler J, Williams M, Joiner R. Cognitive interviewing as a method to inform questionnaire design and validity - Immersive Technology Evaluation Measure (ITEM) for healthcare education. Comput Educ X Real 2023;2:100027. [doi: 10.1016/j.cexr.2023.100027]

21. Mukadam A, Suresh S, Jacobs C. Beyond traditional simulation: an exploratory study on the effectiveness and acceptability of ChatGPT-4o advanced voice mode for communication skills practice among medical students. Cureus 2025 May;17(5):e84381. [doi: 10.7759/cureus.84381] [Medline: 40535400]
22. Winkel AF, Yingling S, Jones AA, Nicholson J. Reflection as a learning tool in graduate medical education: a systematic review. J Grad Med Educ 2017 Aug;9(4):430-439. [doi: 10.4300/JGME-D-16-00500.1] [Medline: 28824754]
23. Calhoun AW, Hui J, Scerbo MW. Quantitative research in healthcare simulation: an introduction and discussion of common pitfalls. In: Nestel D, Hui J, Kunkler K, Scerbo MW, Calhoun AW, editors. Healthcare Simulation Research: A Practical Guide: Springer International Publishing; 2019:153-160. [doi: 10.1007/978-3-030-26837-4_21]
24. Jacobs C, M Rigby J. Developing measures of immersion and motivation for learning technologies in healthcare simulation: a pilot study. J Adv Med Educ Prof 2022 Jul;10(3):163-171. [doi: 10.30476/JAMP.2022.95226.1632] [Medline: 35910517]
25. Jacobs C, Foote G, Williams M. Evaluating user experience with immersive technology in simulation-based education: a modified Delphi study with qualitative analysis. PLoS One 2023;18(8):e0275766. [doi: 10.1371/journal.pone.0275766] [Medline: 37531361]
26. Ryan RM, Deci EL. Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness: The Guilford Press; 2017. [doi: 10.1521/978.14625/28806]
27. Schön DA. The Reflective Practitioner: How Professionals Think in Action: Routledge; 2017. [doi: 10.4324/9781315237473]
28. Boulesteix AL, Lauer S, Eugster MJA. A plea for neutral comparison studies in computational sciences. PLoS One 2013;8(4):e61562. [doi: 10.1371/journal.pone.0061562] [Medline: 23637855]
29. Conover WJ. Practical Nonparametric Statistics, 3rd edition: Wiley; 1999.
30. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
31. Braun V, Clarke V, Boulton E, Davey L, McEvoy C. The online survey as a qualitative research tool. Int J Soc Res Methodol 2020;24(6):641-654. [doi: 10.1080/13645579.2020.1805550]
32. Sridharan K, Sequeira RP. Artificial intelligence and medical education: application in classroom instruction and student assessment using a pharmacology & therapeutics case study. BMC Med Educ 2024 Apr 22;24(1):431. [doi: 10.1186/s12909-024-05365-7] [Medline: 38649959]
33. Arango-Ibanez JP, Posso-Nuñez JA, Díaz-Solórzano JP, Cruz-Suárez G. Evidence-based learning strategies in medicine using AI. JMIR Med Educ 2024 May 24;10:e54507. [doi: 10.2196/54507] [Medline: 38801706]
34. Chiu TKF. Applying the self-determination theory (SDT) to explain student engagement in online learning during the COVID-19 pandemic. J Res Technol Educ 2022 Jan 31;54(sup1):S14-S30. [doi: 10.1080/15391523.2021.1891998]
35. Pan X. Technology acceptance, technological self-efficacy, and attitude toward technology-based self-directed learning: learning motivation as a mediator. Front Psychol 2020;11:564294. [doi: 10.3389/fpsyg.2020.564294] [Medline: 33192838]
36. Deci EL, Ryan RM. The “what” and “why” of goal pursuits: human needs and the self-determination of behavior. Psychol Inq 2000 Oct;11(4):227-268. [doi: 10.1207/S15327965PLI1104_01]
37. Jacobs C, Maidwell-Smith A. Learning from 360-degree film in healthcare simulation: a mixed methods pilot. J Vis Commun Med 2022 Oct;45(4):223-233. [doi: 10.1080/17453054.2022.2097059] [Medline: 35938350]
38. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep;13(3):319-340. [doi: 10.2307/249008]
39. Goodyear P, Ellis RA. University students’ approaches to learning: rethinking the place of technology. Distance Educ 2008 Aug;29(2):141-152. [doi: 10.1080/01587910802154947]
40. Scalese RJ, Obeso VT, Issenberg SB. Simulation technology for skills training and competency assessment in medical education. J Gen Intern Med 2008 Jan;23 Suppl 1(Suppl 1):46-49. [doi: 10.1007/s11606-007-0283-4] [Medline: 18095044]
41. Massoth C, Röder H, Ohlenburg H, et al. High-fidelity is not superior to low-fidelity simulation but leads to overconfidence in medical students. BMC Med Educ 2019 Jan 21;19(1):29. [doi: 10.1186/s12909-019-1464-7] [Medline: 30665397]
42. Mir MM, Mir GM, Raina NT, et al. Application of artificial intelligence in medical education: current scenario and future perspectives. J Adv Med Educ Prof 2023 Jul;11(3):133-140. [doi: 10.30476/JAMP.2023.98655.1803] [Medline: 37469385]
43. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 1;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
44. Bearman M, Palermo C, Allen LM, Williams B. Learning empathy through simulation: a systematic literature review. Simul Healthc 2015 Oct;10(5):308-319. [doi: 10.1097/SIH.0000000000000113] [Medline: 26426561]
45. Holderried F, Stegemann-Philipps C, Herschbach L, et al. A generative pretrained transformer (GPT)-powered chatbot as a simulated patient to practice history taking: prospective, mixed methods study. JMIR Med Educ 2024 Jan 16;10:e53961. [doi: 10.2196/53961] [Medline: 38227363]
46. Vaughn J, Ford SH, Scott M, Jones C, Lewinski A. Enhancing healthcare education: leveraging ChatGPT for innovative simulation scenarios. Clin Simul Nurs 2024 Feb;87:101487. [doi: 10.1016/j.ecns.2023.101487]
47. Rodgers DL, Needler M, Robinson A, et al. Artificial intelligence and the simulationists. Simul Healthc 2023 Dec 1;18(6):395-399. [doi: 10.1097/SIH.0000000000000747] [Medline: 37747487]

48. Alam F, Lim MA, Zulkipli IN. Integrating AI in medical education: embracing ethical usage and critical understanding. Front Med (Lausanne) 2023;10:1279707. [doi: 10.3389/fmed.2023.1279707] [Medline: 37901398]
49. Varkey TC, Varkey JA, Ding JB, et al. Asynchronous learning: a general review of best practices for the 21st century. J Res Innov Teach Learn 2023 Mar 30;16(1):4-16. [doi: 10.1108/JRIT-06-2022-0036]
50. Eg R, Behne DM. Perceived synchrony for realistic and dynamic audiovisual events. Front Psychol 2015;6:736. [doi: 10.3389/fpsyg.2015.00736] [Medline: 26082738]

Abbreviations

AI: artificial intelligence
API: application programming interface
CROSS: Consensus-Based Checklist for Reporting of Survey Studies
GP: general practitioner
ITEM: Immersive Technology Evaluation Measure
LLM: large language model
MITHE: Model for Immersive Technology in Healthcare Education
PSA: prostate-specific antigen
VP: virtual patient

Edited by B Lesselroth; submitted 08.01.25; peer-reviewed by A Altozano, C Br, J Marín-Morales, S Mohanadas, S Ito, V Mavrych; revised version received 30.07.25; accepted 09.09.25; published 24.10.25.

Please cite as:
Jacobs C, Johnson H, Tan N, Brownlie K, Joiner R, Thompson T
Application of AI Communication Training Tools in Medical Undergraduate Education: Mixed Methods Feasibility Study Within a Primary Care Context
JMIR Med Educ 2025;11:e70766
URL: https://mededu.jmir.org/2025/1/e70766
doi:10.2196/70766

Large Language Models for the National Radiological Technologist Licensure Examination in Japan: Cross-Sectional Comparative Benchmarking and Evaluation of Model-Generated Items Study

Toshimune Ito[1,2,3], PhD; Toru Ishibashi[1], PhD; Tatsuya Hayashi[1,2], PhD; Shinya Kojima[1,2,4], PhD; Kazumi Sogabe[1,5], PhD

[1]Department of Radiological Technology, Faculty of Medical Technology, Teikyo University, 2-11-1 Kaga, Itabashi-ku, Tokyo, Japan

[2]Department of Medical Radiology, Graduate School of Medical Technology, Teikyo University, Tokyo, Japan

[3]Department of Medical Radiological Technology, Faculty of Health Sciences, Kyorin University, Tokyo, Japan

[4]Department of Radiology, Tokyo Women's Medical University Adachi Medical Center, Tokyo, Japan

[5]Department of Radiological Sciences, School of Health Sciences, Ibaraki Prefectural University of Health Sciences, Ibaraki, Japan

Corresponding Author:
Toshimune Ito, PhD
Department of Radiological Technology, Faculty of Medical Technology, Teikyo University, 2-11-1 Kaga, Itabashi-ku, Tokyo, Japan

Abstract

Background: Mock examinations are widely used in health professional education to assess learning and prepare candidates for national licensure. However, instructor-written multiple-choice items can vary in difficulty, coverage, and clarity. Recently, large language models (LLMs) have achieved high accuracy in medical examinations, highlighting their potential for assisting item-bank development; however, their educational quality remains insufficiently characterized.

Objective: This study aimed to (1) identify the most accurate LLM for the Japanese National Examination for Radiological Technologists and (2) use the top model to generate blueprint-aligned multiple-choice questions and evaluate their educational quality.

Methods: Four LLMs—OpenAI o3, o4-mini, o4-mini-high (OpenAI), and Gemini 2.5 Flash (Google)—were evaluated on all 200 items of the 77th Japanese National Examination for Radiological Technologists in 2025. Accuracy was analyzed for overall items and for 173 nonimage items. The best-performing model (o3) then generated 192 original items across 14 subjects by matching the official blueprint (image-based items were excluded). Subject-matter experts (≥5 y as coordinators and routine mock examination authors) independently rated each generated item on five criteria using a 5-point scale (1=unacceptable, 5=adoptable): item difficulty, factual accuracy, accuracy of content coverage, appropriateness of wording, and instructional usefulness. Cochran Q with Bonferroni-adjusted McNemar tests compared model accuracies, and one-sided Wilcoxon signed-rank tests assessed whether the median ratings exceeded 4.

Results: OpenAI o3 achieved the highest accuracy overall (90.0%; 95% CI 85.1% - 93.4%) and on nonimage items (92.5%; 95% CI 87.6% - 95.6%), significantly outperforming o4-mini on the full set (P=.02). Across models, accuracy differences on the non-image subset were not significant (Cochran Q, P=.10). Using o3, the 192 generated items received high expert ratings for item difficulty (mean, 4.29; 95% CI 4.11 - 4.46), factual accuracy (4.18; 95% CI 3.98 - 4.38), and content coverage (4.73; 95% CI 4.60 - 4.86). Ratings were comparatively lower for appropriateness of wording (3.92; 95% CI 3.73 - 4.11) and instructional usefulness (3.60; 95% CI 3.41 - 3.80). For these two criteria, the tests did not support a median rating >4 (one-sided Wilcoxon, P=.45 and P≥.99, respectively). Representative low-rated examples (ratings 1 - 2) and the rationale for those scores—such as ambiguous phrasing or generic explanations without linkage to stem cues—are provided in the supplementary materials.

Conclusions: OpenAI o3 can generate radiological licensure items that align with national standards in terms of difficulty, factual correctness, and blueprint coverage. However, wording clarity and the pedagogical specificity of explanations were weaker and did not meet an adoptable threshold without further editorial refinement. These findings support a practical workflow in which LLMs draft syllabus-aligned items at scale, while faculty perform targeted edits to ensure clarity and formative feedback. Future studies should evaluate image-inclusive generation, use Application Programming Interface (API)-pinned model snapshots to increase reproducibility, and develop guidance to improve explanation quality for learner remediation.

(JMIR Med Educ 2025;11:e81807) doi:10.2196/81807

KEYWORDS

large language models; licensing exam; radiology, educational evaluation; medical education; item generation

Introduction

Mock examinations are a key pedagogical tool in training programs for health professionals. These are designed to consolidate the knowledge required for national licensure and to gauge students' achievement [1-3]. In particular, multiple-choice formats are valuable because they enable the systematic, efficient appraisal of the broad foundational knowledge expected in clinical practice, making them integral to the quality of the curriculum. However, most items are written by individual instructors that draw on past examinations or personal clinical experience, and their difficulty and content validity are rarely subjected to systematic review [4,5]. These can result in biases in content coverage, inconsistencies in wording, and variable educational usefulness, which undermine the stability of learning outcome assessments.

Several studies have reported the high accuracy of large language models (LLMs) in health professional licensure examinations, owing to their rapid advancements [6-9]. In text-based multiple-choice questions, models have begun to match or surpass human test-takers while generating rationales and keyword-level explanations that can serve as formative feedback [10-13]. These suggest the potential utility of LLM-assisted item writing during the construction of high-quality question banks. However, most research has centered on the accuracy of LLMs in answering existing licensure items [14-16], while empirical evidence regarding the educational quality of questions authored by LLMs remains scarce [13,17]. A comprehensive appraisal that includes (1) appropriate difficulty, (2) completeness and accuracy of content coverage, (3) clarity of option wording, and (4) usefulness of accompanying explanations is necessary to address this knowledge gap and clarify the practical value of artificial intelligence (AI)-supported mock examinations, as well as its limitations.

This study evaluated the quality of AI-generated multiple-choice questions based on the Japanese National Examination for Radiological Technologists. Several LLMs were used to answer the exam, then the highest-performing model was used to generate a set of mock items. These AI-generated questions were then evaluated across several aspects (ie, item-level difficulty, item-level factual accuracy, accuracy of content coverage, appropriateness of wording, and instructional usefulness) through blinded expert review and statistical analysis. By doing so, this study aims to provide empirical data on the educational soundness of AI-generated items, as well as highlight any emerging challenges.

Methods

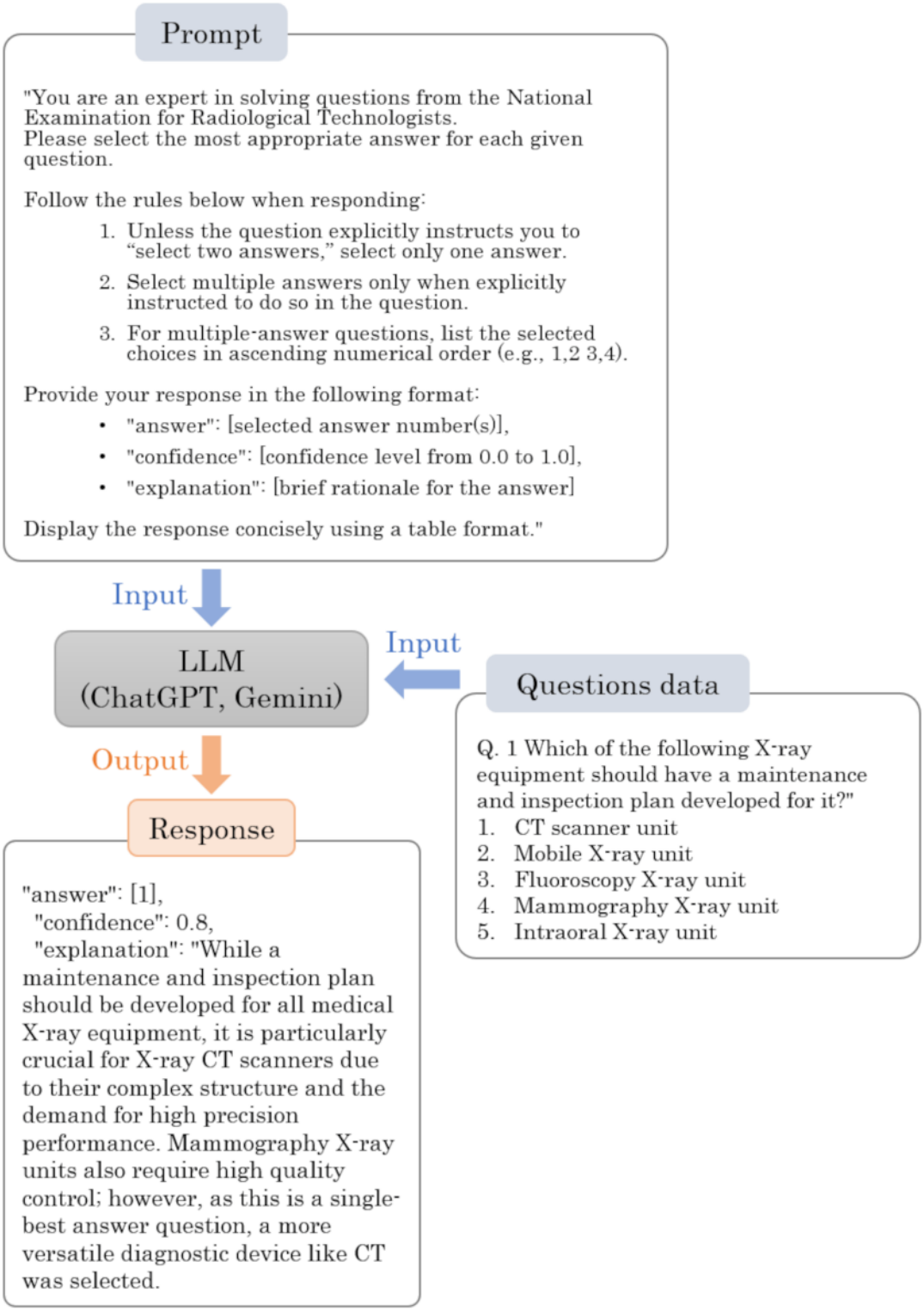
Models and Study Period

Four LLMs released in February 2025 were evaluated: OpenAI o3, OpenAI o4-mini, OpenAI o4-mini-high (all OpenAI), and Gemini 2.5 Flash (Google). The evaluations were conducted from March 14 to May 8, 2025, using the publicly accessible browser interfaces, with the desired engine explicitly selected in each platform's menu. The browser access was chosen to mirror typical educational use and to simplify image I/O (upload, preview, and per-item attachment). The item-generation study was conducted from May 15 to June 28, 2025, using OpenAI o3, the model with the highest answer accuracy. To ensure consistency, we used an identical Japanese prompt template across models. To avoid carryover effects, we started a new session for each 50-item batch with the OpenAI models and used per-item input with Gemini; image files (PNG) were attached when required by an item. As browsing and memory features were disabled, outputs relied solely on pretrained parameters and the provided materials.

Answer Accuracy

Answer accuracy was assessed based on all 200 items of the 77th National Examination for Radiological Technologists, administered on February 20, 2025. All items were multiple-choice, and question stems containing images were presented unchanged. Each model was given the question stem and options in Japanese, then instructed to select the correct answers in single-best or multiple-select format. Multimedia Appendix 1 lists the subjects and the number of items per subject. Due to the differences in each model, the input procedures were adapted accordingly. For OpenAI models, stems and options were pasted from four text files (items 1 - 50, 51 - 100, 101 - 150, and 151 - 200) into separate sessions. PNG files were attached for each image item, with the filenames labeled to match the corresponding item numbers. However, since Gemini permits only one file upload, the stems and options were pasted directly into the prompt while attaching an image file as needed. All inputs were entered manually. A concrete workflow is shown in Figure 1.

Figure 1. Representative interaction with a large language model (LLM). This diagram illustrates the workflow used to evaluate the answer accuracy of large language models. The LLMs were given prompts to answer each question (including text and images when applicable) in Japanese, with specific instructions for answer selection and formatting. The output included the selected answer, a confidence score, and a brief explanation. All actual prompts and inputs were entered in Japanese, but this example is shown in English for illustration purposes. CT: computed tomography.



The outputs of the model were compared to the official answer key issued by the Ministry of Health, Labor and Welfare. The correct and incorrect responses were counted overall for 200 items and separately for the 173 items that did not require image interpretation (ie, nonimage items). Statistical significance was tested across models.

Item Generation

Generation Procedure

The mock items were generated using OpenAI o3, since it had the highest accuracy among all four models. Image-based stems were excluded since all models performed poorly on these. Using the same examination as a blueprint, OpenAI o3 was

used to produce 192 questions across 14 subjects (Table 1), matching the same distribution of items. The model was supplied with text files containing the past 5 years of examination items and the official test specifications, ensuring its alignment with test objectives. Browsing remained disabled. Since Healthcare Safety Management is a new domain introduced in 2025, thereby lacking any historical reference items, it was excluded from the mock item generation. Items were generated separately for each subject in Japanese, and each output included the stem, five options, the key, and a brief rationale.

Table . Distribution of artificial intelligence (AI)-Generated Mock Items.

Subject	Blueprint target (n=200)	Generated (n=192)
Diagnostic Imaging Techniques	20	20
Nuclear Medicine Technology	20	20
Radiation Therapy Technology	20	20
Medical Imaging Informatics	10	10
Basic Medical Sciences	30	30
Radiation Science & Engineering	36	36
X-ray Imaging Equipment	20	20
X-ray Imaging Techniques	20	20
Image Engineering	6	6
Radiation Safety Management	10	10
Healthcare Safety Management[a]	8	0

[a]Since Healthcare Safety Management was only recently introduced as a new subject in the 2025 blueprint, it was excluded from the mock item generation.

Evaluation of Generated Items

All 192 generated questions were reviewed by experts of the subject matter; these were faculty members with at least 5 years of experience as subject coordinators in radiological technology programs and who routinely author mock examinations. Items were assigned to reviewers by discipline, and each question was evaluated by one expert. The reviewers rated each item on a five-point scale: (1) unacceptable; (2) major revision needed; (3) revisable; (4) minor revision; and (5) adoptable across five criteria including, item difficulty, factual accuracy, accuracy of content coverage, appropriateness of wording, and instructional usefulness.

For each criterion, we calculated the median score and tested the statistical significance of the proportion of high ratings (≥4). The evaluation framework, which is based on faculty experience with national examination item writing, is presented in Table 2.

Table . Evaluation of generated items.

Evaluation criterion	Rating scale[a]
Item difficulty	1 - 5
Factual accuracy	1 - 5
Accuracy of content coverage	1 - 5
Appropriateness of wording	1 - 5
Instructional usefulness	1 - 5

[a]Rating scale definition: 1=Unacceptable; 2=Major revision needed; 3=Revisable; 4=Minor revision; 5=Adoptable.

Statistical Analysis

Statistical analysis was performed using JMP (version 18; JMP Statistical Discovery LLC). Cochran Q test was initially used to examine overall differences in answer accuracy; when significant, pairwise differences were probed with McNemar test using Bonferroni correction. The item generation study used a one-sided Wilcoxon signed-rank test (H : median ≤4). Statistical significance was set at $P<.05$ for all analyses.

Ethical Considerations

This study did not involve human participants or patient-identifiable data. The Ethics Committee of Teikyo University reviewed the project and determined that formal ethical approval was not required because the work evaluated the quality of test items and did not constitute human medical research. Accordingly, informed consent was not applicable.

Results

Answer Accuracy

The accuracy of the LLMs on the full 200-item set and the nonimage 173-item set is shown in Table 3. All models consistently scored lower in the full set versus the nonimage set, with OpenAI o3 achieving the best results at 90% and 92.5%, respectively. A significant difference was seen between OpenAI o3 and OpenAI o4-mini on the full set, whereas no significant differences were seen among models on the nonimage set.

Table . Model accuracies and statistical comparisons on 200 benchmark questions and 173 nonimage questions.

Variables	200 questions[a]	173 nonimage questions[a]
Model accuracy		
OpenAI-o4-mini-high, %	86.0 (80.5, 90.1)	88.4 (82.8, 92.4)
OpenAI-o4-mini, %	82.5 (76.6, 87.1)	86.7 (80.8, 91.0)
OpenAI-o3, %	90.0 (85.1, 93.4)	92.5 (87.6, 95.6)
Gemini 2.5 Flash, %	83.0 (77.2, 87.6)	89.6 (84.1, 93.3)
Cochran Q test (*P* value)	.01	.10
Pairwise McNemar test (Bonferroni-adjusted *P* value)		
OpenAI-o4-mini-high versus OpenAI-o4-mini	≥.99	N/A[b]
OpenAI-o4-mini-high versus OpenAI-o3	.44	N/A
OpenAI-o4-mini-high versus Gemini 2.5 Flash	≥.99	N/A
OpenAI-o4-mini versus OpenAI-o3	.02	N/A
OpenAI-o4-mini versus Gemini 2.5 Flash	≥.99	N/A
OpenAI-o3 versus Gemini 2.5 Flash	.06	N/A

[a] Accuracy shown with 95% CIs in parentheses (Wilson score, two-sided, without continuity correction).

[b]Not applicable.

Item Generation

Table 4 presents the scores and statistics for all 192 questions, while Figure 2 illustrates the prompt template and sample outputs. Among item difficulty, factual accuracy, and accuracy of content coverage, the medians and the proportions of scores ≥4 did not differ significantly, although accuracy of content coverage had the highest score. Meanwhile, instructional usefulness had a significantly lower score than appropriateness of wording. The evaluation criteria and evaluation examples of items that scored 1‐2 for the lower-scoring criteria—appropriateness of wording and instructional usefulness—are detailed in Multimedia Appendix 2.

Table . Reviewer ratings by evaluation criterion for the AI-generated items (n=192).

Evaluation criterion[a]	Mean score (95% CI)	*P* value[b]
Item difficulty	4.29 (4.11, 4.46)	<.001
Factual accuracy	4.18 (3.98, 4.38)	.001
Accuracy of content coverage	4.73 (4.60, 4.86)	<.001
Appropriateness of wording	3.92 (3.73, 4.11)	.44
Instructional usefulness	3.60 (3.41, 3.80)	≥.99

[a] “Evaluation criterion” refers to the five evaluation criteria defined in Table 2.

[b]One-sided Wilcoxon signed-rank test against the null hypothesis such that the median score is ≤4.

Figure 2. Prompt summary and representative example of item generation. (A) Summary of the prompts used to instruct the language model to generate original mock questions aligned with the National Examination for Radiological Technologists. The summary outlines the role of the model, input references, specifications of generation, item-creation rules, and output format. (B) The actual prompt and representative response generated by the model. The prompt included specific formatting and content-generation instructions written in Japanese. The response shows the generated item, correct answers, and explanation in Japanese.

A

Role & General Instruction:
Generate practice questions for the National Examination for Radiological Technologists that reflect the official exam's tone, difficulty, and content focus.
Reference materials:
Use items 1–144 from the past five years of the National Examination (72nd to 76th, 2020–2024) as reference material to guide question generation in terms of content and format.
Specifications for Item Generation:
Create the specified number of questions for the designated subject, ensuring coverage of the subtopics and keywords provided in the reference materials.
Instructions for Item Creation:
Generate original five-choice questions that follow the style and level of past exams. Use both single-answer and two-answer formats, marking correct answers and including brief explanations for each.
Output Format :
Present each item in a standardized format with a numbered question, five answer choices, clearly indicated correct answer(s), and a brief explanation for all options.

B

Prompt (Original Japanese)

「あなたは診療放射線技師国家試験の問題作成の専門家です。以下の参考資料を基に、国家試験本番と同等の文体・難易度・出題傾向を持つ練習問題を作成してください。

参考資料：過去5年分（第72回(2020年)～第76回(2024)の診療放射線技師国家試験問題集（問1から問144）

問1　性ホルモンが腫瘍の増殖に関わるのはどれか。2つ選べ。
1. 腟癌
2. 陰茎癌
3. 尿道癌
4. 子宮体癌
5. 前立腺癌
：
：
：
問144　受精が起こるのはどこか。
1. 腟
2. 卵管
3. 卵巣
4. 子宮頸部
5. 子宮内膜

出題内容の指定
- 科目：**[放射線物理学]**
- 問題数：**[10問]**
- 範囲・キーワード：**[出題範囲(PDF_2_放射線物理学)]**

科目 の小項目のみを抽出し表示する。

出題にあたっての条件
- 国試本番と同等の口調・専門性・難易度を再現すること。過去問の典型的な表現（「～はどれか。」など）や形式にならう。
- 問題は**完全新規**に作成すること（参考資料と**同一の問題文にならない**ように留意すること）。過去問をヒントにテーマや表現を変えて出題してよいが解答そのものは新しく考案する。
- 出題形式：**5肢択一形式**（正答1つ）と**5肢複択形式**（正答2つ）を組み合わせること。全体の約**70%**を正答1つの問題、約**30%**を正答2つの問題とする。
- **選択肢**は問題ごとに**1～5の5つ**を用意すること（内容は出題範囲に必ず沿ったもの）。
- 正答が2つある問題では、問題文の末尾に「**2つ選べ。**」と明記し、受験者が2つ選ぶ形式であることを示す。
- **解答表示:** 各問題の選択肢の直後に「正答：○」と正解選択肢を記号で示す（正解が複数ある場合は「正答：2と4」のように列記）。
- **解説:** 各問題に対して、正解の根拠および各選択肢の解説を添えること。他の選択肢が誤りである理由や関連知識も含め、受験者の理解が深まる説明を書く。

出力フォーマット: 以下の形式で問題文、選択肢、正答、解説を記載すること（各問題を通し番号で番号付けする）。

問題1．～（問題文）～はどれか。 ※複数正解の場合、文末に「2つ選べ。」
1. 選択肢1
2. 選択肢2
3. 選択肢3
4. 選択肢4
5. 選択肢5

正答：2 ※複数正解の場合「正答：2と4」

解説：問題文や選択肢に関する解説。正解が正しい理由と他の選択肢が誤りである理由をそれぞれ説明する。」

Response (Generated in Japanese)

練習問題（全10問）
問題1.
水中において 1.25 MeV の光子が物質と相互作用する際、最も卓越する現象はどれか。
光電吸収
コンプトン散乱
電子対生成
レイリー散乱
光核反応
正答: 2
解説: 光電吸収は E^(−3)、電子対生成は E^(+) に支配されるが、低原子番号物質で 1 MeV 付近ではコンプトン散乱が最大となる。レイリー散乱は無電離散乱で寄与が小さく、光核反応は高エネルギー域（数10 MeV 以上）で顕在化する。

問題2.
線エネルギー付与 (LET) が最大となるのはどれか。
5 MeV α粒子
25 MeV 陽子
6 MeV 電子
1 MeV γ線
120 keV X 線
正答: 1
解説: LET は電荷の 2 乗に比例し速度の 2 乗に反比例する。重くて遅いα粒子 (He^(2+)) は密度エネルギー付与が大きく、他の放射線より高LET となる。

問題3.
光電吸収係数を増大させる要因はどれか。2 つ選べ。
低原子番号物質
低エネルギー光子
材料密度の低下
高原子番号物質
高エネルギー光子
正答: 2 と 4
解説: 光電吸収係数は概ね $\propto Z^3 \cdot 4E^{-3}$ に比例する。よって低エネルギー(2)・高 Z (4) が係数を大きくする。密度は質量減弱係数を変えず、(1)(3)(5)はむしろ係数を小さくする。

問題4.
X 線管で発生する制動X 線の最短波長を 1/2 に短縮するには管電圧 (kV) をどのように変化させればよいか。
2 倍にする
√2 倍にする
1/2 にする
1/√2 にする
変化させない
正答: 1
解説: 最短波長 $\lambda_{min} = hc/eV$ であり $\lambda_{min} \propto 1/V$。半減させるには電圧を 2 倍にする必要がある。

問題5.
熱中性子の速度に影響する物理量はどれか。2 つ選べ。
温度
中性子の質量
原子核の原子番号
中性子のスピン
電離密度
正答: 1 と 2
解説: Maxwell-Boltzmann 分布より $v_{rms} = \sqrt{3kT/m}$。速度は温度 (1) と質量 (2) に依存し、核種の Z やスピン、媒質の電離密度とは直接関係しない。

Discussion

Principal Findings

This study compared four LLMs in terms of answer accuracy on the Japanese National Examination for Radiological Technologists. The top performer, OpenAI o3, was used to generate the mock test, which was then evaluated by experts in terms of educational quality. As shown in Table 3, on the full set of items, only the comparison between the OpenAI o3 and OpenAI-o4-mini variant reached statistical significance; when image-based items were excluded, no model differences were observed.

To contextualize the observed accuracy differences, we briefly summarize the multimodal architectures and vision–language pipelines of the evaluated models as they pertain to radiologic image questions. Built on a GPT-4 lineage, OpenAI o3 integrates a high-resolution visual encoder with unified attention over linguistic and visual tokens [18,19], likely enhancing sensitivity to low-contrast findings and subtle anatomical cues typical of radiography and CT. In contrast, OpenAI o4-mini is a lightweight variant with reduced-resolution patch embeddings [20,21], which can yield coarser visual representations and miss subtle image cues. OpenAI o4-mini-high supplements the mini architecture with targeted medical-image fine-tuning and partial recovery of high-resolution inputs [22,23], consistent with improved mapping of relevant visual patterns. Lastly, Gemini 2.5 Flash uses a two-tower design in which an external vision encoder converts images to tags prior to language processing [24,25]; such pipelines may incur information loss for domain-specific anatomical details. In line with these architectural differences, performance gaps emerged on image-based questions but not on text-only items.

The pronounced performance spread on image-based questions could be mainly attributed to the aggressive parameter reduction in OpenAI o4-mini and the information loss inherent in the image-to-tag pipeline in Gemini, both of which weaken visual feature representation. Thus, current systems may not fully capture clinically grounded context and the knowledge required for radiologic image interpretation. This finding is consistent with the results of previous studies reporting similar limitations in specialty radiology examinations [26,27]. However, OpenAI o3 and o4-mini-high have higher resolution encoders and benefit from medical-specific fine-tuning. However, due to the limited sample sizes and proprietary nature of the detailed model architectures, these explanations remain partly hypothetical. Nevertheless, these findings highlight the importance of the visual module scale and the presence of medical-domain training when selecting an LLM for the development of -generated questions in this field.

Building on these findings, the 192 items generated by the top model were reviewed across five educational criteria. Item difficulty, factual accuracy, and content coverage were rated favorably, indicating alignment with national expectations and the official blueprint [26]. By contrast, appropriateness of wording and instructional usefulness were comparatively weaker, with reviewers noting ambiguous phrasing and explanations that did not consistently link stem cues to the correct answer or to distractor misconceptions. These strengths and weaknesses are consistent with observations from related medical-education settings [28-30] and underscore the need for editorial refinement prior to instructional deployment.

This study has several limitations. First, the image-based items were excluded from expert review, thus precluding the assessment of visual tasks. Second, each question was evaluated by a single expert, and thus inter-rater reliability could not be assessed. Third, reproducibility is limited by the use of publicly accessible browser interfaces. All evaluations were conducted through browser UIs with visible labels: OpenAI o3, OpenAI o4-mini, OpenAI o4-mini-high, and Gemini 2.5 Flash. Although this choice mirrors typical educational use and simplifies image I/O, it limits control over versioning and decoding parameters. Prompt delivery also varied across platforms due to UI constraints: OpenAI models received items in 50-question batches per session, whereas Gemini required per-item input, with a single image upload when applicable. Such differences in prompt granularity, context priming, and file-attachment workflows may have influenced outputs and should be considered when interpreting the comparable performance of Gemini Flash and o3. To mitigate these effects, we used an identical Japanese prompt template, disabled memory features, initiated new sessions for each batch, preserved the original exam order, and performed a single pass per item without retries. Input handling is detailed in the Methods section. These input structures reflected platform UI constraints (OpenAI allowed 50-question batches per session, whereas Gemini required per-item prompts and a single image attachment when applicable); although memory features were disabled and each batch began in a new session, processing the OpenAI items in batches could still introduce minor within-session priming; therefore, residual order effects cannot be fully excluded. Application-level temperature settings were not user-configurable. Moreover, because decoding remained stochastic and we performed a single pass per item without retries, run-to-run response variability cannot be fully excluded even with identical prompts. Given that browser-based services can update without notice, outputs may drift over time even when identical prompts and labels are used [31-33]. Thus, to strengthen version control and reproducibility, future studies should standardize prompt injection through Application Programming Interface endpoints with pinned model snapshots, identical per-item wrappers, and fully logged metadata (prompt templates, model identifiers, timestamps, and decoding parameters). In the future, visual encoders are expected to operate at a higher resolution and undergo additional tuning for medical domains. This could enable LLMs to automatically generate image-based items across modalities (eg, computed tomography, magnetic resonance imaging, and ultrasound), thus bringing mock exams closer to clinical reality. Further improvements in the feedback system could also be seen. By delivering adaptive feedback that varies in depth according to each learner's proficiency, students can be provided with on-demand, targeted remediation material. LLMs could also be used to map items to the national blueprint in real time, enabling the detection and correction of domain imbalances while reducing faculty workload. Lastly, aligning these models with overseas licensure frameworks could expand their use to

ultimately support a multilingual, multi-profession, international mock-exam bank.

Conclusions

This study demonstrated that an LLM (OpenAI o3) can attain high accuracy on national radiological technology examination, as well as generate new multiple-choice items with appropriate difficulty, factual correctness, and syllabus coverage, as evaluated by experts. Although the AI-generated questions fell short in terms of wording clarity and pedagogical feedback, these can be mitigated through targeted editorial review. Practically speaking, LLMs can be used to draft content that is eventually refined by the faculty. This workflow could enable the more efficient development of mock examinations and reinforce curriculum alignment without imposing additional burden on instructors. However, performance gaps on image-based items, the absence of inter-rater reliability data, and the inherent volatility of cloud-hosted models underscore the need for cautious implementation and transparent reporting of model metadata. Nevertheless, future advancements in high-resolution visual encoders and medical-specific tuning can close this multimodal gap, while adaptive feedback functions and automated blueprint mapping can further extend the educational value of AI-generated assessments. After overcoming these barriers in terms of technical improvements and reproducibility safeguards, LLMs can be a strong asset in radiological technology education, which can even extend to the licensure preparations of other allied health professionals worldwide.

Acknowledgments

The authors thank Hiroki Ohtani, Hiroki Saito, Tatsuru Ota, Kiyoshi Hishiki, and Masao Fujihara of Teikyo University for their careful evaluation of the problem statements and for the constructive feedback that strengthened this study. Disclosure of generative AI use (language editing only). We used OpenAI o3 solely to assist with language editing (readability, clarity, and minor stylistic consistency). No AI tools were used to generate scientific content, analyze or interpret data, or determine conclusions. All statements and references were verified by the authors, who take full responsibility for the final manuscript. Editing with OpenAI o3 was performed interactively; the manuscript wording was subsequently finalized by Enago Co., Ltd.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The article processing charge for open access publication was supported by Teikyo University's Open Access Publication Support Program. The funder had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during this study.

Authors' Contributions

Conceptualization: T Ito, KS
Data Curation: T Ishibashi
Software: SK
Formal Analysis: TH
Writing – Original Draft: T Ito

Conflicts of Interest

None declared.

Multimedia Appendix 1
Breakdown of the 2025 Japanese National Exam Questions by Subject.
[DOCX File, 16 KB - mededu_v11i1e81807_app1.docx]

Multimedia Appendix 2
Operational definitions and decision rules for item evaluation.
[DOCX File, 18 KB - mededu_v11i1e81807_app2.docx]

References

1. Al-Sheikh MH, Albaker W, Ayub MZ. Do mock medical licensure exams improve performance of graduates? Experience from a Saudi Medical College. Saudi J Med Sci 2022;10(2):157-161. [doi: 10.4103/sjmms.sjmms_173_21]
2. Scott NP, Martin TW, Schmidt AM, Shanks AL. Impact of an online question bank on Resident In-Training exam performance. J Med Educ Curric Dev 2023;10:23821205231206221. [doi: 10.1177/23821205231206221] [Medline: 37822782]

3. Siab F, Morrissey H, Ball P. Pharmacy students' opinions of using mock questions to prepare for summative examinations. Int J Curr Pharm Sci 2020 Jul;12(4):58-65. [doi: 10.22159/ijcpr.2020v12i4.39079]
4. Alawgali SM. An evaluation of a final year multiple choice questions examination at Faculty of medicine-university of Benghazi. Open Access Maced J Med Sci 2024;12(80):1-11. [doi: 10.37376/jsh.vi80.6626]
5. Karthikeyan S, O'Connor E, Hu W. Barriers and facilitators to writing quality items for medical school assessments – a scoping review. BMC Med Educ 2019 Dec;19(1):123. [doi: 10.1186/s12909-019-1544-8]
6. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
7. Tanaka Y, Nakata T, Aiga K, et al. Performance of generative pretrained transformer on the National Medical Licensing Examination in Japan. PLOS Digit Health 2024 Jan;3(1):e0000433. [doi: 10.1371/journal.pdig.0000433] [Medline: 38261580]
8. Saowaprut P, Wabina RS, Yang J, Siriwat L. Performance of large language models on Thailand's National Medical Licensing Examination: a cross-sectional study. J Educ Eval Health Prof 2025;22:16. [doi: 10.3352/jeehp.2025.22.16] [Medline: 40354784]
9. Zhu S, Hu W, Yang Z, Yan J, Zhang F. Qwen-2.5 outperforms other large language models in the Chinese National Nursing Licensing Examination: retrospective cross-sectional comparative study. JMIR Med Inform 2025 Jan 10;13:e63731. [doi: 10.2196/63731] [Medline: 39793017]
10. Tomova M, Roselló Atanet I, Sehy V, Sieg M, März M, Mäder P. Leveraging large language models to construct feedback from medical multiple-choice questions. Sci Rep 2024 Nov 13;14(1):27910. [doi: 10.1038/s41598-024-79245-x] [Medline: 39537899]
11. Kondo T, Okamoto M, Kondo Y. Pilot study on using large language models for educational resource development in Japanese Radiological Technologist Exams. MedSciEduc 2025 Apr;35(2):919-927. [doi: 10.1007/s40670-024-02251-1]
12. Sabaner MC, Hashas ASK, Mutibayraktaroglu KM, Yozgat Z, Klefter ON, Subhi Y. The performance of artificial intelligence-based large language models on ophthalmology-related questions in Swedish proficiency test for medicine: ChatGPT-4 omni vs Gemini 1.5 Pro. AJO International 2024 Dec;1(4):100070. [doi: 10.1016/j.ajoint.2024.100070]
13. Mistry NP, Saeed H, Rafique S, Le T, Obaid H, Adams SJ. Large language models as tools to generate radiology board-style multiple-choice questions. Acad Radiol 2024 Sep;31(9):3872-3878. [doi: 10.1016/j.acra.2024.06.046]
14. Brin D, Sorin V, Konen E, Nadkarni G, Glicksberg BS, Klang E. How large language models perform on the united states medical licensing examination: a systematic review. medRxiv. Preprint posted online on 2023. [doi: 10.1101/2023.09.03.23294842]
15. Zong H, Wu R, Cha J, et al. Large language models in worldwide medical exams: platform development and comprehensive analysis. J Med Internet Res 2024 Dec 27;26:e66114. [doi: 10.2196/66114]
16. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. Sci Rep 2023 Nov 22;13(1):20512. [doi: 10.1038/s41598-023-46995-z] [Medline: 37993519]
17. Kim JK(, Chua M, Lorenzo A, et al. Use of AI (GPT-4)-generated multiple-choice questions for the examination of surgical subspecialty residents. CUAJ 2025 ;19(6):9020. [doi: 10.5489/cuaj.9020]
18. Zhang Y, Pan Y, Zhong T, et al. Potential of multimodal large language models for data mining of medical images and free-text reports. Meta-Radiology 2024 Dec;2(4):100103. [doi: 10.1016/j.metrad.2024.100103]
19. Soni N, Ora M, Agarwal A, Yang T, Bathla G. A review of the opportunities and challenges with large language models in radiology: the road ahead. AJNR Am J Neuroradiol 2025 Jul 1;46(7):ajnr. [doi: 10.3174/ajnr.A8589]
20. Alsabbagh AR, Mansour T, Al-Kharabsheh M, et al. MiniMedGPT: efficient large vision–language model for Medical Visual Question Answering. Pattern Recognit Lett 2025 Mar;189:8-16. [doi: 10.1016/j.patrec.2025.01.001]
21. Chen J, Shen X, Li X, Elhoseiny M. MiniGPT-4: enhancing vision-language understanding with advanced large language models. arXiv. Preprint posted online on Oct 2, 2023. [doi: 10.48550/arXiv.2304.10592]
22. Zhang P, Zang Y, et al. InternLM-xcomposer2-4KHD: a pioneering large vision-language model handling resolutions from 336 pixels to 4K HD. Adv Neural Inf Process Syst. Preprint posted online on Apr 9, 2024. [doi: 10.48550/arXiv.2404.06512]
23. Wang Z, Huang Y, Wu Y, et al. Fusion side tuning: a parameter and memory efficient fine-tuning method for high-resolution medical image classification. : IEEE Presented at: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Dec 3-6, 2024; Lisbon, Portugal. [doi: 10.1109/BIBM62325.2024.10821946] [Medline: 40989005]
24. Gemini Team Google, Georgiev P, Lei VI, et al. Gemini 1.5: unlocking multimodal understanding across millions of tokens of context. arXiv. Preprint posted online on Dec 16, 2024. [doi: 10.48550/arXiv.2403.05530]
25. Boostani M, Bánvölgyi A, Goldust M, et al. Diagnostic performance of GPT-4o and Gemini Flash 2.0 in acne and rosacea. Int J Dermatol 2025 Oct;64(10):1881-1882. [doi: 10.1111/ijd.17729] [Medline: 40064599]
26. Sarangi PK, Datta S, Panda BB, Panda S, Mondal H. Evaluating ChatGPT-4's performance in Identifying Radiological Anatomy in FRCR Part 1 Examination Questions. Indian J Radiol Imaging 2025 Apr;35(02):287-294. [doi: 10.1055/s-0044-1792040]

27. Sarangi PK, Narayan RK, Mohakud S, Vats A, Sahani D, Mondal H. Assessing the capability of ChatGPT, Google Bard, and Microsoft Bing in solving Radiology case vignettes. Indian J Radiol Imaging 2024 Apr;34(2):276-282. [doi: 10.1055/s-0043-1777746] [Medline: 38549897]
28. Morse K, Kumar A, et al. Assessing the potential of USMLE-like exam questions generated by GPT-4. medRxiv. Preprint posted online on Apr 28, 2023. [doi: 10.1101/2023.04.25.23288588]
29. Zhou Z, Rizwan A, Rogoza N, Chung AD, Kwan BY. Differentiating between GPT-generated and human-written feedback for radiology residents. Curr Probl Diagn Radiol 2025;54(5):574-578. [doi: 10.1067/j.cpradiol.2025.02.002] [Medline: 39984362]
30. Kuusemets L, Parve K, Ain K, Kraav T. Assessing AI-generated (GPT-4) versus human created MCQs In Mathematics education: a comparative inquiry into vector topics. IJEMST 2024;12(6):1538-1558. [doi: 10.46328/ijemst.4440]
31. Ma W, Yang C, Kästner C. (Why) is my prompt getting worse? Rethinking regression testing for evolving LLM APIs. Presented at: Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI; Apr 14, 2024; Lisbon Portugal p. 166-171. [doi: 10.1145/3644815.3644950]
32. Schroeder K, Wood-Doughty Z. Can you trust LLM judgments? Reliability of LLM-as-a-judge. arXiv. Preprint posted online on Feb 18, 2025. [doi: 10.48550/arXiv.2412.12509] [Medline: 38076521]
33. Renze M. The effect of sampling temperature on problem solving in large language models. In: Al-Onaizan Y, Bansal M, Chen YN, editors. Findings of the Association for Computational Linguistics: EMNLP 2024: Association for Computational Linguistics; 2024:7346-7356 URL: https://aclanthology.org/2024.findings-emnlp [accessed 2025-08-01] [doi: 10.18653/v1/2024.findings-emnlp.432]

Abbreviations

AI: artificial intelligence
LLM: large language model
UI: user interface

Edited by A Stone, T Leung; submitted 08.08.25; peer-reviewed by PK Sarangi, S Court-Kowalski, Y Fukui; accepted 28.10.25; published 13.11.25.

Please cite as:
Ito T, Ishibashi T, Hayashi T, Kojima S, Sogabe K
Large Language Models for the National Radiological Technologist Licensure Examination in Japan: Cross-Sectional Comparative Benchmarking and Evaluation of Model-Generated Items Study
JMIR Med Educ 2025;11:e81807
URL: https://mededu.jmir.org/2025/1/e81807
doi:10.2196/81807

Novel Blended Learning on Artificial Intelligence for Medical Students: Qualitative Interview Study

Zoe S Oftring[1,2*], MA, Dr med, MD; Kim Deutsch[3*], MA, Dr phil; Daniel Tolks[4,5], Dr rer biol hum; Florian Jungmann[6], PD, Dr med, MD; Sebastian Kuhn[1], Prof Dr, MME, MD

[1]Institute for Digital Medicine, Philipps University Marburg and University Clinic Giessen & Marburg, Baldingerstrasse 1, Marburg, Germany

[2]Department of Paediatrics, University Clinic Giessen & Marburg, Marburg, Germany

[3]Institute of Educational Science, Johannes Gutenberg University, Mainz, Germany

[4]Institute of Anatomy, Rostock University Medical Centre, Rostock, Germany

[5]Professorship in Health Management, International University of Applied Science, Hamburg, Germany

[6]Xcare Group Radiology, Nuclear Medicine and Radiotherapy, Saarlouis, Germany

[*]these authors contributed equally

Corresponding Author:
Sebastian Kuhn, Prof Dr, MME, MD
Institute for Digital Medicine, Philipps University Marburg and University Clinic Giessen & Marburg, Baldingerstrasse 1, Marburg, Germany

Abstract

Background: Artificial intelligence (AI) systems are becoming increasingly relevant in everyday clinical practice, with Food and Drug Administration–approved AI solutions now available in many specialties. This development has far-reaching implications for doctors and the future medical profession, highlighting the need for both practicing physicians and medical students to acquire the knowledge, skills, and attitudes necessary to effectively use and evaluate these technologies. Currently, however, there is limited experience with AI-focused curricular training and continuing education.

Objective: This paper first introduces a novel blended learning curriculum including one module on AI for medical students in Germany. Second, this paper presents findings from a qualitative postcourse evaluation of students' knowledge and attitudes toward AI and their overall perception of the course.

Methods: Clinical-year medical students can attend a 5-day elective course called "Medicine in the Digital Age," which includes one dedicated AI module alongside 4 others on digital doctor-patient communication; digital health applications and smart devices; telemedicine; and virtual/augmented reality and robotics. After course completion, participants were interviewed in semistructured small group interviews. The interview guide was developed deductively from existing evidence and research questions compiled by our group. A subset of interview questions focused on students' knowledge, skills, and attitudes regarding medical AI, and their overall course assessment. Responses were analyzed using Mayring's qualitative content analysis. This paper reports on the subset of students' statements about their perception and attitudes toward AI and the elective's general evaluation.

Results: We conducted a total of 18 group interviews, in which all 35 (100%) participants (female=11, male=24) from 3 consecutive course runs participated. This produced a total of 214 statements on AI, which were assigned to the 3 main categories "Areas of Application," "Future Work," and "Critical Reflection." The findings indicate that students have a nuanced and differentiated understanding of AI. Additionally, 610 statements concerned the elective's overall assessment, demonstrating great learning benefits and high levels of acceptance of the teaching concept. All 35 students would recommend the elective to peers.

Conclusions: The evaluation demonstrated that the AI module effectively generates competences regarding AI technology, fosters a critical perspective, and prepares medical students to engage with the technology in a differentiated manner. The curriculum is feasible, beneficial, and highly accepted among students, suggesting it could serve as a teaching model for other medical institutions. Given the growing number and impact of medical AI applications, there is a pressing need for more AI-focused curricula and further research on their educational impact.

(JMIR Med Educ 2025;11:e65220) doi:10.2196/65220

KEYWORDS

digital transformation; artificial intelligence; clinical AI; chatbot; digital literacy; medical education; medical students; medical curriculum; qualitative content analysis; medical studies

Introduction

Background

The digital transformation in the health care system represents a fundamental process of change and innovation that is altering the roles, competencies, and cooperation of doctors to a large extent [1]. Eric Topol [2] describes an increasing "super-convergence" of technologies that is transforming the existing health care system into a digital health care system. The key characteristics of this new system are individualization, precision, and prevention. It is expected that this will result in data-based health care that will be characterized by a pronounced intensification of interdisciplinary cooperation and a stronger participatory role for patients. Every patient is increasingly becoming a "big data" challenge, with huge amounts of information about previous illnesses and conditions. At the same time, existing medical knowledge is growing exponentially. These two facts cumulate in increasingly complex decision-making processes in patient care. One recent digital transformative technology that can help bridge this complexity gap by preparing, analyzing, and organizing large amounts of data is artificial intelligence (AI). In health care, AI is becoming increasingly important for extracting and interpreting clinically useful information from large volumes of digital data and information sources and, in some cases, deriving recommendations for therapeutic action.

In the following section, the term "AI applications in medicine" refers to medical software, devices, and technologies such as apps whose analytical processes are AI-based and which are used in the health care sector by patients and/or practitioners.

AI Applications in Medicine

Integrating AI applications into medical processes can automate repetitive tasks currently handled by humans. This hybrid working model improves human performance through technology. In 2012, the US Food and Drug Administration (FDA) certified a medical AI application for the first time [3]. Currently, the FDA database comprises 950 applications (as of the last FDA update on August 7, 2024), predominantly in radiology and the cardiovascular field [4]. Clinical AI systems have already demonstrated expert-level performance in radiology [5-7] and equaled the diagnostic performance of health care professionals in medical imaging [6]. Beyond radiology, there are numerous publications on the clinical application of AI [8-16] and large language models such as ChatGPT [17-19]. A scoping review by Han et al [20] generated an overview of all published randomized controlled trials on clinical AI as of November 2023 and found 84 studies. Their review underpins the growing evidence for the use of AI-supported tools in health care. However, from a populational and thus patient perspective, attitudes toward AI in health care are still fluid and demonstrate varying levels of knowledge, acceptance, and skepticism across different countries and demographic groups [21-24].

The Need for Curricular Training About AI in Medicine

This development has far-reaching implications for doctors and requires a fundamental examination of AI systems [25,26]. At present, neither medical professionals already practicing nor the generation currently studying is adequately prepared for the integration of AI in medicine. At the same time, both groups will—or are already—encountering actionable AI in their day-to-day work that is or will be able to predict, diagnose and, if necessary, treat diseases [2]. At a clinical level, doctors require the competencies to critically assess AI applications to use only those tools that have an evidence-based effect on improving clinical workflows or patient outcomes. At the development level, it is also important to ensure that doctors are actively involved in the development and scientific testing of new AI applications. This raises the question of the extent to which these systems can be effectively integrated into the diagnosis and treatment process, as well as how limitations of the systems can be recognized by medical users and how a fallback level can be ensured. In rapidly changing health care systems, it is therefore essential to ensure that doctors have the knowledge, skills, and attitudes to both master current challenges and be prepared for future challenges [1].

The basic competencies required for this must be learned by medical students during studies and continuously developed throughout their careers [27]. There are already various international ideas for this qualification mandate. For example, the Standing Committee of European Physicians addresses this goal in its Policy on Digital Competencies for Doctors and defines digital core competencies [28]. The EU Health Policy Platform has formulated specific instructions for achieving these core competencies [29]. According to these policymakers, educators should consider including content about the following skill sets into their curricula: (1) general digital skills (data and software security, ethical and legal implications), (2) technical digital skills (telemedicine, AI, health apps, smart devices, robotics, virtual reality/augmented reality, data literacy), and (3) the patient-doctor relationship (digital communication and collaboration, digital health literacy). Seth et al [30] created a theoretical framework of topics related to AI that need to be taught to train medical students in this technology. Laupichler et al [31] emphasize the need to assess medical students' AI literacy and attitudes in order to hone medical curricula to the AI educational needs of the next generation.

According to a recent review by Gordon et al [32], a growing number of medical schools are addressing AI throughout medical studies, but this is limited by the fact that only 2 of the 278 included studies focused on educational competencies in AI. For the German landscape, a study on national course programs found that the majority (72%) of surveyed medical schools stated that they offer AI-related learning opportunities [33]. In contrast to this, 70% of German medical students indicated in a survey conducted at the same time that they had never received any education in digital topics [34]. This surprising discrepancy can be explained by the fact that, although most German medical schools report offering such opportunities, they are mostly part of elective or extracurricular courses, with only 2 institutions including a separate subject specifically on AI in the core curriculum. As a result, existing AI curricula are currently only available to a very limited number of students, and large-scale AI education is still lacking. Recent studies consistently highlight knowledge gaps in AI

education from the perspectives of both international medical students [31,35-37] and German medical students [31,34,38-41]. Students displayed low familiarity with AI and limited awareness of its potential applications in health care; they also reported limited or uncertain access to AI education in medical training. However, they believed AI training would be beneficial and showed great interest in working with it. Results differed regarding attitudes. Although Alkhaaldi et al [36], Moldt et al [42], and Laupichler et al [31] found students to be more optimistic and accepting about AI applications, in Boillat et al's [37] survey there was more skepticism among students regarding the potential harm of AI for patients and job safety within the medical profession.

Despite the increasing number of medical schools that include AI-related teaching according to recent literature, current medical curricula struggle to meet the demands of students to equip them with a strong competency base to interact with, integrate, and critically evaluate AI tools in their clinical practice. Integrating education on core AI competencies into the general curriculum on a broader scale could significantly improve students' experience levels with AI, enhance their attitudes toward the technology, and better prepare them to navigate medical AI effectively in clinical practice.

The objective of this paper is twofold. The first part introduces the concept of a novel, multisession elective course, "Medicine in the Digital Age," which integrates AI teaching in the context of digital transformation into the medical curriculum at a German university. The second part presents findings from a qualitative interview evaluation of participants' feedback on the AI module as well as their overall experience of the elective. Our aim was to conduct an explorative prospective study using a semistructured qualitative interview approach to generate a multidimensional insight into students' knowledge, skills, and attitudes in dealing with AI in medical practice and their overall learning experience. For this, we asked students to comment on the following a priori deductive dimensions of interest: "Areas of AI Application," "Future Work," and "Critical Reflection." The interview findings supported the iterative refinement of our teaching concept, as well as the ongoing educational reform processes.

Methods

Ethical Considerations

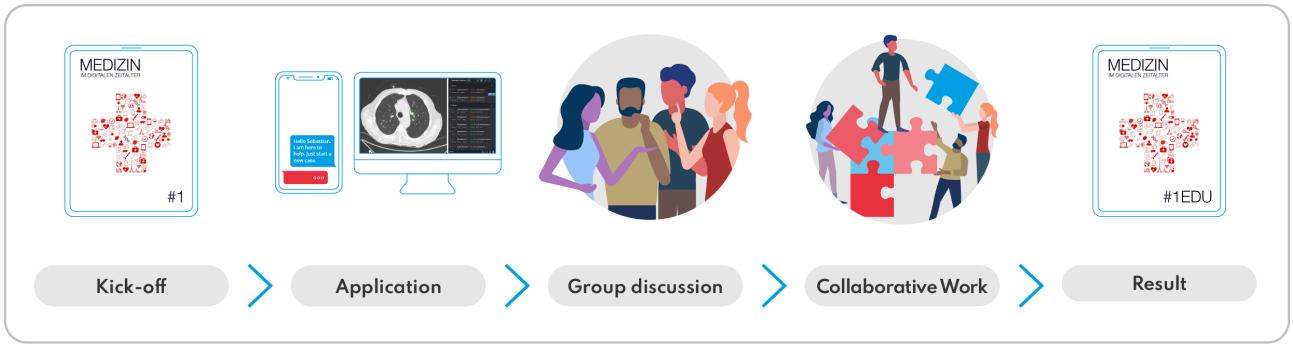
The local ethics committee was consulted during the development of the teaching evaluation for the curriculum presented. Following the consultation, the committee confirmed that the teaching evaluation constituted an additional quality assurance measure for teaching and curriculum development. This evaluation complements the existing concept for quality assurance at the Mainz University Medical Centre [43]. In accordance with the committee's recommendations, it was determined that an ethics vote was not considered necessary. However, participation in the evaluation required informed written consent from all students involved and was strictly voluntary. No compensation was provided to participants. To ensure confidentiality, all responses were pseudonymized prior to analysis.

Structure and Rationale Behind the Module on AI

The teaching module "Artificial Intelligence" is part of the competency-based multisession elective course entitled "Medicine in the Digital Age." The overall course aim is to equip students with digital skills that can be applied in a medically sound, technically feasible, legally compliant, data protection–compliant and ethically responsible manner and thus prepare them for the working environments of the future [2]. It was the first curriculum of its kind in Germany [1].

The "Medicine in the Digital Age" course consists of 5 modules: AI; digital doctor-patient communication; digital health applications and smart devices; telemedicine; and virtual/augmented reality and robotics. The course is offered to medical students in their clinical years as a 5-day elective, with each module comprising an 8-hour face-to-face course day. The didactic concept follows a flipped classroom and blended learning format by combining e-learning (e-book), face-to-face teaching (hands-on workshops, practical exercises, discussion and reflection formats), coproduction, and transfer projects (Figure 1). The different formats alternate over the course and build on each other.

Figure 1. Didactic concept consisting of e-learning, workshops, discussion/reflection formats, and transfer. The results of the learning process are incorporated into the e-learning e-book.



Using an interactive e-book, the participants deal with topics of digital transformation in the preliminary stages of the course. The e-book was created by our working group consisting of experts from medicine, medical education, ethics, media education, data science, and data protection, and also included patient perspectives. Its content mirrors the program of the elective course and contains a dedicated chapter for each module, combining theoretical background, reflective articles,

and stakeholder or patient interviews. Students are expected to read the corresponding chapter in preparation for each course day to independently develop the basics of digital medicine. The entire e-book follows the collaborative concept of "Do it by the book, but be the author" [44]. By incorporating all student transfer projects into an iterative version of the e-book, students are encouraged to actively interact with the course and become the "authors" of their elective course's e-book after completion of the elective.

The thematic breadth and interconnectedness of medical specialties necessitates an interdisciplinary team of lecturers. Therefore, onsite teaching is carried out by various medical disciplines (anesthetists, surgeons, medical informaticians, psychologists, pediatricians, psychosomatics, radiologists, orthopedic and trauma surgeons). In addition, computer scientists, representatives of federal state data protection and medical ethics, and patients complement the team of lecturers in the spirit of a transdisciplinary approach.

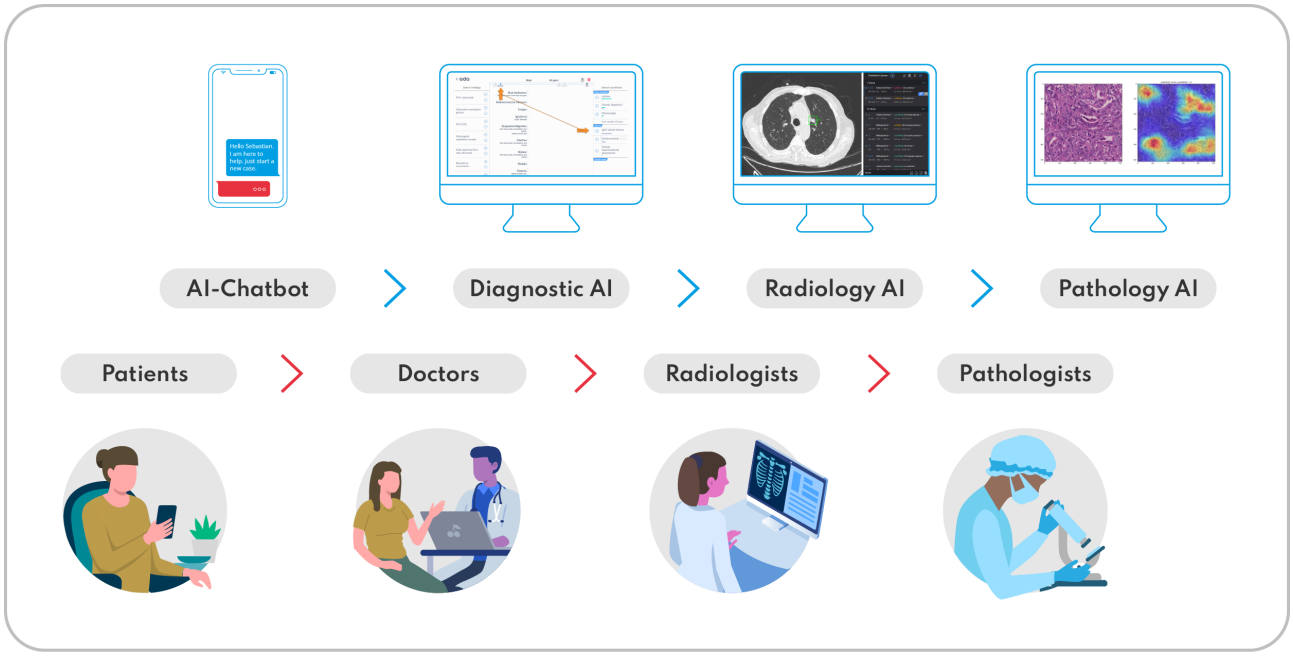
For the AI module, e-learning combined with face-to-face teaching and transfer tasks results in a total of 20 hours of teaching on AI. In accordance with the KSAVE model (Knowledge, Skills, Attitudes, Values, and Ethics) [45,46], the AI module aims at teaching the following overarching competences:

1. The student can describe various areas of application and programs that work with AI and is able to categorize clinical AI assistance systems in medical treatment in an evidence-based manner.
2. The student is able to explain examples of AI-assisted anamnesis, clinical examination, diagnosis, and therapy.
3. The student is able to name limitations of AI applications in current clinical practice and to evaluate the use and benefits of AI within the complex interplay of technical, legal, and ethical principles as well as under sociopolitical framework conditions and to place them in a medical context.
4. The student is able to reflect on how roles in the medical profession will change or evolve in the light of integrated AI assistance systems.

To give an example: a student demonstrates competence mastery by reflecting on the integration of AI-based systems into clinical workflows in oncological imaging in radiology, emphasizing their role in improving early identification of curative versus palliative needs and complementing clinical decision-making. They critically evaluate practical applications across anamnesis, diagnosis, and therapy, addressing limitations such as data quality, clinician acceptance, and ethical concerns. Through this analysis, the student links theoretical knowledge to real-world challenges, demonstrating readiness to apply AI to improve patient care and health care processes.

In the onsite teaching of the AI module, the focus is on practical workshops (Figure 2). These are designed to illustrate the integration of AI into medical treatment processes, followed by discussion and reflection sessions to promote the transfer to the students' own actions.

Figure 2. Workshops within the AI curriculum demonstrate hybrid workflows between humans and AI. The human workflow (lower section) is enhanced by the integration of AI-based narrow intelligence (upper section) along the patient care continuum by various medical specialists. AI: artificial intelligence.



Workshop on Medical History/AI Chatbot

To address the relevant technologies (ie, natural language processing, large language models, and chatbots), students are introduced to an AI-based smartphone app (Ada Health), which acts as a chatbot to take a symptom-based clinical history and make a suspected diagnosis [47-49]. Students then take a clinical history in groups of two from one of the lecturers, who takes on the role of a patient based on a predefined case vignette. First, one of the students takes a classic medical history and

formulates a suspected diagnosis. Thereafter, the second student obtains a medical history of the same patient case by reading out the chatbot's questions. Subsequently, the independently formulated suspected diagnoses are compared with the suspected diagnoses of the chatbot in the entire group. The students discuss which anamnesis questions and diagnoses they did not consider, and which questions and diagnoses the chatbot did not list. The usefulness of the different suspected diagnoses is then discussed and differences in the clinical histories are explored to determine the advantages and disadvantages of each method.

Workshop on Radiology/AI-Supported Radiology

Together with a radiologist, students learn how digitalization is changing the way radiologists work (eg, Picture Archiving and Communication System, radiology information system, speech recognition). Radiological AI applications are demonstrated as examples. Specifically, an AI for the automatic detection of tumor-specific lung foci is demonstrated. The software (InferVision, InferRead CT Lung) provides the user with the size and localization of the lesion as well as an estimate of the malignancy as a percentage. Additionally, an AI application for the automatic diagnosis of conventional X-ray examinations of the thorax is presented (Oxipit, ChestEye). The students learn that a number of published papers have already shown that various AI applications are equivalent to radiologists in individual subtasks [6].

Workshop on Pathology/AI-Supported Pathology

Students are introduced to the influence of digitalization on the field of pathology. For this purpose, a pathologist demonstrates the use of AI as a supporting tool in the diagnosis and detection of malignant changes in histopathological tissue sections [16]. Both the technical and informative background, as well as the process of developing and scientifically evaluating an AI application (AI development life cycle) and its practical application (integration into patient care), are illustrated. Students learn more about the future potential of AI in pathology and can ask questions and contribute their own thoughts.

Discussion and Reflection Formats

For reflection, students and lecturers discuss the following questions together in fishbowl discussions:

- What are the opportunities and risks of using AI in the context of patient treatment?
- How do we deal with probabilities calculated by an AI?
- What will your day-to-day work look like in 2025?
- What new skills will you need in the future?

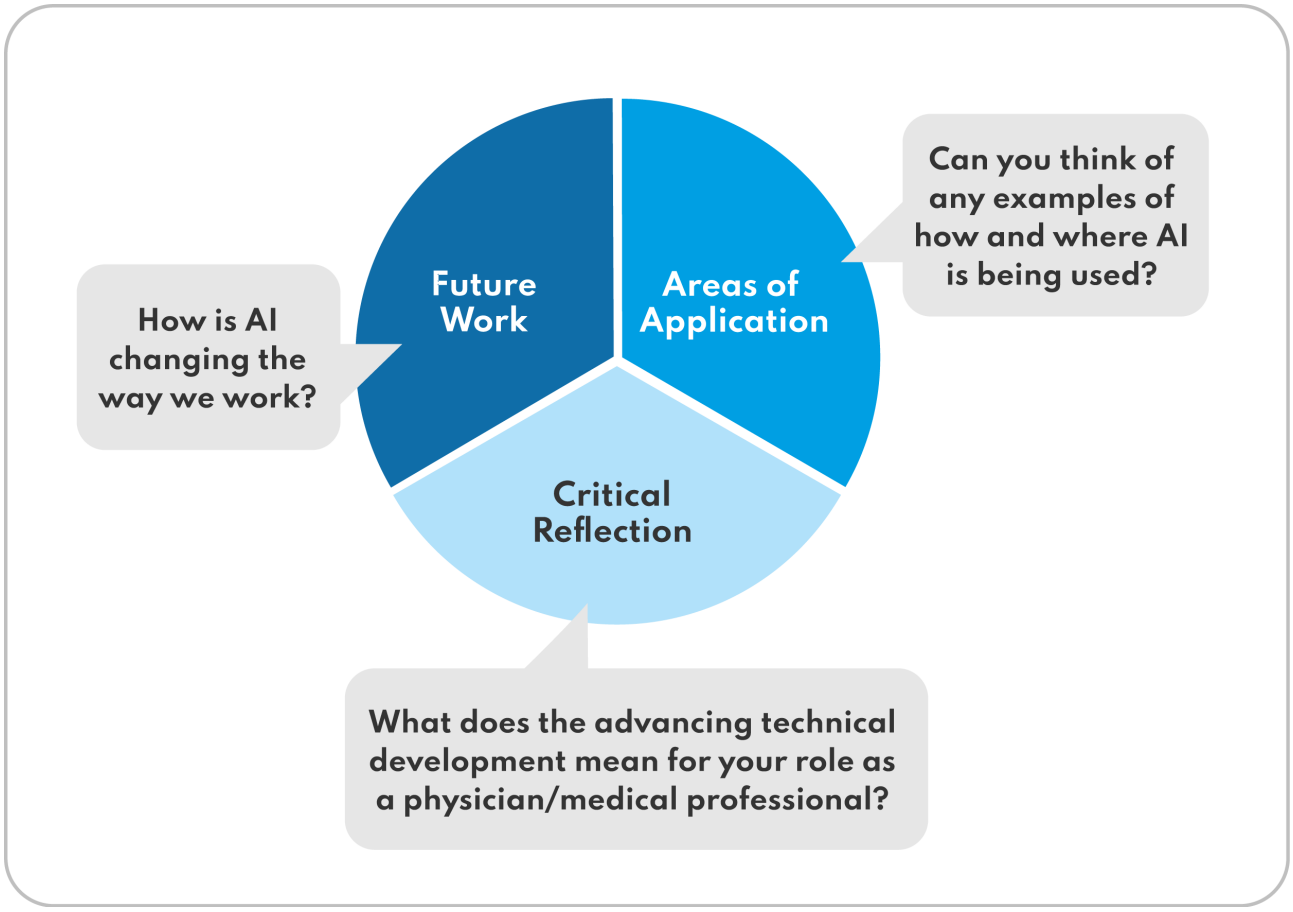
Transfer Projects

Throughout the course, students work in self-selected small groups (groups of 4) on the overarching task of researching a useful medical AI application and presenting it in plenary on the last day of the course. The group presentations are followed by 15-minute discussion rounds with the plenum. The transfer projects students chose reflect the wide range of AI in terms of technology (language, imaging, data procession), medical use cases (conservative medicine, surgical medicine), and different age groups (from AI solutions for pediatrics to palliative care). In addition to their research, part of students' transfer performances also lies in critically analyzing and presenting their various solutions. Among others, students addressed topics such as an AI-supported ultrasound image navigation for regional anesthesia [50], an AI algorithm supporting clinicians in sepsis management [13], or an AI algorithm predicting the end of life developed by Stanford University [51].

Evaluation

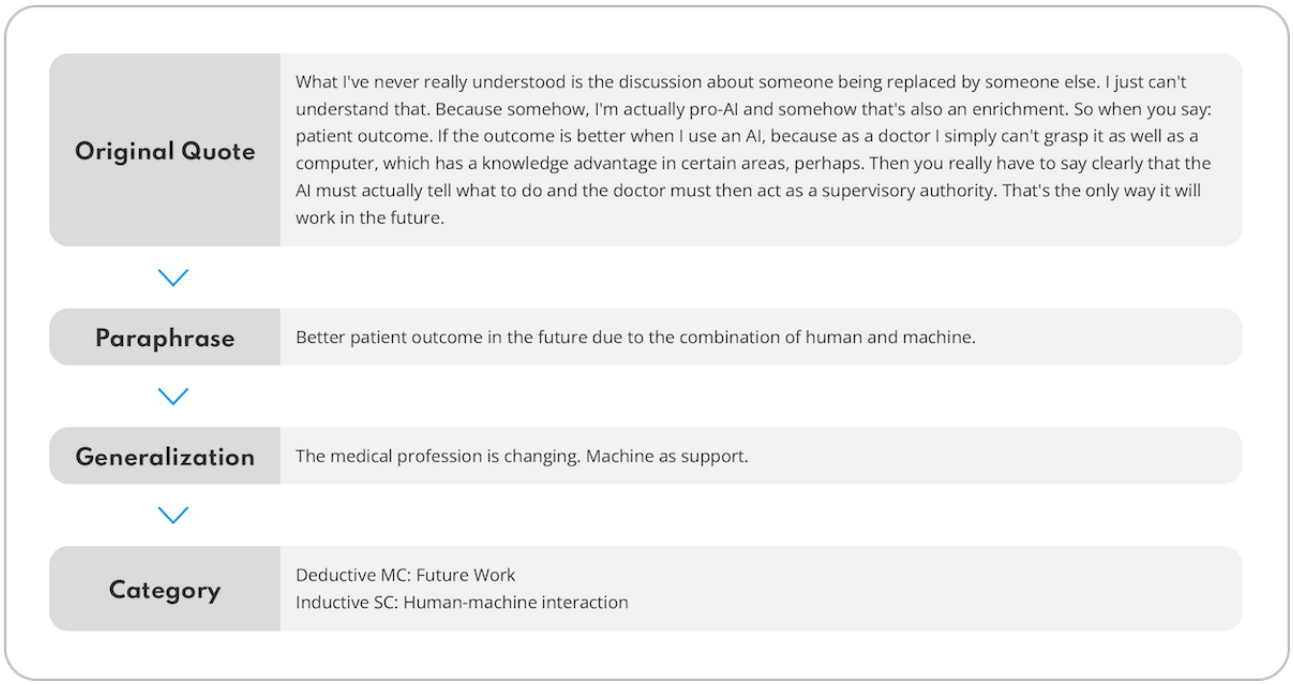
The elective course on Medicine in the Digital Age including the presented AI module was introduced to the medical curriculum at the University Medical Centre of the Johannes Gutenberg-University Mainz. The evaluation of the 3 groups consisting of medical students in their second and third clinical year (ie, years 4 and 5 of the 6-year medical program) was carried out using semistructured, focused, guided group interviews consisting of open and targeted questions based on Merton, Fiske, and Kendall [52,53]. The interview questions were formulated based on the KSAVE model. The interview guide aimed to ascertain the participants' existing competencies in the areas of knowledge, skills, and attitudes in dealing with AI in physicians' practice. This theoretical background resulted in 3 main interview topics: "Areas of Application," "Future Work," and "Critical Reflection" (Figure 3). Additionally, the last section of the interview addressed students' overall assessment of the course. The interview guide was used to evaluate the entire course and is provided in the appendix (Multimedia Appendix 1).

Figure 3. Outline of the main points of the interview guide questions on AI (circle) including guide questions (boxes outside). AI: artificial intelligence.



The focus group interviews took place within 2 weeks of the comprehensive course. Participants were informed about recording, transcription, data usage, storage, and privacy, and consent was obtained beforehand. The interviews were conducted and recorded by a researcher with expertise in qualitative research. The audio files were then transcribed for further analysis by 3 student research assistants (KD, LU, and EK). The interview transcripts were subsequently analyzed using content-structuring qualitative content analysis according to Mayring [54]. This is a text analysis method that follows a logical, systematic pattern and aims at transferring raw text data into structured categories (Figure 4). Categories can be formed deductively based on previous knowledge or hypotheses, or inductively based on new, text-immanent findings. For a structured evaluation of the results, the categories formed in the category system were organized hierarchically into main categories (MCs) and subcategories (SCs). The MCs were deductively derived from the research questions prior to the survey phase. During the analysis process, additional inductive SCs were formed from the interview statements. Inductive coding was used and codings were discussed and agreed upon by the coding research assistants. Saturation was achieved by interviewing all participants and subsequently coding all interview material.

Figure 4. Process of category development through the continuous comparison of the content of the deductive categories with the compiled material. Through the steps of paraphrasing, generalization, and reduction of the content-structuring content analysis, additional inductive categories can be developed, and the statements can then be assigned to an explicit category. MC: main category; SC: subcategory.



The Results section presents the parts of the overall evaluation results that explicitly relate to the AI module and the overall course assessment.

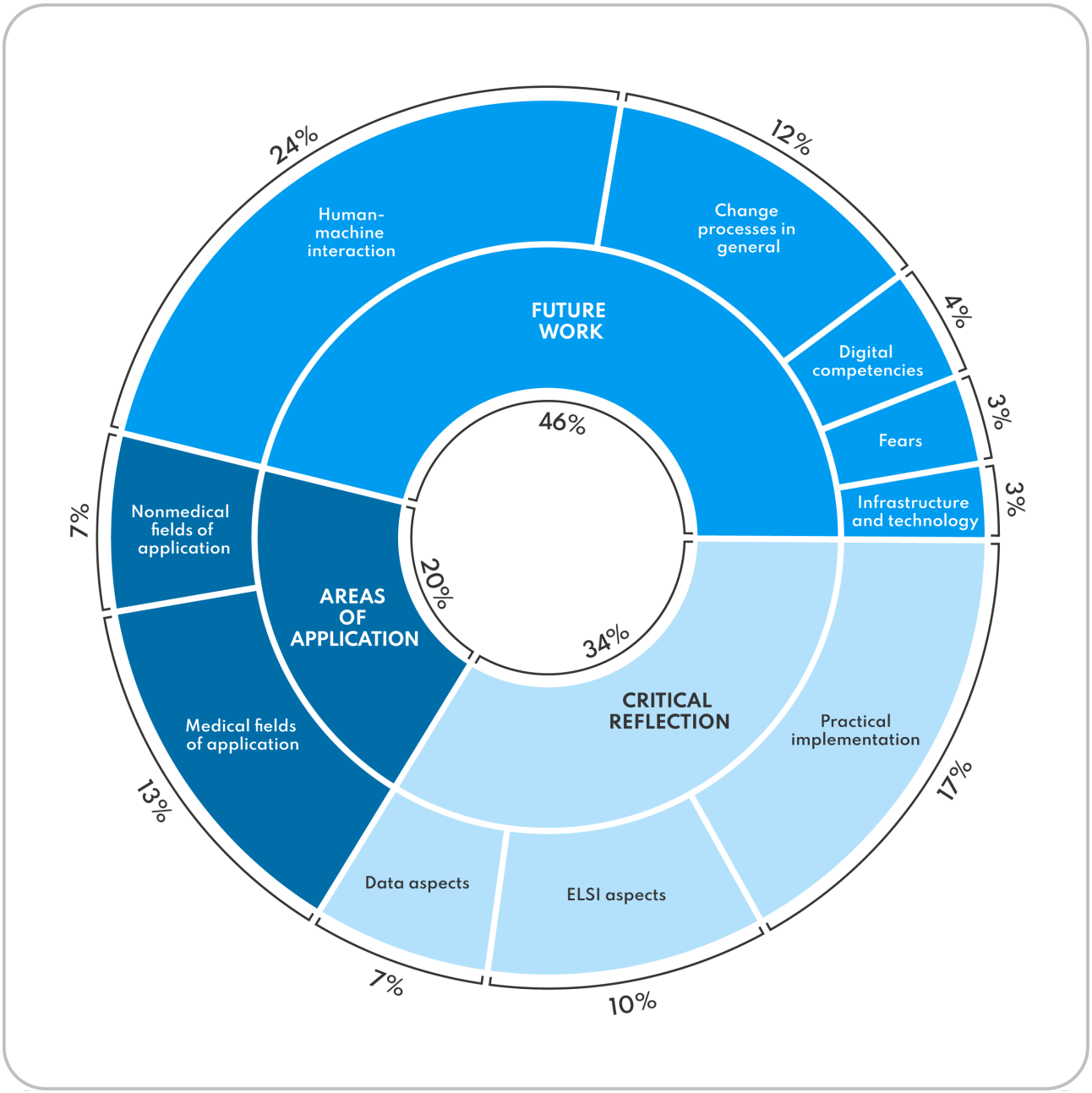
Results

Evaluation Outcomes

From 3 group cycles, 18 semistructured, focused, guided interviews were conducted with all 35 participants (female=11, male=24) from the 3 consecutive courses, which formed the basis for the qualitative evaluation of the course concept. The interviews lasted 24:36 minutes on average. In all interviews, a total of 214 statements were made that could be assigned to the area of “Artificial Intelligence” and 610 statements related to “Overall Course Assessment.”

Following the analysis steps outlined above, the 3 deductive MCs related to AI—“Areas of Application,” “Future Work,” and “Critical Reflection”—were assigned further concrete inductive SCs derived from the content of the text during the evaluation process (Figure 5). Statements in the MC “Overall Course Assessment” were divided into the 6 SCs “Learning Experience,” “Learning Success,” “Structure,” “Content,” “Methods,” and “Conclusion.” For quality assurance purposes, the results report was written in accordance with the Consolidated Criteria for Reporting Qualitative Research (COREQ) checklist [55]. The following section details the qualitative results for each main category. Anchor quotes support the result report for each category. For this, quotes were translated into English by the authors and minor changes were made to improve readability. An extensive overview of anchor quotes for all SCs is provided in the appendix (Multimedia Appendix 2). Identification codes in parentheses accompany each quote to allow allocation to individual participants (the ID code reads as follows: course run:interview number:speaker ID).

Figure 5. Graphical depiction of the qualitative research results on the AI teaching module broken down into main and subcategories, based on 214 coding units (student statements). Percentages are relative to the overall sample of 214 statements on AI. All percentages are rounded. AI: artificial intelligence; ELSI: Ethical, Legal, and Social Implications.



Main Category "Areas of Application"

This category investigated students' knowledge about the existence of different AI applications in health care and beyond and their attitudes toward them after completing the elective course. A total of 20% (43/214) of all statements on AI fall into the category "Areas of Application." Of these, 29 fall into the SC "Medical fields of application" and 14 into the SC "Nonmedical fields of application."

For the SC "Medical fields of application," students listed a variety of medical use cases. The topics that were covered in the course dominated, namely diagnosis in general as well as image diagnosis in radiology, pathology, and dermatology.

> *During the week, for example, I learnt that the radiologists can have the lung round foci assessed by AI during the CT scan. With different probabilities. […] Then we learned in the pathology department that in the future AI will also calculate […] how high the probability is whether it is a tumor or not. Or […] this dermatology AI, which can show whether it's a melanoma or a benign mole. […] So I've definitely learnt a lot, I could go on listing all the examples.* [3:6:F]

In terms of diagnosis, participants rated the use of AI in triage and preliminary anamnesis in the outpatient sector as useful. This could reduce waiting times and counteract missing information. Regarding image diagnosis, the students refer to AI as a "safety net" backing up their own findings with a digital second opinion.

Interviewees expressed disbelief about the status quo on several levels. They discussed the limited usage of AI applications in everyday clinical practice and their own lack of knowledge before attending the course. They identified a general absence of curriculum focused on AI in medicine within their own course of study. Overall, students were surprised by the variety of possible applications, as well as by the rapidly increasing number of market-ready AI medical products. They expressed great interest in the use of AI in their medical practice and see the meaningful and context-specific application of AI as an urgent task of the present.

For the SC "Nonmedical fields of application," at the nonmedical level, students mentioned various possible or already existing application scenarios for AI, some of which have been adopted into everyday life without reflection (eg, voice assistants, navigation, purchase recommendations). The most attention was paid to autonomous driving and transportation.

Main Category "Future Work"

A total of 46% (99/214) of all statements on AI fall into the category "Future Work." Of these, 51 fall into the SC "Human-machine interaction," 26 into the SC "Change process in general," 9 into the SC "Digital competencies," 7 into the SC "Fears" and 6 into the SC "Infrastructure and technology."

For the SC "Human-machine interaction," students focus on the question of how AI can contribute to becoming a better doctor. The students state that they thought intensively about the combination of humans and machines during the course week and consider this combination to be the optimum for the future practice of medicine.

> *Just like the chess player and the computer together (Centaur Chess computer), they are unbeatable. […] and hopefully it will be the same with the doctor [and AI].* [2:1:3]

"Shared decision making" and digital second opinions can counteract incorrect treatment or improve treatment outcomes in the sense of assistive systems and thus increase patient safety. The students describe their experiences with the anamnesis chatbot and characterize the comparison of the two forms of anamnesis as very informative. The differences between the chatbot anamnesis and their own anamnesis practice were thus revealed. Emphasis was placed on the aspect of social anamnesis, which the students carried out more intensively than the chatbot. However, the students rated the chatbot's anamnesis procedure as more structured and systematic than their own. Students predominantly regard AI as an opportunity.

For the SC "Change processes in general," students did not consider the medical profession to be threatened by the use of AI, but the practice of the profession and the subdivision of specialties may change. Students stated that the rapid increase in knowledge means that doctors are even more obliged than before to undergo continuous further training. The preinformed patient will become more of a discussion partner at eye level. The interviewees see this as a great opportunity to improve the doctor-patient relationship, as the inclusion of AI in routine medical procedures could lead to an increase in personnel and time resources, allowing doctors to focus on the traditional core medical activities of consultation, treatment, and care.

For the SC "Digital competencies," the competency profile and requirements for the medical profession are also changing as a result of the transformation. Here, the interviewees speak of a lack of or inadequately trained digital skills, which are also not considered in the standard curriculum. Students would like a safe framework for trying out new technology. Furthermore, the use of AI in everyday medical work requires clear quality criteria that are similar to a review process for evaluating technologies. The ability to correctly classify AI-generated information and to critically question the statements made by the AI is regarded as particularly relevant. According to the students, it is and remains the task of the doctor to assess which application is appropriate for the individual patient and when treating physicians need to actively decide against its use.

> *Where does the AI get the data from? How is it analyzed? And above all, how is it checked to make sure it really is a sensible AI? You need to know that. In a way just like we learn how to read scientific publications. And decide whether they are good or bad.* [2:1:3]

For the SC "Fears," the biggest problem addressed is general ignorance and the resulting fear of, for example, the threat of job losses. Students suspected that this fear is the reason why AI applications are often not developed by health care experts, but by fast-moving commercial companies. Students also considered the combination of human and machine to be problematic if it is not possible for the doctor to understand how the AI operates and reaches decisions.

For the SC "Infrastructure and technology," the students note that the technical change in everyday working life is particularly noticeable through deficits in (technical) equipment.

Main Category "Critical Reflection"

A total of 34% (72/214) of all statements on AI fall into the category "Critical Reflection." Of those, 36 fall into the SC "Practical implementation," 22 into the SC "ELSI aspects" (Ethical, Legal, Social Implications), and 14 into the SC "Data aspects."

For the SC "Practical implementation," students see opportunities for larger-scale, international cooperation. This requires openness and investment in progress and research. They take their practical experience from the "clinical anamnesis" workshop as an example of low-threshold contact, which they want to take out of the course to raise awareness. The limitations of the chatbot, for example, making a misdiagnosis, are critically questioned. In such cases, the intended time saving backfires and becomes extra work.

> *That would unsettle me […] if something completely different comes out as a treatment suggestion or what I see in an MRI image or something like that, then it would make me very insecure and then I would want to make sure. Be it through the senior physician or that I can just have a look: How does this AI come up with this? And if that doesn't work, then it's*

> *unfavorable for the procedure. Then you have to rely on the senior consultant again and maybe the head physician […] and then I'm back where I was before without AI.* [2:1:3]

One problem discussed is that current AIs can only act as "narrow intelligence" in very specific settings, meaning that anomalies that do not correspond to the AI's specific field of action remain undetected. This results in the risk of unconsidered use. In general, all interviewees share the opinion that not everything that is technically possible should also be used in practice and that each use case must be considered individually.

For the SC "ELSI aspects," the question of whether increasing digitalization will reduce or intensify doctor-patient contact is viewed critically. On the one hand, digitalization represents an opportunity to relieve doctors and invest the freed-up capacities in the doctor-patient relationship. On the other hand, there is a risk that AI will impact the interpersonal interaction and thus the patient's individuality.

The possibility of consciously influencing AI is the subject of intense ethical debate. Specifically, it is questioned at what point it is unethical not to use the advancing technology, as this would deliberately deny the patient the best possible treatment.

> *At some point it becomes unethical not to use such things. […] That's actually the point. Why are we always so afraid that we're not important enough? At some point, the doctor is no longer the all-knowing person.* [2:3:1]

The interviewees see a further ethical dilemma in the case of a discrepancy between the diagnosis provided by the doctor and the AI. The right not to know and the handling of probabilities play a decisive role in sensitive areas, such as prenatal or genetic diagnostics and palliative medicine. Students also discussed the unclear legal situation regarding liability issues as a possible cause of rejection of AI applications.

For the SC "Data aspects," regarding data protection, too little regulation violates personal rights. Too much regulation makes it difficult or, in the worst case, prevents access to data for clinical research. In general, students also question the lack of traceability of AI results. They critically note that convenience or lack of time can lead to the results not being checked over time.

Main Category "Overall Course Assessment"

Of all 610 "Overall Course Assessment" statements, 134 fell into the SC "Learning Experience" and 108 into the SC "Learning Success." The remaining statements were categorized into the SCs "Structure" (n=61), "Content" (n=126), "Methods" (n=142), and "Conclusion" (n=39).

For the SCs "Learning Experience" and "Learning Success," students highlighted engaging with AI and digitalization as a significant learning success, given the absence of such topics in the standard curriculum.

> *I'm just glad that I had this week, because it really showed me what we don't learn at university. And how big the topic actually is for us.* [2:2:2]

They described the hands-on interaction with various technologies as "eye-opening" (2:2:4) and the group work on human-AI comparisons as "impressive" (2:1:4). Many students, initially skeptical or ambivalent about AI, reported increased knowledge and awareness of AI technologies and a deeper understanding of their impact as a result of the elective. They felt better prepared for their future careers regarding questioning and categorizing digital tools such as apps or AI, and underlined the gain in competences:

> *I think everyone left with a gain in expertise. Be it in the form of medical expertise, technical expertise, or simply that you've thought about things like data protection and apps and so you've also gained absolute everyday expertise.* [1:3:B1]

For the SCs "Structure," "Content," and "Methods," students appreciated the involvement of diverse experts, valuing the variety of perspectives on the technology.

> *What was outstanding […] was that the input came from the legal side, from the ethical side, from the technical side somehow every time.* [3:8:A]

They praised the active and innovative learning format of the elective, noting that it encouraged reflection and engagement rather than the rote learning typical of other subjects.

> *It's often the case that you're told things and then you have to memorize them. And here it was more the case that you were given information but then had to think about it yourself, for example to discuss it or draw a picture or whatever. And that's a completely different kind of learning, which unfortunately we don't usually do that much of in our degree programs. So I thought it was really good. Because these are actually skills that you should have and not that you can somehow memorize a book.* [2:1:3]

The discussion formats were highlighted as a distinctive feature in comparison with previous teaching experiences, with critical reflection helping students develop a more nuanced understanding of the topic.

> *I think you learnt an incredible amount, especially in the discussions, and you were actually forced to really think about certain theses. I also found this kind of discussion extremely productive.* [2:2:4]

For the SC "Conclusion," students almost unanimously agreed that the elective had broadened their horizons and appreciated the opportunity to participate. They wished that the course would be expanded so that more students could participate. Some expressed a wish for more breaks or even longer discussion sessions. Although students felt that the scope and time commitment of the elective was appropriate, many would have liked it to last longer:

> *"I think the biggest minus is actually the time. It's rare that you leave a course saying: 'Hey, I wish I'd stayed longer.' But […] Tuesday and Wednesday were actually days when I thought: 'Okay. I could have stayed two hours longer'.* [2:2:5]

In summary, the qualitative evaluation showed a high level of acceptance of the course concept and differentiated attitude toward AI among students. The course participants emphasized the increase in their knowledge and competences about the technology as well as the appreciation they felt as a result of the intensive and varied collaboration with each other and with the lecturers. The opportunity for critical discussion, practical interaction, and application was rated particularly positively. All 35 students stated that they would recommend the elective course to peers.

Discussion

Principal Findings

The digitalization of medicine and the use of AI applications is a fundamental process of change that will have a major impact on the future job profile of doctors to an extent that cannot yet be foreseen. What is certain, however, is that we are transitioning from the "information age" to the "age of artificial intelligence" [26] and that the integration of AI into medical treatment processes will redefine human-machine interaction. It is therefore essential to prepare future doctors to use AI in daily practice [27]. At present, although curricula are beginning to change, structured teaching concepts are lacking in terms of curricular mapping, although educators and practitioners emphasize the need to impart such competencies both nationally and internationally. Most students also advocate for AI education in their studies and report limited or no exposure to AI technologies and learning resources [35-37,41]. This does not mean that students must be able to program themselves but they must learn the practical application of AI in line with ELSI principles, data science, biostatistics, and evidence-based medicine during their studies [30,56].

The qualitative results of the AI module show that the embedding of curricular teaching about AI is generally feasible and sensible, that the added value of such a teaching module is recognized by students and acknowledged with great interest and acceptance, and that it leads to an increase in competence among students and promotes a critical and reflective attitude toward new technologies. Regarding the core aspects reflected in the main categories of the analysis (Areas of Application, Future Work, Critical Reflection), several key points can be learned, as detailed in the following sections.

Areas of Application

At present, it is not sufficiently clear how and when AI should be used in clinical diagnostics and therapy. Regarding the "how," students demonstrate a forward-thinking and nuanced examination to potential AI applications in clinical settings.

With regard to the "when," clarification is needed on the specific areas and questions where AI can assist in the clinical workflow [30,57]. Here, students express ambivalence about its integration, acknowledging both benefits and risks.

Future Work

Regarding patient care, students highlighted AI's potential to enhance care through personalized application and resource optimization, aligning with its reported ability to save time and personnel resources amid health care resource scarcity [58]. Students expressed some apprehension about the future impact of AI on the medical profession, though concerns about career choices were less prominent. This aligns with a survey in which 83% of medical students disagreed that AI would render radiologists obsolete [41]. Nonetheless, a minority of participants expressed fears about career prospects. Although AI's full impact remains unpredictable, it is undeniable that medical professions will change. Wartmann and Combs [26] speak of a "reboot" of the health care system and postulate the need to skillfully manage the interface between medicine and machines, as AI will surpass human capabilities in certain tasks [26]. Reflecting this, students stressed the importance of human-machine interaction and corresponding digital skills.

Critical Reflection

Most students underlined the potential of AI for their future career while maintaining a critical perspective, avoiding blind enthusiasm. They emphasized the risk of AI manipulation and its consequences for patient care, underscoring the need for doctors to retain ultimate decision-making authority over AI recommendations. The evaluation presented here thus indicates students' development of a critical attitude due to the module. These findings underscore the importance of future medical curricula teaching students to integrate AI assistance into their decision-making processes [2].

At the industry and developer level, students also acknowledged the need to design AI applications with ELSI aspects in mind. Such recommendations already exist. For example, the multisociety statement on the ethics of AI in radiology [59] and a white paper from the European Society of Radiology outline key ethical and practical considerations for the responsible use of AI in clinical practice [60].

In summary, students acknowledged the evolving nature of AI in health care as well as the necessity for skillful management of the interface between medicine and AI. They emphasized the importance of human-machine interaction as well as the need to develop digital skills while maintaining a reflective mindset toward technology.

Implications

Current literature on medical students' evaluation of their AI competencies demonstrates a relevant knowledge gap and the need for rapid-employment curricula solutions to change this. Overall, both the Medicine in the Digital Age elective as well as its AI teaching module were demonstrated to be feasible and reasonable teaching concepts, which supported maintaining the blended learning approach and the basic content of the modules. Nevertheless, valuable insights for iteration were drawn from the evaluation. First, the course has been updated to reflect technological and regulatory developments, such as the AI Act. Second, insights from students' transfer projects and reflective discussions informed an "agility by design" [61] approach, incorporating noteworthy projects or themes identified through students' input into the subsequent course iterations.

With the didactic framework, course design, and content outlined, this teaching concept can serve as a transferable model for implementation and adaptation in other universities or

training settings. Adjustments may be required to address specific target groups or local circumstances.

Evaluating the AI module's learning objectives—knowledge, skills, and reflection—is critical in both simulated (eg, Observed Structured Clinical Examination exams) and real-world settings. Complex educative interventions like this require robust assessment of efficacy and sustainability. Future research should explore whether a single elective suffices, if refresher courses are needed, or if phased AI education is beneficial. To validate this teaching course, prospective longitudinal trials comparing students who attended the AI module and untrained students are essential.

Methodological Strengths

The "Medicine in the Digital Age" curriculum described here addresses the digital transformation of medicine in an interdisciplinary and interactive way for medical students. AI is one of the 5 teaching modules and the rapid development and adoption of AI technologies in health care requires students and professionals to familiarize themselves with it and develop an attitude toward it. Standard quantitative methods can only inadequately depict the development of a professional attitude. The potential of the qualitative methodology used in teaching research should therefore be emphasized. Qualitative approaches provide insights into the learners' assessment of individual learning success, including gains in the areas of knowledge, skills, and attitude, as well as the content design and methodological structure. They are therefore ideal for the iterative further development of teaching concepts and the evaluation of attitude-oriented teaching content. The application of qualitative methodology represents a distinctive strength and unique contribution of this study. Although most research on medical students' perceptions relies on quantitative questionnaire surveys, this study uses qualitative survey instruments. The 2 existing qualitative studies in Germany are limited by their focus on analyzing free-text survey responses [34] and by their narrow scope, specifically examining students' attitudes toward mental health chatbots [42]. To the best of our knowledge, this study is the first to offer comprehensive, in-depth qualitative insights into German medical students' perceptions and attitudes toward AI.

Limitations

A common limiting factor in qualitative research is the small sample size. Helfferich [62] cites a sample size of between 6 and 120 respondents as appropriate. This means that our sample of 35 students can be assumed to have sufficient result validity. A second limitation could be that the results present a retrospective evaluation. Incorporating a qualitative pre-post analysis might have drawn a more concise picture of students' changes in knowledge, attitudes, and reflection on AI as a result of the course. Third, the findings might not be generalizable to other medical training programs, student attitudes, countries, or demographics. Lastly, group dynamics in the focus groups might have influenced the outcome by introducing social desirability bias.

Conclusions

Digitalization will continue to fundamentally change medicine. Therefore, in line with international appeals, today's education and training curricula must teach students and practicing physicians the basic competencies for using digital tools such as AI applications. It is not enough to simply integrate context-specific AI solutions as teaching examples into existing curricula. The aim of future curricula must be to equip students with the key competencies for their future day-to-day work in the age of AI and enable them to internalize knowledge, skills, and attitudes toward these tools from the beginning of their training. As an outlook for the AI curriculum presented here, it can be said that it addresses this need in a unique way. The qualitative teaching evaluation showed that students were able to deal with the topic in a very differentiated way after the AI teaching unit. The transferability of the curriculum to other university locations can be assumed in principle. The curriculum could therefore serve as an exemplary teaching concept for other universities and contribute to training medical students in two future-oriented skills: AI literacy and its transfer to medical human-machine interaction.

Acknowledgments

Above all, the authors would like to thank all participants of this study for their valuable insights. Furthermore, they would like to thank Björn Hirte for his great support in designing the illustrations in this paper. Open Access funding was provided by the Open Access Publishing Fund of Philipps-Universität Marburg.

Authors' Contributions

All authors were responsible for drafting, conceptualization, methodology, validation, and visualization. KD was responsible for interview curation and analysis. SK was responsible for funding acquisition, project administration, resources, software, and supervision. SK, FJ, KD, and ZSO were responsible for curricular development. DT supported the methodology portion through his expertise as a medical educator. KD, ZSO, and SK wrote the original draft. ZSO edited and finalized the initial draft. All authors reviewed and edited the manuscript.

Conflicts of Interest

SK is the founder and a shareholder of MED.digital. All other authors declare no conflicts of interest.

Multimedia Appendix 1

Interview guide for artificial intelligence in medicine.
[DOC File, 25 KB - mededu_v11i1e65220_app1.doc]

Multimedia Appendix 2
Qualitative results. Anchor quotes for all categories.
[DOC File, 60 KB - mededu_v11i1e65220_app2.doc]

References

1. Kuhn S. Medizin im digitalen Zeitalter: Transformation durch Bildung [Article in German]. Deutsches Ärzteblatt 2018 Jun;115(14):633-638.
2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
3. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med 2020;3(1):118. [doi: 10.1038/s41746-020-00324-0] [Medline: 32984550]
4. Artificial intelligence and machine learning (AI/ML)-enabled medical devices. US Food and Drug Administration. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices?utm_medium=email&utm_source=govdelivery [accessed 2025-02-26]
5. Ting DSW, Liu Y, Burlina P, Xu X, Bressler NM, Wong TY. AI for medical imaging goes deep. Nat Med 2018 May;24(5):539-540. [doi: 10.1038/s41591-018-0029-3] [Medline: 29736024]
6. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health 2019 Oct;1(6):e271-e297. [doi: 10.1016/S2589-7500(19)30123-2] [Medline: 33323251]
7. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. Nat Med 2019 Jan;25(1):24-29. [doi: 10.1038/s41591-018-0316-z] [Medline: 30617335]
8. Annarumma M, Withey SJ, Bakewell RJ, Pesce E, Goh V, Montana G. Automated triaging of adult chest radiographs with deep artificial neural networks. Radiology 2019 Apr;291(1):196-202. [doi: 10.1148/radiol.2018180921] [Medline: 30667333]
9. Arbabshirani MR, Fornwalt BK, Mongelluzzo GJ, et al. Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. NPJ Digit Med 2018;1(1):9. [doi: 10.1038/s41746-017-0015-z] [Medline: 31304294]
10. Murphy A, Skalski M, Gaillard F. The utilisation of convolutional neural networks in detecting pulmonary nodules: a review. Br J Radiol 2018 Oct;91(1090):20180028. [doi: 10.1259/bjr.20180028] [Medline: 29869919]
11. Lindsey R, Daluiski A, Chopra S, et al. Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci U S A 2018 Nov 6;115(45):11591-11596. [doi: 10.1073/pnas.1806905115] [Medline: 30348771]
12. Kitamura G, Chung CY, Moore BE. Ankle fracture detection utilizing a convolutional neural network ensemble implemented with a small sample, de novo training, and multiview incorporation. J Digit Imaging 2019 Aug;32(4):672-677. [doi: 10.1007/s10278-018-0167-7] [Medline: 31001713]
13. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med 2018 Nov;24(11):1716-1720. [doi: 10.1038/s41591-018-0213-5] [Medline: 30349085]
14. Schaffert D, Bibi I, Blauth M, et al. Using automated machine learning to predict necessary upcoming therapy changes in patients with psoriasis vulgaris and psoriatic arthritis and uncover new influences on disease progression: retrospective study. JMIR Form Res 2024 Jun 27;8:e55855. [doi: 10.2196/55855] [Medline: 38738977]
15. Chatterji S, Niehues JM, van Treeck M, et al. Prediction models for hormone receptor status in female breast cancer do not extend to males: further evidence of sex-based disparity in breast cancer. NPJ Breast Cancer 2023 Nov 8;9(1):91. [doi: 10.1038/s41523-023-00599-y] [Medline: 37940649]
16. Försch S, Klauschen F, Hufnagl P, Roth W. Artificial intelligence in pathology. Dtsch Arztebl Int 2021 Mar 26;118(12):194-204. [doi: 10.3238/arztebl.m2021.0011] [Medline: 34024323]
17. Truhn D, Weber CD, Braun BJ, et al. A pilot study on the efficacy of GPT-4 in providing orthopedic treatment recommendations from MRI reports. Sci Rep 2023 Nov 17;13(1):20159. [doi: 10.1038/s41598-023-47500-2] [Medline: 37978240]
18. Griewing S, Gremke N, Wagner U, Lingenfelder M, Kuhn S, Boekhoff J. Challenging ChatGPT 3.5 in senology-an assessment of concordance with breast cancer tumor board decision making. J Pers Med 2023 Oct 16;13(10):1502. [doi: 10.3390/jpm13101502] [Medline: 37888113]
19. Mesko B. The ChatGPT (generative artificial intelligence) revolution has made artificial intelligence approachable for medical professionals. J Med Internet Res 2023 Jun 22;25:e48392. [doi: 10.2196/48392] [Medline: 37347508]

20. Han R, Acosta JN, Shakeri Z, Ioannidis JPA, Topol EJ, Rajpurkar P. Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. Lancet Digit Health 2024 May;6(5):e367-e373. [doi: 10.1016/S2589-7500(24)00047-5] [Medline: 38670745]
21. Miró Catalina Q, Femenia J, Fuster-Casanovas A, et al. Knowledge and perception of the use of AI and its implementation in the field of radiology: cross-sectional study. J Med Internet Res 2023 Oct 13;25:e50728. [doi: 10.2196/50728] [Medline: 37831495]
22. Artificial intelligence: public awareness survey. GOV.UK. 2019. URL: https://www.gov.uk/government/publications/artificial-intelligence-public-awareness-survey [accessed 2025-02-26]
23. Rainie L, Funk C, Anderson M, Tyson A. How Americans think about artificial intelligence. Pew Research Center. 2022. URL: https://www.pewresearch.org/internet/2022/03/17/how-americans-think-about-artificial-intelligence/ [accessed 2025-05-19]
24. Fischer S, Petersen T, Bertelsmann Stiftung. Was Deutschland Über Algorithmen Weiß Und Denkt [Book in German]: Bertelsmann Stiftung; 2018. [doi: 10.11586/2018022]
25. Wartman SA. The empirical challenge of 21st-century medical education. Acad Med 2019 Oct;94(10):1412-1415. [doi: 10.1097/ACM.0000000000002866] [Medline: 31299675]
26. Wartman SA, Combs CD. Medical education must move from the information age to the age of artificial intelligence. Acad Med 2018 Aug;93(8):1107-1109. [doi: 10.1097/ACM.0000000000002044] [Medline: 29095704]
27. Masters K. Artificial intelligence in medical education. Med Teach 2019 Sep;41(9):976-980. [doi: 10.1080/0142159X.2019.1595557] [Medline: 31007106]
28. CPME policy on digital competencies for doctors. Standing Committee of European Doctors (CPME). 2020. URL: https://www.cpme.eu/api/documents/adopted/2020/11/CPME_AD_Board_21112020_100.FINAL_.CPME_.Policy.Digital.Competencies.for_.Doctors.pdf [accessed 2025-02-26]
29. European Commission's Thematic Network on Digital skills for future-proof doctors (Digital Doc). Training future-proof doctors for the digital society. European Junior Doctors. 2020. URL: https://www.juniordoctors.eu/sites/default/files/2021-01/Digital%20Doc_Training%20future-proof%20doctors%20for%20the%20digital%20society.docx.pdf [accessed 2025-02-26]
30. Seth P, Hueppchen N, Miller SD, et al. Data science as a core competency in undergraduate medical education in the age of artificial intelligence in health care. JMIR Med Educ 2023 Jul 11;9:e46344. [doi: 10.2196/46344] [Medline: 37432728]
31. Laupichler MC, Aster A, Meyerheim M, Raupach T, Mergen M. Medical students' AI literacy and attitudes towards AI: a cross-sectional two-center study using pre-validated assessment instruments. BMC Med Educ 2024 Apr 10;24(1):401. [doi: 10.1186/s12909-024-05400-7] [Medline: 38600457]
32. Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Med Teach 2024 Apr;46(4):446-470. [doi: 10.1080/0142159X.2024.2314198] [Medline: 38423127]
33. Mosch L, Back A, Balzer F, et al. Lernangebote zu künstlicher intelligenz in der medizin [Report in German]. : Zenodo; 2021. [doi: 10.5281/ZENODO.5497668]
34. Sorg H, Ehlers JP, Sorg CGG. Digitalization in medicine: are German medical students well prepared for the future? Int J Environ Res Public Health 2022 Jul 7;19(14):8308. [doi: 10.3390/ijerph19148308] [Medline: 35886156]
35. Liu DS, Sawyer J, Luna A, et al. Perceptions of US medical students on artificial intelligence in medicine: mixed methods survey study. JMIR Med Educ 2022 Oct 21;8(4):e38325. [doi: 10.2196/38325] [Medline: 36269641]
36. Alkhaaldi SMI, Kassab CH, Dimassi Z, et al. Medical student experiences and perceptions of ChatGPT and artificial intelligence: cross-sectional study. JMIR Med Educ 2023 Dec 22;9:e51302. [doi: 10.2196/51302] [Medline: 38133911]
37. Boillat T, Nawaz FA, Rivas H. Readiness to embrace artificial intelligence among medical doctors and students: questionnaire-based study. JMIR Med Educ 2022 Apr 12;8(2):e34973. [doi: 10.2196/34973] [Medline: 35412463]
38. Gillissen A, Kochanek T, Zupanic M, Ehlers J. Medical students' perceptions towards digitization and artificial intelligence: a mixed-methods study. Healthcare (Basel) 2022 Apr 13;10(4):723. [doi: 10.3390/healthcare10040723] [Medline: 35455898]
39. Weidener L, Fischer M. Artificial intelligence in medicine: cross-sectional study among medical students on application, education, and ethical aspects. JMIR Med Educ 2024 Jan 5;10(1):e51247. [doi: 10.2196/51247] [Medline: 38180787]
40. Busch F, Hoffmann L, Truhn D, et al. Medical students' perceptions towards artificial intelligence in education and practice: a multinational, multicenter cross-sectional study. medRxiv. Preprint posted online on 2023. [doi: 10.1101/2023.12.09.23299744]
41. Pinto Dos Santos D, Giese D, Brodehl S, et al. Medical students' attitude towards artificial intelligence: a multicentre survey. Eur Radiol 2019 Apr;29(4):1640-1646. [doi: 10.1007/s00330-018-5601-1] [Medline: 29980928]
42. Moldt JA, Festl-Wietek T, Madany Mamlouk A, Nieselt K, Fuhl W, Herrmann-Werner A. Chatbots for future docs: exploring medical students' attitudes and knowledge towards artificial intelligence and medical chatbots. Med Educ Online 2023 Dec;28(1):2182659. [doi: 10.1080/10872981.2023.2182659] [Medline: 36855245]
43. Qualitätssicherung [Website in German]. Universitätsmedizin Mainz. URL: https://www.um-mainz.de/rfl/studium-lehre/informationen-fuer-lehrende-und-einrichtungen/evaluation-qualitaetssicherung.html [accessed 2025-05-19]

44. Kuhn S, Kirchgässner E, Deutsch K. Medizin im digitalen Zeitalter – 'Do it by the book … but be the author!' [Website in German].: Synergie; 2017. URL: https://www.synergie.uni-hamburg.de/de/media/ausgabe04/synergie04-beitrag06-kuhn-kirchgaessner-deutsch.pdf [accessed 2025-02-26]
45. Binkley M, Erstad O, Herman J, et al. Defining twenty-first century skills. In: Griffin P, McGaw B, Care E, editors. Assessment and Teaching of 21st Century Skills: Springer Netherlands; 2012:17-66. [doi: 10.1007/978-94-007-2324-5_2]
46. Seidl T. (Wert-)haltung, als wichtiger bestandteil der entwicklung von 21st century skills an hochschulen. AG curriculum 4.0. diskussionspapier nr. 3 [Report in German]. : Zenodo; 2018. [doi: 10.5281/ZENODO.2634975]
47. Kuhn S, Jungmann SM, Jungmann F. Künstliche Intelligenz für Ärzte und Patienten: “Googeln” war gestern [Article in German]. Dtsch Arztebl Int 2018;115(26):A-1262. [Medline: 30135007]
48. Jungmann SM, Klan T, Kuhn S, Jungmann F. Accuracy of a chatbot (Ada) in the diagnosis of mental disorders: comparative case study with lay and expert users. JMIR Form Res 2019 Oct 29;3(4):e13863. [doi: 10.2196/13863] [Medline: 31663858]
49. Ada Health. URL: https://ada.com/de/ [accessed 2025-02-26]
50. Magoon R, Suresh V. A novel recognition of artificial intelligence in regional anaesthesia. Digit Med 2023 Jun;9(2):e00003. [doi: 10.1097/DM-2023-00003]
51. Hsu J. Stanford’s AI predicts death for better end-of-life care. IEEE Spectrum. 2018. URL: https://spectrum.ieee.org/stanfords-ai-predicts-death-for-better-end-of-life-care [accessed 2025-02-26]
52. Merton RK, Fiske M, Kendall P. The Focused Interview: A Manual of Problems and Procedures, 1st edition: Free Press; 1956.
53. Friebertshäuser B, Langer A, Prengel A, Boller H, Richter S, editors. Handbuch Qualitative Forschungsmethoden in Der Erziehungswissenschaft 3, Vollständig Überarb Aufl, Neuausg [Book in German]: Juventa-Verl; 2010.
54. Mayring P. Qualitative Inhaltsanalyse [Book in German]: Beltz; 2015.
55. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
56. Park SH, Do KH, Kim S, Park JH, Lim YS. What should medical students know about artificial intelligence in medicine? J Educ Eval Health Prof 2019;16:18. [doi: 10.3352/jeehp.2019.16.18] [Medline: 31319450]
57. Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. NPJ Digit Med 2018;1(1):40. [doi: 10.1038/s41746-018-0048-y] [Medline: 31304321]
58. Topol EJ, Verghese A. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again, 1st edition: Basic Books; 2019.
59. Geis JR, Brady AP, Wu CC, et al. Ethics of artificial intelligence in radiology: summary of the Joint European and North American Multisociety Statement. Radiology 2019 Nov;293(2):436-440. [doi: 10.1148/radiol.2019191586] [Medline: 31573399]
60. European Society of Radiology (ESR). What the radiologist should know about artificial intelligence - an ESR white paper. Insights Imaging 2019 Apr 4;10(1):44. [doi: 10.1186/s13244-019-0738-2] [Medline: 30949865]
61. Kuhn S, Jungmann F, Deutsch K, Drees P, Rommens PM. Digitale Transformation der Medizin [Book in German]. OUP 2018;7:453-458. [doi: 10.3238/oup.2018.0453–0458]
62. Helfferich C. Die Qualität Qualitativer Daten [Book in German], 4th edition: VS Verlag für Sozialwissenschaften; 2011.

Abbreviations

AI: artificial intelligence
COREQ: Consolidated Criteria for Reporting Qualitative Research
ELSI: Ethical, Legal, and Social Implications
FDA: US Food and Drug Administration
KSAVE: Knowledge, Skills, Attitudes, Values and Ethics
MC: main category
SC: subcategory

Edited by B Lesselroth; submitted 09.08.24; peer-reviewed by J Aulenkamp, M Laupichler, N Schlicker; revised version received 03.03.25; accepted 06.04.25; published 26.05.25.

Please cite as:
Oftring ZS, Deutsch K, Tolks D, Jungmann F, Kuhn S
Novel Blended Learning on Artificial Intelligence for Medical Students: Qualitative Interview Study
JMIR Med Educ 2025;11:e65220
URL: https://mededu.jmir.org/2025/1/e65220
doi:10.2196/65220

Chatbots' Role in Generating Single Best Answer Questions for Undergraduate Medical Student Assessment: Comparative Analysis

Enjy Abouzeid*, MBChB, MSc, PhD; Rita Wassef*, MBBCh, MSc, MD; Ayesha Jawwad*, BDS, MPH; Patricia Harris*, BSc(Hons), PhD

School of Medicine, University of Ulster, Northland Road, Derry-Londonderry, United Kingdom

*all authors contributed equally

Corresponding Author:
Enjy Abouzeid, MBChB, MSc, PhD
School of Medicine, University of Ulster, Northland Road, Derry-Londonderry, United Kingdom

Abstract

Background: Programmatic assessment supports flexible learning and individual progression but challenges educators to develop frequent assessments reflecting different competencies. The continuous creation of large volumes of assessment items, in a consistent format and comparatively restricted time, is laborious. The application of technological innovations, including artificial intelligence (AI), has been tried to address this challenge. A major concern raised is the validity of the information produced by AI tools, and if not properly verified, it can produce inaccurate and therefore inappropriate assessments.

Objective: This study was designed to examine the content validity and consistency of different AI chatbots in creating single best answer (SBA) questions, a refined format of multiple choice questions better suited to assess higher levels of knowledge, for undergraduate medical students.

Methods: This study followed 3 steps. First, 3 researchers used a unified prompt script to generate 10 SBA questions across 4 chatbot platforms. Second, assessors evaluated the chatbot outputs for consistency by identifying similarities and differences between users and across chatbots. With 3 assessors and 10 learning objectives, the maximum possible score for any individual chatbot was 30. Third, 7 assessors internally moderated the questions using a rating scale developed by the research team to evaluate scientific accuracy and educational quality.

Results: In response to the prompts, all chatbots generated 10 questions each, except Bing, which failed to respond to 1 prompt. ChatGPT-4 exhibited the highest variation in question generation but did not fully satisfy the "cover test." Gemini performed well across most evaluation criteria, except for item balance, and relied heavily on the vignette for answers but showed a preference for one answer option. Bing scored low in most evaluation areas but generated appropriately structured lead-in questions. SBA questions from GPT-3.5, Gemini, and ChatGPT-4 had similar Item Content Validity Index and Scale Level Content Validity Index values, while the Krippendorff alpha coefficient was low (0.016). Bing performed poorly in content clarity, overall validity, and item construction accuracy. A 2-way ANOVA without replication revealed statistically significant differences among chatbots and domains (P<.05). However, the Tukey-Kramer HSD (honestly significant difference) post hoc test showed no significant pairwise differences between individual chatbots, as all comparisons had P values >.05 and overlapping CIs.

Conclusions: AI chatbots can aid the production of questions aligned with learning objectives, and individual chatbots have their own strengths and weaknesses. Nevertheless, all require expert evaluation to ensure their suitability for use. Using AI to generate SBA prompts us to reconsider Bloom's taxonomy of the cognitive domain, which traditionally positions creation as the highest level of cognition.

(JMIR Med Educ 2025;11:e69521) doi:10.2196/69521

KEYWORDS

artificial intelligence; assessment; Bing; ChatGPT; Gemini; medical education; single best answer

Introduction

Across disciplines of education, including medical education, programmatic assessment offers flexible learning modalities that pave the road for individual progression. However, it represents a challenge to educators, as they are required to develop frequent assessments that reflect different competencies, thus necessitating the continuous creation of examination content in a comparatively restricted time [1]. For many years, multiple choice questions (MCQs) have been adopted in medical education for assessing knowledge and clinical reasoning skills

in high-stakes undergraduate and postgraduate medical exams. MCQs are reliable, objective, standardized, equitable, and efficient formats for testing large volumes of content in a limited time. A main problem with MCQs is that producing high-quality questions is time-consuming, from drafting the question that includes a clinical vignette or stem, a lead-in question, a correct answer, and distractors to validation of content and detection of potential flaws [1,2]. To tackle this dilemma, the application of many technological innovations, including artificial intelligence (AI), has been tried [3].

AI refers to machines mimicking the human brain in performing intellectual tasks. This originates from the imitation game developed by the British mathematician Alan Turing, who posed the universally famous question "Can machines think?" [4]. Since then, many AI research laboratories have invested time, effort, and money to answer this question. One particular AI research laboratory known as OpenAI, based in California, United States, has revolutionized our world at the end of 2022 by launching an AI-based large language model (LLM) software (GPT-3.5) that uses natural language processing to engage in human-like conversations and making it freely available for the public [5]. Within a few weeks after its release, the OpenAI chatbot, known as ChatGPT, had gained much attention in many fields, including medical education. It became the fastest-growing app of all time with more than 120 million users in just a few months after its launch [6]. This led competitors to develop and launch other chatbots. Microsoft launched Bing Chat AI in February 2023, followed by Google releasing Gemini in March 2023 [7]. A newer, improved version of ChatGPT (ChatGPT Plus), which uses the GPT-4 Turbo language model, has been developed by OpenAI and launched as a paid subscription version by the end of 2023 [6]

In terms of assessment in medical education, ChatGPT has been the most extensively studied chatbot. It was found to be able to quickly and accurately apply known concepts in medicine to novel problems, including reflection prompts and examination questions, and to mimic human writing styles, introducing a potential threat to the validity of traditional forms of medical student assessment including short answer assessment [8], it even successfully passed the USMLE (United States Medical Licensing Examination) [9]. Similarly, ChatGPT-4 was able to achieve a mean of more than 75% in the newly derived undergraduate medical exit examination: UKMLA (United Kingdom Medical Licensing Assessment) [10]. Its application has been described across multiple areas of academic assessment, for example, developing innovative assessments, grading submitted work, and providing feedback [11]. Nevertheless, concerns persist around the validity of the information provided by all AI tools. Sample [12] argued that if the chatbot response is not properly verified, it can be misleading and result in "junk science."

Additionally, the broad availability of LLMs such as ChatGPT, Gemini, and Bing has facilitated extensive comparative studies across various domains. For example, 1 study evaluated these models using case vignettes in physiology and found that ChatGPT-3.5 outperformed Bing and Google Bard (an old version of Gemini), indicating its superior effectiveness in case-based learning [13]. Another study, using the clinicopathological conferences method, compared the ability of AI chatbots to infer neuropathological diagnoses from clinical summaries. The findings revealed that Google Bard and ChatGPT-3.5 correctly diagnosed 76% of cases, while ChatGPT-4 achieved a higher accuracy rate, correctly identifying 84% of cases [14]. Similarly, a comparison of ChatGPT-3.5, Google Bard, and Microsoft Bing in hematology cases highlighted significant performance differences, with ChatGPT achieving the highest accuracy [15].

Recent studies have explored the use of AI in generating MCQs and single best answer (SBA) questions for medical examinations, highlighting its potential applications and limitations. For instance, Zuckerman et al [16] examined ChatGPT's role in assessment writing, while Kıyak et al [17] and Mistry et al [18] investigated AI-generated MCQs in pharmacotherapy and radiology board exams, respectively.

Despite these contributions, the ability of AI to generate valid SBA questions, an assessment format that better evaluates higher-order cognitive skills such as data interpretation, problem-solving, and decision-making [19], remains an area requiring further exploration. Additionally, a critical consideration is the variation in AI-generated outputs and the potential for examination candidates to predict examination items based on curriculum learning objectives (LOBs). Given the significance of these issues, this study aims to examine the content validity and consistency of different chatbots in generating SBAs for undergraduate medical education.

Methods

Study Context

The Graduate Entry Medical Programme at Ulster University's School of Medicine is a 4-year program. Similar to most UK medical schools, students undergo assessment through a series of SBA papers comprising over 1500 questions across the program. Managing this extensive assessment requirement has prompted the exploration of innovative solutions to support the assessment team.

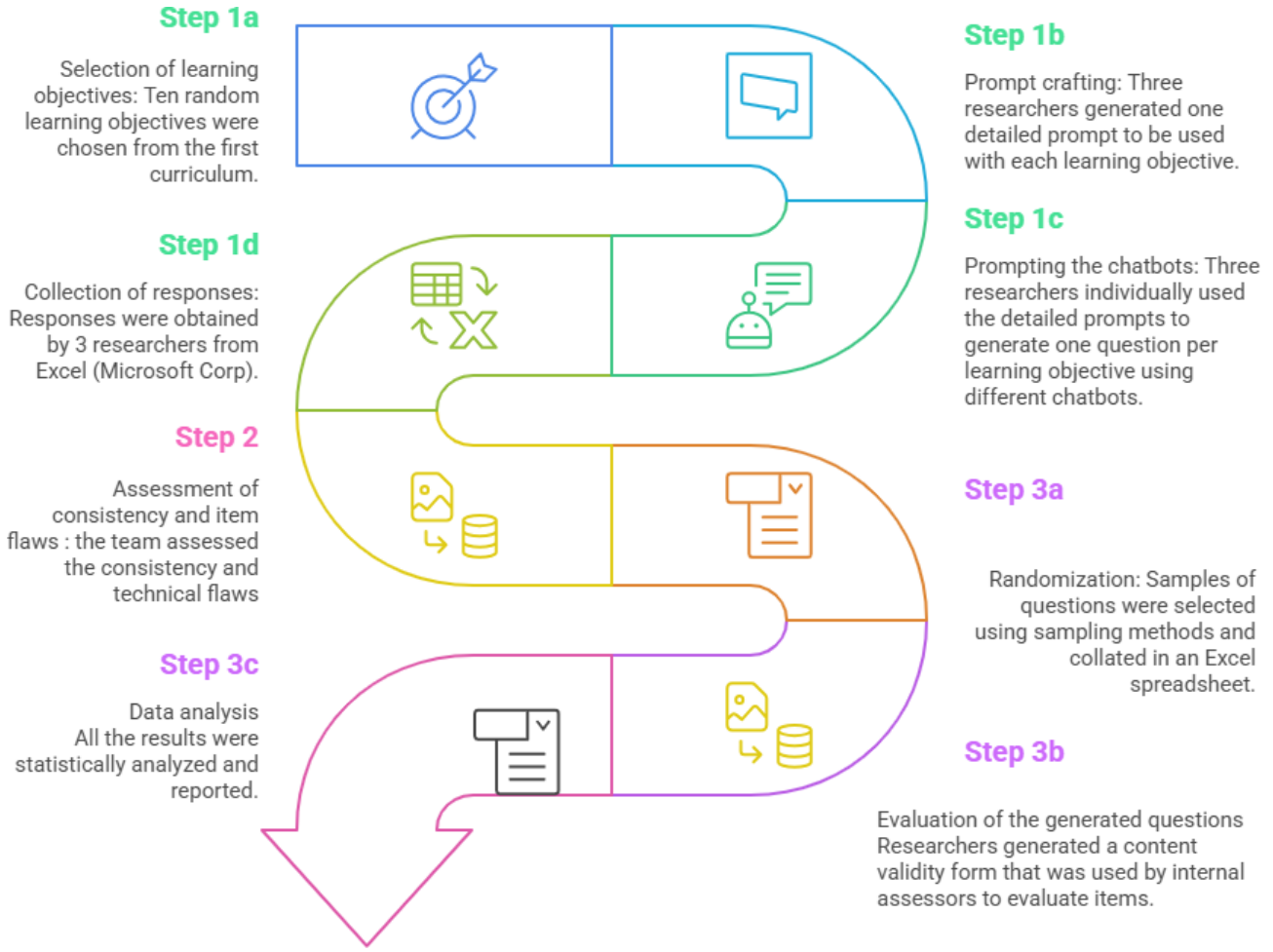
To ensure assessment standards, the school has implemented a rigorous quality assurance process. Questions are first created by designated clinical or academic authors who have been trained and provided with a "house style" to follow. Questions then undergo internal review by other clinical or academic staff before external review by external examiners to ensure they meet rigorous requirements. Post hoc psychometric analysis of question performance is also used to drive evidence-based review and enhancement. This meticulous review process aims to uphold the integrity and effectiveness of assessments used to make high-stakes progression decisions and forms part of a wider suite of quality processes to deliver against the assessment strategy.

Study Design

This exploratory comparative study was conducted between December 2023 and May 2024; we continued to follow the school's established quality assurance process, but the designated first authors of the questions were AI chatbots. This includes 3 versions of AI chatbots: ChatGPT which will be

referred to as ChatGPT-3.5 in this study, Google Gemini, and Microsoft Bing AI, in addition to the subscription-only version of OpenAI: ChatGPT-4 that provides access to GPT-4 Turbo, which is advertised as a more powerful and faster version of GPT-4. During this study, Google changed the name of its platform from Bard to Gemini. For consistency, this paper will refer to the current name: Gemini. Figure 1 depicts the full study design, which included three main phases: (1) Generation of questions using various AI chatbots, (2) Assessment of the consistency of the chatbot outputs, and (3) Evaluation of the quality of the questions generated.

Figure 1. The study design. SBA: single best answer.



Generation of Questions Using Various AI Chatbots

In phase one, the research team randomly selected year 1 curriculum LOBs (n=10) to create SBA questions for. These objectives were selected using stratified random sampling from the official list of LOBs for second-semester educational units. Three researchers were involved, and each one created a new account for each of the 4 chatbot platforms. All researchers used the same predefined prompts (see below) around the same time (end of December 2023) to request 10 questions from each chatbot, one for each LOB. The 10 prompts were entered one by one in the same conversation with each chatbot. All the questions were compiled into a shared Microsoft Excel (Microsoft Corp) spreadsheet for analysis in steps 2 and 3.

To allow a fair comparison, the same prompt was used in each chatbot, which specified SBA features:

- You are a university lecturer in a UK medical school. Generate an MCQ on “the learning objective,” with the following criteria:
  - The question is in a clinical vignette format.
  - The question is designed to assess the knowledge (±clinical judgment) of undergraduate medical students.
  - The question meets the standard for a medical graduate examination.
  - Five choices are allowed for each question.
  - Only 1 correct answer
  - Tag the correct answer.
  - Justify the correct answer.

Assessment of the Consistency and Quality (Item Flaws) of the Chatbot Outputs

In the second phase, researchers involved in the previous step assessed each chatbot’s output consistency and technical flaws. Consistency was evaluated based on the similarity between the outputs generated across the 3 researchers, including any bias in the correct answer allocation (eg, favoring option “A” as the correct answer). Similarity was evaluated based on specific elements of the output and accordingly classified into one of three categories: (1) exact questions: when the outputs contain the same wording, condition, and lead-in question; (2) similar

questions: when the outputs share common elements such as patient characteristics, age, condition, presentation, or lead-in question; (3) different questions: when the outputs do not have any content in common.

Technical item flaws assessed the overall construct and structure of the questions produced by the chatbots using 7 previously published criteria for determining the quality of SBAs [20]. The 7 criteria include judgments on whether the questions: follow the SBA structural format, satisfy the "cover test" rule where the question should be answerable solely from the vignette or stem and lead-in (with the answers "covered"), test the application of knowledge rather than recall isolated facts, have item balance (which ensures a balance in information between the stem, lead-in, and options), tests 1 idea, are dependent on the vignette to reach the correct answer, and have appropriate lead-ins length. The researchers used a defined scale to evaluate how often or to what extent each criterion was met across the 3 researchers' outputs. Each criterion was scored on a scale from 0 to 3 for each of the 10 LOB prompts. In this scale, 0 meant none, 1 meant 1 SBA, 2 meant 2 SBAs, and 3 meant all 3 SBAs, representing the number of questions produced by each chatbot that met the criterion. With 3 assessors and 10 LOBs, the maximum possible score for any individual chatbot was 30.

Assessment of the Content Validity and Accuracy of the Questions Generated

In phase 3, samples of questions generated by the chatbots were distributed to various internal assessors as per our normal quality review process. The questions were selected using stratified random sampling to select 1 of the 3 questions generated by each chatbot for each LOB, yielding a total of 39 questions. Alongside this, a content validation evaluation form, developed by the research team, was used to ensure consistent review between assessors, providing assessors with clear expectations and an understanding of the task. The assessors are faculty members with expertise in the curriculum content. Each question was evaluated by 7 assessors.

Considering published recommendations for content validation [21,22], 20 internal assessors were invited, of which 7 consented to participate. The internal assessors critically reviewed the questions based on several criteria to ensure their quality and alignment with educational objectives. This includes content clarity and validity; accuracy of information, answers, and justification; and educational accuracy. Each of these elements was scored on a Likert scale of 1 to 4 (with 1 representing the lowest level of construct and 4 the highest level of the construct; Multimedia Appendix 1).

Statistical Analysis

Quantitative data was analyzed through scores obtained from the rating scale using IBM SPSS Statistics (version 26; IBM Corp). Subsequently, 2 content validity indexes were computed: the Item Content Validity Index (I-CVI) and the Scale Level Content Validity Index (S-CVI). Percentages and frequencies were calculated for the questions' scores to provide further insights into the data. A 2-way ANOVA without replication was conducted to assess differences in chatbot performance across 6 domains. Post hoc comparisons were performed using the Tukey-Kramer HSD (honestly significant difference) test to identify specific group differences. The average ratings provided by 7 evaluators were used for each chatbot and each criterion. The Krippendorff alpha [23] was used to assess interrater reliability, using the K-Alpha Calculator [24]. A coefficient value of 0.8 is considered satisfactory [23]. However, the low Krippendorff alpha suggested a need for further refinement of the rating scheme or additional training for raters to improve reliability.

Ethical Considerations

Participants were informed that their responses would be anonymized and that they could withdraw from this study at any point without penalty. Informed consent was obtained from all participants before data collection. Only those who provided explicit consent were included in this study. This study received ethical approval from the Ulster University Centre for Higher Education Research and Practice Ethics Committee and the Learning Enhancement Directorate Ethics Filter Committee (LEDEC; formerly CHERP; LEDEC-24-004). All data were anonymized during the analysis phase to ensure confidentiality and to protect participants' identities. Staff members who chose not to participate experienced no disadvantage or impact on their professional standing. No financial or material compensation was offered to participants for their involvement in this research.

Results

Generation of Questions

In response to the predefined prompts provided to the chatbots, 3 of them (free ChatGPT, ChatGPT Plus, and Gemini) generated 10 questions each, for a total of 30 across the 3 researchers. Bing could not respond to the prompt for LOB9 and thus generated 9 questions, for a total of 27 across the 3 researchers. Thus, 117 questions were generated (Multimedia Appendix 2).

Assessment of Consistency Within Chatbots and Technical Item Flaws Among the Outputs

Consistency within chatbots was evaluated based on the similarity of outputs between the 3 researchers and any bias in the allocation of the correct answer option. Bing had the highest degree of similarity between items generated by multiple users (4 exact question matches and 20 similar ones), while ChatGPT-4 had the highest degree of variation (Table 1).

Table . Similarity between the questions generated by different chatbots.

	Gemini (N=30), n (%)	Bing (N=27), n (%)	ChatGPT-3.5 (N=30), n (%)	ChatGPT-4 (N=30), n (%)
Exact questions	0 (0)	4 (14.81)	2 (6.67)	0 (0)
Similar questions	24 (80)	20 (74.07)	22 (73.33)	22 (73.33)
Different questions	6 (20)	3 (11.11)	6 (20)	8 (26.67)

The original predefined prompt did not request answer options to be given in any particular order. Therefore, for assessing potential bias in the correct answer allocation, 3 scenarios were modeled (Table 2):

- Any bias or preference in the correct answer allocation based on the raw chatbot output.
- Any bias or preference in the correct answer allocation based on the chatbot output when the researchers manually ordered answers into alphabetical order.
- Any bias or preference in the correct answer allocation based on a new output, where each chatbot was prompted to produce 30 new SBA questions with answers alphabetically.

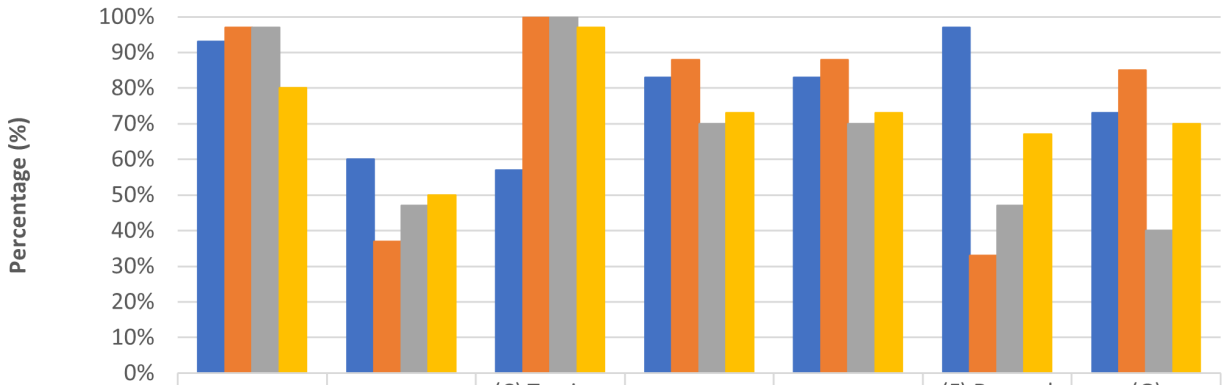
Table . Assessment of possible bias or preference in correct answer allocation.

Options	Gemini (N=30), n (%)	Bing (N=27), n (%)	ChatGPT-3.5 (N=30), n (%)	ChatGPT-4 (N=30), n (%)
Original chatbot output				
A	5 (16.67)	6 (22.22)	9 (30)	11 (36.67)
B	12 (40)	4 (14.81)	10 (33.33)	10 (33.33)
C	6 (20)	10 (37.04)	7 (23.33)	4 (13.33)
D	5 (16.67)	6 (22.22)	3 (10)	4 (13.33)
E	2 (6.67)	1 (3.7)	1 (3.33)	1 (3.33)
Manual reordering of chatbot output into alphabetical order				
A	4 (13.33)	8 (29.63)	8 (26.67)	6 (20)
B	10 (33.33)	3 (11.11)	3 (10)	7 (23.33)
C	3 (10)	5 (18.52)	7 (23.33)	5 (16.67)
D	9 (30)	4 (14.81)	6 (20)	5 (16.67)
E	4 (13.33)	7 (25.93)	6 (20)	7 (23.33)

Gemini, ChatGPT-3.5, and ChatGPT-4 occasionally provided answer options in alphabetical order when not specifically prompted. Gemini consistently demonstrated a preference for the correct answer to be listed as option B. The ChatGPT-3.5 and ChatGPT-4 appeared to favor options A, B, and C. Bing appeared to favor options A and E.

Regarding the technical item flaws among the outputs, the chatbots performed similarly in terms of following an SBA format (Figure 2A) and achieving the "cover test" satisfaction (Figure 2B), although ChatGPT-4 scored slightly lower on satisfying the cover test. Overall, Gemini performed well across most items, except for item balance. Notably, Gemini stood out by creating questions with a lead-in that relied heavily on the vignette for the answer (Figure 2F). Bing scored low across most evaluation items but performed well in generating a lead-in question of appropriate length (Figure 2G). ChatGPT Plus, which required a paid subscription, did not outperform the other chatbots in any item. The evaluation item "questions test the application of knowledge rather than recall of isolated facts" received the lowest scores across all the chatbots (Figure 2C), with Gemini achieving the highest score among them.

Figure 2. Shows technical item flaws among the chatbots: (**A**) single best answer format, (**B**) satisfy the "cover test" rule, (**C**) test the application of knowledge rather than recall isolated facts, (**D**) questions were balanced, (**E**) lead-in question tests one idea, (**F**) questions depend on the vignette to reach an answer, and (**G**) appropriate lead-in question length. The total number of questions generated by Bing was 27.



	(A) Using single-best answer format	(B) Satisfy the "cover test" rule	(C) Testing the application of knowledge	(D) Balanced structure	(E) Test a single concept	(F) Depend on the vignette to reach an answer	(G) Appropriate lead-in question length
Gemini	93%	60%	57%	83%	83%	97%	73%
Bing	97%	37%	100%	88%	88%	33%	85%
ChatGPT-3.5	97%	47%	100%	70%	70%	47%	40%
ChatGPT-4	80%	50%	97%	73%	73%	67%	70%

Assessment of Content Validity and Accuracy

Seven internal assessors evaluated item clarity and relevance, deriving the I-CVI for individual SBA items and the S-CVI (following the Universal Agreement method) to assess the overall content validity for questions from each chatbot (Table 3). Items with I-CVI>0.79 and scales with S-CVI/UA>0.8 can be interpreted as acceptable [20].

Assessors also evaluated items for content clarity and 4 elements of accuracy: vignette information, answers, justifications, and educational accuracy, on a scale from 1 to 4 (Tables 4 and 5). The Krippendorff alpha coefficient was low, 0.016, with a 95% bootstrap CI of −0.066 to 0.116.

As depicted in Tables 3 and 4, SBA questions from 3 chatbots (ChatGPT, Gemini, and ChatGPT Plus) had similar content clarity and S-CVI values. In comparison to the other chatbots, Bing performed worst in content clarity, overall (scale) validity, and all elements of item accuracy. ChatGPT Plus, which required a paid subscription, did not outperform the other chatbots except in the measure of educational accuracy. Further statistical analysis was performed using the 2-way ANOVA without replication, which showed statistically significant differences among chatbots and domains (P<.05). However, the Tukey-Kramer HSD post hoc test revealed no significant pairwise differences between individual chatbots, as all comparisons had P values>.05 and overlapping CIs. Thus, although the chatbots' performance varied overall, specific chatbot differences were not statistically significant.

Table . Item-content validity and scale-content validity across the chatbots.

Item number	Gemini	Bing	ChatGPT-3.5	ChatGPT-4
I-CVI[a]				
1	1	1	1	1
2	1	1	1	1
3	1	1	1	1
4	1	1	1	1
5	0.85	0.85	0.85	0.83
6	0.85	0.85	0.71	0.85
7	0.85	0.85	0.85	0.85
8	0.85	0.85	0.85	0.85
9	0.85	—[b]	0.85	0.85
10	0.85	0.85	0.85	0.85
S-CVI/UA[c]	0.91	0.83	0.9	0.91

[a]I-CVI: Item Content Validity Index.

[b]Not applicable.

[c]S-CVI/UA: Scale Level Content Validity Index.

Table . Average score for content clarity and accuracy of items across the chatbots.

	Content clarity[a]	Accuracy of information[b]	Accuracy of answers[c]	Accuracy of justification[d]	Educational accuracy[e]
Gemini	3.68	3.71	3.8	3.91	3.49
Bing	3.41	3.3	3.49	3.47	3.2
ChatGPT-3.5	3.75	3.71	3.84	3.9	3.5
ChatGPT-4	3.71	3.66	3.81	3.82	3.56

[a]Content clarity refers to the extent to which the question is clearly written, free of ambiguity, and easily understood by the intended audience.

[b]Accuracy of information verifies that the facts, concepts, and explanations presented are scientifically and contextually correct.

[c]Accuracy of answers ensures that the correct response is indeed accurate, while the distractors remain plausible yet distinguishable.

[d]Accuracy of justification evaluates whether the rationale provided for correct and incorrect answers is logically sound, evidence-based, and supports a deeper understanding of the topic.

[e]Educational accuracy assesses whether the question is appropriately challenging to the student level, measures higher cognitive levels (such as application or analysis), and adheres to best practices in assessment design.

Table . Two-way ANOVA table.

Source of variation	Sum of squares due to the source	*df*	Mean sum of squares due to the source	*F* test	*P* value
Average content clarity and accuracy scores	0.304357	2	0.152178	24.26587	<.001
Chatbots	17.9744	4	4.493601	716.5349	<.001
Error	0.05017	8	0.006271	—[a]	—
Total	18.32893	14	—	—	—

[a]Not applicable.

Discussion

Interpretation of Findings

This study was designed to examine the content validity and consistency of SBA questions generated by different chatbots in the context of undergraduate medical education. The findings revealed that no single chatbot excelled in all studied domains nor demonstrated a universal superiority over other chatbots, but rather showed unique strengths of some chatbots in specific areas and highlighted their notable limitations in other ones. This emphasizes the importance of critically assessing the output

of chatbots in a context-sensitive manner. Bing produced items that were least suitable for inclusion in medical student assessment. These findings echo previous studies, which also show Bing to generate less valid MCQs in comparison to other chatbots [25]. ChatGPT-4 showed the greatest variation in responses across users (suggesting higher protection against examination candidates predicting potential assessment items), and had strong performance in content clarity and accuracy, though it also exhibited some less effective question design practices, such as poorer performance in the "cover test" rule. These findings align with the results of Doughty et al [26], who found that GPT-4's ability to generate effective MCQs was nearly on par with human performance, in which 81.7% of the generated MCQs met all evaluation criteria, suggesting that fewer than 1 in 5 questions would need revision by instructors. However, in cases where ChatGPT-4 failed to meet a quality standard, this was typically the only issue with the question. Gemini performed well across all evaluations, matching ChatGPT Plus's strong index score for content validity, and excelled in creating questions where the lead-in tested 1 item and relied heavily on the vignette for the answer. Although slightly behind both ChatGPT versions in content clarity, Gemini scored the highest in providing accurate justifications for the correct answer.

This variation across chatbots is consistent with results from studies where chatbots were asked to answer questions. Kumari et al [15] found significant differences in solving hematology case vignettes using LLMs. ChatGPT achieved the highest score, followed by Google Gemini and then Microsoft Bing. In line with this, Dhanvijay et al [13] reported that ChatGPT-3.5 scored the highest, Bing the lowest, and Bard (Gemini) ranked in the middle when solving case vignettes in physiology. When chatbots were tested on their ability to answer SBA questions, ChatGPT-4 and Microsoft Copilot (Bing) outperformed Google Gemini [27]. Overall, these results suggest that OpenAI's ChatGPT shows strong potential in the medical education field. However, it is worth noting that none of the models were able to answer all questions correctly, and in our study, all platforms had some flaws when generating SBAs.

Additionally, this study's results reveal several key insights and revelations concerning SBA questions produced by AI chatbots. First, we observed that chatbots often exhibit a correct answer bias toward particular options. Recent studies have identified that LLMs tend to display positional bias when handling MCQs [28,29]. Radford et al [30] and Li and Gao [31] found that this susceptibility to positional bias is pronounced in the GPT-2 family however a more recent technical report for GPT-4 suggests AI's performance in MCQ remains susceptible to the position of the correct answer among the choices [32], a pattern referred to as "anchored bias." To minimize this inherent bias that appears to occur across AI platforms, when using AI to generate MCQ or SBA, we would recommend not stipulating an order for answer options in the prompt.

Furthermore, assessment literature emphasizes that high-quality SBA questions should assess the higher levels of Bloom's taxonomy to encourage students' critical thinking and complex problem-solving [33]. Our study revealed that chatbots were not always successful in crafting questions that engaged these advanced cognitive levels, and this was an area of relative weakness when evaluating items. Gemini scored highest, followed by ChatGPT Plus, ChatGPT-3.5, and then Bing. Similar findings regarding ChatGPT's limitations were reported by Herrmann-Werner et al [34]. Likewise, studies by Klang et al [35] and Liu et al [36] also emphasized GPT-4's limited ability to integrate knowledge and apply clinical reasoning, highlighting challenges in logical reasoning, which could limit AI's ability to generate questions that test this concept. However, it should be noted that while human-written questions were rated higher in direct comparisons, the score gap was narrow and largely insignificant, suggesting that AI tools still hold potential as educational aids [2].

Our analysis also revealed some technical flaws, variations, and inconsistencies in item construction within all chatbots. These flaws highlight instances of overconfidence and inadequacies in question design, suggesting an inability of the chatbots to evaluate their output's consistency, relevance, and complexity. Flawed MCQs hinder the accurate and meaningful interpretation of test scores and negatively impact student pass rates. Therefore, identifying and addressing technical flaws in MCQs can enhance their quality and reliability [37]. Similarly, Klang et al [35] reported that approximately 15% of questions generated using detailed prompts required corrections, primarily due to content inaccuracies or methodological shortcomings. These revisions often involved addressing a lack of sensitivity in certain topics, such as failing to include specific details such age, gender, or geographical context in the questions or answers.

Most of the questions tested recall and comprehension levels, but Gemini included some that assessed the application of knowledge. In contrast, Bing struggled to generate questions on specific topics. These findings can be explained as critical thinking at higher levels involves considering evidence, context, conceptualization, methods, and the criteria required for judgment [38]. AI models are trained on large datasets of text, but they may not fully understand the context or underlying concepts behind the content. Higher-order thinking skills, such as application, analysis, and synthesis, require deeper comprehension and reasoning that AI might not be able to simulate effectively.

Thus, using AI to generate SBAs encourages us to reconsider Bloom's taxonomy of the cognitive domains [39,40], which traditionally positions "creation" as the highest level of cognition. In the era of AI, evaluation might be considered the most critical level of cognition [41]. While AI chatbots can often produce well-written questions aligned with LOBs, they still require expert evaluation to ensure their suitability for use. Future research should compare AI-generated outputs with those from subject matter experts to assess accuracy and relevance. Evaluating AI's ability to test higher-order cognition in Bloom's taxonomy is also crucial. As AI evolves, ongoing validation is essential to ensure reliability and effectiveness in assessments.

Despite the methodological rigor and innovative approach of this study, some limitations need to be highlighted to improve the interpretation of the findings presented here. First, the researchers or assessors generated or evaluated only 30 questions per chatbot. Variation was observed in the content validity and

accuracy between the SBAs produced by an individual chatbot. Therefore, this sample may not sufficiently represent the wide range of possible outputs, potentially limiting the generalizability and robustness of the findings. Second, the accuracy of the chatbots' responses may have been compromised by the absence of reference materials, which could have negatively affected their performance. Finally, this study is limited by low interrater reliability and the use of measures are not specifically designed to assess MCQ quality. Future research should consider using validated tools to enhance evaluation accuracy.

Conclusions

Chatbot platforms varied in their ability to generate educational questions. ChatGPT models produced the most variable outputs, reducing predictability while maintaining strong content clarity and accuracy with minimal answer bias. Gemini performed similarly but showed a strong preference for 1 option, while Bing had the least variation and the lowest content clarity and accuracy. ChatGPT-4 did not significantly improve question quality but maximized variability. Technical flaws were present across all platforms, with many questions poorly linked to vignettes. Most tested recall and comprehension, though Gemini included some application-level items, whereas Bing struggled with specific topics.

These findings highlight AI's limitations in generating higher-order thinking questions, reinforcing the need for expert evaluation. This challenges Bloom's taxonomy's traditional cognitive hierarchy, suggesting that "evaluation" may be more critical than "creation" in AI-assisted assessments.

Acknowledgments

The authors extend their gratitude to the internal assessors from the School of Medicine, Ulster University, who dedicated their time to evaluating the questions.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Further data on the assessment of questions generated
[XLSX File, 35 KB - mededu_v11i1e69521_app1.xlsx]

Multimedia Appendix 2
Questions generated.
[XLSX File, 85 KB - mededu_v11i1e69521_app2.xlsx]

References

1. Pugh D, De Champlain A, Gierl M, Lai H, Touchie C. Using cognitive models to develop quality multiple-choice questions. Med Teach 2016 Aug;38(8):838-843. [doi: 10.3109/0142159X.2016.1150989] [Medline: 26998566]
2. Cheung BHH, Lau GKK, Wong GTC, et al. ChatGPT versus human in generating medical graduate exam multiple choice questions-a multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). PLoS One 2023;18(8):e0290691. [doi: 10.1371/journal.pone.0290691] [Medline: 37643186]
3. Rodriguez-Torrealba R, Garcia-Lopez E, Garcia-Cabot A. End-to-end generation of multiple-choice questions using text-to-text transfer transformer models. Expert Syst Appl 2022 Dec;208:118258. [doi: 10.1016/j.eswa.2022.118258]
4. Turing AM. I.—computing machinery and intelligence. Mind 1950 Oct 1;LIX(236):433-460. [doi: 10.1093/mind/LIX.236.433]
5. Rudolph J, Tan S, Tan S. ChatGPT: bullshit spewer or the end of traditional assessments in higher education? JALT 2023;6(1):342-363 [FREE Full text] [doi: 10.37074/jalt.2023.6.1.9]
6. Rudolph J, Tan S, Tan S. War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. JALT 2023;6(1):364-389 [FREE Full text] [doi: 10.37074/jalt.2023.6.1.23]
7. Giannakopoulos K, Kavadella A, Stamatopoulos V, Kaklamanos EG, Salim AA. Evaluation of the performance of generative AI large language models ChatGPT, Google Bard, and Microsoft Bing chat in supporting evidence-based dentistry: comparative mixed methods study. J Med Internet Res 2023 Dec 28;25:e51580. [doi: 10.2196/51580] [Medline: 38009003]
8. Morjaria L, Burns L, Bracken K, et al. Examining the threat of ChatGPT to the validity of short answer assessments in an undergraduate medical program. J Med Educ Curric Dev 2023;10:23821205231204178. [doi: 10.1177/23821205231204178] [Medline: 37780034]
9. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
10. Lai UH, Wu KS, Hsu TY, Kan JKC. Evaluating the performance of ChatGPT-4 on the United Kingdom Medical Licensing Assessment. Front Med (Lausanne) 2023;10:1240915. [doi: 10.3389/fmed.2023.1240915] [Medline: 37795422]

11. O'Connor S. Open artificial intelligence platforms in nursing education: tools for academic progress or abuse? Nurse Educ Pract 2023 Jan;66:103537. [doi: 10.1016/j.nepr.2022.103537] [Medline: 36549229]
12. Sample I. Science journals ban listing of ChatGPT as co-author on papers. The Guardian. 2023. URL: https://www.theguardian.com/science/2023/jan/26/science-journals-ban-listing-of-chatgpt-as-co-author-on-papers [accessed 2025-05-14]
13. Dhanvijay AKD, Pinjar MJ, Dhokane N, Sorte SR, Kumari A, Mondal H. Performance of large language models (ChatGPT, Bing Search, and Google Bard) in solving case vignettes in physiology. Cureus 2023 Aug;15(8):e42972. [doi: 10.7759/cureus.42972] [Medline: 37671207]
14. Koga S, Martin NB, Dickson DW. Evaluating the performance of large language models: ChatGPT and Google Bard in generating differential diagnoses in clinicopathological conferences of neurodegenerative disorders. Brain Pathol 2024 May;34(3):e13207. [doi: 10.1111/bpa.13207] [Medline: 37553205]
15. Kumari A, Kumari A, Singh A, et al. Large language models in hematology case solving: a comparative study of ChatGPT-3.5, Google Bard, and Microsoft Bing. Cureus 2023 Aug;15(8):e43861. [doi: 10.7759/cureus.43861] [Medline: 37736448]
16. Zuckerman M, Flood R, Tan RJB, et al. ChatGPT for assessment writing. Med Teach 2023 Nov;45(11):1224-1227. [doi: 10.1080/0142159X.2023.2249239] [Medline: 37789636]
17. Kıyak YS, Coşkun Ö, Budakoğlu I, Uluoğlu C. ChatGPT for generating multiple-choice questions: evidence on the use of artificial intelligence in automatic item generation for a rational pharmacotherapy exam. Eur J Clin Pharmacol 2024 May;80(5):729-735. [doi: 10.1007/s00228-024-03649-x] [Medline: 38353690]
18. Mistry NP, Saeed H, Rafique S, Le T, Obaid H, Adams SJ. Large language models as tools to generate radiology board-style multiple-choice questions. Acad Radiol 2024 Sep;31(9):3872-3878. [doi: 10.1016/j.acra.2024.06.046] [Medline: 39013736]
19. Tan LT, McAleer JJA, Final FRCR Examination Board. The introduction of single best answer questions as a test of knowledge in the final examination for the fellowship of the Royal College of Radiologists in Clinical Oncology. Clin Oncol (R Coll Radiol) 2008 Oct;20(8):571-576. [doi: 10.1016/j.clon.2008.05.010] [Medline: 18585017]
20. Case SM, Swanson DB. Writing one-best-answer questions for the basic and clinical sciences. In: Constructing Written Test Questions for the Basic and Clinical Sciences: National Board of Medical Examiners; 2016:31-66.
21. Yusoff MSB. ABC of content validation and content validity index calculation. EIMJ 2019;11(2):49-54 [FREE Full text] [doi: 10.21315/eimj2019.11.2.6]
22. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med 2006 Feb;119(2):166. [doi: 10.1016/j.amjmed.2005.10.036] [Medline: 16443422]
23. Krippendorff K. Content Analysis: An Introduction to Its Methodology, 4th edition: SAGE Publications; 2019. [doi: 10.4135/9781071878781]
24. Marzi G, Balzano M, Marchiori D. K-Alpha Calculator-Krippendorff's Alpha Calculator: a user-friendly tool for computing Krippendorff's Alpha inter-rater reliability coefficient. MethodsX 2024 Jun;12:102545. [doi: 10.1016/j.mex.2023.102545] [Medline: 39669968]
25. Agarwal M, Sharma P, Goswami A. Analysing the applicability of ChatGPT, Bard, and Bing to generate reasoning-based multiple-choice questions in medical physiology. Cureus 2023 Jun;15(6):e40977. [doi: 10.7759/cureus.40977] [Medline: 37519497]
26. Doughty J, Wan Z, Bompelli A, et al. A comparative study of AI-generated (GPT-4) and human-crafted MCQs in programming education. Presented at: ACE 2024; Jan 29 to Feb 2, 2024; Sydney, New South Wales, Australia p. 114-123 URL: https://dl.acm.org/doi/proceedings/10.1145/3636243 [accessed 2025-05-14] [doi: 10.1145/3636243.3636256]
27. Rossettini G, Rodeghiero L, Corradi F, et al. Comparative accuracy of ChatGPT-4, Microsoft Copilot and Google Gemini in the Italian entrance test for healthcare sciences degrees: a cross-sectional study. BMC Med Educ 2024 Jun 26;24(1):694. [doi: 10.1186/s12909-024-05630-9] [Medline: 38926809]
28. Pezeshkpour P, Hruschka E. Positional bias in large language models when handling multiple-choice questions. arXiv. Preprint posted online on Aug 22, 2023 URL: https://arxiv.org/abs/2308.11483 [accessed 2025-05-14] [doi: 10.48550/arXiv.2308.11483]
29. Zheng J, Li X, Wang R. Investigating option position biases in large language models. arXiv. Preprint posted online on Sep 7, 2024 URL: https://arxiv.org/abs/2309.03882 [accessed 2025-05-14] [doi: 10.48550/arXiv.2309.03882]
30. Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners. OpenAI Blog 2019;1(8):9 [FREE Full text]
31. Li R, Gao Y. Anchored answers: unravelling positional bias in GPT-2's multiple-choice questions. arXiv. Preprint posted online on May 6, 2024 URL: https://arxiv.org/abs/2405.03205 [accessed 2025-05-14] [doi: 10.48550/arXiv.2405.03205]
32. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. GPT-4 technical report. arXiv. Preprint posted online on Mar 15, 2023 URL: https://arxiv.org/abs/2303.08774 [accessed 2025-05-14] [doi: 10.48550/arXiv.2303.08774]
33. Walsh JL, Harris BHL, Smith PE. Single best answer question-writing tips for clinicians. Postgrad Med J 2017 Feb 1;93(1096):76-81. [doi: 10.1136/postgradmedj-2015-133893]
34. Herrmann-Werner A, Festl-Wietek T, Holderried F, et al. Assessing ChatGPT's mastery of Bloom's taxonomy using psychosomatic medicine exam questions: mixed-methods study. J Med Internet Res 2024 Jan 23;26:e52113. [doi: 10.2196/52113] [Medline: 38261378]

35. Klang E, Portugez P, Gross R, et al. Advantages and pitfalls in utilizing artificial intelligence for crafting medical examinations: a medical education pilot study with GPT-4. BMC Med Educ 2023 Oct 17;23(1):772. [doi: 10.1186/s12909-023-04752-w] [Medline: 37848913]
36. Liu H, Ning R, Teng Z, Liu J, Zhou Q, Zhang Y. Evaluating the logical reasoning ability of ChatGPT and GPT-4. arXiv. Preprint posted online on Apr 7, 2023 URL: https://arxiv.org/abs/2304.03439 [accessed 2025-05-14] [doi: 10.48550/arXiv.2304.03439]
37. Khan HF, Danish KF, Awan AS, Anwar M. Identification of technical item flaws leads to improvement of the quality of single best multiple choice questions. Pak J Med Sci 2013 May;29(3):715-718. [doi: 10.12669/pjms.293.2993] [Medline: 24353614]
38. Dergaa I, Chamari K, Zmijewski P, Ben Saad H. From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing. Biol Sport 2023 Apr;40(2):615-622. [doi: 10.5114/biolsport.2023.125623] [Medline: 37077800]
39. Krathwohl DR. A revision of Bloom's taxonomy: an overview. Theory Pract 2002 Nov 1;41(4):212-218. [doi: 10.1207/s15430421tip4104_2]
40. Tutkun OF, Güzel G, Köroğlu M, Ilhan H. Bloom's revised taxonomy and critics on it. Online J Couns Educ 2012;1(3):23-30.
41. Scheuer-Larsen C, Lauridsen PS. Bloom's taxonomy in the interaction between artificial intelligence and human learning. Viden.AI. URL: https://viden.ai/en/blooms-taxonomy-and-ai [accessed 2025-05-14]

Abbreviations

AI: artificial intelligence
HSD: honestly significant difference
I-CVI: Item Content Validity Index
LLM: large language model
LOB: learning objective
MCQ: multiple choice question
S-CVI: Scale Level Content Validity Index
SBA: single best answer
UKMLA: United Kingdom Medical Licensing Assessment
USMLE: United States Medical Licensing Examination

Edited by A Bahattab, B Lesselroth; submitted 02.12.24; peer-reviewed by U Hin Lai, YS Kiyak; revised version received 22.04.25; accepted 30.04.25; published 30.05.25.

Please cite as:
Abouzeid E, Wassef R, Jawwad A, Harris P
Chatbots' Role in Generating Single Best Answer Questions for Undergraduate Medical Student Assessment: Comparative Analysis
JMIR Med Educ 2025;11:e69521
URL: https://mededu.jmir.org/2025/1/e69521
doi:10.2196/69521

Role of Artificial Intelligence in Surgical Training by Assessing GPT-4 and GPT-4o on the Japan Surgical Board Examination With Text-Only and Image-Accompanied Questions: Performance Evaluation Study

Hiroki Maruyama[1], MD; Yoshitaka Toyama[2], MD, PhD; Kentaro Takanami[2], MD, PhD; Kei Takase[3], MD, PhD; Takashi Kamei[1], MD, PhD

[1]Department of Surgery, Tohoku University Graduate School of Medicine, Sendai, Japan

[2]Department of Diagnostic Radiology, Tohoku University Hospital, 1-1 Seiryo-Machi, Aoba-Ku, Sendai, Japan, Sendai, Japan

[3]Department of Diagnostic Radiology, Tohoku University Graduate School of Medicine, Sendai, Japan

Corresponding Author:
Yoshitaka Toyama, MD, PhD
Department of Diagnostic Radiology, Tohoku University Hospital, 1-1 Seiryo-Machi, Aoba-Ku, Sendai, Japan, Sendai, Japan

Abstract

Background: Artificial intelligence and large language models (LLMs)—particularly GPT-4 and GPT-4o—have demonstrated high correct-answer rates in medical examinations. GPT-4o has enhanced diagnostic capabilities, advanced image processing, and updated knowledge. Japanese surgeons face critical challenges, including a declining workforce, regional health care disparities, and work-hour-related challenges. Nonetheless, although LLMs could be beneficial in surgical education, no studies have yet assessed GPT-4o's surgical knowledge or its performance in the field of surgery.

Objective: This study aims to evaluate the potential of GPT-4 and GPT-4o in surgical education by using them to take the Japan Surgical Board Examination (JSBE), which includes both textual questions and medical images—such as surgical and computed tomography scans—to comprehensively assess their surgical knowledge.

Methods: We used 297 multiple-choice questions from the 2021 - 2023 JSBEs. The questions were in Japanese, and 104 of them included images. First, the GPT-4 and GPT-4o responses to only the textual questions were collected via OpenAI's application programming interface to evaluate their correct-answer rate. Subsequently, the correct-answer rate of their responses to questions that included images was assessed by inputting both text and images.

Results: The overall correct-answer rates of GPT-4o and GPT-4 for the text-only questions were 78% (231/297) and 55% (163/297), respectively, with GPT-4o outperforming GPT-4 by 23% (P=<.01). By contrast, there was no significant improvement in the correct-answer rate for questions that included images compared with the results for the text-only questions.

Conclusions: GPT-4o outperformed GPT-4 on the JSBE. However, the results of the LLMs were lower than those of the examinees. Despite the capabilities of LLMs, image recognition remains a challenge for them, and their clinical application requires caution owing to the potential inaccuracy of their results.

(JMIR Med Educ 2025;11:e69313) doi:10.2196/69313

KEYWORDS

LLM; ChatGPT; Japan Surgical Board Examination; surgical education; large language models; artificial intelligence; Medical Licensing Examination; diagnostic imaging

Introduction

Surgical training requires a considerable time commitment, as it includes various educational activities, on-the-job training, and supervised clinical experience [1]. In Japan, the surgery field is facing many challenges, such as the declining numbers of surgeons, regional health care disparities [2], and working-hour–related challenges [3]. Consequently, it is important to understand whether new technologies such as artificial intelligence (AI) and large language models (LLMs) can augment surgery education and training [4].

LLMs are AI systems trained on billions of words from papers, books, and other internet sources. ChatGPT—released by OpenAI in November 2022—is a generative AI chatbot that supports multimodal inputs and text generation, with a GPT as its backend [5]. ChatGPT has achieved conversational interactivity and human-like or better correct-answer rate across various fields—including the medical field [6]—suggesting that

LLM applications could be beneficial in clinical, educational, and research settings [7].

GPT-4—released in March 2023—achieved an excellent correct-answer rate for United States Medical Licensing Examination (USMLE)-style questions, exceeding the passing threshold of 60% [8]. Moreover, in the field of surgery, GPT-3.5 obtained a 65% correct-answer rate for the US General Surgery Specialist Examination [9], and GPT-4 achieved a 76% correct-answer rate for the Korean Surgical Specialist Examination [10]. However, GPT-4 does not include an image-recognition function; consequently, questions that included images were excluded from both of these studies. To the best of our knowledge, no previous study has yet evaluated the correct-answer rate of LLMs on the Japan Surgical Board Examination (JSBE).

GPT-4-Vision (GPT-4V)—an improved version of GPT-4 with image-processing capabilities [6]—can process and interpret images along with text data, extending its potential application to areas that require image analysis. When both text- and image-based questions from the USMLE were input into GPT-4V, its correct-answer rate improved from 83.6% to 90.7% [11]. However, there have been no reports on the functional evaluation of AI in the field of surgery that includes image evaluations.

GPT-4 Omni (GPT-4o)—released in May 2024—features a considerably faster processing speed than GPT-4 and includes many upgrades, such as its improved non-English-language processing and enhanced visual and speech understanding [12]. Additionally, the GPT-4o knowledge base has been updated with data up to October 2023, enabling it to offer more accurate answers based on recent information and accurate text generation [13]. Several reports have evaluated the performance of Chat-GPT4o using medical examinations, but only 3 reports have evaluated the effectiveness of image input in addition to text [14-16]. Moreover, no study has yet evaluated it in the field of surgery. In many cases, diagnostic imaging plays an important role in surgical treatment plans, and specific images—such as intraoperative imaging findings—are sometimes used. Consequently, evaluating how LLMs handle surgery-specific images is critical for understanding their current capabilities. If LLMs have a high level of knowledge related to surgery-specific images, they have the potential to be effective tools in real clinical practice and surgical education.

There have been few reports evaluating the extent to which LLMs, such as GPT, possess surgical knowledge, particularly in relation to interpreting surgical images—a skill essential for clinical decision-making. This study aims to assess and compare the performance of GPT-4 and GPT-4o on JSBE, focusing not only on general surgical knowledge but also on image recognition and diagnostic accuracy. We examined the models' responses to text-only and text-with-image questions using a retrospective evaluation design. We hypothesized that GPT-4o would outperform GPT-4, particularly on image-based questions. The findings of this study should be useful for medical educators and AI researchers seeking to understand the capabilities and limitations of LLMs in surgical education and training.

Methods

Question Dataset

This study used multiple-choice questions from the 2021 - 2023 JSBE published by the Japan Surgical Society. Each question had five possible choices, with some requiring a single answer and others requiring two. The responses for the two-answer questions were deemed correct only if both correct answers were selected. The number of answers required was specified in the input text. Electronic versions of previous papers that were available for sale were also used. The Japan Surgical Society granted permission to answer these questions.

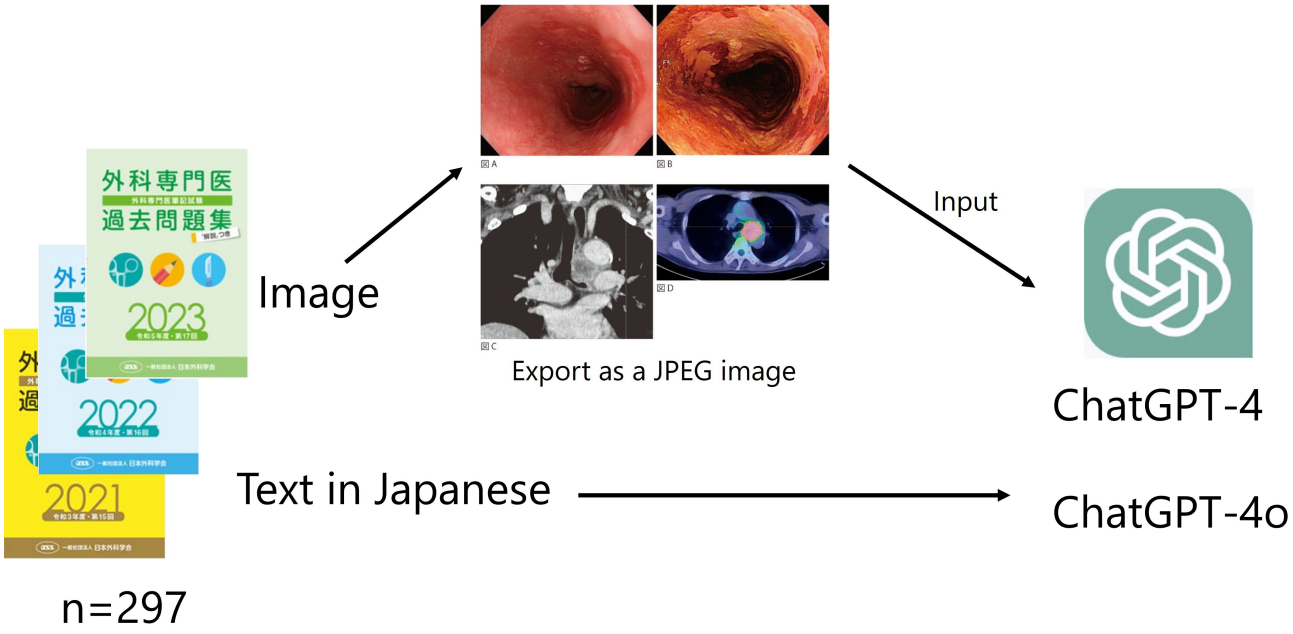
The JSBE is a multidisciplinary surgical knowledge examination designed for senior resident doctors in Japan who have completed a 3-year surgical training program. The number of examinees, successful candidates, and pass rates are listed in Table 1. There were 100 questions in each year, but 1 question in 2022 and 2 questions in 2023 were excluded as inappropriate questions, so a total of 297 questions were used in this study. The questions were presented in Japanese. To evaluate and compare responses to text-with-image questions, text-only questions were also included in the study. The text, obtained from an electronic question booklet, was entered into the models in an extensive markup language (XML) format. Moreover, screenshots of the test images were obtained from the booklet and saved in JPEG format, with their captions also being included.

Table . Annual test results of the Japan Surgical Board Examination.

Year	Examinees	Successful examinees	Pass rate (%)	Correct-answer rate (%)
2021	289	261	90.3	84.2
2022	1594	1534	96.2	92.7
2023	835	814	97.5	92.7

Questions with multiple images were exported and combined into a single image (Figure 1). The correct answers were also obtained from the electronic question booklet. The percentage of correct answers for each topic was calculated based on the number of correct answers provided by actual examinees for each question.

Figure 1. Collection of data from JSBE and input into GPT models. The questions were entered into an electronic booklet in Japanese. The images were saved as screenshots and input into ChatGPT-4o. JSBE: Japan Surgical Board Examination.



Question Classification

The questions were classified based on whether they included images. Of the 297 questions, 104 included images (text-with-image questions), and the remaining 193 were text-only questions. They were grouped into 6 categories—that is, gastrointestinal surgery (134/297, 45.1%), cardiovascular surgery (44/297, 14.8%), thoracic surgery (30/297, 10.1%), pediatric surgery (30/297, 10.1%), breast and endocrine surgery (30/297, 10.1%), and emergency anesthesiology (29/297, 9.8%). The number of image modalities and images per question was as follows: the highest number of questions included computed tomography (CT) images (44/104, 42.3%), followed by endoscopy (15/104, 14.4%), and ultrasound (13/104, 12.5%) images. Additionally, X-ray (10/104, 9.6%), radiofluoroscopy (10/104, 9.6%), magnetic resonance imaging (MRI; 8/104, 7.7%), surface of skin findings (8/104, 7.7%), positron emission tomography (6/104, 5.8%), intraoperative findings (5/104, 4.8%), pathology (5/104, 4.8%), and other modality images (5/104, 4.8%) were also included. Furthermore, 44 out of 104 questions included 1 image (42.3%), 42/104 included 2 images (40.4%), 11/104 included 3 images (10.6%), 6/104 included 4 images (5.8%), and 1/104 included 6 images (which was the maximum; 1.0%). The percentage of correct answers for each topic was calculated based on the percentage of correct answers to the individual questions and was then compared with the percentages of correct answers provided by both GPT-4 and GPT-4o.

Data Collection and Assessment

We used the GPT-4 and GPT-4o models via OpenAI's application programming interface (API) without additional fine-tuning or custom configuration. All parameters were maintained at their default setting. No pretraining or fine-tuning was conducted, and no custom persona was provided. The questions were submitted via the OpenAI API in June 2024, and the GPT-4 and GPT-4o responses were collected. The internal GPT-4o and GPT-4 versions used in this study were gpt-4o-2024-05-13 and gpt-4-turbo-2024-04-09, respectively. GPT-4o was trained on data up to October 2023, whereas GPT-4 was trained on data up to December 2023 [17].

A maximum token limit of 4096 tokens was assumed, consistent with the default for many GPT-based API end points. All other parameters—for example, the temperature, top_p, and frequency penalty—were kept at their default settings.

All test questions were presented in Japanese. To ensure consistent model behavior and clear response formatting, the following English-language prompt was placed before each question: "Please answer the following question. Indicate the symbol of the option selected by you at the end." This prompt was immediately followed by the question text in Japanese. This structure aimed to maintain response consistency across test items. The prompting strategy primarily followed a zero-shot format.

Questions with images were assessed twice—that is, once with and once without images. The answers that matched those in the question booklet were considered correct. Moreover, the percentage of correct answers was calculated for each image modality with the questions, and the percentage of correct answers was calculated based on the number of images included in the question. The percentage of correct answers for GPT-4o and GPT-4 were compared for all questions, text-only questions, and text-with-image questions, by the question category, image modality used, and number of images. For category-by-category comparisons, only the results for questions with image inputs were compared, but for the other items, the results with and without image inputs were also compared.

Statistical Analyses

McNemar test was used to compare the proportion of correct responses between the GPT-4 and GPT-4o. Fisher exact test was used for each category—that is, with or without images, lower-order thinking versus higher-order thinking, and 2 answers

versus 1 answer—to assess the GPT-4o correct-answer rate for each category. Additionally, a chi-square test was conducted to compare grades across topics. All tests were 2-tailed, and *P* values<.05 were considered significant. All *P* values were nominal and were not corrected for multiple comparisons. The statistical analyses were conducted using JMP Pro 17.0 (SAS Institute Inc).

Ethical Considerations

This study did not include human participants or patient data. All the data used in this study are publicly available. Therefore, it was excluded from review by the Institutional Review Board of Tohoku University (IRB number 11000629).

Results

Correct-Answer Rates of GPT-4 and GPT-4o

Of the 297 questions used for the text-input only test, GPT-4 answered 164 (55%) correctly and GPT-4o answered 225 (76%) correctly; thus, GPT-4o outperformed GPT-4 by 21% (*P*<.001). Additionally, when image inputs were performed for the 104 text-with-image questions out of the 297 questions, GPT-4o outperformed GPT-4 by 23% (*P*<.001), providing 231 (78%) and 163 (55%) correct answers, respectively. Comparisons of their correct-answer rates for individual groups showed that GPT-4o provided significantly more correct answers for image-based questions (GPT-4o 67% vs GPT-4 45%; *P*=<.002), text-only questions (83% vs 61%; *P*=<.001), digestive surgery (70% vs 42%; *P*=<.001), cardiovascular surgery (98% vs 68%; *P*<.00031), and breast and endocrine surgery (93% vs 67%; *P*=.0047). However, no significant differences (GPT-4o vs GPT-4) were evident between their correct-answer rates for questions related to thoracic surgery (63% vs 57%; *P*=.41), emergency surgery and anesthesia (86% vs 66%; *P*=.06), and pediatric surgery (73% vs 70%; *P*=.71). Notably, GPT-4o provided more correct responses than GPT-4 (Table 2).

Table . GPT-4 and GPT-4o correct-answer rates for the Japan Surgical Board Examination.[a]

Question type	Number of questions	Image input	Correct-answer rate		*P* value
			GPT-4, n (%)	GPT-4o, n (%)	
All questions	297	–	164 (55)	225 (76)	.001
		+	163 (55)	231 (78)	.001
Text-with-image questions	104	–	46 (44)	64 (62)	.002
		+	47 (45)	70 (67)	.<001
Text-only questions	193	–	118 (61)	161 (83)	.<001
Topic					
Digestive surgery	134	+	56 (42)	94 (70)	.<001
Cardiovascular surgery	44	+	30 (68)	43 (98)	.<001
Thoracic surgery	30	+	17 (57)	19 (63)	.41
Pediatric surgery	30	+	21 (70)	22 (73)	.71
Breast and endocrine surgery	30	+	20 (67)	28 (93)	.005
Emergency and anesthesia	29	+	19 (66)	25 (86)	.06

[a]Data are presented as the number of correct answers.

GPT-4o's Correct-Answer Rate on Text-With-Image Questions Compared With its Rate on Text-Only Questions

Even when image inputs were used for the text-with-image questions, GPT-4o provided 67% correct answers, compared to 83% for the text-only questions, indicating a statistically significant difference (*P*<.002).

Correct-Answer Rate With and Without Image Input

The percentages of correct responses provided by both models with and without image inputs were compared for 104 text-with-image questions—here, GPT-4o provided correct-answer rates of 62% and 67% with and without image inputs (*P*=.2), respectively, whereas GPT-4 provided correct-answer rates of 44% and 45% with and without image inputs (*P*=.86; Table 3).

Table . GPT-4 and GPT-4o correct-answer rates based on image-input and no image-input questions.[a]

Large language model	Input image, n (%)[b]	No input image, n (%)[b]	*P* value
GPT-4	47 (45)	46 (44)	.86
GPT-4o	70 (67)	64 (62)	.20

[a]Data are presented as the number of correct answers. Values in parentheses indicate the percentage of correct responses.

[b]The percentage indicates the percentage of correct answers to the 104 text-with-image questions.

Correct-Answer Rate Comparison of GPT-4 and GPT-4o by Category

GPT-4 provided the highest percentage of correct answers for pediatric-surgery questions (70%) and the lowest for gastrointestinal-surgery questions (19%; *P*=.0027). By contrast, GPT-4o provided the highest percentage of correct answers for cardiovascular-surgery questions (98%) and the lowest for thoracic-surgery questions (63%; *P*=.002). The correct-answer rate for the examinees referred to here is the correct-answer rate for all examinees from 2021 to 2023 (Table 4).

Table . GPT-4o, GPT-4, and examinees' correct-answer rates across various categories.

Topic	Number of questions	Text-with-image questions, n (%)	Correct-answer rate (%)		
			GPT-4	GPT-4o	Examinees
All questions	297	104 (35)	55	78	90
Digestive surgery	134	43 (32)	42	70	89
Cardiovascular surgery	44	19 (43)	68	98	91
Thoracic surgery	30	17 (57)	57	63	88
Pediatric surgery	30	11 (37)	70	73	92
Breast and endocrine surgery	30	7 (23)	67	93	90
Emergency and anesthesia	29	7 (24)	66	86	91
P value	—[a]	—[a]	.003	.<001	

[a]Not applicable.

Comparison of GPT-4 and GPT-4o Responses by Image Modality and Number of Figures

Using the text-with-image questions, the correct-answer rates for GPT-4 and GPT-4o were compared using various imaging modalities and images. GPT-4 provided the highest percentage of correct answers for questions on radiofluoroscopy and inspection (70% and 75%, respectively), whereas GPT-4o provided the highest percentage of correct answers for radiofluoroscopy and ultrasound (80% and 92%, respectively). By contrast, the correct-answer rates of the models were low for questions that included intraoperative and pathological findings—that is, they were 20% and 40% for GPT-4, respectively, and 40% for GPT-4o for both intraoperative and pathological findings. Moreover, a weak negative correlation was evident between the number of images and the percentage of correct answers, but it was not statistically significant (Table 5).

Table . Correct-answer rate comparisons based on imaging modality and number of images.[a]

Variables	n	Correct-answer rate, n (%)			
		GPT-4		GPT-4o	
		Image input +	Image input −	Image input +	Image input −
Imaging modality					
Text-with-image questions	104	47 (45)	46 (44)	70 (67)	64 (62)
XP[b]	10	5 (50)	4 (40)	6 (60)	5 (50)
Radiofluoroscopy	10	7 (70)	6 (60)	8 (80)	6 (60)
Ultrasound	13	6 (46)	8 (62)	12 (92)	10 (77)
CT[c]	60	25 (42)	27 (45)	39 (65)	37 (62)
MRI[d]	8	5 (63)	2 (25)	5 (63)	4 (50)
PET[e]	6	4 (67)	2 (33)	3 (50)	2 (33)
Endoscopy	15	9 (60)	8 (53)	9 (60)	11 (73)
Surface of skin findings	8	6 (75)	4 (50)	5 (63)	6 (75)
Intraoperative findings	5	1 (20)	2 (40)	2 (40)	3 (60)
Pathology	5	2 (40)	2 (40)	2 (40)	2 (40)
Other	5	2 (40)	3 (60)	5 (100)	4 (80)
Number of figures					
1	44	19 (43)	18 (41)	28 (64)	26 (59)
2	42	18 (43)	19 (45)	32 (76)	27 (64)
3	11	7 (64)	8 (73)	10 (91)	9 (82)
4	6	3 (50)	1 (17)	1 (17)	1 (17)
6	1	0 (0)	0 (0)	0 (0)	0 (0)

[a]Data are presented as the number of correct answers. Values in parentheses indicate the percentage of correct responses.

[b]XP: X-ray photograph.

[c]CT: computed tomography.

[d]MRI: magnetic resonance imaging.

[e]PET: positron emission tomography.

Discussion

Principal Findings

GPT-4o significantly outperformed GPT-4 across all evaluated categories. However, neither GPT-4 nor GPT-4o achieved examinee-level accuracy for any question (Table 1). The correct-answer rates for text-only questions were higher than those for text-with-image questions for both models. Moreover, the inclusion of image inputs did not lead to a significant improvement in performance on text-with-image questions (Table 3). Performance varied by image type, with particularly low correct-answer rates for questions involving intraoperative and pathological images. By contrast, the correct-answer rates were relatively higher for radiological images such as CT and MRI images.

The results showed that there was no significant difference in the percentage of correct responses between GPT-4 and GPT-4o for thoracic surgery, emergency and anesthesia, and pediatric surgery. When comparing GPT-4o to the results of examinees, both demonstrated similarly low correct-answer rates for questions related to thoracic and gastrointestinal surgery. There was no consistent pattern evident in terms of which surgical category exhibited the highest correct response rate.

Additional Analysis by Problem Type

In terms of category-specific differences, the percentage of correct responses for both GPT-4 and GPT-4o did not differ significantly for thoracic, emergency and anesthesia, and pediatric surgery. A comparison of GPT-4o and examinee correct-answer rates demonstrated similarly low correct response rates for questions related to the thoracic and gastrointestinal surgeries. However, no consistent trend was evident with the highest percentage of correct responses. The correct-answer rate was low for thoracic surgery because questions in this field comprised a high proportion of text-with-image questions. By contrast, it was high for breast and endocrine surgery and emergency and anesthesiology, which comprised fewer

text-with-image questions. However, the opposite trend was evident for cardiovascular surgery, where no consistent trend was evident in the correct-answer rates for text-with-image questions. This result could be attributed to the fact that many of the questions could be answered correctly without image recognition, or that many of the images were easy to understand even when image recognition was required.

Responses to Intraoperative Imaging Problems

An additional study on the correct-answer rate was conducted to assess the differences in the GPT model responses based on the imaging modality. This is only a hypothesis, but owing to the small sample size, as 1 question contained several types of images, the correct-answer rate was more than 20% lower than the overall GPT-4 correct-answer rates for questions involving intraoperative findings and those of GPT-4o for intraoperative and pathological images compared with those for radiological modalities such as CT and MRI (Table 4). Only questions involving intraoperative findings had a correct-answer rate after image input that was more than 20% lower than the average for both GPT-4 and GPT-4o, which could be considered to be a GPT image-recognition weakness. Additionally, the responses to intraoperative images were evaluated individually (Multimedia Appendix 1). Although liver resection was identified in intraoperative liver-resection images, the actual resection was misidentified. Moreover, in images of mediastinal tumors, the tumor and recurrent nerve were either not mentioned or could not be identified, whereas from the intraoperative inguinal-hernia images, the arteriovenous vein in the inferior abdominal wall was misidentified as the vas deferens.

Implications of Findings

From this research, it is evident that GPT-4o significantly outperformed GPT-4 across all evaluated categories, indicating that OpenAI's model development is progressing steadily. However, despite these improvements, neither GPT-4 nor GPT-4o achieved the correct-answer rates of actual examinees. This highlights that current LLMs, while advancing rapidly, still fall short of the reliability required for high-stakes clinical decision-making or licensing-level assessments. In particular, GPT-4o exhibited lower accuracy on text-with-image questions compared to text-only questions, and the inclusion of image inputs did not significantly improve its performance. This reflects an ongoing limitation in the image recognition capabilities of LLMs, especially for complex visuals such as intraoperative and pathological images, and suggests that caution is warranted when considering these models for clinical use.

However, this study was conducted without pre-tuning, and the accuracy of LLMs could be potentially improved by tuning them in a field-specific manner [7]. Pretuning them on data from medical textbooks and previous examinations could enhance the relevance and accuracy of their responses. Pretraining has the potential to improve model performance, but the process can be complicated and is not supported by some models. The results of the models that did not undergo pretraining can be said to be results that can be applied to general readers and various models.

A "Socratic tutor mode" educational application of GPT-4o has been reported, wherein the complexity of medical questions can be changed during the conversation based on the learner's understanding [7]. In this study, GPT-4o provided a high percentage of correct answers to text-only questions, which could be used for learning guideline content and obtaining general surgical knowledge, where the answers are clear to residents and majors studying surgery. Additionally, if its image-recognition capabilities improve in the future and it becomes possible to diagnose intraoperative images specific to surgery with a high degree of confidence, it could become a useful indicator when making decisions in daily clinical practice.

Comparison to the Literature

Previous studies have shown that the GPT-4 correct-answer rate could be improved for USMLE by using images to complement the text input [11]. However, similar to our study, researchers have reported that inputting image information did not increase the percentage of correct answers [16,18-22]. Additionally, a previous study has suggested that GPT-4 prioritizes verbal information over images [20].

Previous reports [23] have shown that ChatGPT is able to provide more accurate responses when given English-language input than non-English-language input, but it has also been shown that the correct-answer rate of GPT-4o has improved when given Japanese-language inputs [12]. Moreover, the results of this study reflect this fact, which is consistent with the findings of a previous study on radiology [14], highlighting GPT-4o's enhanced reasoning and better responsiveness to Japanese inputs, thereby successfully addressing the limitations of earlier variants.

Strengths and Limitations

It should be noted that, unlike previous studies which relied on recalled questions or researcher-derived answers [9,10], commercially available past examinations were used in this study to ensure a more accurate and reliable assessment, making it a strong point of our research. Nonetheless, this study had several limitations. First, the LLMs were only asked each question once; however, LLMs are generative models, often referred to as "probabilistic parroting" models [24]. This is because they generate answers based on the probability of selecting the most appropriate word from the training data. Consequently, different answers can be returned for the same question with a certain probability when asked multiple times [25]. To address this problem, it is necessary to ask the same question multiple times and assess the degree to which the answers fluctuate. Second, ChatGPT responses can be interspersed with answers based on false evidence or factual errors, commonly referred to as "hallucinations," which is the phenomenon of asserting incorrect content as if it were correct [26,27]. Even though such responses can be determined to be false by specialists, they can confuse doctors during training. This phenomenon occurs even as the correct-answer rate of the model improves—that is, the greater the confidence in responses, the more difficult it can become to identify incorrect information. Third, LLMs—including ChatGPT—are updated periodically, which could alter their correct-answer rate or incur unexpected pretraining as the questions are entered.

Consequently, the reproducibility of test results in future studies remains uncertain. Finally, we used a relatively small number of questions, which could have resulted in an inadequate analysis, particularly for the category-specific correct-answer rate. The differences in the correct-answer rate between GPT-4o and GPT-4 were substantial in the fields of cardiovascular surgery, digestive surgery, and breast and endocrine surgery; however, differences in other fields were minimal, which could be attributed to the limited sample size. If the JSBE is able to obtain more high-quality problems as it continues to hold more events, it could be possible to evaluate models with an even greater correct-answer rate.

In conclusion, GPT-4o outperformed GPT-4 for the JSBE. Although there is still room for improvement in image recognition and clinical applications, which should be approached with caution, the results suggest that improved models and pretraining could provide LLMs with more accurate medical knowledge and enhance their clinical judgment, which could be useful in enhanced learning for surgeons.

Acknowledgments

The authors thank the Japan Surgical Society for granting permission to use the official Japan Surgical Board Examination's well-thought-out and high-quality questions for this study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
LLM's response to intraoperative findings.
[DOCX File, 14 KB - mededu_v11i1e69313_app1.docx]

References

1. Debas HT, Bass BL, Brennan MF, et al. American Surgical Association Blue Ribbon Committee Report on Surgical Education: 2004. Ann Surg 2005 Jan;241(1):1-8. [doi: 10.1097/01.sla.0000150066.83563.52] [Medline: 15621984]
2. Overview of statistics on doctors, dentists [Article in Japanese]. Ministry of Health Labour and Welfare. 2024. URL: https://www.mhlw.go.jp/toukei/saikin/hw/ishi/22/index.html [accessed 2025-07-16]
3. Work style reform for doctors [Article in Japanese]. Ministry of Health, Labour and Welfare. 2024. URL: https://www.mhlw.go.jp/content/10800000/001129457.pdf [accessed 2025-07-16]
4. Varas J, Coronel BV, Villagrán I, et al. Innovations in surgical training: exploring the role of artificial intelligence and large language models (LLM). Rev Col Bras Cir 2023;50:e20233605. [doi: 10.1590/0100-6991e-20233605-en] [Medline: 37646729]
5. ChatGPT. Open AI. 2024. URL: https://openai.com/chatgpt/ [accessed 2025-07-16]
6. Open AI, Achiam J, Adler S, et al. GPT-4 technical report. arXiv. Preprint posted online on Mar 4, 2024. Preprint posted online on Mar 4, 2024 URL: https://arxiv.org/pdf/2303.08774 [accessed 2025-07-25]
7. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
8. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
9. Tran CG, Chang J, Sherman SK, De Andrade JP. Performance of ChatGPT on American Board of Surgery In-Training Examination preparation questions. J Surg Res 2024 Jul;299:329-335. [doi: 10.1016/j.jss.2024.04.060] [Medline: 38788470]
10. Oh N, Choi GS, Lee WY. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. Ann Surg Treat Res 2023 May;104(5):269-273. [doi: 10.4174/astr.2023.104.5.269] [Medline: 37179699]
11. Yang Z, Yao Z, Tasmin M, et al. Performance of multimodal GPT-4V on USMLE with image: potential for imaging diagnostic support with explanations. Radiology and Imaging. Preprint posted online on Nov 5, 2023. [doi: 10.1101/2023.10.26.23297629]
12. Hello GPT-4o. OpenAI. 2024. URL: https://openai.com/index/hello-gpt-4o/ [accessed 2025-07-16]
13. Zhu N, Zhang N, Shao Q, Cheng K, Wu H. OpenAI's GPT-4o in surgical oncology: revolutionary advances in generative artificial intelligence. Eur J Cancer 2024 Jul;206:114132. [doi: 10.1016/j.ejca.2024.114132] [Medline: 38810316]
14. Oura T, Tatekawa H, Horiuchi D, et al. Diagnostic accuracy of vision-language models on Japanese diagnostic radiology, nuclear medicine, and interventional radiology specialty board examinations. Jpn J Radiol 2024 Dec;42(12):1392-1398. [doi: 10.1007/s11604-024-01633-0] [Medline: 39031270]

15. Hayden N, Gilbert S, Poisson LM, Griffith B, Klochko C. Performance of GPT-4 with Vision on text- and image-based ACR diagnostic radiology in-training examination questions. Radiology 2024 Sep;312(3):e240153. [doi: 10.1148/radiol.240153] [Medline: 39225605]
16. Liu CL, Ho CT, Wu TC. Custom GPTs enhancing performance and evidence compared with GPT-3.5, GPT-4, and GPT-4o? A study on the Emergency Medicine Specialist Examination. Healthcare (Basel) 2024 Aug 30;12(17):1726. [doi: 10.3390/healthcare12171726] [Medline: 39273750]
17. OpenAI. Models. 2024. URL: https://platform.openai.com/docs/models/models [accessed 2025-07-16]
18. Nakajima N, Fujimori T, Furuya M, et al. A comparison between GPT-3.5, GPT-4, and GPT-4V: can the large language model (ChatGPT) pass the Japanese Board of Orthopaedic Surgery Examination? Cureus 2024 Mar;16(3):e56402. [doi: 10.7759/cureus.56402] [Medline: 38633935]
19. Takagi S, Koda M, Watari T. The performance of ChatGPT-4V in interpreting images and tables in the Japanese Medical Licensing Exam. JMIR Med Educ 2024 May 23;10:e54283. [doi: 10.2196/54283] [Medline: 38787024]
20. Hirano Y, Hanaoka S, Nakao T, et al. GPT-4 Turbo with vision fails to outperform text-only GPT-4 Turbo in the Japan Diagnostic Radiology Board Examination. Jpn J Radiol 2024 Aug;42(8):918-926. [doi: 10.1007/s11604-024-01561-z] [Medline: 38733472]
21. Ishida K, Arisaka N, Fujii K. Analysis of responses of GPT-4 V to the Japanese National Clinical Engineer Licensing Examination. J Med Syst 2024 Sep 11;48(1):83. [doi: 10.1007/s10916-024-02103-w] [Medline: 39259341]
22. Sawamura S, Kohiyama K, Takenaka T, Sera T, Inoue T, Nagai T. Performance of ChatGPT 4.0 on Japan's National Physical Therapist Examination: a comprehensive analysis of text and visual question handling. Cureus 2024 Aug;16(8):e67347. [doi: 10.7759/cureus.67347] [Medline: 39310431]
23. Harigai A, Toyama Y, Nagano M, et al. Response accuracy of GPT-4 across languages: insights from an expert-level diagnostic radiology examination in Japan. Jpn J Radiol 2025 Feb;43(2):319-329. [doi: 10.1007/s11604-024-01673-6] [Medline: 39466356]
24. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big. Presented at: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency; Mar 1, 2021 p. 610-623. [doi: 10.1145/3442188.3445922]
25. Krishna S, Bhambra N, Bleakney R, Bhayana R. Evaluation of reliability, repeatability, robustness, and confidence of GPT-3.5 and GPT-4 on a radiology board–style examination. Radiology 2024 May;311(2):e232715. [doi: 10.1148/radiol.232715] [Medline: 38771184]
26. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus 2023 Feb;15(2):e35179. [doi: 10.7759/cureus.35179] [Medline: 36811129]
27. Nakaura T, Ito R, Ueda D, et al. The impact of large language models on radiology: a guide for radiologists on the latest innovations in AI. Jpn J Radiol 2024 Jul;42(7):685-696. [doi: 10.1007/s11604-024-01552-0] [Medline: 38551772]

Abbreviations

AI: artificial intelligence
API: application programming interface
CT: computed tomography
JSBE: Japan Surgical Board Examination
LLM: large language model
MRI: magnetic resonance imaging
USMLE: United States Medical Licensing Examination

Edited by B Lesselroth; submitted 02.12.24; peer-reviewed by F Sendra-Portero, K Muraoka, S Pal; revised version received 07.05.25; accepted 01.06.25; published 30.07.25.

Please cite as:
Maruyama H, Toyama Y, Takanami K, Takase K, Kamei T
Role of Artificial Intelligence in Surgical Training by Assessing GPT-4 and GPT-4o on the Japan Surgical Board Examination With Text-Only and Image-Accompanied Questions: Performance Evaluation Study
JMIR Med Educ 2025;11:e69313
URL: https://mededu.jmir.org/2025/1/e69313
doi:10.2196/69313

Performance Evaluation of 18 Generative AI Models (ChatGPT, Gemini, Claude, and Perplexity) in 2024 Japanese Pharmacist Licensing Examination: Comparative Study

Hiroyasu Sato[1,2]; Katsuhiko Ogasawara[3,4], MBA, PhD; Hidehiko Sakurai[2], PhD

[1]Department of Pharmacy, Abashiri-Kosei General Hospital, Abashiri, Japan

[2]Graduate School of Pharmacy, Hokkaido University of Science, 7-Jo 15-4-1 Maeda, Teine, Sapporo, Japan

[3]Graduate School of Health Sciences, Hokkaido University, Sapporo, Japan

[4]Graduate School of Engineering, Muroran Institute of Technology, Muroran, Japan

Corresponding Author:
Hidehiko Sakurai, PhD
Graduate School of Pharmacy, Hokkaido University of Science, 7-Jo 15-4-1 Maeda, Teine, Sapporo, Japan

Abstract

Background: Generative artificial intelligence (AI) has shown rapid advancements and increasing applications in various domains, including health care. Previous studies have evaluated AI performance on medical license examinations, primarily focusing on ChatGPT. However, the availability of new online chat-based large language models (OC-LLMs) and their potential utility in pharmacy licensing examinations remain underexplored. Considering that pharmacists require a broad range of expertise in physics, chemistry, biology, and pharmacology, verifying the knowledge base and problem-solving abilities of these new models in Japanese pharmacy examinations is necessary.

Objective: This study aimed to assess the performance of 18 OC-LLMs released in 2024 in the 107th Japanese National License Examination for Pharmacists (JNLEP). Specifically, the study compared their accuracy and identified areas of improvement relative to earlier models.

Methods: The 107th JNLEP, comprising 345 questions in Japanese, was used as a benchmark. Each OC-LLM was prompted by the original text-based questions, and images were uploaded where permitted. No additional prompt engineering or English translation was performed. For questions that included diagrams or chemical structures, the models incapable of image input were considered incorrect. The model outputs were compared with publicly available correct answers. The overall accuracy rates were calculated based on subject area (pharmacology and chemistry) and question type (text-only, diagram-based, calculation, and chemical structure). Fleiss' κ was used to measure answer consistency among the top-performing models.

Results: Four flagship models—ChatGPT o1, Gemini 2.0 Flash, Claude 3.5 Sonnet (new), and Perplexity Pro—achieved 80% accuracy, surpassing the official passing threshold and average examinee score. A significant improvement in the overall accuracy was observed between the early and the latest 2024 models. Marked improvements were noted in text-only and diagram-based questions compared with those of earlier versions. However, the accuracy of chemistry-related and chemical structure questions remains relatively low. Fleiss' κ among the 4 flagship models was 0.334, which suggests moderate consistency but highlights variability in more complex questions.

Conclusions: OC-LLMs have substantially improved their capacity to handle Japanese pharmacists' examination content, with several newer models achieving accuracy rates of >80%. Despite these advancements, even the best-performing models exhibit an error rate exceeding 10%, underscoring the ongoing need for careful human oversight in clinical settings. Overall, the 107th JNLEP will serve as a valuable benchmark for current and future generative AI evaluations in pharmacy licensing examinations.

(JMIR Med Educ 2025;11:e76925) doi:10.2196/76925

KEYWORDS

generative AI; artificial intelligence; ChatGPT; Gemini; pharmacist; National License Examination

Introduction

Generative artificial intelligence (AI) development has been remarkable in recent years and has been adopted in many fields, including education and health care. There have been reports that generative AI has been used to summarize clinical texts [1-4] and has been introduced into clinical practice [5,6]. Furthermore, the potential benefits of generative AI in medical education have been explored [7-10], and its usefulness has been demonstrated in the writing and publishing of medical research [11].

In the United States, generative AI has been implemented in 86% of health care organizations [12]. Moreover, approximately 40% of health care professionals use generative AI at their workplaces at least once a week [13]. Correspondingly, online chat-based large language models (OC-LLM) have attracted the attention of many users because of their ease of use. In health care, the use of OC-LLMs can have serious consequences if their performance is inadequate. Therefore, verifying the knowledge base and problem-solving capabilities of OC-LLMs in health care settings is essential.

A wealth of information is available on the web in the medical and health care domains, and OC-LLMs acquire a substantial amount of knowledge during pretraining. In addition to general medical knowledge, pharmacists must have expertise in fields, such as physics and chemistry, which differ from those required by other health care professionals. However, few studies have evaluated the performance of OC-LLMs in pharmacies. The performance of ChatGPT (GPT-3.5 and GPT-4V models) in the Japanese National License Examination for Pharmacists (JNLEP) was evaluated by Sato and Ogasawara [14]. Since then, numerous new OC-LLM services and models have been released in 2024. However, the performance of these newly released models in the field of pharmacy has not been sufficiently evaluated. Furthermore, it was hypothesized that each OC-LLM service (ie, ChatGPT, Gemini, Claude, and Perplexity) has distinct strengths and limitations.

Accordingly, the purpose of this study is to evaluate the performance of various OC-LLMs introduced in 2024 in the field of pharmacy using the JNLEP and to assess performance improvements in the latest models.

Methods

Services and Models

The following 18 OC-LLMs, all available as of 2024, were evaluated (Table 1): ChatGPT (7 models), Gemini (4 models), Claude (5 models), and Perplexity (2 models). Claude 3.5 Sonnet (new) was renamed as Claude 3.5 Sonnet in June 2024, and as of January 2025, these models are the most commonly used OC-LLMs. Microsoft Copilot, one of the most popular OC-LLMs [15], was excluded because its underlying engine, GPT-4 (released in 2023), was evaluated in a previous study as the model used in ChatGPT and is mainly used for tasks other than digital browser-based dialogues. Although Copilot has continued to improve in terms of functionality and performance, the details of its current model and update history remain undisclosed. Consequently, this was excluded from the 2024 OC-LLM performance evaluation in this study.

Table . Characteristics, release dates, and evaluation dates of the OC-LLM[a] services and models used in this study.

Service and model	Deprecated or active[b]	Uploadable image	Release date[c]	Evaluation date[d]
ChatGPT				
GPT-3.5	Deprecated	No	November 2022	May 2024
GPT-4	Active	Yes	September 2023	November 2023
GPT-4o mini	Active	No	July 2024	July 2024
GPT-4o	Active	Yes	May 2024	May 2024
o1 mini	Active	No	September 2024	October 2024
o1 preview	Deprecated	No	September 2024	September 2024
o1	Active	Yes	December 2024	December 2024
Gemini				
1.0 Pro	Deprecated	Yes	February 2024	May 2024
1.5 Pro	Active	Yes	May 2024	May 2024
1.5 Flash	Active	Yes	May 2024	August 2024
2.0 Flash Experimental	Active	Yes	December 2024	December 2024
Claude				
3 Haiku	Deprecated	Yes	March 2024	June 2024
3 Sonnet	Deprecated	Yes	March 2024	May 2024
3 Opus	Active	Yes	March 2024	June 2024
3.5 Sonnet	Active	Yes	June 2024	June 2024
3.5 Sonnet (new)	Active	Yes	October 2024	November 2024
Perplexity				
Standard	Active	No	June 2024	November 2024
Pro	Active	Yes	June 2024	December 2024

[a]OC-LLM: online chat-based large language model.

[b]Status of each model, whether deprecated or active as of January 1, 2025.

[c]Release data of each used model in Japan.

[d]Performance evaluation date of each model used in this study.

Japanese National License Examination for Pharmacists

This study used 345 questions from the 107th JNLEP held in February 2022. This dataset is the same as that used by Sato and Ogasawara [14]. The questions in the 107th JNLEP are organized into the following 9 subject categories: physics, chemistry, biology, hygiene, pharmacology, pharmaceuticals, pathophysiology, regulations, and practice. All questions were presented in a multiple-choice format, requiring the selection of 1 or 2 correct answers from the 5 options. The passing criteria for the 107th JNLEP included an overall accuracy of at least 62.9% along with 2 additional conditions. The details of the 107th JNLEP were extensively covered by Sato and Ogasawara [14].

Data Measurement

All OC-LLMs, except for ChatGPT GPT-4, were evaluated for their performance from May to December 2024. The data outcomes for ChatGPT GPT-4 were collected from a preliminary study conducted in November 2023 [14]. For ChatGPT GPT-3.5, a preliminary test was conducted in February 2023. However, a new evaluation was conducted in May 2024 to assess the potential performance improvements in the same model.

For all models, the complete set of questions from the 107th JNLEP was input in Japanese in order of the question numbers. Although response performance can be improved through prompt engineering [16-18], no prompts were used in this study.

For questions that included diagrams or charts, the questions and options were input as text, whereas the diagram or chart portion was input as an image. Some early models could not process the diagrams (Table 1); therefore, these questions were omitted and marked as incorrect.

Data Analysis

The output from each OC-LLM was compared with publicly available correct answers [19] to determine whether the responses were correct or incorrect. An incorrect answer (ie, hallucinations) was defined as a response in which the selected answer differed from the published correct answer, the specified number of answers was not selected, or no answer was provided.

Even when the correct option number could not be explicitly identified in the output by the OC-LLMs, the response was considered correct if the selected content matched the correct answer choice. The accuracy of each model was evaluated based on the total number of subjects and question types (text only, including diagrams, calculations, chemical structures, and graphs). Question-type classification was subjectively determined by the researcher based on the content of the questions. Questions with diagrams were also counted as those containing graphs or chemical structures. The calculated questions included text-only and diagram-based questions.

To assess improvements in model accuracy, statistical comparisons were performed between the 3 model outputs (ChatGPT GPT-4, Gemini 1.0 Pro, and Claude 3 Sonnet) released in early 2024 and those of the latest 4 flagship models (ChatGPT o1, Gemini 2.0 Flash Experimental, Claude 3.5 Sonnet [new], and Perplexity Pro).

Answer consistency was used to validate whether the tasks in which the generative AI model excelled or struggled showed similar trends across models based on the highest accuracy model of each service.

Statistical Analysis

A generalized linear mixed model (GLMM) was used to evaluate the accuracy improvements. The correctness of the responses to each question was set as the dependent variable. The model type (early or latest), question type (text-based or diagram-based), and their interactions were specified as fixed effects. Models and questions were included as random effects. Fleiss' κ [20] was used to assess the consistency of responses. All statistical analyses were performed using R (version 4.4.2; R Foundation for Statistical Computing).

Results

Performance Statistics of AI Models

The performances of 18 generative AI models from 2024 in the pharmaceutical field were evaluated (Table 2). The performance of the 4 flagship models (ChatGPT o1, Gemini 2.0 Flash Experimental, Claude 3.5 Sonnet [new], and Perplexity Pro) was over 80%, which was markedly higher than that of the passing criteria. When reassessed, ChatGPT GPT-3.5 recorded an overall accuracy of 38.8% (134/345), indicating no marked progress from its former performance of 35.4% (122/345), showing no substantial improvement.

Table . Overall accuracy of each OC-LLM[a] on the 107th JNLEP[b].

Service and model	Correct answers[c]	Overall accuracy	Passing criteria[d]
ChatGPT			
GPT-3.5	134	0.388	Failed
GPT-4o mini	215	0.623	Failed
o1 mini	234	0.678	Passed
GPT-4[e]	250	0.724	Passed
o1 preview	272	0.788	Passed
GPT-4o	294	0.852	Passed
o1	299	0.866	Passed
Gemini			
1.0 Pro	171	0.495	Failed
1.5 Flash	242	0.701	Passed
1.5 Pro	246	0.713	Passed
2.0 Flash	288	0.834	Passed
Claude			
3 Sonnet	194	0.562	Failed
3 Haiku	213	0.617	Failed
3 Opus	260	0.753	Passed
3.5 Sonnet	293	0.849	Passed
3.5 Sonnet (new)	297	0.860	Passed
Perplexity			
Standard	228	0.660	Passed
Pro	301	0.872	Passed

[a]OC-LLM: online chat-based large language.

[b]JNLEP: Japanese National License Examination for Pharmacists.

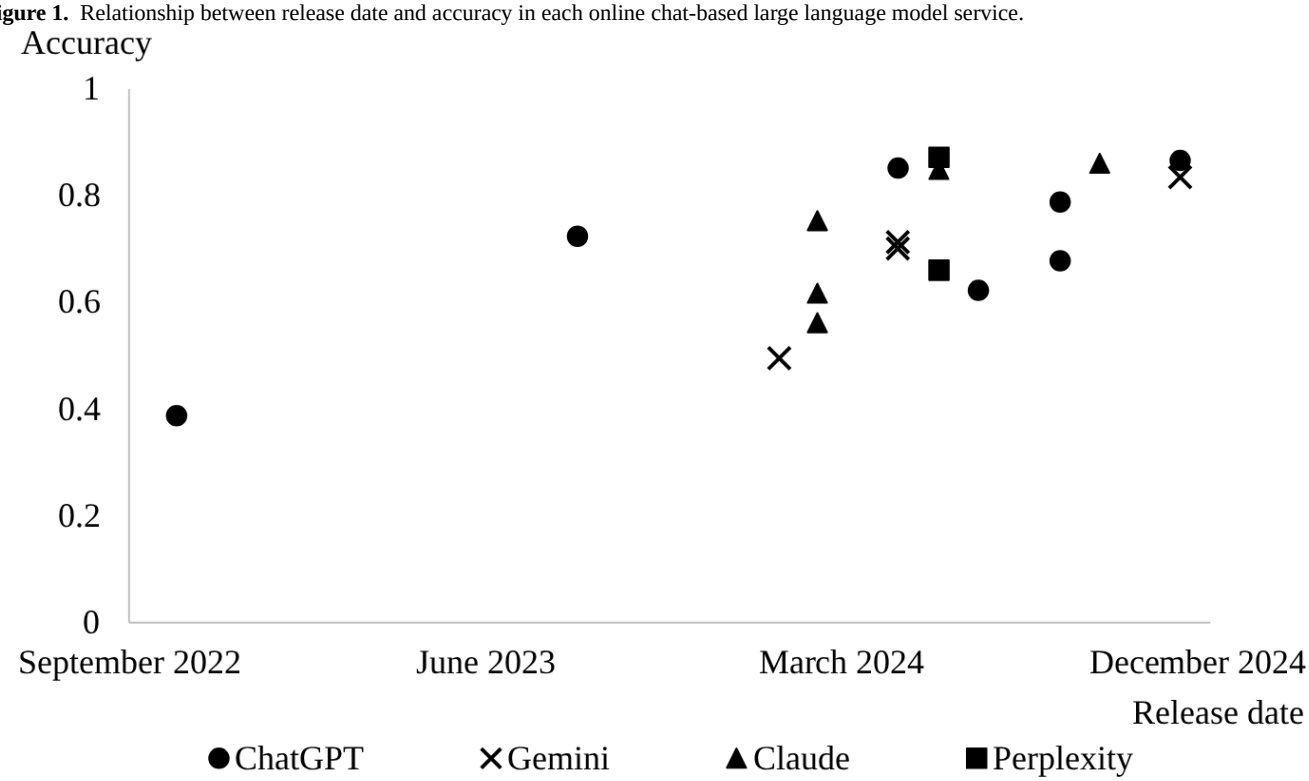
[c]Number of correct answers out of all 345 questions in the 107th JNLEP.

[d]Overall accuracy>62.9%.

[e]GPT-4 results were obtained from Sato and Ogasawara [14].

For all services, the model enhancements were confirmed to result in increased accuracy. All the models released after September 2024, regardless of whether they were light, medium, or high, met the qualification criteria (Figure 1). All GPT-4 results were obtained from Sato and Ogasawara [14]. The raw data of each model's item-by-item correctness are presented in Multimedia Appendix 1.

Figure 1. Relationship between release date and accuracy in each online chat-based large language model service.



Performance of AI Models According to Subject and Question Type

By subject, pathophysiology and pharmacology showed high accuracy for all models except for the ChatGPT GPT-3.5 model. In the most recent models, the accuracy in pharmaceuticals and biology improved substantially, whereas in physics and chemistry, only minor improvements were observed (Table 3). In the 4 flagship models, the average accuracy based on subject was lowest for chemistry (10.3/20, 51.3%), followed by physics (15.3/20, 76.3%). All the other subjects achieved an accuracy exceeding 80%.

For questions that consisted of only text, most models exhibited high accuracy, with a few exceptions. Three models (ChatGPT o1 preview, ChatGPT o1, and Perplexity Pro) showed a correct answer rate of over 90%. The accuracy decreased greatly for questions that included diagrams; the average of all 18 models was 36.7% (22.4/61) and was 50.8% (31.0/61) when models that could not input diagrams were excluded. For questions that included figures, Claude 3.5 Sonnet (new) showed the highest accuracy (47/61, 77%). For the calculation questions, the most recent model showed an improvement in accuracy but did not achieve high accuracy for questions that included chemical structures (Table 4).

Table . Accuracy and number of correct answers according to subject for each OC-LLM[a] in the 107th JNLEP[b,c].

OC-LLM[a]	Subject[d]																	
	Physics (n=20)		Chemistry (n=20)		Biology (n=20)		Hygiene (n=40)		Pharmacology (n=40)		Pharmaceuticals (n=40)		Pathophysiology (n=40)		Regulations (n=30)		Practice (n=95)	
	Correct answers, n	Accuracy	Correct answers, n	Accuracy	Correct answers, n	Accuracy	Correct answers, n	Accuracy	Correct answers, n	Accuracy	Correct answers, n	Accuracy	Correct answers, n	Accuracy	Correct answers, n	Accuracy	Correct answers, n	Accuracy
ChatGPT																		
GPT-3.5	4	0.200	3	0.150	5	0.250	13	0.325	19	0.475	11	0.275	27	0.675	10	0.333	42	0.442
GPT-4[c]	11	0.550	7	0.350	13	0.650	28	0.700	36	0.900	25	0.625	28	0.700	20	0.667	82	0.863
GPT-4o	13	0.650	12	0.600	19	0.950	33	0.825	39	0.975	31	0.775	35	0.875	25	0.833	87	0.916
GPT-4o mini	9	0.450	2	0.100	7	0.350	23	0.575	36	0.900	18	0.450	31	0.775	21	0.700	68	0.716
o1 mini	12	0.600	5	0.250	9	0.450	26	0.650	32	0.800	25	0.625	34	0.850	16	0.533	75	0.789
o1 preview	14	0.700	6	0.300	10	0.500	27	0.675	39	0.975	28	0.700	36	0.900	24	0.800	88	0.926
o1	14	0.700	8	0.400	18	0.900	34	0.850	39	0.975	34	0.850	38	0.950	25	0.833	89	0.937
Gemini																		
1.0 Pro	6	0.300	6	0.300	12	0.600	21	0.525	22	0.550	15	0.375	24	0.600	19	0.633	46	0.484
1.5 Pro	14	0.700	4	0.200	11	0.550	25	0.625	32	0.800	25	0.625	36	0.900	23	0.767	76	0.800
1.5 Flash	11	0.550	9	0.450	17	0.850	29	0.725	32	0.800	24	0.600	32	0.800	22	0.733	66	0.695
2.0 Flash	17	0.850	11	0.550	13	0.650	35	0.875	37	0.925	33	0.825	35	0.875	27	0.900	80	0.842
Claude																		
3 Sonnet	9	0.450	7	0.350	14	0.700	30	0.750	28	0.700	3	0.075	32	0.800	18	0.600	53	0.558
3 Haiku	8	0.400	8	0.400	11	0.550	26	0.650	32	0.800	21	0.525	31	0.775	21	0.700	55	0.579
3 Opus	11	0.550	3	0.150	16	0.800	30	0.750	35	0.875	27	0.675	34	0.850	24	0.800	80	0.842
3.5 Sonnet	15	0.750	8	0.400	16	0.800	36	0.900	37	0.925	32	0.800	36	0.900	28	0.933	85	0.895
3.5 Sonnet (new)	14	0.700	10	0.500	19	0.950	35	0.875	38	0.950	33	0.825	36	0.900	25	0.833	87	0.916
Perplexity																		

OC-LLM[a]	Subject[d]																	
	Physics (n=20)		Chemistry (n=20)		Biology (n=20)		Hygiene (n=40)		Pharmacology (n=40)		Pharmaceuticals (n=40)		Pathophysiology (n=40)		Regulations (n=30)		Practice (n=95)	
	Correct answers, n	Accuracy	Correct answers, n	Accuracy	Correct answers, n	Accuracy	Correct answers, n	Accuracy	Correct answers, n	Accuracy	Correct answers, n	Accuracy	Correct answers, n	Accuracy	Correct answers, n	Accuracy	Correct answers, n	Accuracy
Standard	11	0.550	2	0.100	6	0.300	23	0.575	38	0.950	25	0.625	34	0.850	21	0.700	68	0.716
Pro	16)	0.800	12	0.600	18	0.900	35	0.875	40	1.000	33	0.825	38	0.950	28	0.933	81	0.853

[a]OC-LLM: online chat-based large language model.

[b]JNLEP: Japanese National License Examination for Pharmacists.

[c]GPT-4 results were obtained from Sato and Ogasawara [14].

[d]The mean (SD) for physics is 0.581 (0.172), chemistry is 0.342 (0.162), biology is 0.650 (0.223), hygiene is 0.707 (0.151), pharmacology is 0.849 (0.147), pharmaceuticals is 0.615 (0.211), pathophysiology is 0.829 (0.095), regulations is 0.735 (0.15), and practice is 0.765 (0.157).

Table . Accuracy and number of correct answers based on question type for each model of OC-LLM[a] in the 107th JNLEP[bcd]

OC-LLM[a]	Question type									
	Text (n=284)		Diagram (n=61)		Calculation (n=18)		Graph (n=16)		Chemical structure (n=19)	
	Answers, n	Accuracy	Answers, n	Accuracy	Answers, n	Accuracy	Answers, n	Accuracy	Answers, n	Accuracy
ChatGPT										
GPT-3.5	134	0.472	0	0.000	5	0.278	0	0.000	0	0.000
GPT-4[c]	227	0.799	22	0.361	9	0.500	5	0.313	4	0.211
GPT-4o	254	0.894	39	0.639	12	0.667	7	0.438	10	0.526
GPT-4o mini	215	0.757	0	0.000	8	0.444	0	0.000	0	0.000
o1 mini	231	0.813	0	0.000	15	0.833	0	0.000	0	0.000
o1 preview	271	0.954	0	0.000	15	0.833	0	0.000	0	0.000
o1	260	0.915	39	0.639	16	0.889	9	0.563	7	0.368
Gemini										
1.0 Pro	158	0.556	13	0.213	4	0.222	4	0.250	2	0.105
1.5 Pro	233	0.820	13	0.213	9	0.500	2	0.125	3	0.158
1.5 Flash	211	0.743	31	0.508	7	0.389	7	0.438	7	0.368
2.0 Flash	246	0.866	42	0.689	13	0.722	9	0.563	10	0.526
Claude										
3 Sonnet	182	0.641	28	0.459	9	0.500	9	0.563	5	0.263
3 Haiku	190	0.669	27	0.443	12	0.667	7	0.438	5	0.263
3 Opus	228	0.803	23	0.377	5	0.278	6	0.375	4	0.211
3.5 Sonnet	250	0.880	42	0.689	15	0.833	12	0.750	8	0.421
3.5 Sonnet (new)	250	0.880	47	0.770	16	0.889	11	0.688	11	0.579
Perplexity										
Standard	226	0.796	0	0.000		0.444	0	0.000	0	0.000
Pro	264	0.930	37	0.607		0.556	8	0.500	11	0.579

[a]OC-LLM: online chat-based large language model.

[b]JNLEP: Japanese National License Examination for Pharmacists.

[c]GPT-4 results obtained from Sato and Ogasawara [14].

[d]The mean (SD) for text is 0.788 (0.131), diagram is 0.367 (0.279), calculation is 0.580 (0.219), graph is 0.333 (0.257), and chemical structure is 0.254 (0.213).

Statistical Analysis of Improved Accuracy and Response Consistency

The GLMM analysis demonstrated that the accuracy of the latest flagship models was significantly higher than that of earlier models (P<.001). In addition, questions containing diagrams had significantly lower accuracy than that of text-only questions (P<.001). The interaction term between the flagship status and question type was not significant (P=.53). Therefore, the difference in accuracy between the early and most recent models was consistently observed, regardless of whether the questions included diagrams or were text-based. Moreover, the flagship models did not show a greater improvement in the accuracy of diagram-based questions. The Fleiss' κ value was 0.334, thus verifying the consistency of each question for the 4 flagship models in the 345 questions.

Discussion

Overview

This study evaluated the performances of 18 generative AI models in the pharmacy field by applying the same prompt to an identical input task of the 107th JNLEP. Although previous studies evaluated the performance of generative AI in the health care field using several models, this study is the first to directly compare several OC-LLMs under identical conditions for the same task. Recent meta-analyses have evaluated the performance of generative AI in health care [21,22]. However, as individual

studies differ in language, prompts, and input tasks, inherent limitations exist in terms of interpreting these results.

Among these models, Perplexity Pro achieved the highest overall accuracy (301/345, 87.2%). When restricted to text-only questions, the ChatGPT o1 preview demonstrated the highest accuracy (271/284, 95.4%). For questions including diagrams, Claude 3.5 Sonnet (new) demonstrated the best performance (47/61, 77%). In early multimodal models, such as ChatGPT GPT-4 and Gemini 1.0 Pro, the accuracy for questions with diagrams was low. However, the accuracy of the latest versions of the flagship models has significantly improved. These findings indicate that the ability to recognize diagrams advanced markedly over the past year.

In terms of overall accuracy, the 4 flagship models exceeded not only the passing criteria but also the average examinee score of 68.2% [14]. This suggests that the current generative AI may possess a more extensive knowledge base than that of novice human pharmacists. However, even the best models had over 10% incorrect answers (ie, hallucinations); therefore, these models must be interpreted with caution, especially in health care.

Subject-specific analysis demonstrated accuracy improvements for all subjects when using the latest 4 flagship models. The performance for the subjects of hygiene and regulations tends to be weaker [16,23,24]. This is likely due to the influence of country-specific health care systems and social contexts, which may not be fully covered by pretraining data. In addition, the low accuracy observed in basic science subjects (physics, chemistry, and biology) is consistent with the trends reported in previous studies [25]. However, even in these subjects, improvements in accuracy were observed with the 2024 flagship models; hence, previous weaknesses may have been overcome. This improvement is likely attributable to the enhanced training data, increased model parameters, and the implementation of multimodal and reasoning capabilities. Although improvements in accuracy were observed, the flagship models still showed low accuracy in subjects, such as chemistry (10.3/20, 51.3%) and physics (15.3/20, 76.3%). This may be because these subjects included many questions that required abilities beyond factual knowledge, including calculations and image recognition. Low accuracy in chemistry has also been reported in previous studies [26].

Question type–specific analysis revealed lower accuracy for items that required image recognition or calculation, relative to text-only questions. Considering that image recognition and calculation are abilities that conventional large language models are not designed to handle and are acquired later through multimodal integration, the insufficient performance in this domain may be due to the incomplete maturation of learning.

Among the diagram-based questions, those involving chemical structures exhibited the lowest accuracy. The small mean and SD across all models for chemistry indicate that the performance of the current models did not show a considerable improvement. This may be because of two factors: (1) chemical structures are foundational scientific knowledge needed exclusively by pharmacists, leading to limited web-based availability (ie, reduced opportunities for large language model pretraining); and (2) interpreting chemical structures requires more sophisticated image recognition skills than that required for the interpretation of tables or graphs.

Claude 3.5 Sonnet (new) demonstrated the highest accuracy across all 3 types of questions—computation, graph interpretation, and chemical structure recognition. However, Claude's flagship model showed lower accuracy for text-based questions than that of the ChatGPT and Perplexity flagship models. Therefore, a novel finding of this study is that the top-performing model differed according to the question type.

The GLMM analysis demonstrated a significant increase in overall accuracy by 2024. Although improvements in the accuracy of the questions containing diagrams were observed in the individual models, these differences were not statistically significant. The tendency for lower accuracy on diagram-based questions persisted even in the flagship models.

According to Landis and Koch [27], a Fleiss' κ of 0.344 among the 4 flagship models indicates a certain degree of consistency. This result suggests that although these models handle simpler questions similarly, their incorrect answers differ across more challenging questions, thus indicating variations in their strengths and weaknesses. Initially, it was hypothesized that the types of questions with which each OC-LLM service struggles would differ. Correspondingly, the observation that even the flagship models with high overall accuracy failed to achieve substantial response agreement, as measured by the κ coefficient, supports this hypothesis. Therefore, identifying the specific domains in which each OC-LLM service underperforms remains an important subject for future research, including meta-analysis.

In this study, each model was evaluated using the same task to compare their performance directly. Some models included in this study have been deprecated and are no longer available. Although many new OC-LLMs are expected to emerge in the future, evaluating their performance using the 107th JNLEP will enable their comparison with previous models. Ultimately, the 107th JNLEP can serve as a benchmark for evaluating the performance of generative AI models in the field of pharmacy in Japan.

In this study, questions from the Japanese National Pharmacist Examination were input in Japanese in their original format. Translating non-English tasks into English should improve the accuracy of AI [28-31]. Therefore, this study evaluated the performance of AI models in the pharmaceutical field using Japanese input. However, higher accuracy may be achieved when questions are input using English translations. The accuracy of each model is based on the evaluation time, and the same model may show improved performance due to upgrades. Perplexity has been upgraded multiple times; however, the available models remain as Standard and Pro versions, and the version information is not disclosed to users.

Although the highest-performing model among the 18 OC-LLMs in this study achieved an accuracy of 87.2% (301/345), it also indicated that over 12.8% (n=44 questions) of the responses were incorrect (ie, hallucinations). With the improved performance of the OC-LLMs, it is anticipated that their use by medical and pharmacy students for inputting national

examination questions for self-study will increase. However, as the latest models generate logical and fluent answers, it has become increasingly difficult to identify hallucinations. Even when using flagship models in 2024, the following approaches to reduce the risk of hallucinations are required in medical applications: limit use to cases in which users can independently determine the correctness of the output or confirm the supporting source information through the links provided.

The performance improvements of the OC-LLMs in this study may facilitate their broader integration into routine pharmacy practice in the near future. In clinical pharmacy practice, responding to inquiries from patients and health care professionals regarding drug information is a frequent task. These inquiries include questions about adverse drug reactions, drug interactions, dosage adjustments, or contraindications. Suitably, support from high-performance OC-LLMs is expected to improve the quality of responses and reduce the time required to address such inquiries. The use of OC-LLMs in direct medical support, for example, in selecting personalized pharmacological treatments, requires careful consideration of ethical issues, such as explainability, responsibility, privacy, and patient rights.

Principal Findings

This study evaluated the performance of 18 OC-LLMs available in 2024, based on questions from the National Pharmacist's License Examination in Japan. As the models were upgraded, their accuracy improved. The performance of the flagship models exceeded both the passing criteria and the examinees' average score. In the latest versions of the OC-LLMs, enhancements in multimodal capabilities significantly improved accuracy in both interpreting charts and figures and solving calculation-based questions. Furthermore, the answer consistency of the flagship models was not robust, which suggests that each model had different strengths and weaknesses.

Limitations

In this study, only a single set of examination questions was tested, and each question was entered only once. Generative AI has a characteristic known as temperature, which refers to the inherent variability in its responses. This means that the model can generate different answers even when given the same question. Therefore, if the test is repeated, the accuracy of each OC-LLM method can vary. Several studies have evaluated OC-LLM performance by testing questions over multiple years [32-34] or conducting multiple rounds of testing [35]. However, similar to many previous studies, to evaluate the 18 models within a limited time frame, only 1 set of questions was administered per examination year to each model. Human examinees also underwent the national pharmacist's examination only once, rendering the testing conditions comparable. The 107th JNLEP comprises 345 questions, with multiple items allocated to each subject and question type. Therefore, the examination is considered sufficient to allow for a certain degree of interpretation.

With the progressive improvement of the models over time, the top-performing service shifted from ChatGPT to Claude, and subsequently to Gemini. Across OC-LLM services, such as ChatGPT, Gemini, Claude, and Perplexity, no consistent patterns were observed across subjects or question types. Considering that the key information, such as the volume of pretraining data, number of parameters, and tuning strategies of these OC-LLMs, is not publicly disclosed, fully discussing the factors that contribute to their improved performance in the pharmaceutical field is difficult. These factors include understanding of diagrams, chemical structures, and calculation-based questions.

Comparison With Prior Work

Numerous studies have evaluated the performance of OC-LLM in terms of knowledge of health care license examinations (Table 5). Early OC-LLMs failed the National Medical License Examination; however, the subsequent release of high-performance OC-LLMs met the passing criteria.

The reported OC-LLMs in Table 5 are biased toward ChatGPT, and the challenges and conditions vary according to each report on medical performance. Moreover, the performance of OC-LLM declines in languages other than English because of the smaller volume of training data [18,25,34-36]. Therefore, verifying the performance of OC-LLM in non-English languages is important. An important contribution of this study is its demonstration that multiple flagship OC-LLMs substantially outperform the passing criteria in areas where prior evidence is scarce, specifically in non-English languages and the pharmaceutical field. Achieving high response accuracy from OC-LLMs using non-English prompts has considerable implications for clinical implementation in health care settings in Japan (and other non-English–speaking regions).

One of the major strengths of this study is its systematic evaluation of multiple OC-LLMs released in 2024 under identical input conditions, such as the same prompt text and image resolution or size. To the best of our knowledge, this is the first study to evaluate the longitudinal improvement in generative AI performance in medical examinations.

Table . Studies evaluating the performance of generative AI[a] in health care licensing examinations.

Health care license examination study	Country or region	OC-LLM[b]	Accuracy (%)
Medical license			
Gilson et al (2023) [36]	United States	GPT-3	25.3
Gilson et al (2023) [36]	United States	ChatGPT (unknown)	64.4, 57.8
Flores-Cohaila et al (2023) [37]	Peru	ChatGPT GPT-3.5	77
Flores-Cohaila et al (2023) [37]	Peru	ChatGPT GPT-4	86
Jung et al (2023) [38]	Germany	ChatGPT (unknown)	60.1, 66.7
Shang et al, (2023) [28]	China	ChatGPT GPT-3.5	57
Wang et al, (2023) [39]	China	ChatGPT GPT-3.5	56
Wang et al, (2023) [39]	China	ChatGPT GPT-4	84
Yanagita et al (2023) [40]	Japan	ChatGPT GPT-3.5	42.8
Yanagita et al (2023) [40]	Japan	ChatGPT GPT-4	81.5
Takagi et al (2023) [41]	Japan	ChatGPT GPT-3.5	50.8
Takagi et al (2023) [41]	Japan	ChatGPT GPT-4	79.9
Tanaka et al (2024) [16]	Japan	ChatGPT GPT-3.5	52.9, 63.6
Tanaka et al (2024) [16]	Japan	ChatGPT GPT-4	85.6
Liu et al (2025) [42]	Japan	ChatGPT GPT-4	77
Liu et al (2025) [42]	Japan	ChatGPT GPT-4o	89
Liu et al (2025) [42]	Japan	Gemini 1.5 Pro	80
Liu et al (2025) [42]	Japan	Claude 3 Opus	82
Oztermeli and Oztermeli (2023) [43]	Turkey	ChatGPT GPT-3.5	64.7, 67.1, 70.9, 60.8, 54.3
Siebielec et al (2024) [32]	Poland	ChatGPT GPT-3.5	59.5, 57.5, 63.5, 62.0, 61.0
Wójcik et al (2024) [31]	Poland	ChatGPT GPT-4	67.1
Pharmacist license			
Wang et al (2023) [44]	Taiwan	ChatGPT (unknown)	54.5, 63.5
Wang et al (2025) [45]	Taiwan	ChatGPT GPT-3.5	59
Wang et al (2025) [45]	Taiwan	ChatGPT GPT-4	73
Kunitsu (2023) [46]	Japan	ChatGPT GPT-4	78.2, 75.3
Sato and Ogasawara (2024) [14]	Japan	ChatGPT GPT-3.5	45.5
Sato and Ogasawara (2024) [14]	Japan	ChatGPT GPT-4	72.5
Jin and Kim (2024) [47]	Korea	ChatGPT GPT-3.5	61
Jin and Kim (2024) [47]	Korea	ChatGPT GPT-4	87
Nurse license			
Taira et al (2023) [33]	Japan	ChatGPT GPT-3.5	71, 71, 63, 63, 63
Kaneda et al (2023) [48]	Japan	ChatGPT GPT-3.5	59.9
Kaneda et al (2023) [48]	Japan	ChatGPT GPT-4	79.7
Wu et al (2024) [49]	China	ChatGPT GPT-3.5	51.7
Wu et al (2024) [49]	China	ChatGPT GPT-4	70.5
Wu et al (2024) [49]	China	Google Bard	48.3
Hiwa et al (2024) [50]	Unknown	ChatGPT GPT-3.5	77
Hiwa et al (2024) [50]	Unknown	Gemini (unknown)	75

Health care license examination study	Country or region	OC-LLM[b]	Accuracy (%)
Hiwa et al (2024) [50]	Unknown	Microsoft Copilot	84
Hiwa et al (2024) [50]	Unknown	Llama2	68

[a]AI: artificial intelligence.

[b]OC-LLM: online chat-based large language model.

Conclusions

This study reveals that the performance of OC-LLMs in the pharmaceutical field has greatly improved as of 2024. Particularly, an increase in accuracy was observed for questions with diagrams. In the most recent version of the models, evaluated in 2024, the overall accuracy reached 85%, markedly exceeding the average examinee score, and indicating their potential as valuable support tools. Although caution is necessary due to the potentially serious impact of hallucinations on health care, the benefits of OC-LLMs outweigh the associated risks. Accordingly, health care professionals and medical educators must acquire the skills necessary to effectively use OC-LLMs, particularly the ability to recognize and manage hallucinations.

Data Availability

The datasets analyzed during this study are available in the Ministry of Health, Labour and Welfare in Japan repository [19].

Conflicts of Interest

None declared.

Multimedia Appendix 1
The responses of all the online chat-based large language models to each question.
[XLSX File, 41 KB - mededu_v11i1e76925_app1.xlsx]

References

1. Kruse M, Hu S, Derby N, et al. Zero-shot large language models for long clinical text summarization with temporal reasoning. medRxiv. Preprint posted online on Jul 23, 2025. [doi: 10.1101/2025.07.21.25331947] [Medline: 40766151]
2. Fraile Navarro D, Coiera E, Hambly TW, et al. Expert evaluation of large language models for clinical dialogue summarization. Sci Rep 2025 Jan 7;15(1):1195. [doi: 10.1038/s41598-024-84850-x] [Medline: 39774141]
4. Van Veen D, Van Uden C, Blankemeier L, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med 2024 Apr;30(4):1134-1142. [doi: 10.1038/s41591-024-02855-5] [Medline: 38413730]
4. Lee C, Vogt KA, Kumar S. Prospects for AI clinical summarization to reduce the burden of patient chart review. Front Digit Health 2024;6:1475092. [doi: 10.3389/fdgth.2024.1475092] [Medline: 39575412]
5. UiPath announces AI partnership with Google Cloud to transform medical processes. UiPath. 2025. URL: https://www.uipath.com/newsroom/uipath-announces-medical-summarization-agent-google-cloud [accessed 2025-08-27]
6. OpenBots Gen AI automates patient referral, improves productivity by 30% and reduces errors by 80%. OpenBots. 2024. URL: https://openbots.ai/streamlining-patient-referral-handling-and-emr-integration-through-openbots-gen-ai-automation/ [accessed 2025-03-05]
7. Hale J, Alexander S, Wright ST, Gilliland K. Generative AI in undergraduate medical education: a rapid review. Journal of Medical Education and Curricular Development 2024 Jan;11. [doi: 10.1177/23821205241266697]
8. Parente DJ. Generative artificial intelligence and large language models in primary care medical education. Fam Med 2024 Oct;56(9):534-540. [doi: 10.22454/FamMed.2024.775525] [Medline: 39207784]
9. Janumpally R, Nanua S, Ngo A, Youens K. Generative artificial intelligence in graduate medical education. Front Med (Lausanne) 2024;11:1525604. [doi: 10.3389/fmed.2024.1525604] [Medline: 39867924]
10. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. JMIR Med Educ 2023 Oct 20;9:e48785. [doi: 10.2196/48785] [Medline: 37862079]
11. Biswas S. ChatGPT and the future of medical writing. Radiology 2023 Apr;307(2):e223312. [doi: 10.1148/radiol.223312] [Medline: 36728748]
12. AI adoption in health systems report 2024. Medscape & HIMSS. URL: https://cdn.sanity.io/files/sqo8bpt9/production/68216fa5d161adebceb50b7add5b496138a78cdb.pdf/ [accessed 2025-03-05]
13. Bruce G. How many healthcare employees use generative AI at work. Becker's Hospital Review. 2024. URL: https://www.beckershospitalreview.com/rankings-and-ratings/how-many-healthcare-employees-use-generative-ai-at-work.html/ [accessed 2025-03-05]

14. Sato H, Ogasawara K. ChatGPT (GPT-4) passed the Japanese National License Examination for Pharmacists in 2022, answering all items including those with diagrams: a descriptive study. J Educ Eval Health Prof 2024;21:4. [doi: 10.3352/jeehp.2024.21.4] [Medline: 38413129]
15. Top generative AI chatbots by market share. FirstPageSage. 2025 Jan. URL: https://firstpagesage.com/reports/top-generative-ai-chatbots/ [accessed 2025-03-05]
16. Tanaka Y, Nakata T, Aiga K, et al. Performance of generative pretrained transformer on the National Medical Licensing Examination in Japan. PLOS Digit Health 2024 Jan;3(1):e0000433. [doi: 10.1371/journal.pdig.0000433] [Medline: 38261580]
17. Wada A, Akashi T, Shih G, et al. Optimizing GPT-4 Turbo diagnostic accuracy in neuroradiology through prompt engineering and confidence thresholds. Diagnostics (Basel) 2024 Jul 17;14(14):1541. [doi: 10.3390/diagnostics14141541] [Medline: 39061677]
18. Yan S, Knapp W, Leong A, et al. Prompt engineering on leveraging large language models in generating response to InBasket messages. J Am Med Inform Assoc 2024 Oct 1;31(10):2263-2270. [doi: 10.1093/jamia/ocae172] [Medline: 39028970]
19. The 107th Japanese National License Examination for Pharmacist [Article in Japanese]. The Ministry of Health, Labour and Welfare in Japan. URL: https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/0000198924.html [accessed 2025-08-27]
20. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76(5):378-382. [doi: 10.1037/h0031619]
21. Jin HK, Lee HE, Kim E. Performance of ChatGPT-3.5 and GPT-4 in national licensing examinations for medicine, pharmacy, dentistry, and nursing: a systematic review and meta-analysis. BMC Med Educ 2024 Sep 16;24(1):1013. [doi: 10.1186/s12909-024-05944-8] [Medline: 39285377]
22. Takita H, Kabata D, Walston SL, et al. A systematic review and meta-analysis of diagnostic performance comparison between generative AI and physicians. npj Digit Med 2025;8(1). [doi: 10.1038/s41746-025-01543-z]
23. Hideo D, Hidekazu I, Hiroki N, et al. Performance of generative pretrained transformer on the national licensing examination for medical technologist in Japan [Article in Japanese]. Jpn J Med Technol 2024;73(2):323-331 [FREE Full text]
24. Meo SA, Alotaibi M, Meo MZS, Meo MOS, Hamid M. Medical knowledge of ChatGPT in public health, infectious diseases, COVID-19 pandemic, and vaccines: multiple choice questions examination based performance. Front Public Health 2024;12:1360597. [doi: 10.3389/fpubh.2024.1360597] [Medline: 38711764]
25. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
26. Fergus S, Botha M, Ostovar M. Evaluating academic answers generated using ChatGPT. J Chem Educ 2023 Apr 11;100(4):1672-1675. [doi: 10.1021/acs.jchemed.3c00087]
27. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [doi: 10.2307/2529310] [Medline: 843571]
28. Shang L, Xue M, Hou Y, Tang B. Can ChatGPT pass China's national medical licensing examination? Asian J Surg 2023 Dec;46(12):6112-6113. [doi: 10.1016/j.asjsur.2023.09.089] [Medline: 37775381]
29. Cohen A, Alter R, Lessans N, Meyer R, Brezinov Y, Levin G. Performance of ChatGPT in Israeli Hebrew OBGYN national residency examinations. Arch Gynecol Obstet 2023 Dec;308(6):1797-1802. [doi: 10.1007/s00404-023-07185-4] [Medline: 37668790]
30. Lewandowski M, Łukowicz P, Świetlik D, Barańska-Rybak W. ChatGPT-3.5 and ChatGPT-4 dermatological knowledge level based on the Specialty Certificate Examination in Dermatology. Clin Exp Dermatol 2024 Jun 25;49(7):686-691. [doi: 10.1093/ced/llad255] [Medline: 37540015]
31. Wójcik D, Adamiak O, Czerepak G, Tokarczuk O, Szalewski L. A comparative analysis of the performance of Chatgpt4, Gemini and Claude for the Polish Medical Final Diploma Exam and Medical-Dental Verification Exam. medRxiv. Preprint posted online on Jul 29, 2024. [doi: 10.1101/2024.07.29.24311077]
32. Siebielec J, Ordak M, Oskroba A, Dworakowska A, Bujalska-Zadrozny M. Assessment study of ChatGPT-3.5's performance on the final Polish medical examination: accuracy in answering 980 questions. Healthcare (Basel) 2024 Aug 16;12(16):1637. [doi: 10.3390/healthcare12161637] [Medline: 39201195]
33. Taira K, Itaya T, Hanada A. Performance of the large language model ChatGPT on the National Nurse Examinations in Japan: evaluation study. JMIR Nurs 2023 Jun 27;6:e47305. [doi: 10.2196/47305] [Medline: 37368470]
34. Zong H, Li J, Wu E, Wu R, Lu J, Shen B. Performance of ChatGPT on Chinese National Medical Licensing Examinations: a five-year examination evaluation study for physicians, pharmacists and nurses. BMC Med Educ 2024 Feb 14;24(1):143. [doi: 10.1186/s12909-024-05125-7] [Medline: 38355517]
35. Fujimoto M, Kuroda H, Katayama T, et al. Evaluating large language models in dental anesthesiology: a comparative analysis of ChatGPT-4, Claude 3 Opus, and Gemini 1.0 on the Japanese Dental Society of Anesthesiology Board Certification Exam. Cureus 2024 Sep;16(9):e70302. [doi: 10.7759/cureus.70302] [Medline: 39469383]
36. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]

37. Flores-Cohaila JA, García-Vicente A, Vizcarra-Jiménez SF, et al. Performance of ChatGPT on the Peruvian National Licensing Medical Examination: cross-sectional study. JMIR Med Educ 2023 Sep 28;9:e48039. [doi: 10.2196/48039] [Medline: 37768724]
38. Jung LB, Gudera JA, Wiegand TLT, Allmendinger S, Dimitriadis K, Koerte IK. ChatGPT passes German State Examination in Medicine with picture questions omitted. Dtsch Arztebl Int 2023 May 30;120(21):373-374. [doi: 10.3238/arztebl.m2023.0113] [Medline: 37530052]
39. Wang H, Wu W, Dou Z, He L, Yang L. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: pave the way for medical AI. Int J Med Inform 2023 Sep;177:105173. [doi: 10.1016/j.ijmedinf.2023.105173] [Medline: 37549499]
40. Yanagita Y, Yokokawa D, Uchida S, Tawara J, Ikusaka M. Accuracy of ChatGPT on medical questions in the National Medical Licensing Examination in Japan: evaluation study. JMIR Form Res 2023 Oct 13;7:e48023. [doi: 10.2196/48023] [Medline: 37831496]
41. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: comparison study. JMIR Med Educ 2023 Jun 29;9:e48002. [doi: 10.2196/48002] [Medline: 37384388]
42. Liu M, Okuhara T, Dai Z, et al. Performance of advanced large language models (GPT-4o, GPT-4, Gemini 1.5 pro, Claude 3 opus) on Japanese Medical Licensing Examination: a comparative study. Int J Med Inform 2025:105673. [doi: 10.1101/2024.07.09.24310129] [Medline: 39471700]
43. Oztermeli AD, Oztermeli A. ChatGPT performance in the medical specialty exam: an observational study. Medicine (Abingdon) 2023;102(32):e34673. [doi: 10.1097/MD.0000000000034673]
44. Wang YM, Shen HW, Chen TJ. Performance of ChatGPT on the pharmacist licensing examination in Taiwan. J Chin Med Assoc 2023 Jul 1;86(7):653-658. [doi: 10.1097/JCMA.0000000000000942] [Medline: 37227901]
45. Wang YM, Shen HW, Chen TJ, Chiang SC, Lin TG. Performance of ChatGPT-3.5 and ChatGPT-4 in the Taiwan National Pharmacist Licensing Examination: comparative evaluation study. JMIR Med Educ 2025 Jan 17;11:e56850. [doi: 10.2196/56850] [Medline: 39864950]
46. Kunitsu Y. The potential of GPT-4 as a support tool for pharmacists: analytical study using the Japanese National Examination for Pharmacists. JMIR Med Educ 2023 Oct 30;9:e48452. [doi: 10.2196/48452] [Medline: 37837968]
47. Jin HK, Kim E. Performance of GPT-3.5 and GPT-4 on the Korean Pharmacist Licensing Examination: comparison study. JMIR Med Educ 2024 Dec 4;10:e57451. [doi: 10.2196/57451] [Medline: 39630413]
48. Kaneda Y, Takahashi R, Kaneda U, et al. Assessing the performance of GPT-3.5 and GPT-4 on the 2023 Japanese Nursing Examination. Cureus 2023 Aug;15(8):e42924. [doi: 10.7759/cureus.42924] [Medline: 37667724]
49. Wu Z, Gan W, Xue Z, Ni Z, Zheng X, Zhang Y. Performance of ChatGPT on Nursing Licensure Examinations in the United States and China: cross-sectional study. JMIR Med Educ 2024 Oct 3;10:e52746. [doi: 10.2196/52746] [Medline: 39363539]
50. Hiwa DS, Abdalla SS, Muhialdeen AS, Karim SO. Assessment of nursing skill and knowledge of ChatGPT, Gemini, Microsoft Copilot, and Llama: a comparative study. Barw Med J 2024;2(3). [doi: 10.58742/bmj.v2i2.87]

Abbreviations

AI: artificial intelligence
GLMM: generalized linear mixed model
JNLEP: Japanese National License Examination for Pharmacists
OC-LLM: online chat-based large language model

Edited by J Moen; submitted 04.05.25; peer-reviewed by I Murray, RC Wang Chau; revised version received 16.07.25; accepted 31.07.25; published 18.09.25.

Please cite as:
Sato H, Ogasawara K, Sakurai H
Performance Evaluation of 18 Generative AI Models (ChatGPT, Gemini, Claude, and Perplexity) in 2024 Japanese Pharmacist Licensing Examination: Comparative Study
JMIR Med Educ 2025;11:e76925
URL: https://mededu.jmir.org/2025/1/e76925
doi:10.2196/76925

Perception of Medical Undergraduates on Artificial Intelligence in Medical Education: Qualitative Exploration

Thilanka Seneviratne[1], MBBS, MD; Kaumudee Kodikara[2], MBBS, MMEd, PhD; Isuru Abeykoon[1], MBBS; Wathsala Palpola[1], MBBS

[1]Department of Pharmacology, Faculty of Medicine, University of Peradeniya, Peradeniya, Sri Lanka

[2]Department of Medical Education, Faculty of Medicine, University of Kelaniya, 4 Thalagolla Road, Ragama, Sri Lanka

Corresponding Author:
Kaumudee Kodikara, MBBS, MMEd, PhD
Department of Medical Education, Faculty of Medicine, University of Kelaniya, 4 Thalagolla Road, Ragama, Sri Lanka

Abstract

Background: Artificial intelligence (AI) has revolutionized medical education by delivering tools that enhance and optimize learning. However, there is limited research on the medical students' perceptions regarding the effectiveness of AI as a learning tool, particularly in Sri Lanka.

Objective: The study aimed to explore students' perceived barriers and limitations to using AI for learning as well as their expectations in terms of future use of AI in medical education.

Methods: An exploratory qualitative study was conducted in September 2024, involving focus group discussions with medical students from two major universities in Sri Lanka. Reflexive thematic analysis was used to identify key themes and subthemes emerging from the discussions.

Results: Thirty-eight medical students participated in 5 focus group discussions. The majority of the participants were Sinhalese female students. The perceived benefits included saving time and effort and collecting and summarizing information. However, concerns and limitations centered around inaccuracies of information provided and the negative impacts on critical thinking, social interactions (peer and student teacher), and long-term retention of knowledge. Students were confused about contradictory messages received from educators regarding the use of AI for teaching and learning. However, participants showed an enthusiasm for learning more about the ethical use of AI to enhance learning and indicated that basic AI knowledge should be taught in their undergraduate program.

Conclusions: Participants recognized several benefits of AI-assisted learning but also expressed concerns and limitations requiring further studies for effective integration of AI into medical education. They expressed openness and enthusiasm for using AI while demonstrating confusion and reluctance due to the perspectives and stance of educators. We recommend educating both the educators and learners on the ethical use of AI, enabling a formal integration of AI tools into medical curricula.

(JMIR Med Educ 2025;11:e73798) doi:10.2196/73798

KEYWORDS

medical students; AI in medical education; attitudes; perspectives; learning; qualitative study

Introduction

As the information age wanes with the rise of artificial intelligence (AI), a global trend has risen to implement AI to enhance the effectiveness of the health care system [1]. Notably, diagnosis and treatments for several diseases can now be performed faster and more precisely with the use of AI in clinical medicine [2,3], giving both doctors and patients easier pathways to navigate diseases than ever before. Consequently, attention is being drawn to how the medical workforce can be ready for this transition and, thus, to investigating how the education of future health care professionals can best be delivered to achieve the futuristic goals of clinical practice [4]. As evidenced by the COVID-19 pandemic, AI poses a significant impact on medical education, with the ability to provide medical students with an interactive learning environment, creating virtual simulations allowing learners to practice complex or risky clinical procedures on virtual patients without endangering actual patients [5-8]. The development of ChatGPT, a new AI-driven language model, showcased its potential for assisting learners in self-directed learning while highlighting ethical issues [9].

The World Medical Association and the Standing Committee of European Doctors advocate for the use of AI systems in basic and continuing medical education [10,11], highlighting the need for increasing awareness of the proper use of AI in graduate, postgraduate, and continuing medical education. However, existing literature on AI emphasizes the inability of today's

medical education to meet the needs of AI, proposing a fundamental change in education to achieve the goals stated by the World Medical Association [2,3,11-13]. An in-depth understanding of how the current medical student perceives the use of AI in their education and their comprehension of limitations, challenges, and future projections is vital for integrating AI into existing medical curricula. Many studies have investigated the students' perceptions regarding AI in medicine and medical education in countries such as the Republic of Korea [14], Germany [15,16], the United Kingdom [17,18], Canada [19,20], the United States [21], India [22], Pakistan [23], Australia and New Zealand [24], Malaysia [25], Turkey [26], Palestine [27], Saudi Arabia [28], Egypt [29], Syria [30], Jordan [31], the United Arab Emirates [32], and Kuwait [33], and it is worth noting that such examinations are still notably absent in the context of Sri Lanka, a developing island nation in the Asian continent. This scarcity is noteworthy, especially when considering Sri Lanka's recognized status as a quality health care provider and its contribution to skilled labor migration for developed countries [34]. Recognizing the overwhelming importance of incorporating AI into medical education, this study aims to explore medical students' perceptions toward AI in education, in terms of the barriers and challenges the students face when using AI in their education, and investigate the future projections of AI into medical education in the eye of the medical student. The findings will aid in making beneficial decisions regarding the inclusion of AI in medical curricula in the future while filling a niche in AI literature on undergraduate medical education.

Methods

Study Design

An exploratory qualitative study was conducted in September 2024 using focus group discussions to gather medical undergraduates' perspectives on using AI in their education. Our objective was to identify barriers and challenges the students face when using AI in their education and to investigate the future projections as perceived by learners to improve their learning experience. We adopted a constructivist approach with the understanding that meaning is constructed through dialog between the researcher and the researched [35]. Semi-structured focus group discussions enabled an in-depth exploration of the student participants' subjective reality and experiences [36] in using AI in their education and the freedom to consider different issues [37] as the discussions were co-constructed by the researcher and participant [38].

Ethical Considerations

After obtaining the ethics approval from the Ethics Review Committee, Faculty of Medicine, University of Peradeniya (ERC No: 2024/EC/27), general information about the study was shared with potential participants via WhatsApp groups. Third- and fourth-year medical students of the Faculty of Medicine, University of Peradeniya, and the Faculty of Medicine, University of Kelaniya, were recruited to the study. Students were informed via the information sheet of the voluntary nature of participation, their right to refuse to participate or answer any specific questions, and to withdraw from the study at any time. Informed consent was obtained from all participants. No personally identifiable information was collected, aside from participants' ethnicity and year of study. All data were securely stored on a password-protected platform, accessible only to the research team. All data collected was anonymized. Purposive sampling was carried out at this stage. To ensure maximal variation sampling, we recruited student volunteers of different genders and ethnic groups, including students from the foreign quota. The interviews were conducted in compliance with the standards for reporting qualitative research set by the Consolidated Criteria for Reporting Qualitative Research (COREQ) [39]. Participants did not receive any incentives for participating in the focus group discussions.

Designed Focus Group Discussion Protocol

The focus group discussions were facilitated by two trained researchers (KK and WP) in September 2024. The objectives of the present study were explained at the beginning of each focus group, and verbal consent was obtained from each participant. The participants were provided with essential background information. No extra questions were asked after the focus group discussions, no additional information was given by the facilitators, and no other questions were requested subsequently. Only the facilitator and the focus group participants were present for the focus group discussions. Other individuals were not allowed to attend the focus group discussions to ensure the privacy and confidentiality of the discussion. Before recording the responses, permission for audio recording was requested by the facilitator verbally and via the written consent form.

Data Collection

The participants in the focus group discussions met with a trained research assistant to provide their details and to hand over the consent forms. The time and venue of the focus group discussions were decided at this meeting. We conducted focus group discussions in English using a semi-structured interview guide, which explored how medical students felt about using AI in their education. The questions for the interview guide were extracted from previous similar research [19,27].

We investigated medical students' perceived barriers and limitations to using AI for learning, as well as their expectations in terms of future use of AI in medical education. The open-ended questions served to guide, but not constrain, the interview. We encouraged the participants to describe their experiences. We emboldened them to react to each other's opinions and generate new ideas from different points of view. We arranged and conducted focus group discussions, commencing with the first responders. When two successive focus group discussions yielded no novel themes and only repeated ideas, indicating thematic redundancy, we considered that data saturation had been reached [40]. Two trained researchers (KK and WP) conducted all the focus group discussions. Each focus group consisted of 7-9 students. Each focus group discussion lasted 1-1.5 hours and was held at the Faculties of Medicine, University of Peradeniya and University of Kelaniya. With the consent of the students, we audio taped the discussions for later transcription. We informed participants that their identities would remain confidential, and their views

and opinions would be anonymized. We removed all identifiable features during transcription. Participants were informed that the focus group discussions constituted part of a research project, and the findings might be published and used to improve medical education.

Data Analysis

Reflexive thematic analysis (TA) [41-43] was used to analyze the data from the focus group discussions. The qualitative data analysis coincided with the focus group discussions, allowing the researchers to gather information until data saturation. We removed all identifying features during transcription. In analyzing the focus group data, the reflexive element meant that we interpreted the data through our lens of experience as educators, centrally recognizing our subjectivity and seeking to develop a sense of meaning from our responses. As a result, our reflexive analysis of the data led to the output of codes, subthemes, and overarching themes through the coding process [43] and thus represents the reflexive TA approach. This method was structured around the common six-step process for TA as described by Braun and Clarke [41]. Using a sentence-by-sentence process, we manually coded each transcript and sorted the talk into categories and subcategories. All transcripts were coded by two authors (IA and WP). An open coding scheme was used for the first interview. After achieving consensus among the authors, this coding frame was used to code the remaining four transcripts. The authors compared coding for consistency to ensure a common language. We identified commonalities and differences across all interviews before regrouping the codes into themes. We compared interpretations and discussed them among all authors until there were no discrepancies. The authors reached a consensus regarding verbatim remarks selected to highlight the relevant subthemes arising from the analysis.

Research Team and Reflexivity

Project conception and survey design were performed by TS, who is a consultant pediatrician and a senior lecturer in Pharmacology who believes that AI training in medical education is important and has published similar work. KK is a lecturer in medical education who holds a PhD, with a qualitative and quantitative research background in medical education. Some of the focus group participants were personally known to TS, while all participants were unfamiliar to the second facilitator (KK) and the rest of the research team (IA and WP).

To enhance the credibility of this study, TS, who personally knew some participants and held a senior academic position, refrained from conducting the focus group discussions to minimize the influence of positional authority. KK and WP, to whom the participants were unknown, conducted the focus group discussions. Moreover, member checking on the accuracy of transcriptions was done. All transcripts were coded independently by two authors (IA and WP) who were not directly involved in prior AI-related research. Regular meetings were held to discuss coding decisions and reconcile discrepancies through consensus. We discussed the developed themes and subthemes with KK, an experienced qualitative researcher. TS participated in later stages of theme development, in order to enhance the neutrality and trustworthiness of the analysis. All four authors discussed our own biases to become aware of and be transparent about our perspectives, personal feelings, and preconceptions, and we considered these critically concerning the research being conducted [44]. All authors discussed and resolved any disagreements regarding coding or developing themes. All interpretations were critically reviewed by the entire research team to ensure they were grounded in the raw data. To contribute to the dependability of the data, we kept a reflexivity diary to reflect on the process [45]. We did this because TS is a senior lecturer who had previously contributed to the development of an AI tool used for assessing answers to short answer questions and who believes that AI training in medical education is important. We were cognizant, therefore, that TS may have a propensity to seek the positive elements of the data. In the focus group discussions, KK and WP encouraged participants to express both their positive and negative perceptions, and we consciously sought divergent opinions within the data during analysis. KK and WP emphasized to participants that whatever they mentioned in this study would not affect them in any way in their assessments. To further improve the validity of the findings, all coauthors cross-checked the analysis of all five transcripts.

Results

Participants' Characteristics

In the current study, 38 undergraduate medical students participated in the five focus group discussions. Twenty-two out of 38 (60%) participants were Sinhalese, complying with the composition of the student cohorts in the universities where the study was conducted. Table 1 shows the composition of the focus group discussion participants.

Table . Composition of focus group discussion participants.

FGD[a] no	Total number of participants	Gender		Year of study		Ethnicity			
		Male	Female	3	4	Sinhalese	Tamil	Muslim	Other
1	7	4	3	7	0	5	0	2	0
2	8	3	5	0	8	5	1	0	2
3	7	3	4	0	7	5	1	1	0
4	9	4	5	9	0	5	1	2	1
5	7	2	5	7	0	3	0	3	1

[a]FGD: focus group discussion.

Three major themes emerged from the qualitative analysis: satisfaction with the perceived benefits of using AI for learning, negative attitude toward AI for learning due to perceived limitations, and optimism about the future use of AI to enhance student learning. The themes, subthemes, and initial codes are shown in Table 2.

Table . Identified themes, subthemes, and codes.

Themes	Subthemes	Codes
Satisfaction with the perceived benefits of using AI[a] for learning,	Improvement of knowledge	AI for supplementary learning, AI enable identification of knowledge gaps, using AI for problem solving, Using AI for in-depth learning, AI as an enabler of exploration and knowledge expansion, using AI for learning new words/unfamiliar content, AI for preliminary learning, AI for foundational learning.
	Enhanced efficiency	Ease of use of AI tools, using AI is time-saving, AI enables ease of access to information, AI provides streamlined information, AI gives focused answers, AI is useful for organizing information, AI helps directed learning, AI gives quick solutions, AI helps faster reading, using AI requires fewer resources for learning, AI simplifies content.
Negative attitude towards AI due to perceived limitations	Issues with relevance and accuracy	AI gives contradictory information, reliability concerns, no personalization of AI-generated responses, selective applicability of AI, contextual limitations of AI data, AI is fact-centered, AI generates less precise information, AI provides over-generalized information, teachers' preference for authenticity, need for cross-checking AI information, AI emphasizes minor issues.
	Impact on critical thinking skills	AI promotes surface-level thinking, use of AI causes passive learning, use of AI reduces reasoning ability, use of AI reduces generative ability, AI hinders independent thinking, use of AI reduces creativity.
	Impact on knowledge retention	AI enables short-term memory; using AI reduces long-term memory.
	Impact on collaborative learning and motivation	Use of AI reduces engagement, use of AI reduced peer interactions and fear of social isolation, using AI limits active participation, AI reduces learner motivation for engagement, AI reduces motivation for learning, using AI undermines learner preparation.
Optimism about the future use of AI to enhance student learning	Emerging awareness to guide for better use of AI for learning	AI for objective tasks, selectivity of AI for proper use, not using AI to obtain visual content, use AI for self-evaluation, AI as preparation for lecturer encounters, selective applicability of AI.
	Experiential learning to expand use of AI for learning	Learning from experience, learning AI boundaries, lesser concerns with experience.
	Desire for formal education of AI in curricula	Including AI in teaching, formal teaching of AI, inclusion of AI in curricula, teaching how best to use AI by lecturers.

[a]AI: artificial intelligence.

Satisfaction With the Perceived Benefits of Using AI for Learning

The theme of satisfaction with the perceived benefits of using AI for learning developed over many references across all student participants. The participants explained that their motivation to use AI was driven by factors such as simplicity, time-saving qualities, and efficient access to information. The participants unanimously agreed that AI tools improve their knowledge. AI served as a starting point for understanding concepts and generating initial ideas, which participants refined or expanded later using textbooks or other sources:

> *AI is a good way to get a basic idea. Once we get that basic idea it is easy to build upon that base by referring to textbooks and other resources.* [P04]

Participants valued AI for its assistance in answering follow-up questions, clarifying concepts, and exploring unfamiliar theoretical concepts until the students understood thoroughly. As one participant pointed out:

> *We can ask the question and get an answer. We can follow up until we understand it. We can ask again and again.* [P22]

AI helped most students to learn new words and expand their knowledge by identifying gaps or what they may have overlooked when learning or answering questions. Students used ChatGPT to compare the answers and identify missing points:

> *Then I look at the ChatGPT answer and realize that I have missed some points, so then I can go back and find these points and add those as well.* [P03]

All the participants appreciated AI tools for their efficiency and ease of use. AI was seen as an easily accessible, easy-to-use tool that enabled obtaining new ideas and simplified learning, which created an appeal to engage with it for learning. As one participant expressed:

> *AI is a very easy method to get an idea about our answers and it's very interesting. I like to use it.* [P23]

AI tools were perceived to provide concise, filtered, and summarized information, which provided organized and structured answers to questions. The participants preferred AI tools like ChatGPT over other search engines such as Google because AI was identified to deliver organized and to-the-point answers to their queries:

> *Rather than Googling a question, we prefer to use AI...coz we usually get an organized, structured answer to our questions there.* [P35]

Moreover, all the student participants appreciated the time-saving nature of AI. AI helped direct learners to essential concepts, saving time by offering streamlined information. According to the participants, AI significantly reduced the time required to accomplish tasks such as answering questions:

> *It is just simple because it's time saving when we are writing answers to a theory question.* [P27]

> *It showed you the direction of where you should go when you are writing an answer or when trying to read up something more.* [P15]

The participants elaborated that the time invested in perusing multiple sources such as videos or scientific reports can now be foregone, with the use of AI that gives the required information much faster than before they used AI for learning:

> *ChatGPT gives precise and summarized answers, so it takes less time compared to things like going through Youtube to learn something….No need to refer to many resources to formulate an answer also. Therefore, it is very convenient.* [P33]

AI was used by almost all the study participants mostly as a shortcut to avoid effort in deeper understanding or problem solving. The quick solutions given by AI with focused answers were viewed as a time-saving method of learning:

> *If I have the AI tool I always just look it up and then I get the answer straight away.* [P19]

Negative Attitude Toward AI for Learning Due to Perceived Limitations

While all participants confirmed the regular use of AI tools for academic purposes, a recurring element was a lack of certainty regarding the information obtained through AI tools such as ChatGPT. All the participants described a hesitancy to use AI for obtaining factual information, stating concerns over reliability. The students described instances where the information provided by the AI tool differed significantly from textbooks and lecture notes.

> *Sometimes the facts they (AI tool) give are different from the facts in lecture notes or reference books.* [P12]

The participants were discouraged from using AI tools to learn about infections or diseases. The students did not want AI assistance to explain the clinical reasoning processes. The students noted that AI failed to account for geographic, regional, and contextual data, particularly in clinical and epidemiological aspects, and provided contradictory information when compared with trusted sources.

> *After a few tries, I have stopped using AI for clinical work. It is not that much fitting to our setting. Mostly I go according to the textbook but sometimes I clear it out with a lecturer.* [P28]

Some study participants stressed that AI-generated answers sometimes include irrelevant or rare information, which emphasized minor issues, which were viewed as unhelpful and misleading.

> *AI gives very minor details at times which are not taught during lectures, and which were not included as objectives in classes.* [P32]

Hence, the AI outputs were not accepted at face value, and the discussion revealed that students frequently resorted to cross-checking information obtained from AI tools with trusted sources such as Medscape, research articles, or lecture notes to ensure accuracy.

> *Most of the time we also cross-check with textbooks or with other articles on the internet like Medscape or research articles.* [P13]

Moreover, some students were highly dissatisfied with the lack of personalization of AI-generated answers to questions. Students found that the general, less precise, and mechanistic nature of AI-generated information is of lesser value for learning. This feeling intensified as the students felt that AI failed to account for additional details or their individual thought processes, which made students lean toward their teachers to obtain information. As one participant pointed out:

> *It is not very personified…might be a bit of a negative thing. Like a teacher would understand what you are trying to say and tell you how to get about it…. but AI can't recognize that.* [P04]

Another issue that dissatisfied some students about using AI for their learning was the perceived negative impact on their critical thinking skills. The students were reluctant to use AI as they felt that AI promoted passive learning as it merely encouraged surface-level thinking. Overreliance on AI was felt by students as leading to decreased effort in deeper thought processes, hindering independent thinking, creativity, and reasoning ability:

> *Sometimes I feel like we don't think enough, especially during clinicals. Then I get the answer straight away so when I get used to that I stop starting to think about why something is what it is… like that.* [P07]

> *Like it disturbs our thinking ability a lot because you get used to just get it from AI without doing anything of our own. We probably won't be developing…or doing any thinking of our own.* [P26]

A few participants were highly worried regarding the perceived impact AI had on their knowledge retention. Students felt that the use of AI promoted only short-term memory and worried that relying on AI would affect their long-term retention as they compared learning from AI with that of peers or their teachers:

> *When we are discussing it with others, we feel we can remember it more. If we discuss it with my friends or lecturers, it goes to long-term memory and help to remember things more clearly.* [P11]

> *Sometimes I feel reluctant to use because we don't think enough. It's like a flash in the pan.* [P33]

Another significant barrier to using AI for learning among most medical students was the perceived negative impact on collaborative learning and motivation. Use of AI by students within the classroom for readily obtaining information meant that it undermined the prior preparation of enthusiastic students, which led to demotivation, reduced engagement, and active participation in the classroom.

> *We don't have to prepare and come anymore. It (using AI) actually decreased our participation and motivation.* [P35]

The participants saw the overreliance on AI as a potential threat to peer interaction and teacher–student relationships, which culminated in fear of social isolation.

> *As of now, we still interact with all the students and the lecturers so it hasn't affected us that much yet… but if we entirely rely on those things only, it will reduce our social interactions it's highly probable that it will affect us.* [P29]

Most students were discouraged from using AI tools for their learning by the attitude the teachers showed. As one participant discussed:

> *Lectures are very unhappy if we use ChatGPT or anything of the sort. Like, if they feel like we used it for…maybe an essay or whatever, they will shout basically. It's really easy to just be away from it (AI assistance) than get shouted at you know.* [P05]

Optimism About the Future Use of AI to Enhance Student Learning

However, most of the study participants expressed an optimistic view toward the use of AI in the future. This sentiment emerged due to learning from experience that happened over time, which developed an emerging awareness about mitigating issues they encountered before. The students identified the selective applicability of AI, stating various ways in which AI can be used and where AI has failed.

> *AI is good to get answers for most MCQs and also for structured answers like one-word answers. But it can't give pictures… for example, things like that AI can't give us.* [P21]

For most of the participants, their initial concerns regarding AI reliability diminished as they gained experience with determining what to ask and how to interpret AI responses.

> *Previously we were worried. But it's not there now. Much. Like we know what should be asked from AI and what should not be asked of AI. It's good for self-evaluation because we can practice with it.* [P34]

Some participants of the study used optimism-building to prepare for lecturer encounters, which resulted in clarifying doubts and enabling deeper learning.

> *There's a limit to what you can learn from AI. But if you learn it and then come to a class, then you understand more with what the lecturer says.* [P01]

The students acknowledged that the age of AI is upon them and acutely felt the need to use AI for academic purposes.

> *It's like you can't not use it at all right? Everybody use AI for something or other so you would lose if you don't know how to use it right? So you have to learn one way or another.* [P13]

Building awareness for better use of AI resulted in a desire among the medical students for formal teaching of the ethical use of AI. Almost all students felt that if the teachers "taught" them how to use it to improve their learning, they would use it better in a more reliable manner. As one participant expressed not so eloquently:

> *If they (lecturers) taught us how to use AI properly rather than shouting at us for doing it, I think we would understand how to do it okay. Some of them (lecturers) talk sort of very… I mean look down on us completely, if they feel like we have used it in like in a presentation or anything. Is it such a big deal? Should we not use AI at all?* [P22]

Most participants of the study felt that formalizing AI learning by various means would facilitate ethical use of AI for both teaching by the teachers and learning by the students.

> *Teaching how to give the commands to ChatGPT, and like what can be done and what can't be done or things like what are the things that are okay to ask from AI… Also like what is okay to use AI for and what not to. Is it ok to get a title or a picture from AI*

or not? If you get that early on from the lecturers, then we can use it more effectively I think. [P09]

Discussion

Principal Results

AI has become increasingly integrated into educational settings, offering both opportunities and challenges. This qualitative analysis reveals a complex and nuanced picture of both enthusiasm and concerns. The majority of participating medical students believe that AI is important for enhanced learning and desire to learn more about AI for better use. We also found that despite these attitudes, there remains a reserve for using AI in education brought on by concerns about the reliability, privacy, and ethical issues and a lack of integration of AI into medical education in Sri Lanka. The clinical, scientific, economic, and ethical future of health care will be significantly impacted by AI [46]. The current generation of medical doctors is set to enter an industry that is significantly more AI - driven than it was when their training commenced. The current study opens the dialog for an inevitable yet successful implementation of AI in medical education.

The participants viewed AI tools as supplementary learning aids, helping students identify knowledge gaps and facilitating problem-solving. Additionally, AI facilitated exploration and aided learning new vocabulary and unfamiliar content, acting as both a preliminary and foundational learning resource. These capabilities align with findings that AI can provide personalized learning experiences, thereby enhancing knowledge acquisition and skill development [47]. The efficiency was identified as another significant benefit of AI tools. This was documented in previous studies where efficiency allowed students to allocate more time to higher-order thinking tasks, thereby enhancing the overall learning experience [48]. The enthusiasm shown by medical students in the global arena is reflected in the current study, who also report limited AI literacy and a wish for formal learning of AI to enhance student learning [15,49,50].

Despite its benefits, study participants were reluctant to use AI for learning mainly due to inaccuracies and contextually irrelevant information that arises from AI. Studies have revealed the necessity of learning how to manage AI-driven misinformation [51,52]. Although AI-assisted learning has been shown to promote active learning and learner engagement [53], the participants of this study perceived that relying on AI would affect long-term retention of knowledge negatively. Moreover, the study demonstrated concerns that reliance on AI may diminish critical thinking abilities. Consistently depending on AI tools has led to deteriorating basic foundational skills, critical thinking, and problem-solving skills [54]. Overdependence on AI tools has been shown to lead to cognitive offloading, where individuals rely on technology for tasks that require analytical thinking, potentially hindering the development of critical thinking skills [47]. This reliance may result in a superficial understanding of content, as students might accept AI-generated information without thorough evaluation.

Contrary to the evidence supporting enhanced collaborative learning in AI platforms [55], the present study reveals the negative impact of AI-assisted learning on collaborative learning. The participants of the study revealed a demotivation to engage with the peers and tutors for learning, due to the ease of accessing required information through AI platforms. Hence, students feared that the use of AI for learning might result in social isolation. However, students were adamant that the envisioned social isolation had only marginally affected them at present. It is important to note that these sentiments could negatively impact collaborative learning experiences and the development of communication skills, which are essential for critical thinking and knowledge application. Moreover, the participants were also concerned about the lack of personalization that human educators provide. Impersonal feedback may not address individual student needs effectively, leading to decreased motivation and engagement in learning activities [47]. These findings of this study demonstrate the importance of equipping medical students as well as educators with the tools necessary to enhance learner engagement in AI-assisted learning platforms [56].

Interestingly, the study participants appeared hopeful regarding the future use of AI to enhance their learning. The students emphasized the importance of emerging and building awareness and experiential learning to expand the use of AI for learning. The students identified the areas where AI tools fail, showing their growing awareness of the boundaries of AI-based assistance. However, they expected formal teaching of how best to use AI for learning by the faculty. This was amidst contradictory messages students received from educators that spanned a ban on incorporating AI into education to embracing AI-assisted learning. In a background where the evidence is overwhelmingly supportive of the ability of AI to enhance student learning [57], the overt reluctance of educators to accept that students are embracing the inevitable is questionable. This attitude may be because of regional factors or due to the traditional teaching/learning methods largely utilized in the education systems in Sri Lanka [58]. This sentiment may also stem from the apparent mistrust the educators may have on the reliability of AI-generated information [59] and doubts regarding students' misuse of AI tools for assignments and exams, affecting critical thinking and information retrieval skills [60]. Another study revealed that the educators were concerned about students' self-reliance on AI applications at the cost of traditional teaching methods, which might deprive them of skills best learned in person or group teaching [61]. However, overcoming such reservations is critical to keep abreast of the advancing technologies and ensure that educators facilitate medical students' learning in the age of technological revolution [62].

A recurring finding in this study was the strong student desire for guidance from faculty on how to use AI effectively for learning. However, this raises an important question: Should all medical educators, including clinicians like surgeons or pediatricians, be expected to teach AI-related content? Given the complex and evolving nature of AI and the already demanding responsibilities of clinical faculty, assigning this task universally may be unrealistic and potentially problematic. The concern is that untrained educators could unintentionally provide inaccurate or conflicting guidance, which was reflected in the confusion expressed by students in our study. Rather than

placing this responsibility on individual subject-area faculty, institutions should adopt a more structured and strategic approach. Drawing on Chan's AI Ecological Education Policy Framework, AI education should be delivered through dedicated units or trained personnel within the institution, ideally through interdisciplinary collaboration. This would ensure consistency, ethical alignment, and adaptability as AI technologies evolve [63]. Formal training and centralized resources for both faculty and students can support responsible AI use while relieving clinical educators of the expectation to independently master and teach AI tools.

While this study briefly referenced ethical concerns related to AI, it is important to more fully acknowledge the complex ethical landscape surrounding its use in medical education. Beyond data privacy and accuracy, there are deeper concerns regarding student development and equity. As Masters outlines in AMEE Guide No. 158, the use of AI tools may inadvertently undermine core competencies in learners, such as critical appraisal, summarization, and ethical decision-making, if overused or uncritically adopted. Furthermore, algorithmic bias, lack of transparency in AI-generated outputs, and passive data surveillance present significant risks to fairness, autonomy, and human dignity in the educational process. These tools can obscure the need for empathy, reflection, and professional judgment—essential traits of a medical practitioner [64].

Similarly, Alam et al argue that medical education must move beyond basic AI literacy to include a critical understanding of the ethical implications of AI-assisted learning and publishing. This includes interrogating how AI might alter authorship norms, contribute to inequities in access and performance, or even reinforce existing systemic biases if not carefully regulated [65]. The path forward, therefore, requires not just training students to use AI responsibly but also preparing institutions to adopt clear ethical guidelines and foster a culture of reflective, values-based technology use. As AI becomes increasingly integrated into medical education and health care, ethical fluency must be considered as essential as digital literacy for both students and educators.

Limitations

The present study has some limitations. The study did not quantitatively evaluate the impact of AI-assisted learning on achieving learning outcomes, satisfaction, or other measurable aspects of medical education, which could supplement the qualitative findings of this study. Additionally, since the study's focus was on understanding the perception of students, the perspectives of other stakeholders, such as the faculty and health care professionals, were not captured, and this could be explored in future research.

Conclusions

In a region where integrating AI into medical curricula is lacking, this study adds to our understanding of what medical students think about the challenges of AI tools in medical education. The present study highlights the improvement of knowledge and enhanced efficiency as the two primary advantages of AI-assisted learning as perceived by medical students. However, several concerns were interwoven within the widespread adoption of AI among the study participants. The reliance on AI was felt to have a negative impact on critical thinking and cognitive engagement, as students prioritized convenience over deep learning. Interestingly, students feared social isolation as a possible future impact of integrating AI into their learning. Students' lack of understanding of how much AI assistance is accepted at an institutional level and the need to "play" according to teacher preferences confused and hindered the students despite their enthusiasm to use AI for learning. A balanced approach is warranted to maximize the benefits of AI in education while mitigating its limitations. AI should complement, rather than replace, traditional learning methods, with educators guiding students on how to use AI effectively. Future research to explore the contextual factors and policy efforts is critical to refine AI's role in education, ensuring it enhances learning outcomes while preserving essential cognitive and social development.

Acknowledgments

All authors declared that they had insufficient funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee (APF) support for the publication of this article.

Data Availability

The transcripts generated and analyzed during this study are not publicly available but can be obtained from the corresponding author on reasonable request.

Authors' Contributions

TS and KK contributed equally to conceptualization, designing, implementation, data analysis, writing, and revising the manuscript. IA and WP contributed equally to implementation, analysis, and revising the manuscript. All authors reviewed and approved the final version of the manuscript.

Conflicts of Interest

None declared.

References

1. Pearce C, McLeod A, Rinehart N, Whyte R, Deveny E, Shearer M. Artificial intelligence and the clinical world: a view from the front line. Med J Aust 2019 Apr;210 Suppl 6:S38-S40. [doi: 10.5694/mja2.50025] [Medline: 30927469]
2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
3. Matheny M, Israni ST, Ahmed M, Whicher D. Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril: National Academy of Medicine; 2019. URL: https://nam.edu/wp-content/uploads/2021/07/4.3-AI-in-Health-Care-title-authors-summary.pdf [accessed 2025-09-16]
4. Civaner MM, Uncu Y, Bulut F, Chalil EG, Tatli A. Artificial intelligence in medical education: a cross-sectional needs assessment. BMC Med Educ 2022 Nov 9;22(1):772. [doi: 10.1186/s12909-022-03852-3] [Medline: 36352431]
5. Suh I, McKinney T, Siu KC. Current perspective of metaverse application in medical education, research and patient care. Virtual Worlds 2023;2(2):115-128. [doi: 10.3390/virtualworlds2020007]
6. Malhotra K, Wong BNX, Lee S, et al. Role of artificial intelligence in global surgery: a review of opportunities and challenges. Cureus 2023 Aug;15(8):e43192. [doi: 10.7759/cureus.43192] [Medline: 37692604]
7. Dave M, Patel N. Artificial intelligence in healthcare and education. Br Dent J 2023 May;234(10):761-764. [doi: 10.1038/s41415-023-5845-2] [Medline: 37237212]
8. Gomes RFT, Schmith J, Figueiredo RD, et al. Use of artificial intelligence in the classification of elementary oral lesions from clinical images. Int J Environ Res Public Health 2023 Feb 22;20(5):3894. [doi: 10.3390/ijerph20053894] [Medline: 36900902]
9. Kooli C. Chatbots in education and research: a critical examination of ethical implications and solutions. Sustainability 2023;15(7):5614. [doi: 10.3390/su15075614]
10. Roda S. Digital skills for doctors—explaining European doctors' position. J Eur CME 2021;10(1):2014097. [doi: 10.1080/21614083.2021.2014097] [Medline: 34912589]
11. WMA Statement on Augmented Intelligence in Medical Care. URL: https://www.wma.net/policies-post/wma-statement-on-augmented-intelligence-in-medical-care/ [accessed 2025-01-20]
12. Wartman SA, Combs CD. Medical education must move from the information age to the age of artificial intelligence. Acad Med 2018 Aug;93(8):1107-1109. [doi: 10.1097/ACM.0000000000002044] [Medline: 29095704]
13. Lee J, Wu AS, Li D, Kulasegaram KM. Artificial intelligence in undergraduate medical education: a scoping review. Acad Med 2021 Nov 1;96(11S):S62-S70. [doi: 10.1097/ACM.0000000000004291] [Medline: 34348374]
14. Oh S, Kim JH, Choi SW, Lee HJ, Hong J, Kwon SH. Physician confidence in artificial intelligence: an online mobile survey. J Med Internet Res 2019 Mar 25;21(3):e12422. [doi: 10.2196/12422] [Medline: 30907742]
15. Pinto Dos Santos D, Giese D, Brodehl S, et al. Medical students' attitude towards artificial intelligence: a multicentre survey. Eur Radiol 2019 Apr;29(4):1640-1646. [doi: 10.1007/s00330-018-5601-1] [Medline: 29980928]
16. McLennan S, Meyer A, Schreyer K, Buyx A. German medical students´ views regarding artificial intelligence in medicine: a cross-sectional survey. PLOS Digit Health 2022 Oct;1(10):e0000114. [doi: 10.1371/journal.pdig.0000114] [Medline: 36812635]
17. Castagno S, Khalifa M. Perceptions of artificial intelligence among healthcare staff: a qualitative survey study. Front Artif Intell 2020;3:578983. [doi: 10.3389/frai.2020.578983] [Medline: 33733219]
18. Bisdas S, Topriceanu CC, Zakrzewska Z, et al. Artificial intelligence in medicine: a multinational multi-center survey on the medical and dental students' perception. Front Public Health 2021;9:795284. [doi: 10.3389/fpubh.2021.795284] [Medline: 35004598]
19. Pucchio A, Rathagirishnan R, Caton N, et al. Exploration of exposure to artificial intelligence in undergraduate medical education: a Canadian cross-sectional mixed-methods study. BMC Med Educ 2022 Nov 28;22(1):815. [doi: 10.1186/s12909-022-03896-5] [Medline: 36443720]
20. Mehta N, Harish V, Bilimoria K, et al. Knowledge of and attitudes on artificial intelligence in healthcare: a provincial survey study of medical students. Medical Education. Preprint posted online on Jan 15, 2021. [doi: 10.1101/2021.01.14.21249830]
21. Collado-Mesa F, Alvarez E, Arheart K. The role of artificial intelligence in diagnostic radiology: a survey at a single radiology residency training program. J Am Coll Radiol 2018 Dec;15(12):1753-1757. [doi: 10.1016/j.jacr.2017.12.021] [Medline: 29477289]
22. Jackson P, Ponath Sukumaran G, Babu C, et al. Artificial intelligence in medical education—perception among medical students. BMC Med Educ 2024 Jul 27;24(1):804. [doi: 10.1186/s12909-024-05760-0] [Medline: 39068482]
23. Ahmed Z, Bhinder KK, Tariq A, et al. Knowledge, attitude, and practice of artificial intelligence among doctors and medical students in Pakistan: a cross-sectional online survey. Ann Med Surg (Lond) 2022 Apr;76:103493. [doi: 10.1016/j.amsu.2022.103493] [Medline: 35308436]
24. Scheetz J, Rothschild P, McGuinness M, et al. A survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology and radiation oncology. Sci Rep 2021 Mar 4;11(1):5193. [doi: 10.1038/s41598-021-84698-5] [Medline: 33664367]

25. Tung AYZ, Dong LW. Malaysian medical students' attitudes and readiness toward AI (artificial intelligence): a cross-sectional study. J Med Educ Curric Dev 2023;10:23821205231201164. [doi: 10.1177/23821205231201164] [Medline: 37719325]
26. Yüzbaşıoğlu E. Attitudes and perceptions of dental students towards artificial intelligence. J Dent Educ 2021 Jan;85(1):60-68. [doi: 10.1002/jdd.12385] [Medline: 32851649]
27. Jebreen K, Radwan E, Kammoun-Rebai W, et al. Perceptions of undergraduate medical students on artificial intelligence in medicine: mixed-methods survey study from Palestine. BMC Med Educ 2024 May 7;24(1):507. [doi: 10.1186/s12909-024-05465-4] [Medline: 38714993]
28. Aboalshamat K, Alhuzali R, Alalyani A, et al. Medical and dental professionals readiness for artificial intelligence for Saudi Arabia Vision 2030. Int J Pharm Res Allied Sci 2022;11(4):52-59. [doi: 10.51847/NU8y6Y6q1M]
29. Khater AS, Zaaqoq AA, Wahdan MM, Ashry S. Knowledge and attitude of Ain Shams University medical students towards artificial intelligence and its application in medical education and practice. Educ Res Innov J 2023 Jul 1;3(10):29-42. [doi: 10.21608/erji.2023.306718]
30. Swed S, Alibrahim H, Elkalagi NKH, et al. Knowledge, attitude, and practice of artificial intelligence among doctors and medical students in Syria: a cross-sectional online survey. Front Artif Intell 2022;5:1011524. [doi: 10.3389/frai.2022.1011524] [Medline: 36248622]
31. Al-Qerem W, Eberhardt J, Jarab A, et al. Exploring knowledge, attitudes, and practices towards artificial intelligence among health professions' students in Jordan. BMC Med Inform Decis Mak 2023 Dec 14;23(1):288. [doi: 10.1186/s12911-023-02403-0] [Medline: 38098095]
32. Boillat T, Nawaz FA, Rivas H. Readiness to embrace artificial intelligence among medical doctors and students: questionnaire-based study. JMIR Med Educ 2022 Apr 12;8(2):e34973. [doi: 10.2196/34973] [Medline: 35412463]
33. Buabbas AJ, Miskin B, Alnaqi AA, et al. Investigating students' perceptions towards artificial intelligence in medical education. Healthcare (Basel) 2023 May 1;11(9):1298. [doi: 10.3390/healthcare11091298] [Medline: 37174840]
34. De Silva AP, Liyanage IK, De Silva STG, Jayawardana MB, Liyanage CK, Karunathilake IM. Migration of Sri Lankan medical specialists. Hum Resour Health 2013 May 21;11:1-6. [doi: 10.1186/1478-4491-11-21] [Medline: 23693092]
35. Mann K, MacLeod A. Constructivism: learning theories and approaches to research. Res Med Edu 2015 Jul(15):49-66. [doi: 10.1002/9781118838983]
36. Saunders MNK, Lewis P, Thornhill A, Bristow A. Understanding Research Philosophy and Approaches to Theory Development: University of Birmingham; 2015. URL: https://www.researchgate.net/publication/309102603_Understanding_research_philosophies_and_approaches [accessed 2025-09-19]
37. Kvale S. The 1,000-page question. Qual Inq 1996 Sep;2(3):275-284. [doi: 10.1177/107780049600200302]
38. Cohen L, Manion L, Morrison K. Research Methods in Education, 5th edition: Published online alone; 2013. [doi: 10.4324/9780203224342]
39. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
40. Braun V, Clarke V. To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales. Qual Res Sport Exerc Health 2021 Mar 4;13(2):201-216. [doi: 10.1080/2159676X.2019.1704846]
41. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
42. Braun V, Clarke V. Conceptual and design thinking for thematic analysis. Qual Psychol 2022;9(1):3-26. [doi: 10.1037/qup0000196]
43. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health 2019 Aug 8;11(4):589-597. [doi: 10.1080/2159676X.2019.1628806]
44. Yardley L. Demonstrating validity in qualitative psychology. In: Qualitative Psychology: A Practical Guide to Research Methods, 2nd edition: Sage; 2008, Vol. 2:235-251.
45. Shaw R. Embedding reflexivity within experiential qualitative psychology. Qual Res Psychol 2010 Aug 26;7(3):233-243. [doi: 10.1080/14780880802699092]
46. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019 Jun;6(2):94-98. [doi: 10.7861/futurehosp.6-2-94] [Medline: 31363513]
47. Zhai C, Wibowo S, Li LD. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. Smart Learn Environ 2024;11(1). [doi: 10.1186/s40561-024-00316-7]
48. St-Hilaire F, Do VD, Frau A, et al. A new era: intelligent tutoring systems will transform online learning for millions. arXiv. Preprint posted online on Mar 3, 2022. [doi: 10.48550/arXiv.2203.03724]
49. Wood EA, Ange BL, Miller DD. Are we ready to integrate artificial intelligence literacy into medical school curriculum: students and faculty survey. J Med Educ Curric Dev 2021;8:23821205211024078. [doi: 10.1177/23821205211024078] [Medline: 34250242]
50. Kimmerle J, Timm J, Festl-Wietek T, Cress U, Herrmann-Werner A. Medical students' attitudes toward AI in medicine and their expectations for medical education. J Med Educ Curric Dev 2023;10:23821205231219346. [doi: 10.1177/23821205231219346] [Medline: 38075443]

51. Park SH, Do KH, Kim S, Park JH, Lim YS. What should medical students know about artificial intelligence in medicine? J Educ Eval Health Prof 2019;16:18. [doi: 10.3352/jeehp.2019.16.18] [Medline: 31319450]
52. Ochodo EA, de Haan MC, Reitsma JB, Hooft L, Bossuyt PM, Leeflang MMG. Overinterpretation and misreporting of diagnostic accuracy studies: evidence of "spin". Radiology 2013 May;267(2):581-588. [doi: 10.1148/radiol.12120527] [Medline: 23360738]
53. Aluko HA, Aluko A, Offiah GA, et al. Exploring the effectiveness of AI-generated learning materials in facilitating active learning strategies and knowledge retention in higher education. IJOA 2025. [doi: 10.1108/IJOA-07-2024-4632]
54. Basha JY, IJSSC. The negative impacts of AI tools on students in academic and real-life performance. J Yunus Basha,/IJSSC 2024;1(3):1-16. [doi: 10.51470/IJSSC.2024.01.03.01]
55. Msambwa MM, Wen Z, Daniel K. The impact of AI on the prsonal and collaborative learning environments in higher education. Euro J Educ 2025 Mar;60(1):e12909 [FREE Full text] [doi: 10.1111/ejed.12909]
56. Tan SC, Lee AVY, Lee M. A systematic review of artificial intelligence techniques for collaborative learning over the past two decades. Comput Educ: Artif Intell 2022;3:100097. [doi: 10.1016/j.caeai.2022.100097]
57. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. JMIR Med Educ 2019 Jun 15;5(1):e13930. [doi: 10.2196/13930] [Medline: 31199295]
58. Kodikara K, Seneviratne T, Godamunne P, Premaratna R. Challenges in learning procedural skills: student perspectives and lessons learned for curricular design. Teach Learn Med 2024;36(4):435-453. [doi: 10.1080/10401334.2023.2226633] [Medline: 37350450]
59. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019 Dec 3;5(2):e16048. [doi: 10.2196/16048] [Medline: 31793895]
60. Gülhan Güner S, Yiğit S, Berşe S, Dirgar E. Perspectives and experiences of health sciences academics regarding ChatGPT: a qualitative study. Med Teach 2025 Mar;47(3):550-559. [doi: 10.1080/0142159X.2024.2413425] [Medline: 39392461]
61. Banerjee M, Chiew D, Patel KT, et al. The impact of artificial intelligence on clinical education: perceptions of postgraduate trainee doctors in London (UK) and recommendations for trainers. BMC Med Educ 2021 Aug 14;21(1):429. [doi: 10.1186/s12909-021-02870-x] [Medline: 34391424]
62. Abouammoh N, Alhasan K, Aljamaan F, et al. Perceptions and earliest experiences of medical students and faculty with ChatGPT in medical education: qualitative study. JMIR Med Educ 2025 Feb 20;11:e63400. [doi: 10.2196/63400] [Medline: 39977012]
63. Chan CKY. A comprehensive AI policy education framework for university teaching and learning. Int J Educ Technol High Educ 2023;20(1):38. [doi: 10.1186/s41239-023-00408-3]
64. Masters K. Ethical use of artificial intelligence in health professions education: AMEE Guide No. 158. Med Teach 2023 Jun;45(6):574-584. [doi: 10.1080/0142159X.2023.2186203] [Medline: 36912253]
65. Alam F, Lim MA, Zulkipli IN. Integrating AI in medical education: embracing ethical usage and critical understanding. Front Med (Lausanne) 2023;10:1279707. [doi: 10.3389/fmed.2023.1279707] [Medline: 37901398]

Abbreviations

AI: artificial intelligence
COREQ: Consolidated Criteria for Reporting Qualitative Research
TA: thematic analysis

Edited by J Gentges; submitted 12.03.25; peer-reviewed by C Ma, RA Muaygil; revised version received 09.07.25; accepted 16.07.25; published 19.09.25.

Please cite as:
Seneviratne T, Kodikara K, Abeykoon I, Palpola W
Perception of Medical Undergraduates on Artificial Intelligence in Medical Education: Qualitative Exploration
JMIR Med Educ 2025;11:e73798
URL: https://mededu.jmir.org/2025/1/e73798
doi:10.2196/73798

Impact of Prompt Engineering on the Performance of ChatGPT Variants Across Different Question Types in Medical Student Examinations: Cross-Sectional Study

Ming-Yu Hsieh[1,2], MD, PhD; Tzu-Ling Wang[3], MSc, RNC; Pen-Hua Su[2,4*], MD, PhD; Ming-Chih Chou[2,5*], MD, PhD

[1]Division of Pediatric Surgery, Department of Surgery, Chung Shan Medical University Hospital, Taichung City, Taiwan

[2]Institute of Medicine, School of Medicine, Chung Shan Medical University, No. 110, Sec. 2, Jiang Kuo South Road, South district, Taichung City, Taiwan

[3]Department of Nursing, School of Medicine, Chung Shan Medical University, Taichung City, Taiwan

[4]Department of Pediatrics, Chung Shan Medical University Hospital, Taichung City, Taiwan

[5]Division of Thoracic Surgery, Department of Surgery, Chung Shan Medical University Hospital, Taichung City, Taiwan

[*]these authors contributed equally

Corresponding Author:
Ming-Chih Chou, MD, PhD
Institute of Medicine, School of Medicine, Chung Shan Medical University, No. 110, Sec. 2, Jiang Kuo South Road, South district, Taichung City, Taiwan

Abstract

Background: Large language models such as ChatGPT (OpenAI) have shown promise in medical education assessments, but the comparative effects of prompt engineering across optimized variants and relative performance against medical students remain unclear.

Objective: This study aims to systematically evaluate the impact of prompt engineering on five ChatGPT variants (GPT-3.5, GPT-4.0, GPT-4o, GPT-4o1-mini, and GPT-4o1) and benchmark their performance against fourth-year medical students in midterm and final examinations.

Methods: A 100-item examination dataset covering multiple choice questions, short answer questions, clinical case analysis, and image-based questions was administered to each model under no-prompt and prompt-engineering conditions over 5 independent runs. Student cohort scores (N=143) were collected for comparison. Responses were scored using standardized rubrics, converted to percentages, and analyzed in SPSS Statistics (v29.0) with paired *t* tests and Cohen *d* ($P<.05$).

Results: Baseline midterm scores ranged from 59.2% (GPT-3.5) to 94.1% (GPT-4o1), and final scores ranged from 55% to 92.4%. Fourth-year students averaged 89.4% (midterm) and 80.2% (final). Prompt engineering significantly improved GPT-3.5 (10.6%, $P<.001$) and GPT-4.0 (3.2%, $P=.002$) but yielded negligible gains for optimized variants ($P=.07$ - $.94$). Optimized models matched or exceeded student performance on both exams.

Conclusions: Prompt engineering enhances early-generation model performance, whereas advanced variants inherently achieve near-ceiling accuracy, surpassing medical students. As large language models mature, emphasis should shift from prompt design to model selection, multimodal integration, and critical use of artificial intelligence as a learning companion.

(JMIR Med Educ 2025;11:e78320) doi:10.2196/78320

KEYWORDS

ChatGPT; prompt engineering; medical education; large language models; assessment performance

Introduction

The integration of large language models (LLMs), such as OpenAI's ChatGPT series, into medical education has generated considerable interest due to their potential for automating and enhancing assessment processes [1]. Initial investigations focused on foundational models—ChatGPT 3.5 and ChatGPT 4.0—to benchmark baseline performance in answering clinical and basic science questions representative of medical student examinations [2,3]. These early studies demonstrated that, compared to ChatGPT 3.5, ChatGPT 4.0 exhibited significant improvements in overall accuracy, reasoning ability, and contextual understanding, particularly in complex scenario-based items [3]. Prompt engineering—providing structured guidance within the input prompt—was shown to further augment performance for both models, yielding notable score increases and reducing variance across repeated trials [4].

Since the release of ChatGPT 4.0, OpenAI has introduced enhanced variants optimized for performance and generalization: GPT-4o (optimized for multimodal tasks), GPT-4o1-mini (a compact, latency-reduced iteration), and GPT-4o1 (the full-capacity optimized version) [5,6]. However, the comparative effects of prompt engineering on these advanced models remain unexplored. Given their architectural refinements and improved instruction-following capabilities, it is plausible that newer variants may naturally internalize prompt structures, diminishing the incremental benefit of explicit guidance.

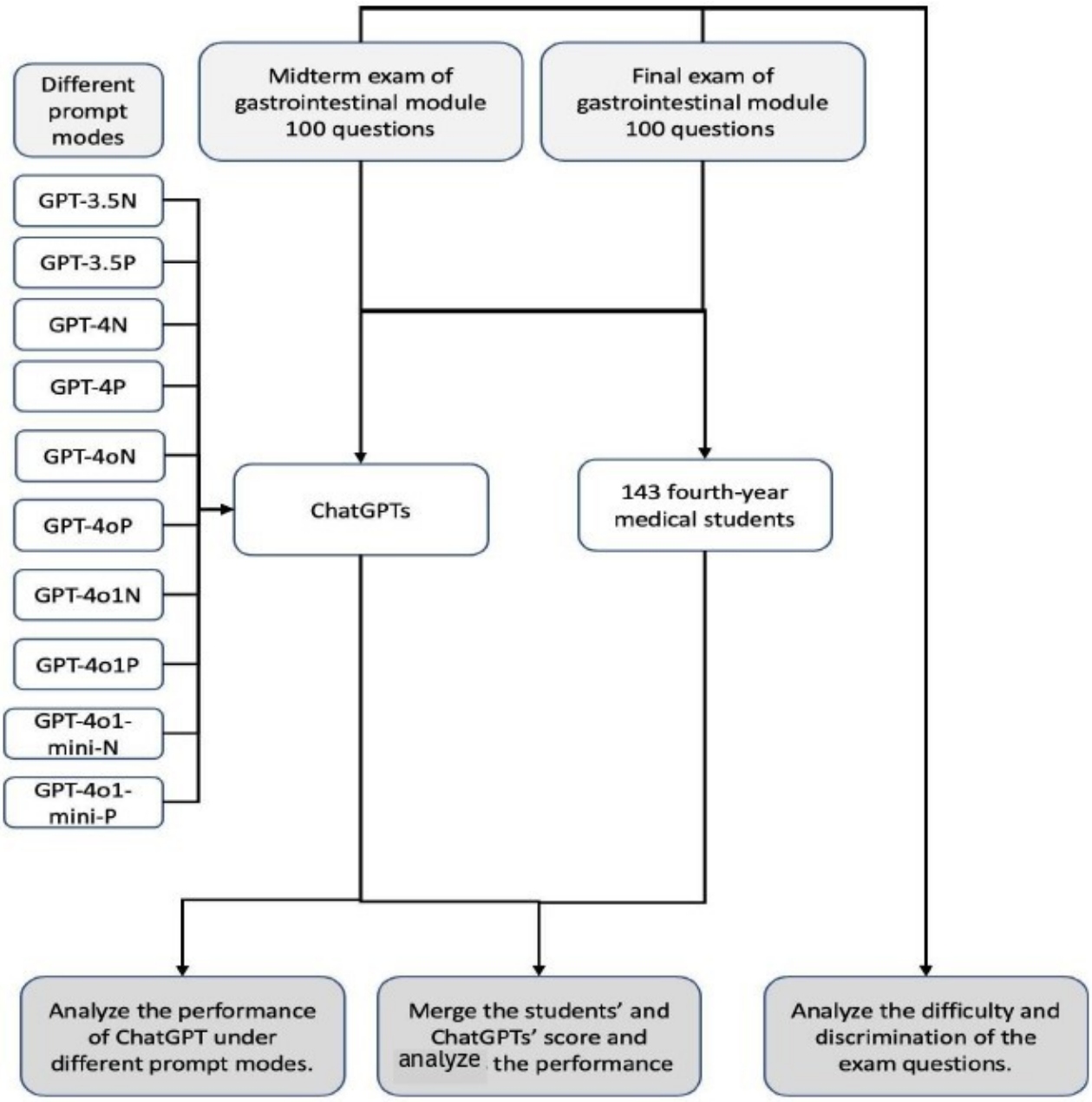
This study expands upon prior work by systematically evaluating the impact of prompt engineering across 5 ChatGPT variants—3.5, 4.0, 4o, 4o1mini, and 4o1—using a robust examination framework comprising multiple question types (multiple choice questions, MCQs; short answer questions, SAQs; clinical case analysis, CCA; and image-based interpretation, IBI). By comparing performance with and without structured prompts, we aim to quantify the degree to which prompt dependency has evolved alongside model iterations. The findings will inform best practices for leveraging LLMs in high-stakes medical education settings and contribute to understanding the maturation of prompt engineering as a methodology.

Methods

Study Design and Model Selection

This cross-sectional evaluation compared the performance of 5 OpenAI GPT variants: ChatGPT 3.5 (GPT-3.5-turbo), ChatGPT 4.0 (GPT-4), GPT-4o (optimized for multimodal tasks), GPT-4o1-mini (compact and latency-reduced), and GPT-4o1 (full-capacity optimized). Each variant was assessed under 2 prompting conditions: without a structured prompt (N) and with prompt engineering (P) (Figure 1).

Figure 1. Workflow of the study. Midterm and final gastrointestinal-module exams (100 items each) were administered to 143 fourth-year medical students and to 5 ChatGPT variants (GPT-3.5, GPT-4, GPT-4o, GPT-4o1-mini, and GPT-4o1) under 2 prompting conditions (no prompt vs prompt engineering). Each model or condition combination was run 5 times, responses were scored against the official key by 2 blinded reviewers, and performance metrics (mean, SD, *P* values, and effect sizes) as well as exam item difficulty and discrimination were analyzed.



Selecting the Student Cohort

Our curriculum uses modular teaching from the second semester of year 3 through the first semester of year 4. We selected fourth-year medical students to ensure participants had completed the modular clinical instruction, avoiding the variability introduced by students newly exposed to clinical modules. This choice ensures that the student cohort had comparable training before exam administration.

Examination Dataset and Question Types

We curated a 100-item question set drawn from the official medical student midterm and final examinations administered at Chung Shan Medical University in the 2024‐2025 academic year. The items encompassed 4 question types:

- MCQs: single-best-answer format (n=40)
- SAQs: 1‐2 sentence responses (n=20)
- CCA: open-ended diagnostic and management scenarios (n=20)
- IBI: radiographic or histologic images requiring identification or explanation (n=20)

Prompt Engineering

For the P condition, each item was prefaced with a standardized instruction prompt:

"You are an expert medical educator. Answer the following question concisely and justify your reasoning step by step."

In the N condition, only the question stem was presented.

Evaluation Procedure

For each model and condition, we conducted 5 independent runs (rounds) to account for stochastic variations. In each run, the model received the full 100-item set sequentially via the OpenAI API (v1) with default temperature settings (temperature=0.7, max_tokens=512). Outputs were recorded and collated.

Scoring and Outcome Measures

Responses were scored against the official answer key by 2 independent reviewers blinded to model and condition. MCQs were scored dichotomously (1 point for correct answer, otherwise 0). SAQs and IBI were scored on a 0‐2 scale (0=incorrect or missing, 1=partially correct, and 2=fully correct). CCA responses were scored on a 0‐3 rubric evaluating diagnostic accuracy, management plan, and justification. Total raw scores were converted to percentages of maximum possible scores.

Statistical Analysis

Statistical analyses were conducted in SPSS Statistics (version 29, IBM Corp). For each model and condition, descriptive statistics (mean scores and SD) were derived using the frequencies and descriptives procedures. Paired 2-tailed *t* tests (paired-samples *t* test) compared scores between the no-prompt (N) and prompt-engineering (P) conditions within each variant. Effect sizes (Cohen *d*) were calculated based on mean differences and pooled SDs. Statistical significance was set at $P<.05$, and all tests were 2-sided.

Ethical Considerations

The study protocol was reviewed and approved by the Institutional Review Board (IRB) of Chung Shan Medical University Hospital (approval number CSMU-2024-075), in accordance with institutional policies and the Declaration of Helsinki. The IRB granted a waiver of written informed consent because the research analyzed routinely collected deidentified examination records, involved no direct interaction or intervention with students, and posed minimal risk. Before transfer to the study team, all records were deidentified by the medical school; no direct identifiers (eg, names, student IDs, email, and IP addresses) were accessed. Analyses were performed on files labeled with random study codes, and only aggregate results are reported. Data were stored on password-protected institutional servers with access restricted to the research team and will be retained per institutional policy; no individual-level raw data will be publicly shared. Participants received no compensation, and inclusion in the dataset had no impact on grades, course standing, or academic evaluation.

Results

Overall Performance on Midterm and Final Examinations

As summarized in Table 1, baseline performance of ChatGPT 3.5 on the midterm examination was 59.2% (SD 2.1), whereas ChatGPT 4.0 achieved 81.4% (SD 1.8). The optimized variants—GPT-4o, GPT-4o1-mini, and GPT-4o1—further improved mean midterm scores to 91.3% (SD 0.8), 86.1% (SD 1), and 94.1% (SD 0.5), respectively (Table 1). A similar trend was observed for the final examination (Table 1), where GPT-3.5 scored 55% (SD 2.4) and GPT-4.0 scored 84.2% (SD 1.7), with GPT-4o, GPT-4o1-mini, and GPT-4o1 achieving 90.6% (SD 0.9), 82.1% (SD 0.6), and 92.4% (SD 0.6), respectively.

Table . Basic information of the exams: overall performance of ChatGPT variants on midterm and final examinations. Mean percentage scores (SD) for 5 GPT models (GPT-3.5, GPT-4.0, GPT-4o, GPT-4o1-mini, and GPT-4o1) under no-prompt and prompt-engineering conditions are listed, illustrating baseline accuracy and comparative gains across both examinations.

Exams	Midterm exams	Final exams
Total questions, N	100	100
Overall discrimination, mean (SD)	0.25 (0.18)	0.32 (0.21)
Overall difficulty level, mean (SD)	0.82 (0.14)	0.72 (0.18)
Memorization questions, n	66	63
Discrimination, mean (SD)	0.27 (0.16)	0.34 (0.20)
Difficulty level, mean (SD)	0.84 (0.12)	0.74 (0.15)
Application questions, n	34	37
Discrimination, mean (SD)	0.21 (0.20)	0.29 (0.22)
Difficulty level, mean (SD)	0.78 (0.16)	0.69 (0.21)

Comparison With Medical Student Performance

A cohort of 143 fourth-year medical students took the identical midterm and final examinations, achieving a mean midterm score of 89.4% (SD 7.13) and a mean final score of 80.2% (SD 8.73) (Table 2). GPT-3.5 underperformed relative to students (59.2% vs 89.4%, $P<.001$; 55% vs 80.2%, $P<.001$), whereas advanced variants such as GPT-4o1 matched or exceeded student performance on both the midterm (94.1% vs 89.4%, $P<.001$) and final exams (92.4% vs 80.2%, $P<.001$) (Table 3).

Table . GPTs' performance in different question types.

ChatGPT versions	GPT-3.5N	GPT-3.5P	GPT-4N	GPT-4P	GPT-4oN	GPT-4oP	GPT-o1miniN	GPT-o1miniP	GPT-o1N	GPT-o1P	Students
Midterm exams											
Memorization questions											
Correct rate (%)	63.55	73.23	87.74	90.97	91.94	91.29	91.29	91.61	97.42	95.81	89.79
Application questions											
Correct rate (%)	56.57	69.71	77.14	80.57	91.43	90.29	78.86	80	88	91.43	92.78
Final exams											
Memorization questions											
Correct rate (%)	56.62	64.31	86.46	91.08	89.54	91.08	85.23	88	94.15	95.08	89.79
Application questions											
Correct rate (%)	67.57	64.86	89.19	94.59	89.19	91.89	75.68	78.38	91.89	89.19	92.78

Table . Comparison of ChatGPT variants and student cohort performance. Mean percentage scores (SD) are listed for each GPT model and a cohort of 143 fourth-year medical students on midterm and final examinations, with statistical significance (*P* values) for model versus student differences.

ChatGPT versions	GPT-3.5N	GPT-3.5P	GPT-4N	GPT-4P	GPT-4oN	GPT-4oP	GPT-4o1-mini-N	GPT-4o1-mini-P	GPT-4o1N	GPT-4o1P	Students
Midterm exams											
Original score, mean (SD)	61.03 (0.84)	71.96 (1.64)	83.92 (1.14)	87.22 (1.14)	91.75 (0.71)	90.93 (0.84)	86.80 (0.84)	87.42 (1.10)	94.02 (0.45)	94.23 (0.55)	89.4 (7.13)
Standardized score	−2.81	−1.72	−0.52	−0.19	0.26	0.18	−0.23	−0.17	0.49	0.51	0.03
Final exams											
Original score, mean (SD)	60.59 (0.45)	64.31 (0.55)	87.84 (0.55)	92.35 (0.45)	90.20 (0.71)	91.18 (0.71)	81.57 (0.45)	84.71 (0.55)	92.75 (0.55)	91.57 (0.55)	80.2 (8.73)
Standardized score	−2.77	−1.85	0.76	1.26	1.02	1.13	0.06	0.41	1.31	1.17	−0.01

Performance by Question Type

Table 2 presents model and student accuracy across 4 question types. All variants and students achieved the highest accuracy on MCQs, with GPT-4o1 reaching 98.5% (SD 1.2), students 92.3% (SD 5), and GPT-3.5 the lowest at 70.4% (SD 3). SAQs followed a similar pattern, ranging from 62.3% (SD 2.7) for GPT-3.5% to 92.1% (SD 1.5) for GPT-4o1, with students at 85.6% (SD 6.8). CCA yielded the greatest variability: GPT-3.5 scored 48.7% (SD 3.5) versus 88.4% (SD 2) for GPT-4o1 and 75.2% (SD 8.1) for students. IBI performance ranged from 55.2% (SD 3.1) to 90.2% (SD 1.8) for GPT-4o1, with student IBI at 78.5% (SD 7.5).

Error Analysis by Question Type

To further elucidate model performance nuances, we analyzed error rates across question types. CCA questions exhibited an error rate approximately 3 times higher than memory recall items when answered by early ChatGPT models (GPT-3.5 and GPT-4.0). Representative examples include:

- Memory recall: describing regulators of gastric acid secretion—GPT-3.5 misidentified pancreatic enzymes as inhibitory due to misinterpreting "major."

- CCA: a 65-year-old man with acute abdominal pain—GPT-3.5 attributed findings to pancreatitis; GPT-4o1-mini (no prompt) misdiagnosed cholecystitis.
- Short answer: advantages of laparoscopic appendectomy—GPT-4.0 only cited "smaller incision," omitting recovery and complication benefits.
- Image interpretation: abdominal X-ray—GPT-3.5 confused free air with pneumatosis intestinalis; GPT-4o (with prompt) correctly identified small-bowel obstruction. Subsequent optimized variants (GPT-4o, GPT-4o1-mini, and GPT-4o1) trained on broader multilingual corpora reduced CCA error rates, with GPT-4o1 achieving 88.4% accuracy (Table 2), approaching student performance (75.2%). This analysis highlights specific failure modes and improvements in reasoning and language comprehension.

Effect of Prompt Engineering

As shown in Table 4, prompt engineering significantly enhanced performance for early models. For the midterm, GPT-3.5 improved from 59.2% to 69.8% (Cohen d=1.5; P<.001), and GPT-4.0 from 81.4% to 84.6% (Cohen d=0.7; P=.002). In contrast, advanced variants exhibited no significant benefit: GPT-4o (91.3% vs 91.6%; P=.07), GPT-4o1-mini (86.1% vs 87.4%; P=.69), and GPT-4o1 (94.1% vs 94.2%; P=.55). Similar patterns were observed in the final exam (Table 4), where prompt-engineering scores for GPT-3.5 and GPT-4.0 increased significantly (P<.01), but not for GPT-4o (P=.94), GPT-4o1-mini (P=.58), or GPT-4o1 (P=.24).

Table . Investigation of the scores in different prompt modes: effect of prompt engineering on ChatGPT performance. Paired comparison of mean (SD) scores and *P* values is listed for each model variant under no-prompt versus prompt-engineering conditions, highlighting the variable benefit of structured prompts across model generations.

Rounds	1	2	3	4	5	Mean (SD)	*P* value
Midterm exams							
GPT-3.5N	60	59	58	60	59	59.2 (0.84)	<.001
GPT-3.5P	68	72	71	69	69	69.8 (1.64)	—[a]
GPT-4N	80	81	81	82	83	81.4 (1.14)	.002
GPT-4P	83	84	85	85	86	84.6 (1.14)	—
GPT-4oN	88	89	88	90	88	88.6 (0.89)	.07
GPT-4oP	89	90	90	89	90	89.6 (0.55)	—
GPT-4o1-miniN	91	90	92	90	91	90.8 (0.84)	.69
GPT-4o1-miniP	91	91	91	92	90	91 (0.71)	—
GPT-4o1N	92	92	91	92	92	91.8 (0.45)	.55
GPT-4oP	91	92	92	92	91	91.6 (0.55)	—
Final exams							
GPT-3.5N	54	56	55	54	56	55(1)	<.01
GPT-3.5P	61	60	60	60	60	60.2 (0.45)	—
GPT-4N	85	84	84	85	83	84.2 (0.84)	<.01
GPT-4P	89	87	87	88	88	87.8 (0.84)	—
GPT-4oN	89	90	90	90	90	89.8 (0.45)	.94
GPT-4oP	90	91	90	91	90	90.4 (0.55)	—
GPT-4o1-miniN	91	92	92	91	91	91.4 (0.55)	.58
GPT-4o1-miniP	92	91	92	92	91	91.6 (0.55)	—
GPT-4o1N	91	92	91	91	92	91.4 (0.55)	.24
GPT-4oP	91	92	92	92	92	91.8 (0.45)	—

[a]Not applicable.

Stability Across Runs

Coefficient of variation (CV) across the 5 independent runs decreased with model version. GPT-3.5 midterm CV was 3.5%, whereas GPT-4o1 recorded a CV of 0.6%. Prompt engineering reduced CV by an average of 0.4 percentage points for GPT-3.5 and GPT-4.0, but had a negligible impact on optimized variants.

These findings demonstrate not only a clear progression in raw performance and stability from GPT-3.5 to GPT-4o1, but also that the top-tier optimized models can match or surpass human student performance (Table 3), indicating their potential as both assessment tools and educational companions.

Discussion

Principal Findings

This study provides the first systematic evaluation of prompt engineering across multiple ChatGPT variants, highlighting the evolution of LLM capabilities in medical education settings. We observed that GPT-4 variants (GPT-4o, GPT-4o1-mini, and GPT-4o1) significantly outperformed earlier models, consistent with the findings by Kung et al [7] that ChatGPT 4.0 surpassed ChatGPT 3.5 on the United States Medical Licensing Examination, achieving accuracy at or near the passing threshold. In our work, advanced models not only exhibited higher baseline scores but also demonstrated greater stability across repeated runs, underscoring architectural improvements in reasoning and context retention.

Prompt engineering yielded substantial performance gains for early-generation models—GPT-3.5 and GPT-4—mirroring reports that structured guidance can boost LLM accuracy [4], but its use diminished for optimized variants. Safrai and Azaria [8] found that GPT-4 maintained performance even when confronted with extraneous "small talk" inserted into medical prompts, whereas GPT-3.5's performance degraded under similar conditions. Our findings extend this observation, showing that GPT-4o and its successors exhibit minimal dependency on explicit prompt structures, suggesting that these models have internalized reasoning scaffolds natively.

Our analysis by question type aligns with the in-depth evaluation by Knoedler et al [9], who reported variable ChatGPT performance across categories and a negative correlation with question difficulty (r_s=−0.306; P<.001) in the United States Medical Licensing Examination step 1 practice items. Similarly, we noted that CCA and IBI posed the greatest challenges for all models, although advanced variants narrowed the gap. These parallels reinforce the generalizability of LLM behavior across diverse educational assessment formats.

Our error analysis underscores the importance of evaluating LLMs not only by overall scores but also by question-type vulnerabilities. Early models' difficulties with multistep reasoning and complex Chinese phrasing, especially in clinical scenarios and image-based tasks, point to inherent limitations in contextual understanding. The marked reduction of these errors in optimized variants demonstrates progress but also indicates areas where artificial intelligence (AI) may still mislead learners. Educators should therefore integrate error-focused feedback loops when deploying LLMs: by exposing students to AI-generated mistakes in controlled settings, learners can develop critical appraisal skills and better discern AI hallucinations. This approach transforms AI from a mere answer engine into a pedagogical tool that actively fosters analytical thinking and deep learning.

Comparison With Medical Student Performance

The cohort of 143 fourth-year medical students achieved a mean midterm score of 89.4% (SD 7.13) and a mean final score of 80.2% (SD 8.73) (Table 3). GPT-3.5 underperformed relative to students (59.2% vs 89.4%; P<.001 and 55% vs 80.2%; P<.001), whereas advanced variants such as GPT-4o1 matched or exceeded student performance on both the midterm (94.02% vs 89.4%; P<.001) and final exams (92.75% vs 80.2%; P<.001). This indicates that top-tier LLMs can approach or surpass human proficiency in standardized medical assessments.

AI as a Learning Companion Beyond Assessment

Advanced LLMs show promise as AI-enabled educational tools, capable of rapidly synthesizing complex medical knowledge to aid student understanding. Studies have demonstrated AI's use in generating personalized explanations and feedback that enhance learning efficiency [2,4]. However, LLMs may still produce errors and "hallucinations" [10], underscoring the importance of maintaining critical appraisal and scholarly rigor when integrating AI into medical education.

Educational Value for Diagnosing Student Weaknesses

Although AI does not achieve 100% accuracy, its overall correctness surpassed that of most medical students in our cohort. Students spend significant time retrieving correct answers and understanding explanations; AI can serve as a learning companion by rapidly aggregating and summarizing complex medical knowledge, guiding step-by-step reasoning. Integrating our AI system into adaptive learning platforms could help students quickly identify weak areas, practice targeted question types, and maintain critical appraisal to avoid overreliance on AI outputs. This targeted approach not only enhances learning efficiency but also empowers students to become more self-directed learners, using AI as a diagnostic tool to identify knowledge gaps and focus their study efforts where they are most needed.

Strategies to Reduce AI Hallucinations

To mitigate risks of LLM-generated misinformation, future implementations should consider several evidence-based strategies. Cross-model consensus approaches—querying multiple LLMs (eg, GPT-4o1 and open-source alternatives) and adopting majority-vote answers—can increase reliability and reduce single-model biases. Expert fine-tuning using annotated medical datasets would strengthen domain-specific accuracy, particularly for specialized clinical scenarios. Integration of real-time evidence retrieval through literature and guideline search APIs would ensure responses include verifiable reference citations, enhancing transparency and trustworthiness. In addition, implementing confidence scoring systems coupled with human-AI collaboration frameworks would route low-confidence responses to human experts for review, creating a safety net against potential hallucinations while maintaining efficiency.

Collectively, these results underscore a maturation of LLMs: as model architectures advance, the marginal benefit of prompt engineering declines, and the potential educational role shifts from prompt design to strategic integration and fine-tuning. For practitioners and educators, this suggests a shift from elaborate

prompt design toward focusing on model selection and integration strategies—such as multimodal input handling and curriculum-specific fine-tuning—to maximize efficacy in high-stakes assessments. Future research should explore adaptive prompting frameworks that tailor AI guidance to learner needs and investigate real-world clinical scenario applications, while prioritizing the development of robust safeguards against AI hallucinations to ensure safe and effective integration into medical education curricula.

Conclusions

This comprehensive evaluation across 5 ChatGPT variants demonstrates a progressive enhancement in performance and stability in medical examination tasks. Notably, optimized LLMs (GPT-4o, GPT-4o1-mini, and GPT-4o1) not only matched but significantly exceeded the mean scores of fourth-year medical students on both midterm and final exams, underscoring their capacity to approach—or surpass—human proficiency in standardized assessments.

While prompt engineering substantially improved outcomes for early-generation models (GPT-3.5 and GPT-4.0), optimized variants achieved near-ceiling accuracy with negligible gains from structured prompts, indicating that these models inherently internalize contextual guidance. These findings suggest a strategic pivot for educators and assessment designers: from intricate prompt crafting toward thoughtful model selection, multimodal integration, and domain-specific fine-tuning.

Furthermore, the ability of advanced LLMs to rapidly synthesize and organize complex medical knowledge positions them as valuable AI-enabled learning companions. Educators should leverage AI's strengths in personalized explanation and feedback while maintaining rigorous critical appraisal to identify potential errors or "hallucinations."

Future research should investigate adaptive prompting frameworks tailored to individual learner needs and assess the educational impact of AI-augmented tools in real-world clinical training environments.

Acknowledgments

This work was supported by grant CSH2019A019.

Authors' Contributions

M-YH conceived and designed the study, curated the data, conducted the statistical analysis (IBM SPSS Statistics; version 29), and drafted the manuscript. T-LW contributed to data curation, examination, administration, and manuscript review. P-HS and M-CC supervised the project, contributed to study design and interpretation, and critically revised the manuscript. M-CC is the corresponding author and P-HS is the co-corresponding author.

Conflicts of Interest

None declared.

References

1. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. 2020 Presented at: Neural Information Processing Systems (NeurIPS) 2020; Dec 6-12, 2020; Vancouver, BC, Canada p. 1877-1901 URL: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf [accessed 2025-09-25]
2. GPT technical report. OpenAI. 2022. URL: https://cdn.openai.com/papers/gpt-4.pdf [accessed 2025-09-25]
3. GPT-4 technical report. OpenAI. 2023 Mar 14. URL: https://openai.com/research/gpt-4 [accessed 2025-09-25]
4. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput Surv 2023 Sep 30;55(9):1-35. [doi: 10.1145/3560815]
5. GPT-4o technical report: multimodal capabilities. OpenAI. 2024. URL: https://openai.com/index/gpt-4o-system-card/ [accessed 2025-09-25]
6. GPT-4o1mini release notes. OpenAI. 2025. URL: https://help.openai.com/en/articles/6825453-chatgpt-release-notes [accessed 2025-09-25]
7. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
8. Safrai M, Azaria A. Performance of ChatGPT-35 and GPT-4 on the United States Medical Licensing Examination with and without distractions. arXiv. Preprint posted online on Sep 12, 2023. [doi: 10.48550/arXiv.2309.08625]
9. Knoedler L, Knoedler S, Hoch CC, et al. In-depth analysis of ChatGPT's performance based on specific signaling words and phrases in the question stem of 2377 USMLE step 1 style questions. Sci Rep 2024 Jun 12;14(1):13553. [doi: 10.1038/s41598-024-63997-7] [Medline: 38866891]
10. Beutel G, Geerits E, Kielstein JT. Artificial hallucination: GPT on LSD? Crit Care 2023 Apr 18;27(1):148. [doi: 10.1186/s13054-023-04425-6] [Medline: 37072798]

Abbreviations

CCA: clinical case analysis
CV: coefficient of variation
IBI: image-based interpretation
LLM: large language model
MCQ: multiple choice question
SAQ: short answer question

Edited by T Sian; submitted 30.05.25; peer-reviewed by G Raut, H Tsai; revised version received 06.08.25; accepted 06.08.25; published 01.10.25.

Please cite as:
Hsieh MY, Wang TL, Su PH, Chou MC
Impact of Prompt Engineering on the Performance of ChatGPT Variants Across Different Question Types in Medical Student Examinations: Cross-Sectional Study
JMIR Med Educ 2025;11:e78320
URL: https://mededu.jmir.org/2025/1/e78320
doi:10.2196/78320

ChatGPT in Medical Education: Bibliometric and Visual Analysis

Yuning Zhang[1], MM; Xiaolu Xie[2], ME; Qi Xu[3], MD

[1]School of Basic Medical Sciences, Gannan Medical University, Ganzhou, China

[2]School of Medical and Information Engineering, Gannan Medical University, Ganzhou, China

[3]School of Public Health and Health Management, Gannan Medical University, 1 Harmony Avenue, Rongjiang New District, Ganzhou, China

Corresponding Author:
Qi Xu, MD
School of Public Health and Health Management, Gannan Medical University, 1 Harmony Avenue, Rongjiang New District, Ganzhou, China

Abstract

Background: ChatGPT is a generative artificial intelligence–based chatbot developed by OpenAI. Since its release in the second half of 2022, it has been widely applied across various fields. In particular, the application of ChatGPT in medical education has become a significant trend. To gain a comprehensive understanding of the research developments and trends regarding ChatGPT in medical education, we conducted an extensive review and analysis of the current state of research in this field.

Objective: This study used bibliometric and visualization analysis to explore the current state of research and development trends regarding ChatGPT in medical education.

Methods: A bibliometric analysis of 407 articles on ChatGPT in medical education published between March 2023 and June 2025 was conducted using CiteSpace, VOSviewer, and Bibliometrix (RTool of RStudio). Visualization of countries, institutions, journals, authors, keywords, and references was also conducted.

Results: This bibliometric analysis included a total of 407 studies. Research in this field began in 2023, showing a notable surge in annual publications until June 2025. The United States, China, Türkiye, the United Kingdom, and Canada produced the most publications. Networks of collaboration also formed among institutions. The University of California system was a core research institution, with 3.4% (14/407) of the publications and 0.17 betweenness centrality. *BMC Medical Education*, *Medical Teacher*, and the *Journal of Medical Internet Research* were all among the top 10 journals in terms of both publication volume and citation frequency. The most prolific author was Yavuz Selim Kiyak, who has established a stable collaboration network with Isil Irem Budakoglu and Ozlem Coskun. Author collaboration in this field is usually limited, with most academic research conducted by independent teams and little communication between teams. The most frequent keywords were “AI,” “ChatGPT,” and “medical education.” Keyword analysis further revealed “educational assessment,” “exam,” and “clinical practice” as current research hot spots. The most cited paper was “Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models,” and the paper with the strongest citation burst was “Are ChatGPT’s Knowledge and Interpretation Ability Comparable to Those of Medical Students in Korea for Taking a Parasitology Examination?: A Descriptive Study.” Both papers focus on evaluating ChatGPT’s performance in medical exams.

Conclusions: This study reveals the significant potential of ChatGPT in medical education. As the technology improves, its applications will expand into more fields. To promote the diversification and effectiveness of ChatGPT in medical education, future research should strengthen interregional collaboration and enhance research quality. These findings provide valuable insights for researchers to identify research perspectives and guide future research directions.

(JMIR Med Educ 2025;11:e72356) doi:10.2196/72356

KEYWORDS

ChatGPT; medical education; bibliometric; VOSviewer; CiteSpace; artificial intelligence; AI

Introduction

Background

Large language models (LLMs) represent major progress in artificial intelligence (AI), especially for computational linguistics and natural language processing. These generative AI models are fundamentally based on the transformer neural network architecture [1]. Training is conducted using extensive text datasets, including books, documents, and website content. LLMs have been developed to predict subsequent words or tokens. Through this process, they learn to recognize complex language patterns, including vocabulary, grammar, semantics, and even specialized knowledge such as medicine [2].

ChatGPT, an AI chatbot developed by OpenAI [3], was launched in November 2022 as an LLM [4,5]. Built on the generative pretrained transformer architecture, it uses tens of billions of

parameters trained on massive internet text datasets [6]. ChatGPT excels at understanding and generating humanlike language, conducting natural dialogues, and delivering high-quality responses to user queries [7,8]. Its advanced text processing capabilities have driven unprecedented adoption: more than 1 billion monthly users within 4 months of release, demonstrating rapid societal integration [9].

ChatGPT demonstrates significant potential across diverse fields, such as translation, text summarization, and programming assistance [10]. Its effectiveness extends to specialized domains such as medical education [11]. In preclinical education, students use ChatGPT for medical knowledge acquisition and personalized learning [12,13]. Conversely, educators are able to use ChatGPT to implement innovative teaching methodologies and cultivate interactive learning environments [14-16]. In clinical education [17], ChatGPT simulates clinical environments to help students improve clinical skills [18-21]. Furthermore, the pass rates and accuracy in medical licensing exams and professional subject tests [22,23] in countries such as the United States [24-26], China [27,28], Japan [29,30], and Italy [31] have attracted significant attention [32]. ChatGPT is regarded as a significant instrument for promoting innovation and enhancing efficiency in the domain of medical education.

Objectives

While existing literature reviews have explored ChatGPT's applications and limitations in medical education, important questions remain unanswered. These include the collaboration networks among countries, institutions, and authors; the most influential journals; and the most cited publications. This study used bibliometric analysis to map collaboration networks and thematic evolution and provide a comprehensive understanding of the development of ChatGPT in medical education.

A bibliometric analysis is a rigorous scientific method that provides researchers across various fields with comprehensive guidance and support [33]. It allows researchers to gain in-depth insights into prevailing issues, key trends, and research limitations within their disciplines [34-36]. On the basis of recommendations from previous studies, this study proposes the following research questions (RQs):

1. Who are the most productive researchers and which are the most productive institutions and countries or regions in the field of ChatGPT in medical education?
2. What is the status of academic collaboration among researchers, countries, or regions in the field of ChatGPT in medical education?
3. Which are the most influential journals and articles in the field of ChatGPT in medical education?
4. What are the main research themes in the field of ChatGPT in medical education?
5. What are the research trends for ChatGPT in medical education?

Methods

Literature Sources and Search Strategy

The Web of Science database was chosen for this research due to its extensive coverage of more than 12,000 academic journals. When compared to other databases, including PubMed, MEDLINE, and Scopus, Web of Science offers a robust and reliable framework for bibliometric analysis [37]. After determining pertinent title keywords, a comprehensive bibliographic search was conducted online via the Web of Science database. The search was carried out in accordance with the following format:

> *((TS=(ChatGPT)) OR TS=(Chatbot*)) OR TS=(Chat Generative Pre-trained Transformer) and ((((((((((TS=(medic* educat*)) OR TS=(medic* student*)) OR TS=(clinical clerkship*)) OR TS=(medic* school*)) OR TS=(medic* learner*)) OR TS=(medic* trainee*)) OR TS=(medic*clerk*)) OR TS=(medical education)) OR TS=(medical student)) OR TS=(medical school)) OR TS=(medical student education)) OR TS=(healthcare). NOT ALL=(retracted)—Time: Tue Jul 01, 2025, 19:18:42 GMT+0800 (CST)*

A total of 1817 documents were retrieved. These documents were screened according to the inclusion and exclusion criteria. The inclusion criteria were as follows: (1) original research articles and review articles related to ChatGPT in medical education and (2) English-language articles. After screening, of the 1817 retrieved articles, 1610 (88.61%) were retained. Following application of the exclusion criteria (articles unrelated to ChatGPT in medical education and duplicate articles), of the remaining 1610 articles, 1203 (74.72%) were excluded. The research topics of the 1203 excluded articles are summarized as follows: 370 (30.76%) were non-ChatGPT studies, 298 (24.77%) involved ChatGPT and patients, 263 (21.86%) involved ChatGPT and clinical treatment, 144 (11.97%) involved ChatGPT and hospitals, 52 (4.32%) involved ChatGPT and nonmedicine, 36 (2.99%) were ChatGPT non–medical education review articles, 20 (1.66%) involved ChatGPT and health care professional perspectives, 14 (1.16%) involved ChatGPT and nonmedical interactions with students, and 5 (0.42%) involved ChatGPT and veterinary medicine. This resulted in 407 publications being selected for bibliometric analysis. A comprehensive dataset, along with the corresponding references, was subsequently extracted from the relevant publications and organized in plain-text format for future research endeavors. This process was conducted independently by 2 authors, who cross-verified their work. Any discrepancies were resolved by a senior author.

Data Collection and Statistics

The data were exported in plain-text file format using CiteSpace (version 6.3.R1; 64 bits; advanced) and R (version 4.5.0; R Foundation for Statistical Computing) with the *bibliometrix* package [38]. The data included the full record and cited references and were stored in the download format (.txt). The data extracted from the Bibliometric online platform [39] were exported in tab-delimited file format, with content and storage format identical to those described above.

CiteSpace, a bibliometric analysis software developed by Chaomei Chen, has achieved widespread use [40,41]. The software has been proven to provide feasible and reliable text mining and knowledge visualization methods. These methods

have been used to explore research performance allocation and collaboration, research status and frontiers, and future trends. In this study, CiteSpace was used to detect parameters and visually analyze institution distribution, the dual-map overlay of journals, and burst detection.

Burst detection, based on the Kleinberg algorithm, uses an infinite state automaton to model document streams, thereby extracting meaningful structures [42]. These analyses can reveal themes exhibiting rapid growth over extended periods, as well as those that are inherently more transient.

VOSviewer, released in 2010 by Nees Jan van Eck and Ludo Waltman (Leiden University), is mainly used for bibliometric network graph analysis [43]. We used VOSviewer version 1.6.20 to visualize and analyze the country distribution and collaboration, journal distribution, author distribution and collaboration, and keyword distribution.

In addition, Bibliometrix (RTool of RStudio; Posit PBC) was used to visualize the distribution of keywords over time in the form of a heat map [44]. Microsoft Excel 2024 was used to analyze the monthly publication trends of literature from March 2023 to June 2025.

Ethical Considerations

The study did not involve human participants, therefore the ethical approval was not required.

Results

Overview of Publication Status

Our search and screening efforts yielded 407 articles (Figure 1). The release of the AI chatbot ChatGPT in November 2022 was immediately followed by 2 review articles on ChatGPT published in March 2023, which provided an overview of ChatGPT in medical education and prediction of potential future applications for ChatGPT. The analysis of publication trends for this topic was conducted using Microsoft Excel 2024 and presented the number of publications in tabular form (Table 1). The number of articles showed a gradual increasing trend. In early 2023, the number of published articles per month was less than 10, and starting in September 2023, this number increased significantly. By 2024, the number of articles per month remained at more than 10, and in 2025, it was more than 20 articles per month. In May 2025, the number of articles per month reached 33. This implies that, over time, an increasing number of scholars are focusing their attention on this domain.

Figure 1. Flowchart of data collection and bibliometric analysis.

Table . Number of articles per month and cumulative number.

Month and year	Monthly publications, n	Cumulative publications, n
March 2023	2	2
April 2023	1	3
May 2023	2	5
June 2023	2	7
July 2023	2	9
August 2023	8	17
September 2023	14	31
October 2023	13	44
November 2023	7	51
December 2023	12	63
January 2024	14	77
February 2024	15	92
March 2024	9	101
April 2024	15	116
May 2024	19	135
June 2024	17	152
July 2024	15	167
August 2024	15	182
September 2024	17	199
October 2024	20	219
November 2024	16	235
December 2024	24	259
January 2025	23	282
February 2025	23	305
March 2025	26	331
April 2025	22	353
May 2025	33	386
June 2025	21	407

Analysis of National Publication Counts

Publication counts by country were used to analyze contributions in this field. According to the results, the publications originated from 66 countries. Visualizing the geographic distribution of the 66 countries using VOSviewer revealed that Asia, Europe, Africa, North America, South America, and Oceania were all represented and that the countries were mainly concentrated in the northern hemisphere (Figure 2). In total, 38% (25/66) of the countries were in Asia, and 33% (22/66) were from Europe, the 2 continents with the highest number of countries in this study. It is noteworthy that the linkages between countries or regions were concentrated between East Asia and North America, East Asia and Europe, North America and Europe, and North America and Oceania.

Figure 2. Countries and regions involved in the research in this field. The links between the countries and regions indicate their collaborations and connections.

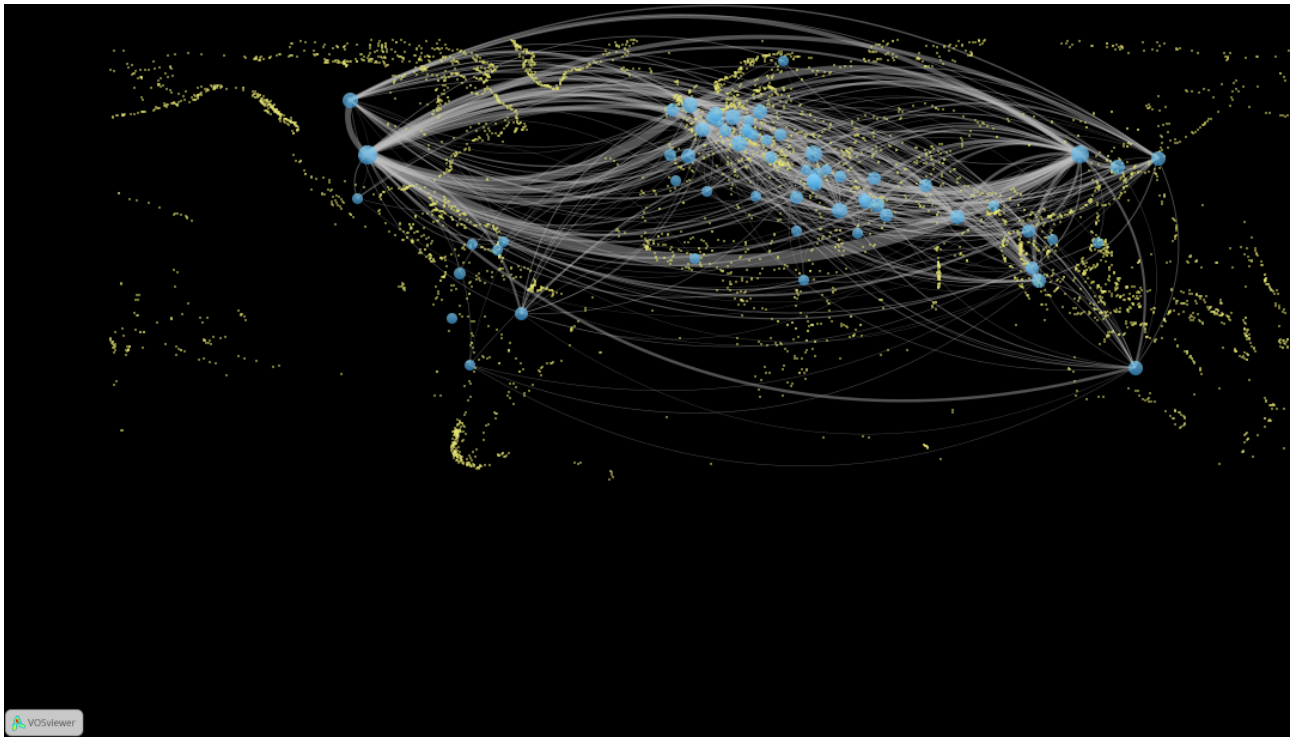


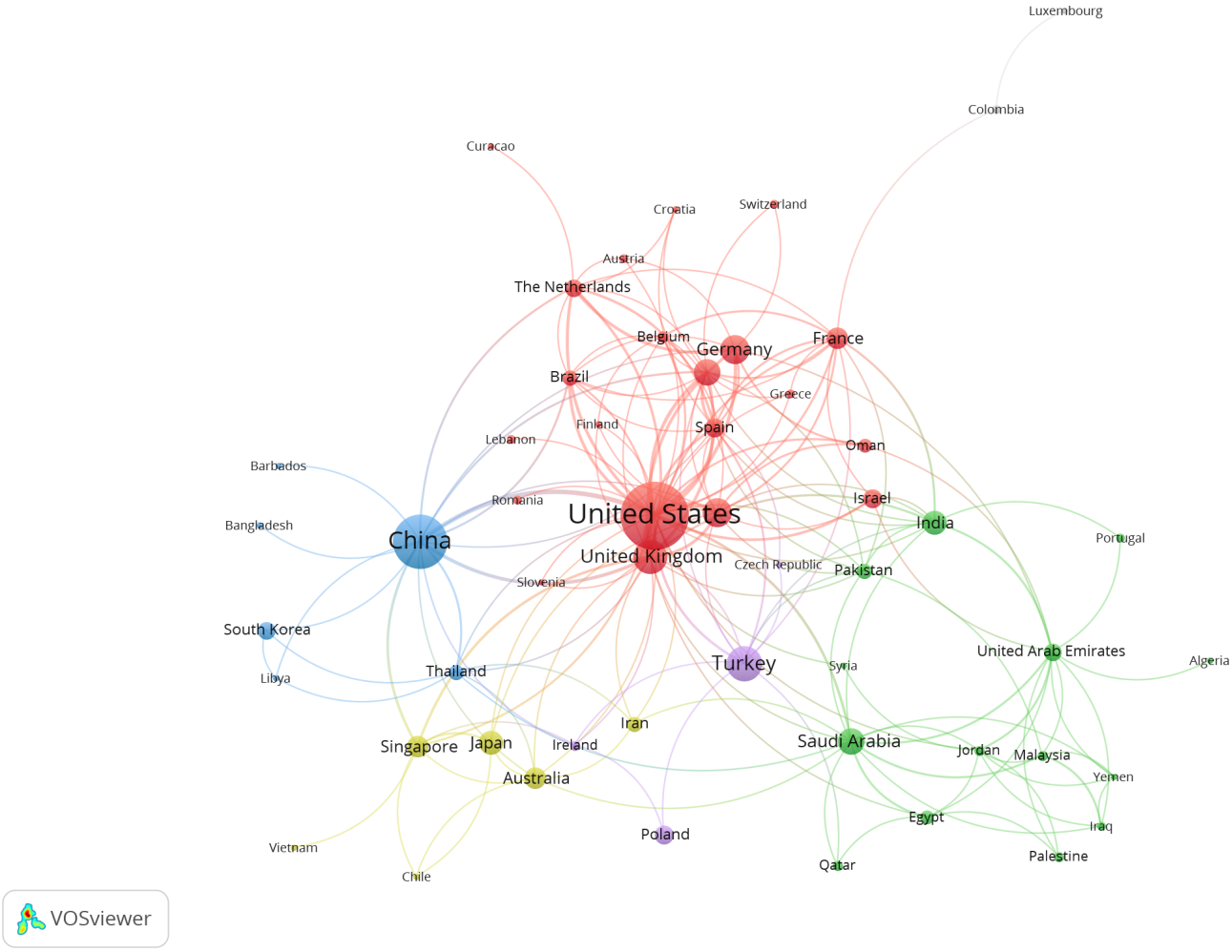
Table 2 shows the top 10 countries or regions in terms of number of publications and their corresponding citation frequency and centrality. The United States was the most prominent country with 31% (126/407) of the publications, closely followed by China, Türkiye, the United Kingdom, Canada, and Germany.

Table . Top 10 countries by number of publications and their number of citations and centrality (N=407).

Rank	Country	Publications, n (%)	Citations, n	Centrality
1	United States	126 (31)	3227	0.42
2	China	81 (19.9)	1026	0.19
3	Türkiye	35 (8.6)	215	0.08
4	United Kingdom	31 (7.6)	1844	0.17
5	Canada	24 (5.9)	668	0.05
6	Germany	22 (5.4)	463	0.06
7	Italy	19 (4.7)	730	0.06
8	Saudi Arabia	18 (4.4)	227	0.11
9	India	15 (3.7)	121	0.07
10	Japan	15 (3.7)	155	0.01

The results of the global collaboration network analysis show that countries and regions were roughly divided into 5 clusters in VOSviewer based on the closeness of collaboration and are indicated by different colors in Figure 3. The United States, China, Türkiye, the United Kingdom, and Canada were the top 5 countries in terms of the number of publications, and there were cooperative relationships between them. The betweenness centrality (BC) was calculated when analyzing the national and regional collaboration networks using CiteSpace, which represents the strength of association between nodes. Among the top 10 countries, the United States, China, the United Kingdom, and Saudi Arabia were the main research centers in this field.

Figure 3. Analysis of collaborative network visualization of countries and regions in VOSviewer.



Analysis of Publication Institutions

The scientific output came from 847 institutions. CiteSpace identified 147 institutions with 346 cooperative networks (Figure 4). The most productive institutions regarding research in this field were the University of California system (14/407, 3.4% of publications), Harvard University (11/407, 2.7% of publications), National University of Singapore (11/407, 2.7% of publications), the Commonwealth System of Higher Education (11/407, 2.7% of publications), the University of Toronto (10/407, 2.5% of publications), the University of London (8/407, 2% of publications), Gazi University (8/407, 2% of publications), the University of Pittsburgh (8/407, 2% of publications), Central South University (7/407, 1.7% of publications), and Stanford University (7/407, 1.7% of publications). Five of the top 10 institutions were from the United States. The remaining institutions were from Singapore, Canada, the United Kingdom, Türkiye, and China (Table 3).

Figure 4. Analysis of collaborative network visualization of institutions in CiteSpace.

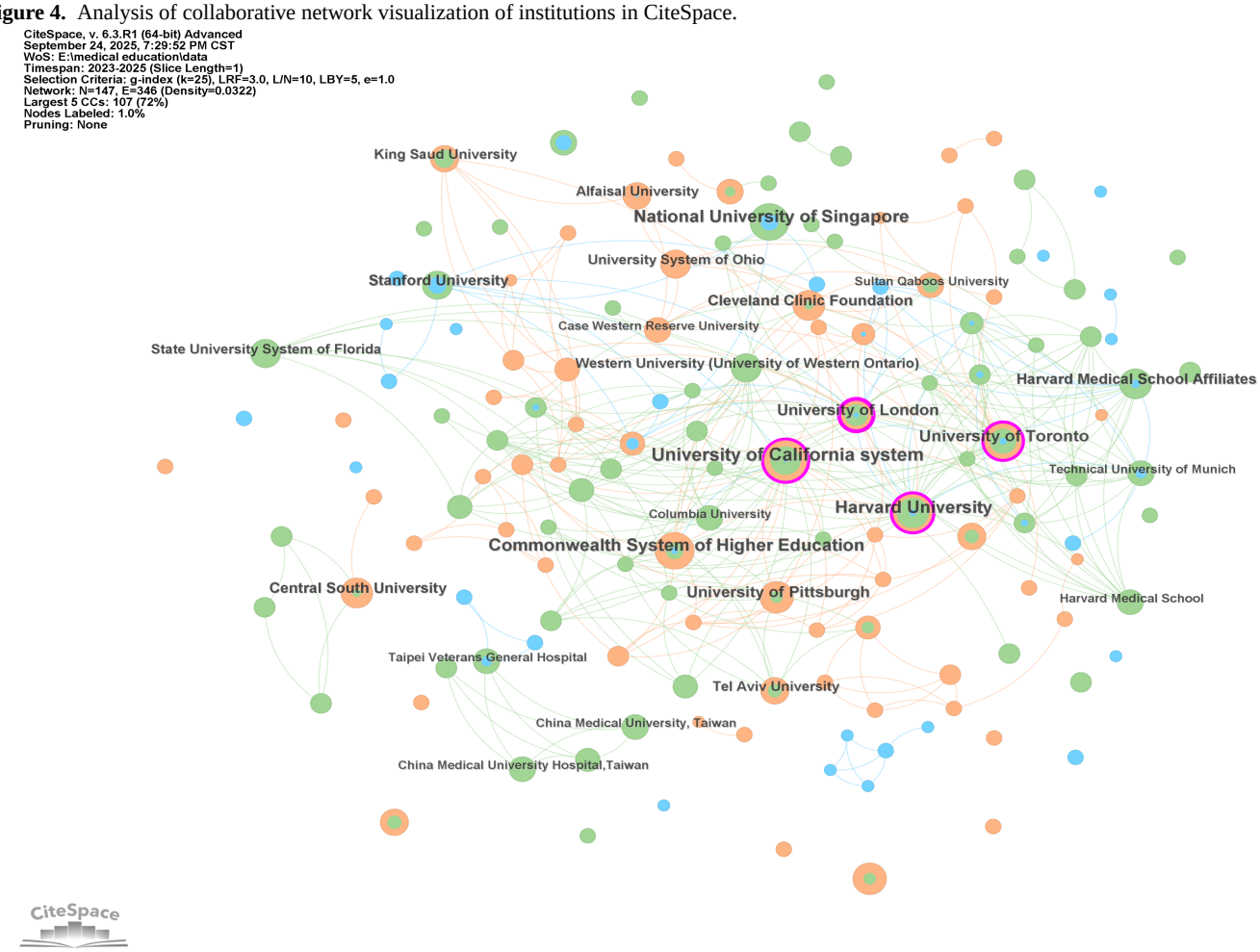


Table . Top 10 institutions and their centrality in CiteSpace (N=407).

Rank	Institution	Publications, n (%)	Centrality
1	University of California system	14 (3.4)	0.17
2	Harvard University	11 (2.7)	0.14
3	National University of Singapore	11 (2.7)	0
4	Commonwealth System of Higher Education	11 (2.7)	0.03
5	University of Toronto	10 (2.5)	0.1
6	University of London	8 (2)	0.22
7	Gazi University	8 (2)	0
8	University of Pittsburgh	8 (2)	0.01
9	Central South University	7 (1.7)	0
10	Stanford University	7 (1.7)	0.08

In CiteSpace, each node represents a institution, and the radius of a node increases with its contribution to research in the field, whereas the BC is proportional to the size of the purple ring around the nodes. The larger the purple circle, the larger the value of the betweenness centrality. Network visualization revealed that there were 4 central institutions: the University of California system (BC=0.17), Harvard University (BC=0.14), the University of Toronto (BC=0.10), and the University of London (BC=0.22; Figure 4). This reflects the significant bridging role of these institutions in the research on ChatGPT in medical education.

Analysis of Publication Quantity and Journal Impact

This study encompassed 407 articles published across 197 sources and journals. Table 4 lists the 10 most prolific sources and journals ranked by publication volume, along with their 2024 impact factor (IF). The top 10 journals published a total of 139 papers. Among these, *BMC Medical Education* (IF of

3.2; quartile 1; 40/407, 9.8% of publications) had the highest number of publications, followed by *Medical Teacher* (IF of 4.4; quartile 1; 32/407, 7.9% of publications), the *Journal of Medical Internet Research* (IF of 6.0; quartile 1; 11/407, 2.7% of publications), *Scientific Reports* (IF of 3.9; quartile 1; 11/407, 2.7% of publications), and *PLOS ONE* (IF of 2.6; quartile 2; 10/407, 2.5% of publications). Eight of the top 10 journals in terms of publications were distributed in quartile 1 of the Journal Citation Reports (Figure 5). The source or journal with the highest cocitation frequency was *arXiv*, followed by *JMIR Medical Education*, *Cureus*, *Medical Teacher*, *BMC Medical Education*, and the *Journal of Medical Internet Research* (Figure 6). Seven of the top 10 sources or journals in terms of cocitation frequency were distributed in quartile 1 of the Journal Citation Reports (Table 5). It is important to note that 3 of the top 10 journals in terms of publications were also among the top 10 journals in terms of cocitation frequency: *Medical Teacher*, *BMC Medical Education*, and the *Journal of Medical Internet Research*.

Table . Top 10 sources by number of publications and their corresponding journal impact factor (IF; Journal Citation Reports [JCR] 2024) and JCR quartile (N=407).

Rank	Source	Publications, n (%)	IF (JCR 2024)	JCR quartile
1	*BMC Medical Education*	40 (9.8)	3.2	1
2	*Medical Teacher*	32 (7.9)	4.4	1
3	*Journal of Medical Internet Research*	11 (2.7)	6.0	1
4	*Scientific Reports*	11 (2.7)	3.9	1
5	*PLOS ONE*	10 (2.5)	2.6	2
6	*Frontiers in Medicine*	9 (2.2)	3.0	1
7	*Digital Health*	7 (1.7)	3.3	1
8	*Healthcare*	7 (1.7)	2.7	2
9	*Nurse Education Today*	6 (1.5)	4.2	1
10	*Postgraduate Medical Journal*	6 (1.5)	2.7	1

Figure 6. Analysis of collaborative network visualization of journals' citations in VOSviewer.

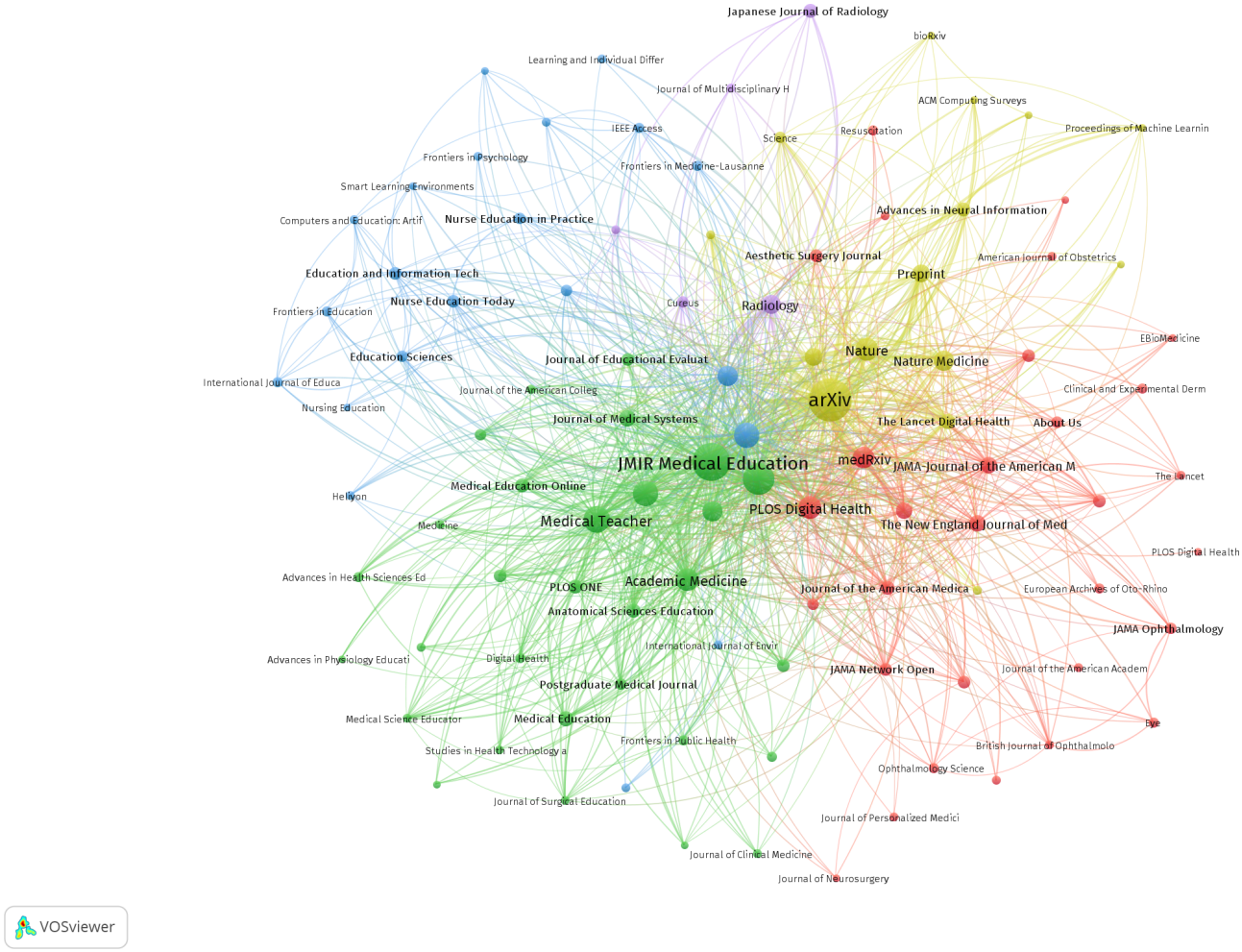


Figure 5. Analysis of collaborative network visualization of journals in VOSviewer.

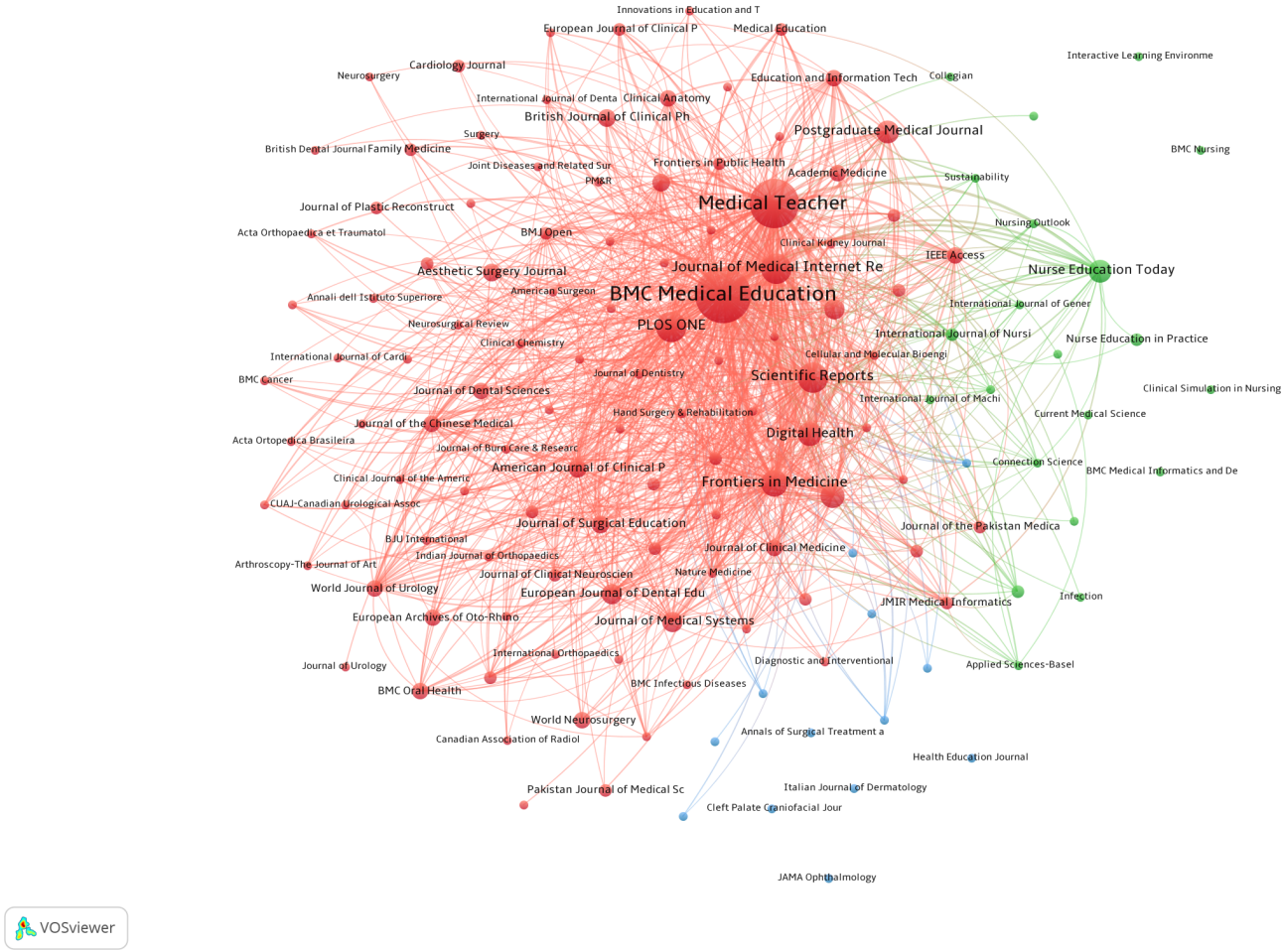


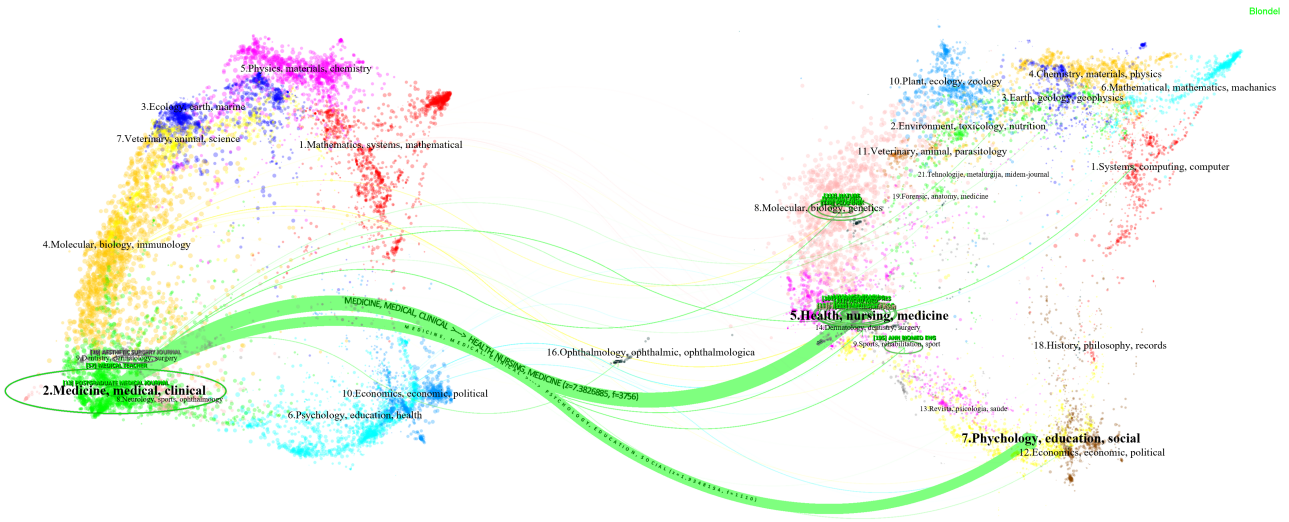
Table . Top 10 sources by number of cocitations and their corresponding journal impact factor (IF; Journal Citations Report [JCR] 2024) and JCR quartile.

Rank	Source	Number of cocitations	IF (JCR 2024)	JCR quartile
1	*arXiv*	692	None	None
2	*JMIR Medical Education*	542	3.2	1
3	*Cureus Journal of Medical Science*	357	1.3	2
4	*Medical Teacher*	252	4.4	1
5	*BMC Medical Education*	237	3.2	1
6	*Journal of Medical Internet Research*	216	6.0	1
7	*PLOS Digital Health*	191	7.7	1
8	*Nature*	188	48.5	1
9	*Academic Medicine*	187	5.2	1
10	*medRxiv*	164	None	None

The visualization in VOSviewer showed the journals that have published articles on ChatGPT in medical education and the relationship between them. On the basis of the similarity between journals, they were divided into 3 categories: the red cluster focused on educational, technical, and basic research (eg, *BMC Medical Education* and *Medical Teacher*); the green cluster focused on the discipline of nursing and extended to nursing education, clinical simulation training, and health care informatics (eg, *Nurse Education Today* and *International Journal of Nursing Studies*); and the blue cluster focused on specialized clinical practice, particularly in surgery, obstetrics and gynecology, orthopedics, and ophthalmology (eg, *American Journal of Obstetrics and Gynecology* and *Cleft Palate Craniofacial Journal*).

Journals were among the most common sources for publishing research results. Figure 7 presents a dual-map overlay of all academic journals, illustrating the citation paths of various subject areas. This dual-map overlay consists of 2 base maps: one on the left for citing journals and one on the right for cited journals [45]. The disciplines represented by the citing journals are indicated by the labels on the left side of the dual map, whereas the disciplines of the cited journals are shown on the right [46].

Figure 7. The dual-map overlay of journals.



Notably, 2 primary citation trajectories can be identified on the map. They are from health, nursing, and medicine to medicine, medical, and clinical (green) and from psychology, education, and social to medicine, medical, and clinical (green). In addition, the leading-edge research results were mainly distributed in 2 (medicine, medical, and clinical). This suggests that research on ChatGPT in medical education is particularly active within these disciplines. In contrast, the knowledge base that frontier researchers rely on primarily stems from 5 (health, nursing, and medicine) and 7 (psychology, education, and social). Citing journals and cited journals were both from the field of medicine. This implies that the application of ChatGPT in fields where medical education intersects with other disciplines is still limited.

Analysis of the Author Collaboration Network Graph

Analyzing the coauthorship network of this study helped identify potential collaborators and authoritative figures in the field. The author with the highest number of publications was Yavuz Selim Kiyak (Table 6). He has formed a stable core collaboration group with Isil Irem Budakoglu and Ozlem Coskun, who are from the same institution (Figure 8). All 3 authors ranked among the top 10 in terms of publication volume. However, among the top 10 most prolific authors, the remaining 7 have each formed independent teams. The analysis of author collaboration suggests that most academic research was conducted in independent teams without cross-team communication. Therefore, large interinstitutional collaboration networks have yet to be established.

Table . Top 10 authors by number of publications and their institutions and total link strength (N=407).

Rank	Author	Publications, n (%)	Institution	Total link strength
1	Yavuz Selim Kiyak	8 (2)	Gazi University (Türkiye)	1372
2	Ken Masters	5 (1.2)	Sultan Qaboos University (Oman)	945
3	Isil Irem Budakoglu	4 (1)	Gazi University (Türkiye)	1081
4	Ozlem Coskun	4 (1)	Gazi University (Türkiye)	837
5	Chia-Hung Kao	4 (1)	China Medical University (China)	641
6	Michael Alfertshofer	3 (0.7)	Ludwig Maximilian University of Munich (Germany)	774
7	Shuji Awano	3 (0.7)	Kyushu Dental University (Japan)	748
8	Olena Bolgova	3 (0.7)	Alfaisal University (Saudi Arabia)	735
9	Tzeng-Ji Chen	3 (0.7)	Taipei Veterans General Hospital, Hsinchu Branch (China)	1041
10	Wisit Cheungpasitporn	3 (0.7)	Mayo Clinic (United States)	1087

Figure 8. Collaborative network visualization of authors in VOSviewer.

Co-citation refers to the situation in which different authors are cited by the same article. These authors then form a co-citation relationship.The increase in co-citation counts indicates a greater degree of similarity among different authors' research, with the analysis itself reflecting the research strength of the respective authors. Table 7 lists the top 10 authors in terms of cocitation

frequency. The most frequently cocited authors were Tiffany H Kung (n=176), Aidan Gilson (n=141), Malik Sallam (n=109), and Arun James Thirunavukarasu (n=58). It is noteworthy that Kung is highly influential in the field of research on ChatGPT in medical education.

Table . Top 10 authors by number of citations and their institutions and total link strength.

Rank	Author	Number of citations	Institution	Total link strength
1	Tiffany H Kung	176	Harvard Medical School (United States)	1945
2	Aidan Gilson	141	Yale School of Medicine (United States)	1577
3	Malik Sallam	109	The University of Jordan (Jordan)	1264
4	Arun James Thirunavukarasu	58	University of Cambridge (United Kingdom)	773
5	Gunther Eysenbach	53	JMIR Publications (Canada)	645
6	Hyunsu Lee	49	Keimyung University (South Korea)	775
7	Karan Singhal	49	Google Research (United States)	826
8	Yavuz Selim Kıyak	42	Gazi University (Türkiye)	389
9	Andrew Mihalache	42	University of Western Ontario (Canada)	549
10	Rehan Ahmed Khan	40	Riphah International University (Pakistan)	502

Keyword Analysis of Global Research

To provide an overview of the primary content of the articles, it is possible to use keywords to analyze the frontiers of research on ChatGPT in medical education. Table 8 lists the top 20 keywords by frequency. The most frequent keyword was “artificial intelligence (AI),” followed by “ChatGPT,” “medical education,” “large language models,” and “generative AI.” The keyword co-occurrence network was visualized using VOSviewer, where the connecting lines between different keywords indicate that they have co-occurrence relationships (Figure 9). The keywords that make up this network were categorized into 4 clusters. The keywords in the red cluster in Figure 9 were related to the foundational elements of medical education, core disciplines, and ethical issues, such as “academic writing,” “radiology,” “ethics-medical,” and “machine learning.” The keywords in the green cluster focused on assessment methods, exam systems, and clinical decision-making processes in medical education, such as “medical exam,” “clinical decision-making,” and “teaching and learning.” The keywords in the blue cluster focused on specific applications, challenges, and practical effects of generative AI in medical practice, medical education, and specialties, such as “clinical practice,” “nursing education,” and “clinical skills.” The keywords in the yellow cluster were related to the evaluation of different generative AI models and multiple research methods and practices, such as “google bard,” “meta-analysis,” and “diagnosis.”

Table . Top 20 keywords with the highest frequency of occurrence and their corresponding total link strength.

Rank	Keyword	Number of occurrences	Total link strength
1	“AI”	225	715
2	“ChatGPT”	216	676
3	“Medical education”	107	336
4	“Large language models”	101	338
5	“Generative AI”	29	108
6	“Education”	28	102
7	“Chatbot”	26	86
8	“Natural language processing”	22	110
9	“Medical student”	20	56
10	“Machine learning”	18	74
11	“Healthcare”	12	62
12	“Ethics”	9	48
13	“Gemini”	9	39
14	“Clinical decision-making”	9	26
15	“Medical exam”	9	48
16	“Bard”	8	40
17	“Nursing”	8	31
18	“Assessment”	8	40
19	“Clinical reasoning”	8	20
20	“Exam”	8	34

Figure 9. The co-occurrence of keywords in VOSviewer.

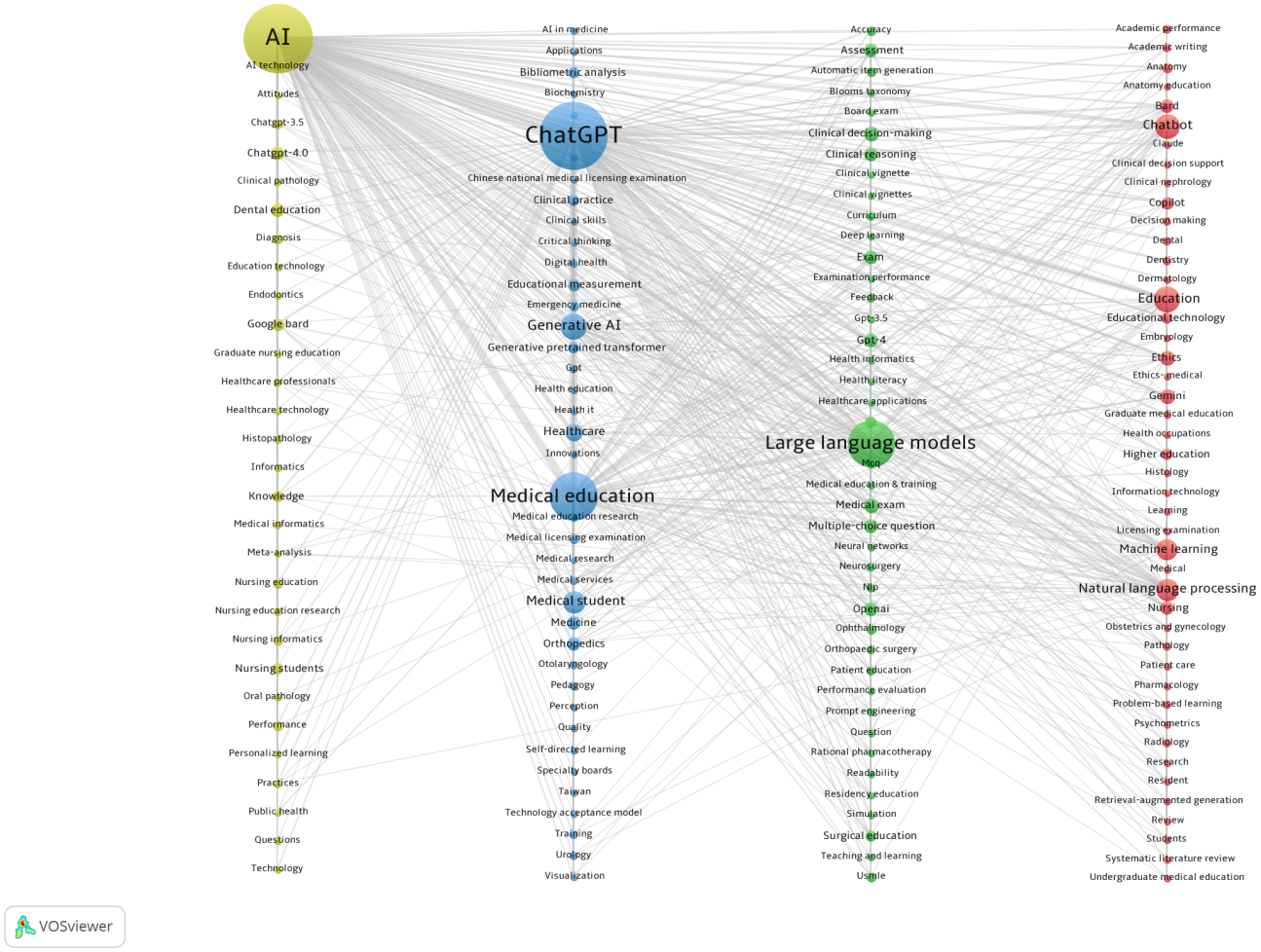


Figure 10 shows the monthly prevalence of the keywords from March 2023 to June 2025. Keywords such as “educational evaluation” and “medical disciplines” were research hot spots in 2023. Studies on GPT-4 continued throughout 2024. By 2025, research expanded to other LLMs such as Google Bard, Gemini, and Copilot. Research on medical exams continued from 2024 to 2025. Figure 11 shows the cumulative frequency of keywords between March 2023 and June 2025. Although research on clinical decision-making began in April 2023, related studies increased nearly in 2025. Ethics, starting in early 2024, rapidly became a research hot spot in a relatively short period.

Figure 10. Monthly distribution heat map of keywords in Bibliometrix.

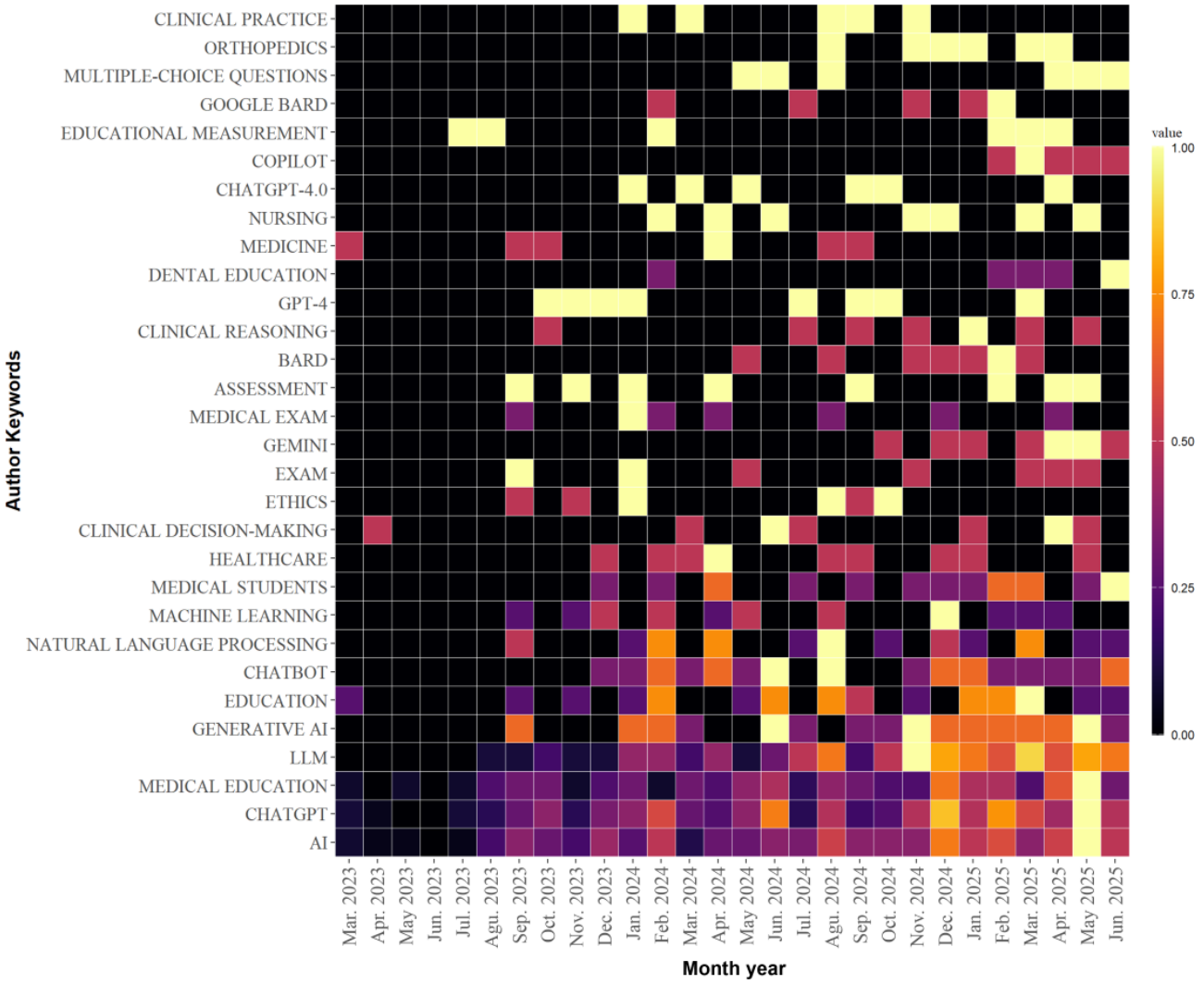
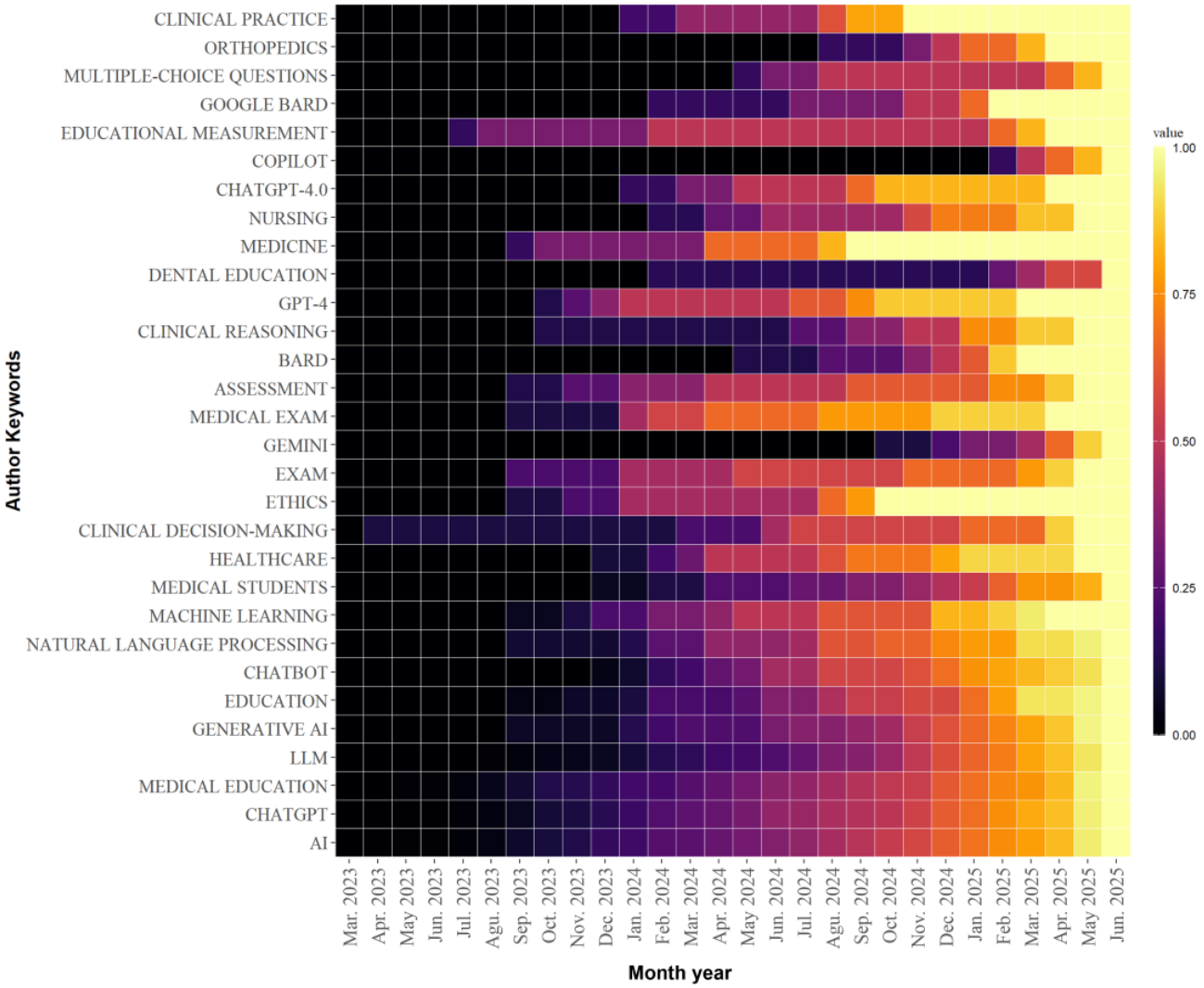


Figure 11. Cumulative distribution heat map of keywords in Bibliometrix.



Characteristics of Cited Research Articles

Table 9 lists the top 10 articles in terms of citation frequency. The most frequently cited article was “Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models” [47] (n=169). The second most cited article was “How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment” [48] (n=125). Both articles were about ChatGPT’s participation in the United States Medical Licensing Examination (USMLE). They evaluated ChatGPT’s performance on the USMLE, reflecting strong researcher interest in AI’s exam capabilities during the 2023 to 2025 study period.

Table . Top 10 most cited references.

Rank	Article title	Source	Authors	Year	Number of citations
1	"Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models"	*PLOS Digital Health*	Kung et al [47]	2023	169
2	"How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment"	*JMIR Medical Education*	Gilson et al [48]	2023	125
3	"ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns"	*Healthcare (Basel)*	Malik Sallam [49]	2023	63
4	"The Rise of ChatGPT: Exploring Its Potential in Medical Education"	*Anatomical Sciences Education*	Hyunsu Lee [50]	2024	51
5	"The Role of ChatGPT, Generative Language Models, and Artificial Intelligence in Medical Education: A Conversation With ChatGPT and a Call for Papers"	*JMIR Medical Education*	Gunther Eysenbach [19]	2023	51
6	"Large Language Models in Medicine"	*Nature Medicine*	Thirunavukarasu et al [51]	2023	44
7	"ChatGPT - Reshaping Medical Education and Clinical Management"	*Pakistan Journal of Medical Sciences*	Khan et al [52]	2023	39
8	"Artificial Hallucinations in ChatGPT: Implications in Scientific Writing"	*Cureus Journal of Medical Science*	Alkaissi et al [53]	2023	36
9	"ChatGPT in Medicine: An Overview of Its Applications, Advantages, Limitations, Future Prospects, and Ethical Considerations"	*Frontiers in Artificial Intelligence*	Dave et al [54]	2023	35
10	"Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine"	*New England Journal of Medicine*	Lee et al [55]	2023	34

Table 10 shows the top 20 references with the strongest citation bursts. The first citation burst occurred in 2023. This was a study comparing ChatGPT and Korean medical students on a parasitology exam. An article published in 2023, titled "Chat Generative Pretrained Transformer Fails the Multiple-Choice American College of Gastroenterology Self-Assessment Test" [56], experienced a citation burst in 2024, with the burst lasting until 2025. Researchers have continued to study ChatGPT's ability to pass medical exams; the types of exams have ranged from the USMLE to basic subject exams. Furthermore, researchers have begun to evaluate ChatGPT's performance on different types of exam questions.

Table . Top 20 references with the strongest citation bursts.

Title	First author	Source	IF2024	JCR	Publication type	Publication year	Strength	Begin	End	2023 - 2025
Are ChatGPT's Knowledge and Interpretation Ability Comparable to Those of Medical Students in Korea for Taking a Parasitology Examination?: A Descriptive Study	Sun Huh [57]	Journal of Educational Evaluation for Health Professions	3.7	Q1	Article	2023	3.97	2023	2023	
Will ChatGPT Transform Healthcare?	No authors listed	Nature Medicine	50	Q1	Editorial	2023	3.91	2023	2023	
ChatGPT Passing USMLE Shines a Spotlight on the Flaws of Medical Education	Amarachi B Mbakwe [58]	PLOS Digital Health	7.7	Q1	Editorial	2023	3.79	2023	2023	
ChatGPT: the Future of Discharge Summaries?	Sajan B Patel [59]	The Lancet Digital Health	24.1	Q1	Comment	2023	3.54	2023	2023	
ChatGPT for Clinical Vignette Generation, Revision, and Evaluation	James RA Benoit [60]	medRxiv	None	None	Article	2023	3.35	2023	2023	
Abstracts Written by ChatGPT Fool Scientists	Holly Else [61]	Nature	48.5	Q1	Article	2023	3.09	2023	2023	
Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine	Peter Lee [55]	The New England Journal of Medicine	78.5	Q1	Review	2023	3.01	2023	2023	

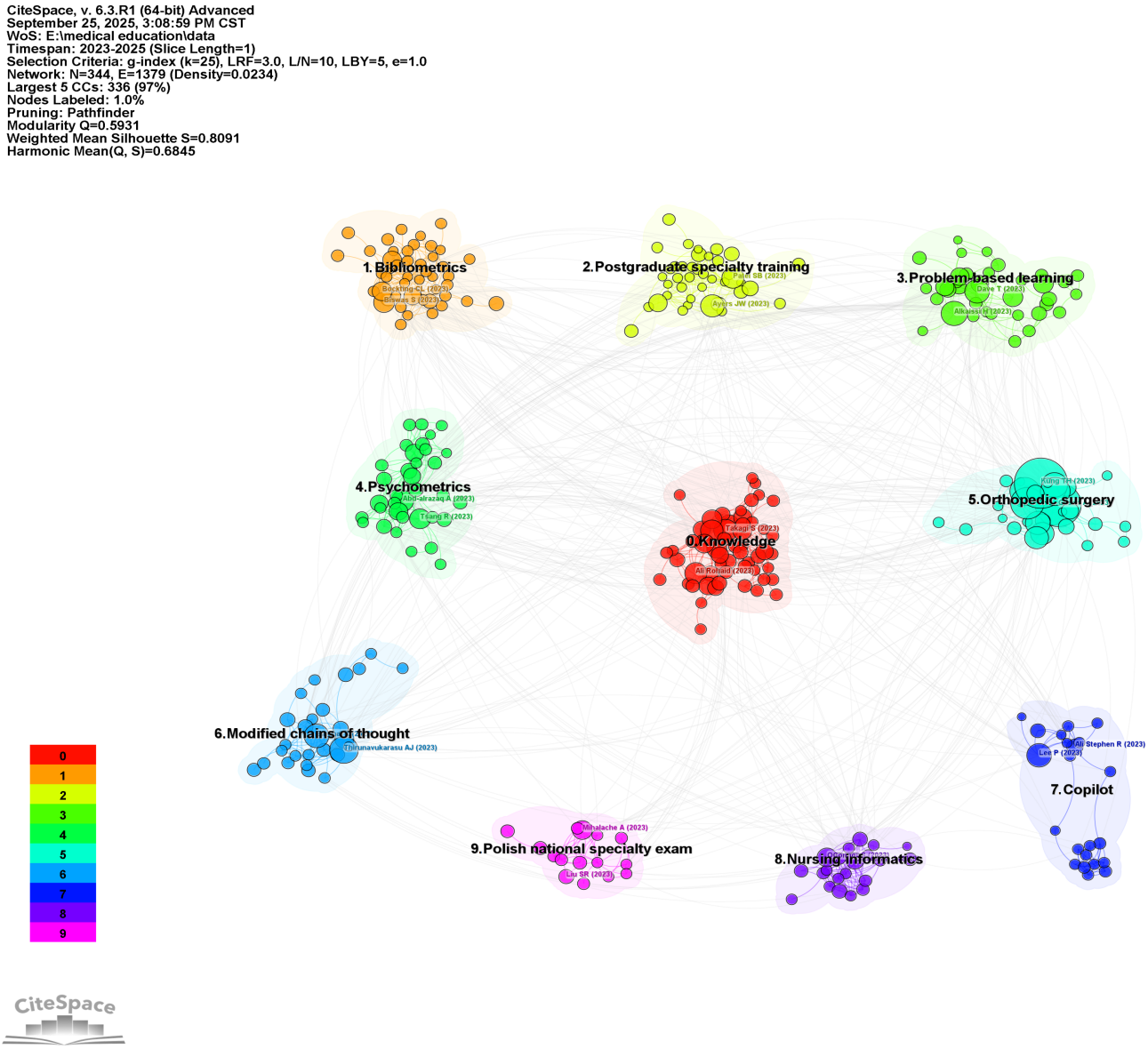
Title	First author	Source	IF2024	JCR	Publication type	Publication year	Strength	Begin	End	2023 - 2025
Tools Such as ChatGPT Threaten Transparent Science; Here Are Our Ground Rules for Their Use	No authors listed	Nature	48.5	Q1	Editorial	2023	2.23	2023	2023	
Could AI Help You to Write Your Next Paper?	Matthew Hutson [62]	Nature	48.5	Q1	Review	2022	2.23	2023	2023	
Appropriateness of Cardiovascular Disease Prevention Recommendations Obtained From a Popular Online Chat-Based Artificial Intelligence Model	Ashish Sarraju [63]	JAMA-Journal of the American Medical Association	55	Q1	Article	2023	2.23	2023	2023	
Medical Education Trends for Future Physicians in the Era of Advanced Technology and Artificial Intelligence: An Integrative Review	Eui-Ryoung Han [64]	BMC Medical Education	3.2	Q1	Review	2019	2.23	2023	2023	
The Exciting Potential for ChatGPT in Obstetrics and Gynecology	Amos Grünebaum [65]	American Journal of Obstetrics and Gynecology	8.4	Q1	Article	2023	2.23	2023	2023	
ChatGPT: Not All Languages Are Equal	Mohamed L Seghier [66]	Nature	48.5	Q1	Comment	2023	2.23	2023	2023	
GPT Takes the Bar Exam	Michael James Bommarito [67]	arXiv	None	None	Article	2022	2.23	2023	2023	

Title	First author	Source	IF2024	JCR	Publication type	Publication year	Strength	Begin	End	2023 - 2025
Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models	Tiffany H Kung [47]	PLOS Digital Health	7.7	Q1	Article	2023	2.13	2023	2023	
ChatGPT: Five Priorities for Research	Eva AM van Dis [8]	Nature	48.5	Q1	Comment	2023	2.13	2023	2023	
Chat Generative Pre-trained Transformer Fails the Multiple-Choice American College of Gastroenterology Self-Assessment Test	Kelly Suchman [56]	American Journal of Gastroenterology	7.6	Q1	Article	2023	1.8	2024	2025	
Capabilities of GPT-4 on Medical Challenge Problems	Harsha Nori [68]	arXiv	None	None	Article	2023	1.67	2023	2023	
Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing	Yu Gu [69]	ACM Transactions on Computing for Healthcare	8	Q1	Article	2022	1.67	2023	2023	
Natural Language Processing: State of the Art, Current Trends and Challenges	Diksha Khurana [70]	Multimedia Tools and Applications	3	Q3	Review	2023	1.67	2023	2023	

The process of clustering analysis was conducted based on the relevance between documents, with the result that the literature was divided into 9 categories (Figure 12), each of which was identified with a different color. The category with the highest number of publications was cluster 0. The term “knowledge” was a frequently occurring keyword in this category, indicating a concentration of studies evaluating ChatGPT’s medical knowledge. This finding is related to the research themes of the highly cited articles and burst articles. The cluster evolution shows that cluster 0 (knowledge) originated from clusters 2 (postgraduate specialty training) and 3 (problem-based learning), later developing into cluster 5 (orthopedic surgery). This progression reflects a research shift from foundational knowledge toward clinical applications. It is noteworthy that,

after Microsoft released Copilot in May 2023, studies expanded to cluster 7 (Copilot).

Figure 12. Clustering of references based on similarity.



Discussion

Principal Findings

We analyzed Web of Science literature on ChatGPT in medical education using VOSviewer and CiteSpace. The bibliometric results showed that researchers from 66 countries have participated in this field. The United States was the most prolific contributor, with 31% (126/407) of the published papers included in this study. Furthermore, the University of California system was ranked first with 3.4% (14/407) of the publications, reflecting the sustained investment of the United States in this area. Networks of countries and institutions have been established. This confirms active global communication and cooperation on the application of ChatGPT in medical education. Current research collaborations facilitate deeper international exchange in this field.

Notable contributors in this field include Yavuz Selim Kiyak, Ken Masters, Isil Irem Budakoglu, and Ozlem Coskun, all of whom are from academic institutions. Notably, 2 of the top 10 authors with the most publications are from hospitals or clinics. This indicates that clinicians are increasingly paying attention to the application of ChatGPT in clinical education. The results of the journal analysis showed that both citing and cited journals were from the field of medicine, and there was no emerging trend toward interdisciplinary research. The results of keyword analysis showed a shift in research focus from broader topics such as medical education assessment and medical exams to specific clinical disciplines such as dentistry and nephrology.

Comparison to the Literature

Our research findings are consistent with the existing literature. Research on ChatGPT in medical education primarily focuses on “medical knowledge,” “educational assessment,” “clinical decision-making,” and “exam performance” [50,71,72], with increasing attention to ethics concerns [73]. Previous studies have documented the increase in publications in the field, as well as the central role played by the United States [74].

Analyzing collaboration networks and journals may provide researchers in this field with a comprehensive understanding of institutional collaboration and journal publication information. Moreover, through thematic analysis, the evolution of research topics and recent research hot spots in this field were revealed.

Implications of the Findings

The findings of this study have a number of implications. The considerable increase in research on ChatGPT in medicine indicates its extensive integration into medical education and clinical practice. The strong international collaboration suggests that research outcomes related to ChatGPT in medical education are being shared worldwide. Authors conduct research in multiple small and isolated groups. This means that researchers in this field tend to work in independent teams and lack communication across different teams. The results of the dual-map overlay of the journals showed that both the citing and cited journals were from the field of medicine, indicating that research on ChatGPT in medical education has not yet been integrated with other disciplines, failing to form a cross-disciplinary trend.

Research themes evolved from preclinical education to clinical practice simulation, confirmed via keyword and citation analysis. As a learning tool, ChatGPT has passed medical licensing exams in the United States [75,76], India [77], the United Kingdom [78], and South Korea [79]. Its applications have expanded from basic exams to specialty tests, including the American Orthopaedic In-Training Examination [80], the Membership of the Royal Colleges of Physicians of the United Kingdom exam [81], and the Chinese Critical Care Examination [82]. Researchers test ChatGPT's medical knowledge through multiple exams. Nowadays, assessing the feasibility of ChatGPT as a learning tool is a research hot spot. Concurrently with the release of LLMs such as Microsoft Copilot, Google Gemini, and China's DeepSeek, studies have begun to compare different models' exam performance [83-85]. However, as LLMs keep evolving, we urgently need rigorous evidence of their reliability in medical testing. This proof remains essential before medical students fully adopt these learning tools.

In clinical teaching, ChatGPT has been used to emulate a range of clinical scenarios, including undiagnosed diabetes, kidney injury, and ophthalmic diseases [18-20]. This creates an interactive clinical reasoning environment for students, enhancing engagement during learning. However, it is important to note that ChatGPT sometimes generates inaccurate or fabricated information [86,87]. Medical educators and students need a clear understanding of ChatGPT's capabilities and limitations across medical specialties to effectively use AI tools for teaching and learning.

ChatGPT's integration into medical education raises ethical issues, highlighted by our keyword analysis. Researchers are concerned that it could unintentionally reveal a patient's personal information [88]. However, little research has been conducted on ChatGPT in medical ethics education (eg, medical ethics courses) and its educational impact [89]. While many studies have evaluated ChatGPT's performance in medical licensing exams across various countries, research on its ability to address medical ethics issues remains limited. The 2024 study by Danehy et al [90] showed that GPT-3.5 and ChatGPT-4 performed worse on ethics questions than on medical knowledge questions. This suggests that ChatGPT's training emphasizes medical knowledge over medical ethics. This training bias may be a potential trigger for ethical controversies involving ChatGPT in clinical practice.

Study Strengths and Limitations

This study has both strengths and limitations. To our knowledge, this is the first study to use bibliometric analysis to study the use of ChatGPT in medical education rather than general medicine. Furthermore, the visualization of quantitative results provides a comprehensive understanding of the current status of publications, research hot spots, and development trends related to ChatGPT in medical education.

Despite best efforts to include all the relevant terms and terminology in the literature search, some relevant papers may have been omitted. The search was confined to Web of Science, and only research articles written in English were included, with articles in other languages not being considered. In addition, due to the ongoing nature of the research, recent high-quality studies may not have been included.

Subsequently, the discussion focus on providing strong evidence to demonstrate the feasibility of ChatGPT as a learning tool, evaluating ChatGPT's medical ethics awareness in medical education, and offering evidence to support the application of ChatGPT in medical ethics.

Conclusions

In conclusion, this bibliometric analysis of ChatGPT in medical education reveals characteristics such as rapid publication growth, concentrated contributions from leading countries and institutions, decentralized author networks, and evolving thematic focuses. It will be crucial to enhance institution collaboration and cross-team partnerships in the future. This will promote the application potential of ChatGPT in various fields of medical education. Improving the effectiveness of ChatGPT is expected to provide educators and students with a more efficient medical teaching and learning process.

Acknowledgments

This study was supported by the Jiangxi Province Education Science 14th Five-Year Plan Project (grant 22QN048).

Authors' Contributions

YZ designed the study, collected and analyzed the data, and wrote the manuscript. XX assisted with data collection and analysis. QX contributed to the methodology, provided writing and editing support, and supervised the project.

Conflicts of Interest

None declared.

References

1. Floridi L, Chiriatti M. GPT-3: its nature, scope, limits, and consequences. Minds Mach 2020 Nov;30:681-694. [doi: 10.1007/s11023-020-09548-1]
2. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA 2025 Jan 28;333(4):319-328. [doi: 10.1001/jama.2024.21700] [Medline: 39405325]
3. ChatGPT. OpenAI. URL: https://chat.openai.com/ [accessed 2025-09-26]
4. Ali K, Barhom N, Tamimi F, Duggal M. ChatGPT-a double-edged sword for healthcare education? Implications for assessments of dental students. Eur J Dent Educ 2024 Feb;28(1):206-211. [doi: 10.1111/eje.12937] [Medline: 37550893]
5. Tian S, Jin Q, Yeganova L, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Brief Bioinform 2023 Nov 22;25(1):bbad493. [doi: 10.1093/bib/bbad493]
6. Sharma S, Pajai S, Prasad R, et al. A critical review of ChatGPT as a potential substitute for diabetes educators. Cureus 2023 May 1;15(5):e38380. [doi: 10.7759/cureus.38380] [Medline: 37265899]
7. Jansen BJ, Jung SG, Salminen J. Employing large language models in survey research. Nat Lang Proc J 2023 Sep;4:100020. [doi: 10.1016/j.nlp.2023.100020]
8. van Dis EA, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. Nature 2023 Feb;614(7947):224-226. [doi: 10.1038/d41586-023-00288-7] [Medline: 36737653]
9. Haleem A, Javaid M, Singh RP. An era of ChatGPT as a significant futuristic support tool: a study on features, abilities, and challenges. BenchCouncil Trans Benchmarks Stand Eval 2022 Oct;2(4):100089. [doi: 10.1016/j.tbench.2023.100089]
10. Ignjatović A, Stevanović L. Efficacy and limitations of ChatGPT as a biostatistical problem-solving tool in medical education in Serbia: a descriptive study. J Educ Eval Health Prof 2023;20:28. [doi: 10.3352/jeehp.2023.20.28] [Medline: 37840252]
11. Abujaber AA, Abd-Alrazaq A, Al-Qudimat AR, Nashwan AJ. A strengths, weaknesses, opportunities, and threats (SWOT) analysis of ChatGPT integration in nursing education: a narrative review. Cureus 2023 Nov 11;15(11):e48643. [doi: 10.7759/cureus.48643] [Medline: 38090452]
12. Liu J, Liu F, Fang J, Liu S. The application of chat generative pre-trained transformer in nursing education. Nurs Outlook 2023;71(6):102064. [doi: 10.1016/j.outlook.2023.102064] [Medline: 37879261]
13. Wu Y, Zheng Y, Feng B, Yang Y, Kang K, Zhao A. Embracing ChatGPT for medical education: exploring its impact on doctors and medical students. JMIR Med Educ 2024 Apr 10;10:e52483. [doi: 10.2196/52483] [Medline: 38598263]
14. Jeyaraman M, K SP, Jeyaraman N, Nallakumarasamy A, Yadav S, Bondili SK. ChatGPT in medical education and research: a boon or a bane? Cureus 2023 Aug 29;15(8):e44316. [doi: 10.7759/cureus.44316] [Medline: 37779749]
15. Wang C, Li S, Lin N, et al. Application of large language models in medical training evaluation-using ChatGPT as a standardized patient: multimetric assessment. J Med Internet Res 2025 Jan 1;27:e59435. [doi: 10.2196/59435] [Medline: 39742453]
16. Wu Z, Li S, Zhao X. The application of ChatGPT in medical education: prospects and challenges. Int J Surg 2025 Jan 1;111(1):1652-1653. [doi: 10.1097/JS9.0000000000001887] [Medline: 38935099]
17. Scherr R, Halaseh FF, Spina A, Andalib S, Rivera R. ChatGPT interactive medical simulations for early clinical education: case study. JMIR Med Educ 2023 Nov 10;9:e49877. [doi: 10.2196/49877] [Medline: 37948112]
18. Chatterjee S, Bhattacharya M, Pal S, Lee SS, Chakraborty C. ChatGPT and large language models in orthopedics: from education and surgery to research. J Exp Orthop 2023 Dec 1;10(1):128. [doi: 10.1186/s40634-023-00700-1] [Medline: 38038796]
19. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ 2023 Mar 6;9:e46885. [doi: 10.2196/46885] [Medline: 36863937]
20. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. Ophthalmol Sci 2023 May 5;3(4):100324. [doi: 10.1016/j.xops.2023.100324] [Medline: 37334036]
21. Gonzalez-Garcia A, Bermejo-Martinez D, Lopez-Alonso AI, Trevisson-Redondo B, Martín-Vázquez C, Perez-Gonzalez S. Impact of ChatGPT usage on nursing students education: a cross-sectional study. Heliyon 2024 Dec 31;11(1):e41559. [doi: 10.1016/j.heliyon.2024.e41559] [Medline: 39850430]
22. Gencer A, Aydin S. Can ChatGPT pass the thoracic surgery exam? Am J Med Sci 2023 Oct;366(4):291-295. [doi: 10.1016/j.amjms.2023.08.001] [Medline: 37549788]
23. Soulage CO, Van Coppenolle F, Guebre-Egziabher F. The conversational AI "ChatGPT" outperforms medical students on a physiology university examination. Adv Physiol Educ 2024 Dec 1;48(4):677-684. [doi: 10.1152/advan.00181.2023] [Medline: 38991037]
24. Benítez TM, Xu Y, Boudreau JD, et al. Harnessing the potential of large language models in medical education: promise and pitfalls. J Am Med Inform Assoc 2024 Feb 16;31(3):776-783. [doi: 10.1093/jamia/ocad252] [Medline: 38269644]

25. Kim TW. Application of artificial intelligence chatbots, including ChatGPT, in education, scholarly work, programming, and content generation and its prospects: a narrative review. J Educ Eval Health Prof 2023;20:38. [doi: 10.3352/jeehp.2023.20.38] [Medline: 38148495]
26. Sabry Abdel-Messih M, Kamel Boulos MN. ChatGPT in clinical toxicology. JMIR Med Educ 2023 Mar 8;9:e46876. [doi: 10.2196/46876] [Medline: 36867743]
27. Ming S, Guo Q, Cheng W, Lei B. Influence of model evolution and system roles on ChatGPT's performance in Chinese medical licensing exams: comparative study. JMIR Med Educ 2024 Aug 13;10:e52784. [doi: 10.2196/52784] [Medline: 39140269]
28. Huang CH, Hsiao HJ, Yeh PC, Wu KC, Kao CH. Performance of ChatGPT on stage 1 of the Taiwanese medical licensing exam. Digit Health 2024 Feb 16;10:20552076241233144. [doi: 10.1177/20552076241233144] [Medline: 38371244]
29. Ishida K, Hanada E. Potential of ChatGPT to pass the Japanese medical and healthcare professional national licenses: a literature review. Cureus 2024 Aug 6;16(8):e66324. [doi: 10.7759/cureus.66324] [Medline: 39247019]
30. Kawahara T, Sumi Y. GPT-4/4V's performance on the Japanese National Medical Licensing Examination. Med Teach 2025 Mar;47(3):450-457. [doi: 10.1080/0142159X.2024.2342545] [Medline: 38648547]
31. Scaioli G, Lo Moro G, Conrado F, Rosset L, Bert F, Siliquini R. Exploring the potential of ChatGPT for clinical reasoning and decision-making: a cross-sectional study on the Italian Medical Residency Exam. Ann Ist Super Sanita 2023;59(4):267-270. [doi: 10.4415/ANN_23_04_05] [Medline: 38088393]
32. Torres-Zegarra BC, Rios-Garcia W, Ñaña-Cordova AM, et al. Performance of ChatGPT, Bard, Claude, and Bing on the Peruvian National Licensing Medical Examination: a cross-sectional study. J Educ Eval Health Prof 2023;20:30. [doi: 10.3352/jeehp.2023.20.30] [Medline: 37981579]
33. Arruda H, Silva ER, Lessa M, Proença Jr D, Bartholo R. VOSviewer and Bibliometrix. J Med Libr Assoc 2022 Jul 1;110(3):392-395. [doi: 10.5195/jmla.2022.1434] [Medline: 36589296]
34. Zhao X, Nan D, Chen C, Zhang S, Che S, Kim JH. Bibliometric study on environmental, social, and governance research using CiteSpace. Front Environ Sci 2023;10. [doi: 10.3389/fenvs.2022.1087493]
35. Zhou F, Zhang T, Jin Y, et al. Worldwide tinnitus research: a bibliometric analysis of the published literature between 2001 and 2020. Front Neurol 2022 Jan 31;13:828299. [doi: 10.3389/fneur.2022.828299] [Medline: 35173675]
36. Zhou F, Zhang T, Jin Y, et al. Unveiling the knowledge domain and emerging trends of olfactory dysfunction with depression or anxiety: a bibliometrics study. Front Neurosci 2022 Sep 8;16:959936. [doi: 10.3389/fnins.2022.959936] [Medline: 36161166]
37. Zhou Q, Pei J, Poon J, et al. Worldwide research trends on aristolochic acids (1957-2017): suggestions for researchers. PLoS ONE 2019 May 2;14(5):e0216135. [doi: 10.1371/journal.pone.0216135] [Medline: 31048858]
38. Bibilometrix. URL: https://www.bibliometrix.org/home/ [accessed 2025-09-29]
39. Bibliometrc. URL: https://bibliometric.com/ [accessed 2025-09-29]
40. Synnestvedt MB, Chen C, Holmes JH. CiteSpace II: visualization and knowledge discovery in bibliographic databases. AMIA Annu Symp Proc 2005;2005:724-728. [Medline: 16779135]
41. Chen C. Searching for intellectual turning points: progressive knowledge domain visualization. Proc Natl Acad Sci U S A 2004 Apr 6;101 Suppl 1(Suppl 1):5303-5310. [doi: 10.1073/pnas.0307513100] [Medline: 14724295]
42. Xu S, Xu D, Wen L, et al. Integrating unified medical language system and Kleinberg's burst detection algorithm into research topics of medications for post-traumatic stress disorder. Drug Des Devel Ther 2020 Sep 24;14:3899-3913. [doi: 10.2147/DDDT.S270379] [Medline: 33061296]
43. van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 2010 Aug;84(2):523-538. [doi: 10.1007/s11192-009-0146-3] [Medline: 20585380]
44. Aria M, Cuccurullo C. bibliometrix: an R-tool for comprehensive science mapping analysis. J Informetr 2017 Nov;11(4):959-975. [doi: 10.1016/j.joi.2017.08.007]
45. Hou J, Yang X, Chen C. Emerging trends and new developments in information science: a document co-citation analysis (2009–2016). Scientometrics 2018 May 1;115(2):869-892. [doi: 10.1007/s11192-018-2695-9]
46. Li Q, Long R, Chen H, Chen F, Wang J. Visualized analysis of global green buildings: development, barriers and future directions. J Clean Prod 2020 Feb 1;245:118775. [doi: 10.1016/j.jclepro.2019.118775]
47. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Sep;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
48. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
49. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887. [doi: 10.3390/healthcare11060887] [Medline: 36981544]
50. Lee H. The rise of ChatGPT: exploring its potential in medical education. Anat Sci Educ 2024;17(5):926-931. [doi: 10.1002/ase.2270] [Medline: 36916887]

51. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
52. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT - Reshaping medical education and clinical management. Pak J Med Sci ;39(2). [doi: 10.12669/pjms.39.2.7653]
53. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus 2023 Feb;15(2):e35179. [doi: 10.7759/cureus.35179] [Medline: 36811129]
54. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023;6:1169595. [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
55. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023 Mar 30;388(13):1233-1239. [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
56. Suchman K, Garg S, Trindade AJ. Chat generative pretrained transformer fails the multiple-choice American College of Gastroenterology Self-Assessment Test. Am J Gastroenterol 2023 Dec 1;118(12):2280-2282. [doi: 10.14309/ajg.0000000000002320] [Medline: 37212584]
57. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. J Educ Eval Health Prof ;20:1. [doi: 10.3352/jeehp.2023.20.1]
58. Mbakwe AB, Lourentzou I, Celi LA, Mechanic OJ, Dagan A. ChatGPT passing USMLE shines a spotlight on the flaws of medical education. PLOS Digit Health 2023 Feb;2(2):e0000205. [doi: 10.1371/journal.pdig.0000205] [Medline: 36812618]
59. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health 2023 Mar;5(3):e107-e108. [doi: 10.1016/S2589-7500(23)00021-3] [Medline: 36754724]
60. Benoit JRA. ChatGPT for clinical vignette generation, revision, and evaluation. Medical Education. Preprint posted online on Feb 8, 2023. [doi: 10.1101/2023.02.04.23285478]
61. Else H. Abstracts written by ChatGPT fool scientists. Nature New Biol 2023 Jan 19;613(7944):423. [doi: 10.1038/d41586-023-00056-7]
62. Hutson M. Could AI help you to write your next paper? Nature New Biol 2022 Nov 3;611(7934):192-193. [doi: 10.1038/d41586-022-03479-w]
63. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. JAMA 2023 Mar 14;329(10):842-844. [doi: 10.1001/jama.2023.1044] [Medline: 36735264]
64. Han ER, Yeo S, Kim MJ, Lee YH, Park KH, Roh H. Medical education trends for future physicians in the era of advanced technology and artificial intelligence: an integrative review. BMC Med Educ 2019 Dec;19(1). [doi: 10.1186/s12909-019-1891-5]
65. Grünebaum A, Chervenak J, Pollet SL, Katz A, Chervenak FA. The exciting potential for ChatGPT in obstetrics and gynecology. Am J Obstet Gynecol 2023 Jun;228(6):696-705. [doi: 10.1016/j.ajog.2023.03.009]
66. Seghier ML. ChatGPT: not all languages are equal. Nature New Biol 2023 Mar 9;615(7951):216. [doi: 10.1038/d41586-023-00680-3]
67. Bommarito MJ, Katz DM. GPT takes the bar exam. SSRN Journal. 2022 Dec 29. [doi: 10.2139/ssrn.4314839]
68. Capabilities of GPT-4 on medical challenge problems. arXiv. Preprint posted online on Apr 12, 2023. [doi: 10.48550/arXiv.2303.13375]
69. Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthcare 2022 Jan 31;3(1):1-23. [doi: 10.1145/3458754]
70. Khurana D, Koli A, Khatter K, Singh S. Natural language processing: state of the art, current trends and challenges. Multimed Tools Appl 2023 Jan;82(3):3713-3744. [doi: 10.1007/s11042-022-13428-4]
71. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. Acad Med 2024 Jan 1;99(1):22-27. [doi: 10.1097/ACM.0000000000005439] [Medline: 37651677]
72. Tan S, Xin X, Wu D. ChatGPT in medicine: prospects and challenges: a review article. Int J Surg 2024 Jun 1;110(6):3701-3706. [doi: 10.1097/JS9.0000000000001312] [Medline: 38502861]
73. Cheng Y, Zhu L. A review of ChatGPT in medical education: exploring advantages and limitations. Int J Surg 2025 Jul 1;111(7):4586-4602. [doi: 10.1097/JS9.0000000000002505] [Medline: 40465793]
74. Wu J, Ma Y, Wang J, Xiao M. The application of ChatGPT in medicine: a scoping review and bibliometric analysis. J Multidiscip Healthc 2024 Apr 18;17:1681-1692. [doi: 10.2147/JMDH.S463128] [Medline: 38650670]
75. Bicknell BT, Butler D, Whalen S, et al. ChatGPT-4 omni performance in USMLE disciplines and clinical skills: comparative analysis. JMIR Med Educ 2024 Nov 6;10:e63430. [doi: 10.2196/63430] [Medline: 39504445]
76. Alfertshofer M, Knoedler S, Hoch CC, et al. Analyzing question characteristics influencing ChatGPT's performance in 3000 USMLE®-style questions. Med Sci Educ 2024 Sep 28;35(1):257-267. [doi: 10.1007/s40670-024-02176-9] [Medline: 40144074]
77. Surapaneni KM. Assessing the performance of ChatGPT in medical biochemistry using clinical case vignettes: observational study. JMIR Med Educ 2023 Nov 7;9:e47191. [doi: 10.2196/47191] [Medline: 37934568]

78. Lai UH, Wu KS, Hsu TY, Kan JK. Evaluating the performance of ChatGPT-4 on the United Kingdom Medical Licensing Assessment. Front Med (Lausanne) 2023 Sep 19;10:1240915. [doi: 10.3389/fmed.2023.1240915] [Medline: 37795422]
79. Oh N, Choi GS, Lee WY. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. Ann Surg Treat Res 2023 May;104(5):269-273. [doi: 10.4174/astr.2023.104.5.269] [Medline: 37179699]
80. Kung JE, Marshall C, Gauthier C, Gonzalez TA, Jackson JB 3rd. Evaluating ChatGPT performance on the orthopaedic in-training examination. JB JS Open Access 2023;8(3):e23.00056. [doi: 10.2106/JBJS.OA.23.00056] [Medline: 37693092]
81. Maitland A, Fowkes R, Maitland S. Can ChatGPT pass the MRCP (UK) written examinations? Analysis of performance and errors using a clinical decision-reasoning framework. BMJ Open 2024 Mar 15;14(3):e080558. [doi: 10.1136/bmjopen-2023-080558] [Medline: 38490655]
82. Wang X, Tang J, Feng Y, Tang C, Wang X. Can ChatGPT-4 perform as a competent physician based on the Chinese critical care examination? J Crit Care 2025 Apr;86:155010. [doi: 10.1016/j.jcrc.2024.155010] [Medline: 40023616]
83. Thesen T, Tuan RL, Blumer J, Lee MW. LLM-based generation of USMLE-style questions with ASPET/AMSPC knowledge objectives: all RAGs and no riches. Br J Clin Pharmacol 2025 Jun 8. [doi: 10.1002/bcp.70119] [Medline: 40483567]
84. Camarata T, McCoy L, Rosenberg R, Temprine Grellinger KR, Brettschnieder K, Berman J. LLM-generated multiple choice practice quizzes for preclinical medical students. Adv Physiol Educ 2025 Sep 1;49(3):758-763. [doi: 10.1152/advan.00106.2024] [Medline: 40516963]
85. Yang H, Li M, Zhou H, et al. Large language model synergy for ensemble learning in medical question answering: design and evaluation study. J Med Internet Res 2025 Jul 14;27:e70080. [doi: 10.2196/70080] [Medline: 40658884]
86. Barrington NM, Gupta N, Musmar B, et al. A bibliometric analysis of the rise of ChatGPT in medical research. Med Sci (Basel) 2023 Sep 17;11(3):61. [doi: 10.3390/medsci11030061] [Medline: 37755165]
87. Ang TL, Choolani M, See KC, Poh KK. The rise of artificial intelligence: addressing the impact of large language models such as ChatGPT on scientific publications. Singapore Med J 2023 Apr;64(4):219-221. [doi: 10.4103/singaporemedj.SMJ-2023-055] [Medline: 37006087]
88. Naik N, Hameed BM, Shetty DK, et al. Legal and ethical consideration in artificial intelligence in healthcare: who takes responsibility? Front Surg 2022 Mar 14;9:862322. [doi: 10.3389/fsurg.2022.862322] [Medline: 35360424]
89. Weidener L, Fischer M. Teaching AI ethics in medical education: a scoping review of current literature and practices. Perspect Med Educ 2023 Oct 16;12(1):399-410. [doi: 10.5334/pme.954] [Medline: 37868075]
90. Danehy T, Hecht J, Kentis S, Schechter CB, Jariwala SP. ChatGPT performs worse on USMLE-style ethics questions compared to medical knowledge questions. Appl Clin Inform 2024 Oct;15(5):1049-1055. [doi: 10.1055/a-2405-0138] [Medline: 39209308]

Abbreviations

AI: artificial intelligence
BC: betweenness centrality
IF: impact factor
LLM: large language model
RQ: research question
USMLE: United States Medical Licensing Examination

Edited by B Lesselroth; submitted 08.02.25; peer-reviewed by C br, D Nan, G Gill, MK Ghanta, W Yang; revised version received 11.08.25; accepted 09.09.25; published 07.10.25.

Please cite as:
Zhang Y, Xie X, Xu Q
ChatGPT in Medical Education: Bibliometric and Visual Analysis
JMIR Med Educ 2025;11:e72356
URL: https://mededu.jmir.org/2025/1/e72356
doi:10.2196/72356

AI's Accuracy in Extracting Learning Experiences From Clinical Practice Logs: Observational Study

Takeshi Kondo[1,2], MD, MHPE, PhD; Hiroshi Nishigori[1], MD, MMEd, PhD

[1]Center for Medical Education, Nagoya University Graduate School of Medicine, 65, Tsurumai-cho, Showa-ku, Nagoya city, Aichi, Japan

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

Corresponding Author:
Takeshi Kondo, MD, MHPE, PhD
Center for Medical Education, Nagoya University Graduate School of Medicine, 65, Tsurumai-cho, Showa-ku, Nagoya city, Aichi, Japan

Abstract

Background: Improving the quality of education in clinical settings requires an understanding of learners' experiences and learning processes. However, this is a significant burden on learners and educators. If learners' learning records could be automatically analyzed and their experiences could be visualized, this would enable real-time tracking of their progress. Large language models (LLMs) may be useful for this purpose, although their accuracy has not been sufficiently studied.

Objective: This study aimed to explore the accuracy of predicting the actual clinical experiences of medical students from their learning log data during clinical clerkship using LLMs.

Methods: This study was conducted at the Nagoya University School of Medicine. Learning log data from medical students participating in a clinical clerkship from April 22, 2024, to May 24, 2024, were used. The Model Core Curriculum for Medical Education was used as a template to extract experiences. OpenAI's ChatGPT was selected for this task after a comparison with other LLMs. Prompts were created using the learning log data and provided to ChatGPT to extract experiences, which were then listed. A web application using GPT-4-turbo was developed to automate this process. The accuracy of the extracted experiences was evaluated by comparing them with the corrected lists provided by the students.

Results: A total of 20 sixth-year medical students participated in this study, resulting in 40 datasets. The overall Jaccard index was 0.59 (95% CI 0.46-0.71), and the Cohen κ was 0.65 (95% CI 0.53-0.76). Overall sensitivity was 62.39% (95% CI 49.96%-74.81%), and specificity was 99.34% (95% CI 98.77%-99.92%). Category-specific performance varied: symptoms showed a sensitivity of 45.43% (95% CI 25.12%-65.75%) and specificity of 98.75% (95% CI 97.31%-100%), examinations showed a sensitivity of 46.76% (95% CI 25.67%-67.86%) and specificity of 98.84% (95% CI 97.81%-99.87%), and procedures achieved a sensitivity of 56.36% (95% CI 37.64%-75.08%) and specificity of 98.92% (95% CI 96.67%-100%). The results suggest that GPT-4-turbo accurately identified many of the actual experiences but missed some because of insufficient detail or a lack of student records.

Conclusions: This study demonstrated that LLMs such as GPT-4-turbo can predict clinical experiences from learning logs with high specificity but moderate sensitivity. Future improvements in AI models, providing feedback to medical students' learning logs and combining them with other data sources such as electronic medical records, may enhance the accuracy. Using artificial intelligence to analyze learning logs for assessment could reduce the burden on learners and educators while improving the quality of educational assessments in medical education.

(JMIR Med Educ 2025;11:e68697) doi:10.2196/68697

KEYWORDS

large language models; ChatGPT; workplace-based assessment; artificial intelligence; AI

Introduction

Background

To improve the quality of education in clinical settings, it is important to understand what learners experience and how they learn [1,2]. Various workplace-based assessment tools have been developed and used to enable educators to track learners' progress and provide feedback [3]. However, the rigorous management of learners' progress requires frequent observation of learners, frequent evaluations, and feedback from educators. This can impose a high burden on both learners and educators, potentially hindering learning [4,5]. Thus, the challenge is accurately monitoring learning in clinical settings without burdening learners or educators.

Learners in clinical settings often document their learning and practice experiences. If these records can be analyzed to

understand learners' contexts, monitoring their learning without imposing additional burdens may be possible. One such record kept by learners during clinical clerkship is a logbook. The logbook documents the cases encountered, procedures performed, and learners' reflections. It serves as a tool for prompting student reflections and facilitating feedback and dialogue between educators and learners [6-8]. Evaluating these records against curriculum competencies and goals without adding an extra burden on learners can help monitor their progress [9]. However, educators may have to manually match and analyze these records, which may be a significant burden [5].

Artificial intelligence (AI)–assisted text extraction and standard matching could be useful in this context. Previous studies have successfully used natural language processing, a branch of AI, to analyze supervisory feedback comments and predict student performance against competency standards [10]. AI models that integrate multiple information sources to represent student performance have also been developed [11]. Among AI technologies, large language models (LLMs) have gained attention in medical education because of their extensive pretraining on large datasets, allowing them to handle various situations, including multilingual support, with minimal adjustment [12]. Research using ChatGPT, an LLM, has shown that it can apply codes to interview texts using a codebook, suggesting its potential for extracting competency-based evaluations from student descriptions [13]. However, owing to a lack of such research, aggregation accuracy remains uncertain. Determining the extent to which LLMs can aggregate items related to curriculum goals from learner descriptions may open up opportunities to leverage LLMs to monitor learner progress and enhance education quality.

In Japanese undergraduate education, the Model Core Curriculum for Medical Education (MCC) [14] was established to define two-thirds of the undergraduate curriculum and is used as a guideline for undergraduate medical education. The MCC outlines the experiences that medical students should have by the time they graduate, focusing primarily on clinical clerkships [14]. In Japanese clinical clerkships, medical students are partially observed directly by supervisors [15], but it is difficult for busy supervisors to grasp the full scope of experiences that medical students encounter [14]. If experiences could be understood through analysis of learning logs kept by medical students, valuable information for improving the learning environment could be obtained.

Objectives

Therefore, this study focused on undergraduate clinical clerkships in Japan to investigate the accuracy with which LLMs can aggregate goals from records kept for learning. Our research question was as follows: how accurately can an LLM predict experiences related to the goals defined by the MCC from the records that students keep for learning during clinical clerkships?

Methods

Context

This study was conducted as part of the participatory clinical clerkship at the Nagoya University School of Medicine, a program designed to provide medical students with practical experience in clinical settings. During the final year of medical school (sixth year), students participate in this program for 4 weeks, recording their daily experiences and learning activities. A trial to transform these records into an electronic portfolio began in 2024. This study was part of this trial.

Dataset

This study used learning log data from sixth-year medical students to extract their experiences related to core curriculum goals. Learning log data consisted of daily records of experiences and learning activities entered by medical students into an electronic portfolio during a clinical clerkship from April 22, 2024, to May 24, 2024. The data were treated as weekly datasets.

Extraction of Experiences

The template for extracting experiences from the dataset was the MCC [14]. This study used a table of symptoms, examinations, and procedures that medical students are expected to encounter in patients during their clinical clerkship at Nagoya University School of Medicine as the template for experience extraction (Textbox 1).

Textbox 1. Symptoms, examinations, and procedures that medical students are expected to encounter in patients.

Symptoms

- Fever
- General malaise
- Anorexia
- Weight loss
- Weight gain
- Altered mental status
- Syncope
- Seizure
- Vertigo and dizziness
- Edema
- Rash
- Cough and sputum production
- Blood in sputum and hemoptysis
- Dyspnea
- Chest pain
- Palpitations
- Dysphagia
- Abdominal pain
- Nausea and vomiting
- Hematemesis
- Melena
- Constipation
- Diarrhea
- Jaundice
- Abdominal distention and abdominal mass
- Lymphadenopathy
- Abnormal urine output or urination
- Hematuria
- Menstrual abnormality
- Anxiety or depression
- Cognitive dysfunction
- Headache
- Skeletal muscle paralysis or muscle weakness
- Gait disturbance
- Sensory disturbance
- Back pain
- Arthralgia or joint swelling

Examinations

- Full blood count
- Blood biochemistry
- Coagulation or fibrinolysis

- Immunoserology tests
- Urinalysis
- Stool (fecal) examination
- Blood typing (ABO, and RhHD), blood compatibility test (cross-matching), and atypical antibody screening
- Arterial blood gas analysis
- Pregnancy test
- Microbiological tests (bacterial smear, culture, identification, and antibiotic sensitivity test)
- Cerebrospinal fluid
- Pleural fluid analysis
- Peritoneal fluid analysis
- Histopathology and cytology (including intraoperative rapid diagnosis)
- Genetic testing and chromosome analysis
- Electrocardiography (ECG)
- Lung function tests
- Endocrine and metabolic function tests
- Electroencephalography
- Ultrasound
- X-ray
- Computed tomography
- Magnetic resonance imaging
- Nuclear medicine examination
- Endoscopy

Procedures

- Position change and transfer
- Skin antisepsis
- Application of topical medications
- Airway suction
- Nebulizer
- Venous blood sampling
- Peripheral venous catheterization
- Insertion and extraction of nasogastric tube
- Insertion and extraction of urinary catheter
- Intradermal injection
- Subcutaneous injection
- Intramuscular injection
- Intravenous injection
- Urinalysis (including pregnancy test)
- Microbiological testing (including Gram staining)
- Recording of a 12-lead ECG
- Rapid bedside ultrasound (including focused assessment with sonography for trauma [FAST]) for clinical decision-making
- Rapid antigen or pathogen testing
- Blood glucose test
- Aseptic technique

- Surgical hand washing
- Gowning techniques in the operating room
- Basic sutures and suture removal

OpenAI's ChatGPT, Google's Gemini, and Anthropic's Claude were considered for the LLMs used in experience extraction. Trial prompts and randomly selected student records were entered into each web platform, and the extracted results were compared in terms of validity. Validity was evaluated from the perspective of whether the output followed the expected format, whether the output matched the experience items expected from the text, and whether the output was reproducible. ChatGPT by OpenAI produced the most valid outputs, so it was selected for this study.

LLMs, including ChatGPT, receive text data as input and generate subsequent text based on these data. Therefore, the prompt given to the LLM is crucial. In this study, prompts were created using medical students' learning log data, which were provided to ChatGPT to extract their experiences from the logs. Experiences were extracted based on a table of symptoms, examinations, and procedures that students were expected to experience, with ChatGPT outputting a list of symptoms, examinations, and procedures inferred from the text data. To automate this process, a web application using GPT-4-turbo was developed, which allowed medical students to input learning log data and receive the extracted experiences as a list output from GPT-4-turbo (gpt-4-0125-preview). The prompt used for GPT-4-turbo and the web application code are provided in Multimedia Appendix 1.

Evaluation of Extracted Experiences

The extracted experience goals were presented to the medical students via email. Students were asked to compare the list with their actual experiences, including those not recorded in their reflections, and submit a corrected list. The corrected lists were compared with the original learning log data to evaluate the accuracy of the extracted experiences.

Data Analysis

The accuracy of the extracted experience goals was evaluated using the R software (version 4.1.2; R Foundation for Statistical Computing). The agreement rate between the extracted and corrected experience goals was calculated, and the accuracy of the extracted experience goals was assessed based on this agreement rate.

Ethical Considerations

This study was approved by the ethics committee of Nagoya University Graduate School of Medicine (approval 2023-0451 31742). All participants were informed about the study's purpose, methods, risks, and benefits and were allowed to opt out. All data were fully anonymized and handled to prevent the identification of individuals. No compensation was provided to participants in this study.

Results

Study Period, Participants, and Data Characteristics

During the clinical participation-based clerkship at Nagoya University Hospital from April 22, 2024, to May 24, 2024, a total of 61% (20/33) of the sixth-year students who made entries in the e-portfolio participated in the study, yielding 40 data points. All records were written in Japanese, with an average letter count of 446.2 (SD 353.52; range 72-1473). The predicted and actual experiences are shown in Table 1.

Table . Predicted and actual experience items.

Record index	Predicted item	Actual item	Number of matches	Number of experienced items extracted by GPT-4-turbo	Number of items that the students marked as experiences they had
1	Skeletal muscle paralysis or muscle weakness, gait disturbance, and sensory disturbance	Skeletal muscle paralysis or muscle weakness, gait disturbance, and sensory disturbance	3	3	3
2	Endocrine and metabolic function tests	Endocrine and metabolic function tests	1	1	1
3	Fever	Fever	1	1	1
4	Basic sutures and suture removal	Basic sutures and suture removal	1	1	1
5	Seizure, electroencephalography, and MRI[a]	Aseptic technique, electroencephalography, MRI, weight gain, and seizure	3	3	5
6	Skeletal muscle paralysis or muscle weakness, gait disturbance, and sensory disturbance	Skeletal muscle paralysis or muscle weakness, gait disturbance, and sensory disturbance	3	3	3
7	Anorexia, abdominal distention and abdominal mass, and ultrasound	Palpitations and skeletal muscle paralysis or muscle weakness	0	3	2
8	Venous blood sampling	Venous blood sampling	1	1	1
9	Rapid bedside ultrasound (including FAST[b]) for clinical decision-making and ultrasound	Skin antisepsis, rapid bedside ultrasound (including FAST) for clinical decision-making, aseptic technique, surgical handwashing, gowning techniques in the operating room, basic sutures and suture removal, ultrasound, fever, and diarrhea	2	2	9
10	Basic sutures and suture removal	Basic sutures and suture removal	1	1	1
11	Basic sutures and suture removal	Surgical handwashing, gowning techniques in the operating room, basic sutures and suture removal, full blood count, blood biochemistry, coagulation and fibrinolysis, histopathology and cytology (including intraoperative rapid diagnosis), x-ray, CT[c], MRI, general malaise, and weight loss	1	1	12
12	Venous blood sampling and pregnancy test	Venous blood sampling	1	2	1

Record index	Predicted item	Actual item	Number of matches	Number of experienced items extracted by GPT-4-turbo	Number of items that the students marked as experiences they had
13	Surgical handwashing, gowning techniques in the operating room, and basic sutures and suture removal	Surgical handwashing, gowning techniques in the operating room, and basic sutures and suture removal	3	3	3
14	Pregnancy test and basic sutures and suture removal	Position change and transfer, insertion and extraction of a urinary catheter, surgical handwashing, gowning techniques in the operating room, basic sutures and suture removal, histopathology and cytology (including intraoperative rapid diagnosis), MRI, and abdominal distention and abdominal mass	1	2	8
15	Surgical handwashing	Surgical handwashing, gowning techniques in the operating room, basic sutures and suture removal, full blood count, blood biochemistry, histopathology and cytology (including intraoperative rapid diagnosis), ultrasound, and x-ray	1	1	8
16	Microbiological tests (bacterial smear, culture, identification, and antibiotic sensitivity test), nuclear medicine examination, general malaise, cough and sputum production, dyspnea, abdominal pain, nausea and vomiting, and abnormal urine output or urination	General malaise and edema	1	8	2
17	Surgical handwashing	Aseptic technique, surgical handwashing, gowning techniques in the operating room, basic sutures and suture removal, and cough and sputum production	1	1	5
18	Fever, urinalysis, microbiological tests (bacterial smear, culture, identification, and antibiotic sensitivity test), nausea and vomiting, and hematuria	Full blood count, blood biochemistry, immunoserology tests, urinalysis, microbiological tests (bacterial smear, culture, identification, and antibiotic sensitivity test), edema, palpitations, hematuria, and back pain	3	5	9

Record index	Predicted item	Actual item	Number of matches	Number of experienced items extracted by GPT-4-turbo	Number of items that the students marked as experiences they had
19	Blood glucose test and endocrine and metabolic function tests	Blood glucose test and endocrine and metabolic function tests	2	2	2
20	Cognitive dysfunction	Cognitive dysfunction	1	1	1
21	Chest pain	Chest pain	1	1	1
22	Surgical handwashing and basic sutures and suture removal	Aseptic technique, surgical handwashing, gowning techniques in the operating room, and basic sutures and suture removal	2	2	4
23	Cognitive dysfunction, abnormal urine output or urination, and urinalysis	Cognitive dysfunction and abnormal urine output or urination	2	3	2
24	Dyspnea	Dyspnea	1	1	1
25	Gowning techniques in the operating room	Position change and transfer, skin antisepsis, aseptic technique, surgical handwashing, gowning techniques in the operating room, and basic sutures and suture removal	1	1	6
26	Full blood count and blood biochemistry	Full blood count, blood biochemistry, immunoserology tests, and edema	2	2	4
27	CT, MRI, and x-ray	Position change and transfer, full blood count, arterial blood gas analysis, ultrasound, x-ray, CT, MRI, skeletal muscle paralysis or muscle weakness, gait disturbance, and back pain	3	3	10
28	Endocrine and metabolic function tests	Endocrine and metabolic function tests	1	1	1
29	Basic sutures and suture removal	Position change and transfer, surgical handwashing, gowning techniques in the operating room, and basic sutures and suture removal	1	1	4
30	Weight loss and skeletal muscle paralysis or muscle weakness	Blood glucose test, weight loss, and skeletal muscle paralysis or muscle weakness	2	2	3
31	Ultrasound and endoscopy	Ultrasound and endoscopy	2	2	2

Record index	Predicted item	Actual item	Number of matches	Number of experienced items extracted by GPT-4-turbo	Number of items that the students marked as experiences they had
32	Basic sutures and suture removal	Skin antisepsis, aseptic technique, surgical handwashing, gowning techniques in the operating room, basic sutures and suture removal, full blood count, blood biochemistry, coagulation or fibrinolysis, immunoserology tests, histopathology and cytology (including intraoperative rapid diagnosis), ultrasound, x-ray, and headache	1	1	13
33	Skin antisepsis and position change and transfer	Skin antisepsis and position change and transfer	2	2	2
34	Back pain	Weight loss, cognitive dysfunction, skeletal muscle paralysis or muscle weakness, sensory disturbance, and back pain	1	1	5
35	Arterial blood gas analysis, peripheral venous catheterization, insertion and extraction of a nasogastric tube, insertion and extraction of a urinary catheter, aseptic technique, surgical handwashing, gowning techniques in the operating room, and basic sutures and suture removal	Peripheral venous catheterization, aseptic technique, full blood count, blood biochemistry, coagulation or fibrinolysis, arterial blood gas analysis, pleural fluid analysis, ultrasound, x-ray, CT, and endoscopy	3	8	11
36	Weight gain, endocrine and metabolic function tests, and blood glucose test	Blood glucose test, full blood count, blood biochemistry, urinalysis, stool (fecal) examination, endocrine and metabolic function tests, ultrasound, CT, and weight gain	3	3	9
37	Endoscopy	Endoscopy	1	1	1
38	X-ray	X-ray and cough and sputum production	1	1	2
39	Abdominal pain	ID not found	0	1	1
40	Skin antisepsis	Skin antisepsis	1	1	1

[a]MRI: magnetic resonance imaging.

[b]FAST: focused assessment with sonography for trauma.

[c]CT: computed tomography.

The predicted items were experience items extracted using GPT-4-turbo from the students' practice records. The actual items were those that the students marked as experiences they had during that period. The English-translated version of the students' records used by GPT-4-turbo to extract experiences can be found in Multimedia Appendix 1, with the "Index" column in Table 1 corresponding to the "Index" column in Multimedia Appendix 1.

Agreement Between LLM Predictions and Student-Reported Experiences

The Jaccard index was 0.59 (95% CI 0.46-0.71), indicating moderate agreement, and the Cohen κ was 0.65 (95% CI 0.53-0.76), indicating substantial agreement. Sensitivity and specificity were 62.39% (95% CI 49.96%-74.81%) and 99.34% (95% CI 98.77%-99.92%), respectively. The sensitivity and specificity of the LLM for each category were as follows: 45.43% (95% CI 25.12%-65.75%) and 98.75% (95% CI 97.31%-100%) for symptoms, 46.76% (95% CI 25.67%-67.86%) and 98.84% (95% CI 97.81%-99.87%) for examinations, and 56.36% (95% CI 37.64%-75.08%) and 98.92% (95% CI 96.67%-100%) for procedures, respectively. There was no significant variation among the categories. However, when calculating by category, the sensitivity tended to be lower than the overall calculation, likely due to the influence of items that were not extracted at all. The correlation between the number of characters in the students' records and sensitivity and specificity was 0.04 and –0.64, respectively, indicating a negligible correlation with sensitivity and a moderate negative correlation with specificity. The correlation coefficients for the Jaccard index and the Cohen κ were 0.06 and –0.07, respectively, showing negligible correlations with record length.

Patterns of Missed Experiences

There were several patterns in experiences that were not captured by GPT-4-turbo's analysis even though students considered to have had those experiences. In this paragraph, we explain these patterns with examples corresponding to specific entries in Table 1. Due to the large volume of student records, the full texts are provided in Multimedia Appendix 2 rather than Table 1. One pattern was when predictable experiences were not picked up by GPT-4-turbo's analysis. For example, a student (index 19 in Table 1) described encountering a case of hereditary amyotrophic lateral sclerosis, but GPT-4-turbo's analysis failed to capture the student's experience with muscle weakness, a symptom of amyotrophic lateral sclerosis. Another pattern was when insufficient description made prediction difficult. In total, 20% (8/40) of the students (indexes 9, 11, 15, 17, 22, 25, 29, and 32 in Table 1) recorded observing surgery, but it was unclear from the description whether they assisted in the surgery or merely observed, making it difficult for GPT-4-turbo to extract related procedures such as surgical handwashing and gowning techniques. A third pattern was when experiences were not recorded by the students, making prediction impossible. For instance, a student recorded observing a surgery (index 15 in Table 1) but actually performed suturing, an experience not captured by GPT-4-turbo due to lack of record. Similarly, a student (index 30 in Table 1) noted examining a patient with diabetes but did not record performing computed tomography or ultrasound examinations.

Discussion

Principal Findings

In this study, we analyzed the records kept by medical students during their clinical clerkship for learning purposes using GPT-4-turbo to predict the clinical procedures they experienced. The experiences extracted by GPT-4-turbo were evaluated for accuracy after being revised by the medical students. The extraction of experiences by GPT-4-turbo showed a sufficient level of agreement with the items that students actually experienced and demonstrated high specificity. The high specificity suggests that the extracted experiences likely mirror what the students actually encountered. However, the low sensitivity indicates that some experiences that students actually had were not captured by GPT-4-turbo's analysis of the records. There were three main reasons why certain experiences could not be extracted: (1) experiences that could have been predicted by GPT-4-turbo's analysis were not identified; (2) the descriptions were insufficient, making prediction difficult; and (3) there were experiences that students did not record at all.

Implications of Findings

The results of this study suggest that LLMs such as GPT-4-turbo are able to extract experiences from learning records with sufficient accuracy. On the other hand, when the content of the learning records is insufficient or when students do not record their experiences, experience extraction becomes difficult, indicating that improving the accuracy of LLMs alone may not be sufficient.

Comparison to the Literature

Comparison with previous studies suggests that LLMs are making it easier and more accurate to extract experiences from learning records. Unlike previous studies [10,11], which required extensive pretraining on large text datasets, this study was able to extract experiences from learning records using only prompt engineering without additional training. A related study using LLMs investigated how well GPT-3 could extract predefined codes from documents and compared its results to those of human coders [13]. In that study, providing 5 examples for each code resulted in a Cohen κ of 0.61 for some codes, although the Cohen κ for most codes was lower. In contrast, our study used GPT-4-turbo to extract experiences from learning records without providing specific examples, achieving a Cohen κ of 0.65. Although direct comparison is difficult due to differences in study targets, GPT-4-turbo may have achieved higher extraction accuracy. These findings indicate that, with the advent and evolution of LLMs, extracting experiences from learning records is becoming easier and potentially more accurate.

In addition, this study demonstrates that performance monitoring is possible by analyzing narrative records primarily intended for student learning using LLMs rather than aggregating list-based records mainly used for evaluation, as in some previous studies. Previous studies have explored the use of logbooks to monitor learners' progress of learning. Attempts have been made to monitor skills and experiences using logbooks [16], track the progress of entrustable professional activities [8], and count the cases encountered [7]. However, the "logbooks" used in these studies were lists of cases experienced or evaluations rather than detailed descriptions of experiences [7,8]. This format is more useful for evaluation purposes rather than for recording learning, which ultimately adds to the burden on learners. Our study suggests that analyzing reflections purely recorded for learning purposes can also extract

experiences, offering a technique that monitors learning situations while reducing the burden on learners and educators.

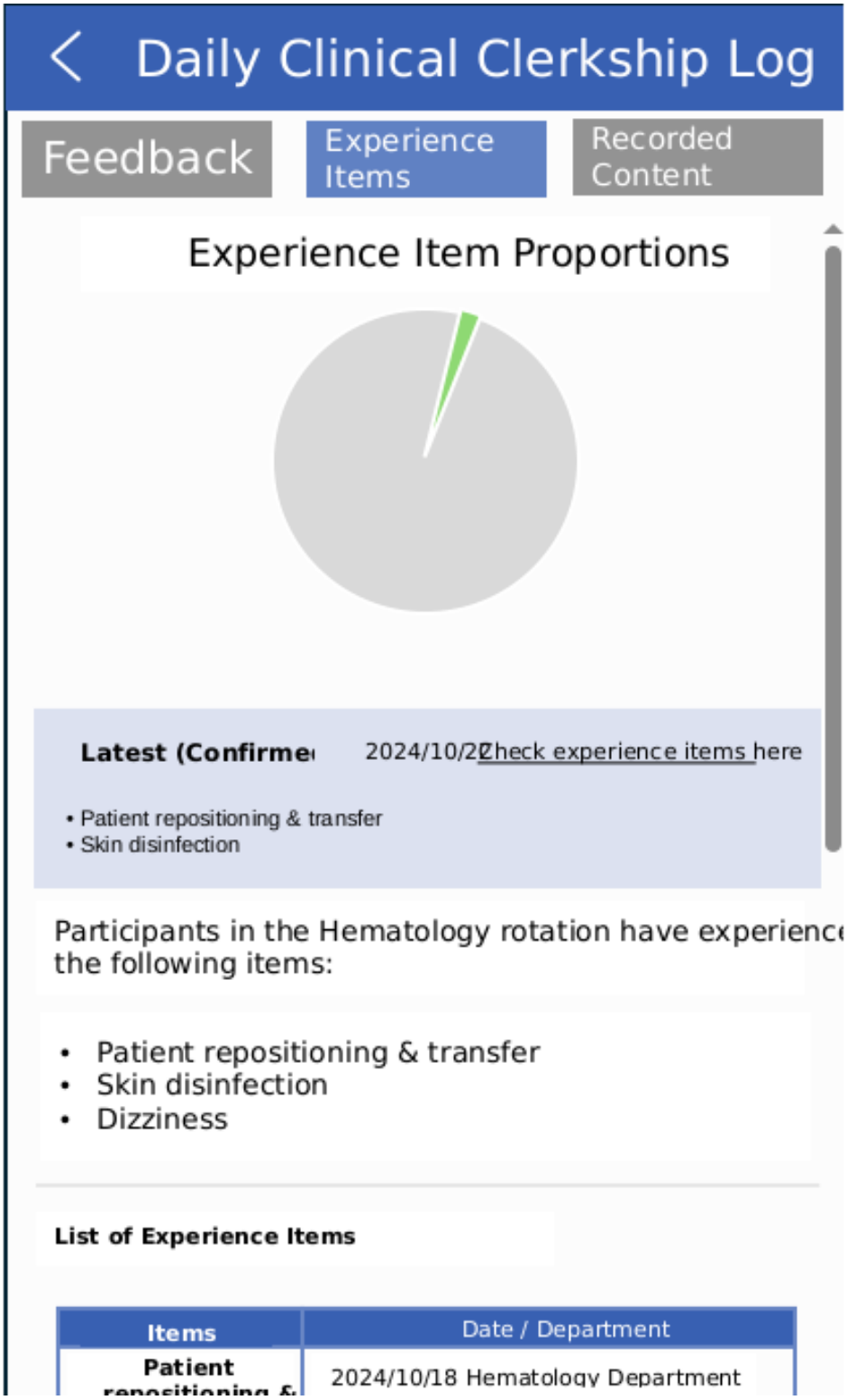
Future Directions

While this study demonstrated the usefulness of experience extraction by LLMs, it also highlighted new challenges. GPT-4-turbo failed to extract some experiences that students actually had. The first pattern involved experience items that could not be extracted despite being predictable from the learning log content. The second pattern involved cases in which descriptions were ambiguous, making inference difficult. The third pattern involved experiences that medical students believed they had but were not recorded in learning logs, making inference impossible. Regarding the first pattern, insufficient reasoning ability of GPT-4-turbo is considered the cause. However, the reasoning ability of LLMs is improving with model evolution [17], and future improvements in LLM accuracy may partially address this issue. Regarding the second and third patterns, insufficient content in medical students' learning logs appears to be the cause, resulting in inadequate information to infer students' experiences. To address these challenges, it may be necessary to enrich students' learning records or extract experiences from other sources.

To enhance the quality of medical students' learning records, providing feedback using the list of experiences extracted by LLMs may be beneficial. Previous studies have shown that logbooks are useful for performance monitoring and improving educational quality, but they have also pointed out that the quality of the records is often insufficient and that feedback is needed [18]. In this study, medical students reviewed the list of experiences extracted from their learning records by GPT-4-turbo and added items that they had actually experienced but were not extracted. Since missing or incomplete records can be a reason for experiences not being extracted, this review process may serve as feedback for students, helping them reflect on what they failed to document in their records. As shown in the development version of the web application in Figure 1, displaying experience items extracted from learning logs might motivate students to improve their learning log documentation.

Combining other data such as electronic health records written by the students might be effective for more accurate monitoring of medical students' performance. Feeding both learning logs and electronic health record descriptions into GPT-4-turbo could enhance the accuracy of experience extraction. Such an approach could lead to more accurate assessment of medical students without increasing the burden on students or faculty. However, since many LLMs, including GPT-4-turbo, are cloud-based, privacy concerns may arise [19,20]. Therefore, new approaches will need to be developed to address these privacy issues in the future.

Figure 1. Development version of the web application.



Limitations

This study has several limitations. First, this study used learning log data from clinical participation-based clerkships at a single university; therefore, its generalizability to learning log data from other universities or clinical clerkships is not guaranteed. In addition, the data collection period was limited to 1 month, which may not capture the full range of experiences or seasonal variations in clinical activities. While the accuracy of the extracted experience content was evaluated by using learning log data recorded by medical students and asking them to make corrections, the quality and quantity of the learning log data recorded by the students could affect the accuracy of the extracted experience content. Large-scale collaborative studies across multiple institutions and over longer periods are needed to ensure broader generalizability. Furthermore, this study used a list of symptoms, examinations, and procedures in the MCC as a template for extracting experience content; however, the

results of using other templates were not examined. Future research is needed to assess performance using other evaluation criteria. Although we confirmed the correlation between record length and extraction sensitivity and specificity, we did not quantitatively evaluate the quality of the records. Future work should investigate the relationship between record quality and extraction performance. In this study, the accuracy of the extracted experience content was evaluated by using learning log data recorded by medical students and asking them to make corrections, but no strict criteria were set for what constitutes "experience" when students made corrections. Moreover, students' judgments about whether they had actually experienced a procedure are subjective, and they may have overreported certain experiences or overlooked ones they truly had. In the clinical clerkship that served as this study's setting, supervising physicians did not continuously monitor students, so only the students themselves could verify their experiences. Therefore, we had to rely on students' subjective reports. In future work, it will be desirable to establish more objective evaluation criteria to reduce potential bias.

Conclusions

In this study, records kept by medical students for learning during clinical clerkships were analyzed using GPT-4-turbo to predict experienced clinical activities. The high specificity of the GPT-4-turbo predictions suggests that the extracted experiences are likely what students actually encountered. However, the low sensitivity indicates that some actual student experiences were not captured by the GPT-4-turbo analysis. Future improvements in AI model performance, providing feedback to medical students on their records and combining learning logs with other data sources such as electronic medical records, may enhance accuracy. Analyzing records using AI may enable detailed assessments while avoiding excessive burdens on learners and educators.

Acknowledgments

ChatGPT (OpenAI) was used in part to create an initial English translation of the Japanese version of this manuscript. This work was supported by Japan Society for the Promotion of Sciences Grants-in-Aid for Scientific Research 23K27816 and 25K06542.

Authors' Contributions

TK was responsible for study planning, data collection and analysis, and manuscript writing. HN collaborated with TK on study planning and provided supervision and advice on data analysis and manuscript writing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The prompts for the OpenAI application programming interface (API) and GitHub repository include an experience extraction API, prompt used in API, R code to analyze the data, and the data themselves.
[DOCX File, 17 KB - mededu_v11i1e68697_app1.docx]

Multimedia Appendix 2
Student records.
[XLSX File, 23 KB - mededu_v11i1e68697_app2.xlsx]

References

1. AlHaqwi AI, Taha WS. Promoting excellence in teaching and learning in clinical education. J Taibah Univ Med Sci 2015 Mar;10(1):97-101. [doi: 10.1016/j.jtumed.2015.02.005]
2. Vanka A, Hovaguimian A. Teaching strategies for the clinical environment. Clin Teach 2019 Dec;16(6):570-574. [doi: 10.1111/tct.12928] [Medline: 30178546]
3. Liu C. An introduction to workplace-based assessments. Gastroenterol Hepatol Bed Bench 2012;5(1):24-28. [Medline: 24834194]
4. Ott MC, Pack R, Cristancho S, Chin M, Van Koughnett JA, Ott M. "The Most Crushing Thing": understanding resident assessment burden in a competency-based curriculum. J Grad Med Educ 2022 Oct;14(5):583-592. [doi: 10.4300/JGME-D-22-00050.1] [Medline: 36274774]
5. Szulewski A, Braund H, Dagnone DJ, et al. The assessment burden in competency-based medical education: how programs are adapting. Acad Med 2023 Nov 1;98(11):1261-1267. [doi: 10.1097/ACM.0000000000005305] [Medline: 37343164]
6. Alotaibi HM, Alharithy R, Alotaibi HM. Importance of the reflective logbook in improving the residents' perception of reflective learning in the dermatology residency program in Saudi Arabia: findings from a cross-sectional study. BMC Med Educ 2022 Dec 13;22(1):862. [doi: 10.1186/s12909-022-03948-w] [Medline: 36514091]

7. Alabbad J, Abdul Raheem F, Almusaileem A, Almusaileem S, Alsaddah S, Almubarak A. Medical students' logbook case loads do not predict final exam scores in surgery clerkship. Adv Med Educ Pract 2018;9:259-265. [doi: 10.2147/AMEP.S160514] [Medline: 29713211]
8. Berberat PO, Rotthoff T, Baerwald C, et al. Entrustable professional activities in final year undergraduate medical training - advancement of the final year training logbook in Germany. GMS J Med Educ 2019;36(6):Doc70. [doi: 10.3205/zma001278] [Medline: 31844642]
9. AbdulAzeem Abdullah Omer A. Using logbooks to enhance students' learning: lessons from a mixed-methods study in an undergraduate surgical rotation. Sudan J Med Sci 2021;16(3):409-429. [doi: 10.18502/sjms.v16i3.9701]
10. Gin BC, Ten Cate O, O'Sullivan PS, Hauer KE, Boscardin C. Exploring how feedback reflects entrustment decisions using artificial intelligence. Med Educ 2022 Mar;56(3):303-311. [doi: 10.1111/medu.14696] [Medline: 34773415]
11. Millán E, Loboda T, Pérez-de-la-Cruz JL. Bayesian networks for student model engineering. Comput Educ 2010 Dec;55(4):1663-1683. [doi: 10.1016/j.compedu.2010.07.010]
12. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887. [doi: 10.3390/healthcare11060887] [Medline: 36981544]
13. Xiao Z, Yuan X, Liao QV, Abdelghani R, Oudeyer PY. Supporting qualitative analysis with large language models: combining codebook with GPT-3 for deductive coding. Presented at: Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23); Mar 27-31, 2023; Sydney, Australia. [doi: 10.1145/3581754.3584136]
14. The Model Core Curriculum for Medical Education (2022 revision). Medical Education Model Core Curriculum Expert Research Committee. 2022. URL: http://jsme.umin.ac.jp/eng/activities/index.html [accessed 2025-10-03]
15. Nishigori H. Medical education in Japan. Med Teach 2024 Jun 4;46(sup1):S4-S10. [doi: 10.1080/0142159X.2024.2372108]
16. Levine RB, Kern DE, Wright SM. The impact of prompted narrative writing during internship on reflective practice: a qualitative study. Adv Health Sci Educ Theory Pract 2008 Dec;13(5):723-733. [doi: 10.1007/s10459-007-9079-x] [Medline: 17899421]
17. Kosinski M. Evaluating large language models in theory of mind tasks. Proc Natl Acad Sci U S A 2024 Nov 5;121(45):e2405460121. [doi: 10.1073/pnas.2405460121] [Medline: 39471222]
18. Paydar S, Esmaeeli E, Ameri F, Sabahi A, Meraji M. Investigating the advantages and disadvantages of electronic logbooks for education goals promotion in medical sciences students: a systematic review. Health Sci Rep 2023 Dec;6(12):e1776. [doi: 10.1002/hsr2.1776] [Medline: 38125281]
19. Samsi S, Zhao D, McDonald J, et al. From words to watts: benchmarking the energy costs of large language model inference. Presented at: Proceedings of the 2023 IEEE High Performance Extreme Computing Conference (HPEC '23); Sep 25-29, 2023; Boston, MA. [doi: 10.1109/HPEC58863.2023.10363447]
20. Madaan A, Aggarwal P, Anand A, et al. AutoMix: automatically mixing language models. arXiv. Preprint posted online on Oct 19, 2023. [doi: 10.48550/arXiv.2310.12963]

Abbreviations

AI: artificial intelligence
LLM: large language model
MCC: Model Core Curriculum for Medical Education

Edited by B Lesselroth; submitted 17.11.24; peer-reviewed by F Shojaei, V Mavrych, Y Asada; revised version received 29.07.25; accepted 22.09.25; published 15.10.25.

Please cite as:
Kondo T, Nishigori H
AI's Accuracy in Extracting Learning Experiences From Clinical Practice Logs: Observational Study
JMIR Med Educ 2025;11:e68697
URL: https://mededu.jmir.org/2025/1/e68697
doi:10.2196/68697

ChatGPT's Performance on Portuguese Medical Examination Questions: Comparative Analysis of ChatGPT-3.5 Turbo and ChatGPT-4o Mini

Filipe Prazeres[1,2,3], MD, MSc, PhD

[1]Faculty of Health Sciences, University of Beira Interior, Av. Infante D. Henrique, Covilhã, Portugal

[2]Family Health Unit Beira Ria, Gafanha da Nazaré, Portugal

[3]CINTESIS@RISE, Department of Community Medicine, Information and Health Decision Sciences, Faculty of Medicine of the University of Porto, Porto, Portugal

Corresponding Author:
Filipe Prazeres, MD, MSc, PhD
Faculty of Health Sciences, University of Beira Interior, Av. Infante D. Henrique, Covilhã, Portugal

Abstract

Background: Advancements in ChatGPT are transforming medical education by providing new tools for assessment and learning, potentially enhancing evaluations for doctors and improving instructional effectiveness.

Objective: This study evaluates the performance and consistency of ChatGPT-3.5 Turbo and ChatGPT-4o mini in solving European Portuguese medical examination questions (2023 National Examination for Access to Specialized Training; Prova Nacional de Acesso à Formação Especializada [PNA]) and compares their performance to human candidates.

Methods: ChatGPT-3.5 Turbo was tested on the first part of the examination (74 questions) on July 18, 2024, and ChatGPT-4o mini on the second part (74 questions) on July 19, 2024. Each model generated an answer using its natural language processing capabilities. To test consistency, each model was asked, "Are you sure?" after providing an answer. Differences between the first and second responses of each model were analyzed using the McNemar test with continuity correction. A single-parameter *t* test compared the models' performance to human candidates. Frequencies and percentages were used for categorical variables, and means and CIs for numerical variables. Statistical significance was set at $P<.05$.

Results: ChatGPT-4o mini achieved an accuracy rate of 65% (48/74) on the 2023 PNA examination, surpassing ChatGPT-3.5 Turbo. ChatGPT-4o mini outperformed medical candidates, while ChatGPT-3.5 Turbo had a more moderate performance.

Conclusions: This study highlights the advancements and potential of ChatGPT models in medical education, emphasizing the need for careful implementation with teacher oversight and further research.

(JMIR Med Educ 2025;11:e65108) doi:10.2196/65108

KEYWORDS

ChatGPT-3.5 Turbo; ChatGPT-4o mini; medical examination; European Portuguese; AI performance evaluation; Portuguese; evaluation; medical examination questions; examination question; chatbot; ChatGPT; model; artificial intelligence; AI; GPT; LLM; NLP; natural language processing; machine learning; large language model

Introduction

Generative artificial intelligence (AI) represents a branch of AI dedicated to the development of systems that can autonomously generate high-quality digital content on demand, and it can do so across various modalities, such as written text, images, audio, and video [1-3]. Generative AI tools are trained on large datasets, enabling them to produce work that mirrors human-created content [2]. Nowadays, there are several examples of generative AI tools, including ChatGPT (OpenAI Inc), Runway, Gemini (Google Inc), DALL-E (OpenAI Inc), Copilot (Microsoft Inc), Midjourney, NovelAI (Anlatan), Claude (Anthropic), and Jasper AI, among others. ChatGPT, the large language model (LLM) chatbot, developed by OpenAI [4], that started the AI boom in November 2022, became the most popular AI tool of 2023, accounting for over 60.2% of visits between September 2022 and August 2023, with a total of 14.6 billion website visits [5]. ChatGPT's availability as a free-to-use, low-bandwidth service may reduce disparities compared to paid versions or models by making advanced AI technology accessible to a broader and more diverse global population [6], contributing to making it the most popular generative AI tool [7].

Recent literature reviews regarding AI have shown that this type of technology has potential applications in several fields, spanning from the architecture, engineering, and construction industry to health care [8-11]. The possible applications in

medicine are substantial, ranging from diagnostic and treatment support (eg, clinical imaging improvement, classification of diseases, prediction of disease onset, development of treatment, and medication prescriptions) [12] to facilitate communication and engagement between medical professionals and their patients [13], and also improving medical education and its accessibility [10,14,15]. For example, ChatGPT can be used as a study tool to clearly explain complex medical concepts [16,17] (eg, radiology reports [18]), create memory aids for challenging topics, clarify medical practice questions, summarize research articles, compile lists of differential diagnoses [17], generate medical examination questions [19], and simulate physician-patient interactions [14].

Medical written examinations are an important part in evaluating the competence and knowledge of medical students and graduates (eg, access of physicians to specialized training, such is the case in Portugal). These examinations not only test factual knowledge but also evaluate the critical thinking and problem-solving skills of the candidates. With the recent growing interest in AI, an important question arises: Can AI, specifically ChatGPT, perform at a level comparable to human candidates in medical written examinations? By evaluating ChatGPT's ability to correctly answer medical questions, its medical proficiency and its potential role as an educational tool can be assessed. Successfully completing this task can demonstrate ChatGPT's capability to serve as a resource for medical students by providing continuous access to information, particularly benefiting students in remote or under-resourced areas [6].

ChatGPT is known for having the capability of performing near the passing threshold of 60% accuracy of the United States Medical Licensing Examination (USMLE) [20] and for approximately having the knowledge equivalent to a third-year medical student [21]. ChatGPT's performance on medical examinations has been analyzed across different countries and questions. A 2023 systematic review with a meta-analysis of 19 articles found a mean performance of ChatGPT of around 61% [22], and a more recent review published in 2024 concluded that, despite ChatGPT's satisfactory performance in examinations, further studies are necessary to fully explore its potential in medical education [23].

Furthermore, ChatGPT struggles with non-English language assessments possibly due to a limited understanding of linguistic nuances and Western-centric internet data, which may not fully represent the clinical and disease differences in some countries, like African and Asian populations [24], warranting more studies in other languages to ensure better understanding of ChatGPT's accuracy in diverse cultural contexts. For example, ChatGPT performed considerably lower on a medical examination in Chinese (45.8% correct answers on the Chinese National Medical Licensing Examination) [25], and even worse in the French examination with 22% correct answers [26].

In July 2024, OpenAI launched GPT-4o mini, a smaller version of its latest GPT-4o ("o" for "omni") AI language model. This new model replaced GPT-3.5 Turbo in ChatGPT, making this an ideal time to study the performance of both free models in resolving written medical examinations.

This study aims to evaluate the performance and consistency of 2 AI models, ChatGPT-3.5 Turbo and ChatGPT-4o mini, in solving the questions of a non-English language (European Portuguese) written medical examination, with a format of multiple-choice with one best answer—the 2023 National Examination for Access to Specialized Training (Prova Nacional de Acesso à Formação Especializada [PNA])—and compare their performance to that of human candidates.

Methods

Study Design

The PNA examination is part of the requirements for entering specialized medical training in Portugal. Its purpose is to rank candidates for accessing specialized training vacancies, so no minimum passing grade is needed [27].

The PNA questions used in this study were from the actual 2023 Portuguese PNA examination, which is publicly available on the web [27]. This examination includes 150 questions with 5 multiple-choice answers each, with only a single best answer, similar to the USMLE. The questions are based on clinical vignettes and divided into 2 parts with 75 questions each. The examination emphasizes clinical reasoning and the application and integration of clinical knowledge and is scored on a scale from 0 to 150 points, with no penalties for blank or incorrect answers. It covers various medical disciplines, including medicine, surgery, pediatrics, gynecology and obstetrics, and psychiatry. The examination duration is 240 minutes, divided into 2 parts of 120 minutes each [27].

ChatGPT-3.5 Turbo was provided with the first part of the examination (74 no image-based multiple-choice questions [MCQs]) on July 18, 2024, and ChatGPT-4o mini with the second part of the examination (74 no image-based MCQs) on July 19, 2024. The questions were entered into the models in European Portuguese and in a format similar to how they are presented to human candidates, and each model was requested to provide a single-letter answer, just like human candidates. For each question, the models generated an answer using their natural language processing capabilities. Following each model's response, a follow-up question, "Are you sure?" was asked to test for consistency—this technique was previously used by Brin et al [28]. An example of the input format of the questions and the respective responses by ChatGPT in European Portuguese is depicted in Table 1, with corresponding translations to English performed by ChatGPT-4o mini. Each question was addressed in a new chat session to reduce the potential influence of memory retention bias of ChatGPT.

Table . Example of the input format of the questions and the respective responses by ChatGPT.

	Examination question in European Portuguese	Examination question translated to English (performed by ChatGPT-4o mini)
Question	Um homem de 73 anos vem à consulta hospitalar para reavaliação de doença pulmonar obstrutiva crónica. Na consulta prévia, há seis meses, apresentava-se em estadio GOLD B. Refere agora, desde há três meses, agravamento da dispneia para esforços médios, sem alteração do padrão habitual de tosse ou de expetoração. Nega febre, perda de apetite ou outras queixas de novo. A história médica revela ainda obesidade. A medicação habitual inclui brometo de tiotrópio e salmeterol. É ex-fumador de 40 UMA desde há 10 anos. Os sinais vitais são temperatura 36 °C, frequência respiratória 18/min, frequência cardíaca 78/min e pressão arterial 115/89 mm Hg; SpO2 94% (ar ambiente). Ao exame físico apresenta cianose labial, com aparência confortável e atrofia muscular na área temporal. A auscultação pulmonar revela crepitações raras dispersas, sem prolongamento do tempo expiratório. O restante exame físico, incluindo a auscultação cardíaca, encontra-se dentro dos parâmetros de normalidade. Qual das seguintes alternativas mais provavelmente reduz o risco de morte relacionada com esta doença? A. Azitromicina. B. Montelucaste. C. Oxigenoterapia. D. Prednisolona. E. Teofilina. Indique apenas a letra da resposta, sem justificação.	A 73-year-old man comes to the hospital for reevaluation of chronic obstructive pulmonary disease (COPD). At the previous visit six months ago, he was in GOLD stage B. He now reports worsening shortness of breath with moderate exertion over the past three months, without changes in his usual cough or sputum production. He denies fever, loss of appetite, or any new complaints. His medical history also reveals obesity. His usual medications include tiotropium bromide and salmeterol. He is a former smoker with a 40 pack-year history, quit 10 years ago. Vital signs are temperature 36 °C, respiratory rate 18/min, heart rate 78/min, and blood pressure 115/89 mm Hg; SpO2 94% (room air). On physical examination, he shows lip cyanosis, appears comfortable, and has muscle atrophy in the temporal area. Lung auscultation reveals scattered rare crackles, with no prolonged expiratory phase. The rest of the physical examination, including cardiac auscultation, is within normal limits. Which of the following is most likely to reduce the risk of death related to this condition? A. Azithromycin. B. Montelukast. C. Oxygen therapy. D. Prednisolone. E. Theophylline. Please indicate only the letter of the answer, without justification.
ChatGPT-3.5 Turbo answer	A resposta correta é: (C) Oxigenoterapia.	The correct answer is: (C) Oxygen therapy.
Follow-up question	Tem a certeza?	Are you sure?
ChatGPT-3.5 Turbo answer	Peço desculpa pelo erro anterior. Vamos rever a situação: Dadas as características do paciente com DPOC grave e hipoxemia (SpO2 de 94% em ar ambiente), a intervenção que mais provavelmente reduzirá o risco de morte relacionada com esta doença é a oxigenoterapia. Portanto, a resposta correta é: (C) Oxigenoterapia.	I apologize for the previous error. Let's review the situation: Given the characteristics of the patient with severe COPD and hypoxemia (SpO2 of 94% on room air), the intervention that is most likely to reduce the risk of death related to this condition is oxygen therapy. Therefore, the correct answer is: (C) Oxygen therapy.

Human candidates' scores were obtained from the official records of the PNA 2023 examination, which are also publicly available on the web [27].

To compare the mean scores of human candidates to those of the language models, all questions of the PNA 2023 examination had to be answered. Since the examination included 2 questions using images (one in the first part and another one in the second part; both with electrocardiogram strips), these questions were answered by GPT-4o, as it can handle images in addition to text.

Ethical Considerations

This study exclusively used data that had been previously published online and did not involve direct interaction with human participants. As a result, ethical guidelines pertaining to human participants are not applicable.

Statistical Analysis

Analyses were performed using IBM SPSS Statistics (Version 21). The McNemar test [29] with continuity correction [30] was used to determine differences between the first and second responses of ChatGPT-3.5 Turbo and ChatGPT-4o mini. Single-parameter *t* test was used to compare the performance of ChatGPT-3.5 Turbo and ChatGPT-4o mini with that of human candidates. Frequencies and percentages were used for

categorical variables and means and CIs for numerical variables. Statistical significance was considered at $P<.05$.

Results

Overall Performance and Consistency

In the initial response with ChatGPT-3.5 Turbo, of the 74 questions, 40 (54%) answers were correct and 34 (46%) answers were incorrect. After the follow-up question, "Are you sure?," the number of correct answers decreased to 28 (38%), while the number of incorrect answers increased to 46 (62%). This change occurred because ChatGPT-3.5 Turbo corrected 12 originally incorrect answers, but also changed 24 originally correct answers to incorrect. This pattern of change approached, but did not reach, significance ($\chi^2_1=3.361$, $P=.067$).

Initially, of the 74 questions, ChatGPT-4o mini produced 48 (65%) correct answers and 26 (35%) incorrect answers. After being asked, "Are you sure?," the correct answers dropped to 42 (57%), while incorrect answers rose to 32 (43%). This change occurred because ChatGPT-4o mini fixed 12 previously wrong answers but also changed 18 previously correct answers to incorrect. This pattern of change was not statistically significant ($\chi^2_1=0.833$, $P=.361$).

The 2 questions using images (one in the first part and another one in the second part) were answered correctly by GPT-4o.

LLM Chatbot Versus Human

When evaluating AI capabilities in relation to human abilities, LLM responses in part 1 of PNA (74 questions resolved by ChatGPT-3.5 Turbo plus 1 by GPT-4o) showed lower accuracy than human respondents. The human mean score was statistically significantly higher by 6.04 (95% CI 5.65-6.43) than the LLM score of 41 ($P<.001$).

In part 2 of PNA (74 questions resolved by ChatGPT-4o mini added to 1 question by GPT-4o), the LLM score showed higher accuracy than human respondents. The human mean score was statistically significantly lower by 5.58 (95% CI 5.25-5.9) than the LLM score of 49 ($P<.001$).

Discussion

Principal Findings

This study analyzes the performance of 2 ChatGPT models (ChatGPT-3.5 Turbo and ChatGPT-4o mini) on the Portuguese medical written examination: 2023 National Examination for Access to Specialized Training, revealing important differences in accuracy and consistency. Although, both ChatGPT-3.5 Turbo and ChatGPT-4o mini answered correctly in the majority of the questions, ChatGPT-4o mini achieved a higher accuracy rate of 65% (48/74) compared to ChatGPT-3.5 Turbo's 54% (40/74), demonstrating a superior capability in handling medical questions. Additionally, ChatGPT-4o mini showed greater consistency in confirming answers, highlighting its reliability. When evaluated against human respondents, ChatGPT-4o mini outperformed the average human accuracy, while ChatGPT-3.5 Turbo fell short.

Strengths

This study stands out for its innovative approach in analyzing the performance of ChatGPT-3.5 Turbo and ChatGPT-4o mini in a medical examination context. It is the first to evaluate these models using an examination conducted in a less commonly studied language, Portuguese, thereby broadening the scope of language-specific AI assessments. By incorporating the actual scores of human candidates for comparison, the study provides a robust benchmark against real-world performance. Furthermore, the research examines the stability of the AI's answers by repeatedly asking "Are you sure?," offering valuable insights into the consistency of the responses.

Comparison to Prior Work

A recent study evaluated ChatGPT's performance on medical licensing examinations across multiple countries (United States, Italy, France, Spain, United Kingdom, and India) and determined a variable accuracy, ranging from 22% on the French examination to 73% on the Italian examination [26]. In this study, ChatGPT answered correctly in more than 50% of the Portuguese medical examination questions, positioning it next to the countries with better performance. For example, in a Turkish study, ChatGPT reached 70.9% accuracy in the medical specialty examination [31]. In the Iranian medical licensing examination, ChatGPT performed with 68.5% of the questions answered correctly [32]. And in Poland, ChatGPT achieved a 67.1% correct response rate on the Polish medical specialization licensing examination [33].

When analyzing the differences between the 2 ChatGPT versions, ChatGPT-4o mini outperformed ChatGPT-3.5 Turbo in this study: 65% (48/74) vs 54% (40/74) correct response rate. This suggests that advancements in the underlying architecture and training data of ChatGPT-4o mini (knowledge up to October 2023) have improved its capability to understand and respond to medical questions with more accuracy. Previous studies evaluating the performance of different ChatGPT models found that ChatGPT-4 consistently performed better compared to ChatGPT-3.5. For example, ChatGPT-4 outperformed ChatGPT-3.5 on the Polish Medical Final Examination [34], the Spanish Medical Residency Entrance Examination (Médico Interno Residente) [35], the 2023 Japanese Nursing Examination [36], the Peruvian National Licensing Medical Examination (Examen Nacional de Medicina) [37], and in the USMLE soft skill assessments [28], to name a few. Nonetheless, ChatGPT-4 is a paid model and thus not accessible to everyone, which is not the case for the most recent free-to-use ChatGPT-4o mini.

Another important aspect is consistency. The results of this study revealed that ChatGPT-3.5 Turbo was less stable when asked to confirm its original answers. These results are consistent with those of Brin et al [28], who found that ChatGPT-3.5 altered its answers 82.5% of the time in the USMLE assessments [28]. Unfortunately, in this study, it was not shown that by changing the original answers, ChatGPT-3.5 Turbo improves its accuracy. This contrasts with studies on human students, which have shown that changing their answers usually improves their test scores [38]. One can wonder, since the "awareness of what one knows and does not know depends in part on how much one knows" [39], does ChatGPT-3.5 Turbo

change its answers because it does not know, or does it simply change answers to satisfy the user when prompted?

When evaluating the AI models against human respondents, it was found that in part 2 of the PNA examination (74 questions resolved by ChatGPT-4o mini plus 1 question by GPT-4o), the LLM outperformed the average accuracy of human participants. In contrast, in part 1 of the PNA examination (74 questions resolved by ChatGPT-3.5 Turbo plus 1 question by GPT-4o), LLM showed lower accuracy than human respondents. This indicates that while earlier versions, like ChatGPT-3.5 Turbo, may have required a high degree of human oversight, more recent and advanced versions, like ChatGPT-4o mini, have the potential to match or exceed human performance in medical domains. Although no previous studies have analyzed the performance of ChatGPT-4o mini, and no direct comparisons can be made, some studies have already noted that LLMs outperformed human candidates in several medical examinationinations (eg, the German Medical State Examinations of 2022 [40], part 1 of the Fellowship of the Royal College of Ophthalmologists MCQ examination [41], and the University of Toronto Family Medicine Residency Progress Test [42]).

Limitations

This study has several limitations regarding the performance evaluation of ChatGPT-3.5 Turbo and ChatGPT-4o mini. The analysis was based solely on ChatGPT's indication of the correct answer, which, while aligning with expectations for human candidates, does not consider other aspects of examination performance. Additionally, the grading did not account for the complexity or length of the questions, providing an incomplete assessment of the models' performance. Further studies should incorporate a more comprehensive evaluation framework that considers the reasoning process and evaluates performance across a broader range of question types and difficulties.

Future Perspectives

This study highlights the importance of continuous improvement in ChatGPT models to further enhance their reliability and accuracy. The superior performance of ChatGPT-4o mini compared to its predecessor offers promising applications in medical education. Its higher accuracy and consistency suggest that it could serve as an effective tool for training medical students. However, a broader assessment of ChatGPT-4o mini across various tests and real-world scenarios is required, as good performance on a specific test may not indicate abilities for general and reliable medical education usage. Additionally, there are known drawbacks and ethical considerations when using AI applications, including the potential for fabricated, incorrect, or biased information [43]. Other issues include limited training periods and the possibility of providing different answers to the same question depending on how the question is phrased [43]. A recent systematic scoping review by Xu et al [44] advises medical students to use ChatGPT cautiously, cross-checking information with reliable sources and disclosing AI-generated content in their work. Teachers should guide students on the effective and ethical use of ChatGPT, assess its reliability, and explore mixed assessment methods to evaluate student abilities while considering its impact on traditional assignments [44].

Conclusion

On the 2023 Portuguese National Examination for Access to Specialized Training, ChatGPT-4o mini achieved an accuracy rate of 65% (48/74), surpassing ChatGPT-3.5 Turbo. This demonstrates a superior capability in handling medical questions. ChatGPT-4o mini outperformed medical candidates, while ChatGPT-3.5 Turbo had a more moderate performance. This study highlights the advancements and potential of ChatGPT models in medical education, emphasizing the importance of careful implementation with teacher oversight and further research.

Acknowledgments

This study was supported by national funds through FCT (Fundação para a Ciência e a Tecnologia, I.P.) within CINTESIS R&D Unit (reference UIDB/4255/2020) and within the scope of the project RISE, Associated Laboratory (reference LA/P/0053/2020). During the preparation of this manuscript, the author used ChatGPT-4o mini in order to improve the language of the manuscript and correct grammatical errors. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

Conflicts of Interest

None declared.

References

1. Feuerriegel S, Hartmann J, Janiesch C, Zschech P. Generative AI. Bus Inf Syst Eng 2024 Feb;66(1):111-126. [doi: 10.1007/s12599-023-00834-7]
2. Ramdurai B, Adhithya P. The impact, advancements and applications of generative AI. Int J Comput Sci Eng 2023;10(6):1-8. [doi: 10.14445/23488387/IJCSE-V10I6P101]
3. Cao Y, Li S, Liu Y, et al. A comprehensive survey of AI-generated content (AIGC): a history of generative AI from GAN to ChatGPT. arXiv. Preprint posted online on Mar 7, 2023. [doi: 10.48550/arXiv.2303.04226]
4. Introducing ChatGPT. OpenAI. 2022 Nov 30. URL: https://openai.com/index/chatgpt [accessed 2024-07-30]

5. Conte N. Ranked: the most popular AI tools. Visual Capitalist. 2024 Jan 24. URL: https://www.visualcapitalist.com/ranked-the-most-popular-ai-tools [accessed 2025-02-19]
6. Wang X, Sanders HM, Liu Y, et al. ChatGPT: promise and challenges for deployment in low- and middle-income countries. Lancet Reg Health West Pac 2023 Dec;41:100905. [doi: 10.1016/j.lanwpc.2023.100905] [Medline: 37731897]
7. Aydin Ö, Karaarslan E. Is ChatGPT leading generative AI? What is beyond expectations? Acad Platform J Eng Smart Sys 2023;11(3):118-134. [doi: 10.21541/apjess.1293702]
8. BuHamdan S, Alwisy A, Bouferguene A. Generative systems in the architecture, engineering and construction industry: a systematic review and analysis. Int J Archit Comput 2021 Sep;19(3):226-249. [doi: 10.1177/1478077120934126]
9. Ali O, Abdelbaki W, Shrestha A, Elbasi E, Alryalat MAA, Dwivedi YK. A systematic literature review of artificial intelligence in the healthcare sector: benefits, challenges, methodologies, and functionalities. J Innov Knowl 2023 Jan;8(1):100333. [doi: 10.1016/j.jik.2023.100333]
10. Younis HA, Eisa TAE, Nasser M, et al. A systematic review and meta-analysis of artificial intelligence tools in medicine and healthcare: applications, considerations, limitations, motivation and challenges. Diagnostics (Basel) 2024 Jan 4;14(1):109. [doi: 10.3390/diagnostics14010109] [Medline: 38201418]
11. Ruksakulpiwat S, Thorngthip S, Niyomyart A, et al. A systematic review of the application of artificial intelligence in nursing care: where are we, and what's next? J Multidiscip Healthc 2024;17:1603-1616. [doi: 10.2147/JMDH.S459946] [Medline: 38628616]
12. Bitkina OV, Park J, Kim HK. Application of artificial intelligence in medical technologies: a systematic review of main trends. Digit Health 2023;9. [doi: 10.1177/20552076231189331] [Medline: 37485326]
13. Sedaghat S. Early applications of ChatGPT in medical practice, education and research. Clin Med (Lond) 2023 May;23(3):278-279. [doi: 10.7861/clinmed.2023-0078] [Medline: 37085182]
14. Gandomani HS. ChatGPT in medical education: how we can use in medical education: challenges and opportunities. J Multidiscip Care 2023;12(1):1-2. [doi: 10.34172/jmdc.1232]
15. Sani I. Enhancing medical education with ChatGPT: a promising tool for the future. Can J Med 2024 Apr 1;6(1):1-4. [doi: 10.33844/cjm.2024.6032]
16. Hosseini M, Gao CA, Liebovitz DM, et al. An exploratory survey about using ChatGPT in education, healthcare, and research. PLoS One 2023;18(10):e0292216. [doi: 10.1371/journal.pone.0292216] [Medline: 37796786]
17. Guo AA, Li J. Harnessing the power of ChatGPT in medical education. Med Teach 2023 Sep;45(9):1063. [doi: 10.1080/0142159X.2023.2198094] [Medline: 37036161]
18. Jeblick K, Schachtner B, Dexl J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol 2024 May;34(5):2817-2825. [doi: 10.1007/s00330-023-10213-1] [Medline: 37794249]
19. Cheung BHH, Lau GKK, Wong GTC, et al. ChatGPT versus human in generating medical graduate exam multiple choice questions—a multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). PLoS One 2023;18(8):e0290691. [doi: 10.1371/journal.pone.0290691] [Medline: 37643186]
20. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
21. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
22. Levin G, Horesh N, Brezinov Y, Meyer R. Performance of ChatGPT in medical examinations: a systematic review and a meta-analysis. BJOG 2024 Feb;131(3):378-380. [doi: 10.1111/1471-0528.17641] [Medline: 37604703]
23. Sumbal A, Sumbal R, Amir A. Can ChatGPT-3.5 pass a medical exam? A systematic review of ChatGPT's performance in academic testing. J Med Educ Curric Dev 2024;11. [doi: 10.1177/23821205241238641] [Medline: 38487300]
24. Cherif H, Moussa C, Missaoui AM, Salouage I, Mokaddem S, Dhahri B. Appraisal of ChatGPT's aptitude for medical education: comparative analysis with third-year medical students in a pulmonology examination. JMIR Med Educ 2024 Jul 23;10:e52818. [doi: 10.2196/52818] [Medline: 39042876]
25. Wang X, Gong Z, Wang G, et al. ChatGPT performs on the Chinese National Medical Licensing Examination. J Med Syst 2023 Aug 15;47(1):86. [doi: 10.1007/s10916-023-01961-0] [Medline: 37581690]
26. Alfertshofer M, Hoch CC, Funk PF, et al. Sailing the seven seas: a multinational comparison of ChatGPT's performance on medical licensing examinations. Ann Biomed Eng 2024 Jun;52(6):1542-1545. [doi: 10.1007/s10439-023-03338-3] [Medline: 37553555]
27. Prova Nacional de Acesso à Formação Especializada 2023 Perguntas Frequentes. ACSS. 2023. URL: https://www.acss.min-saude.pt/wp-content/uploads/2018/09/FAQ_PNA2023.pdf [accessed 2024-08-01]
28. Brin D, Sorin V, Vaid A, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep 2023 Oct 1;13(1):16492. [doi: 10.1038/s41598-023-43436-9] [Medline: 37779171]
29. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 1947 Jun;12(2):153-157. [doi: 10.1007/BF02295996] [Medline: 20254758]

30. Edwards AL. Note on the correction for continuity in testing the significance of the difference between correlated proportions. Psychometrika 1948 Sep;13(3):185-187. [doi: 10.1007/BF02289261] [Medline: 18885738]
31. Oztermeli AD, Oztermeli A. ChatGPT performance in the medical specialty exam: an observational study. Medicine (Baltimore) 2023 Aug 11;102(32):e34673. [doi: 10.1097/MD.0000000000034673] [Medline: 37565917]
32. Ebrahimian M, Behnam B, Ghayebi N, Sobhrakhshankhah E. ChatGPT in Iranian medical licensing examination: evaluating the diagnostic accuracy and decision-making capabilities of an AI-based model. BMJ Health Care Inform 2023 Dec 11;30(1):e100815. [doi: 10.1136/bmjhci-2023-100815] [Medline: 38081765]
33. Wójcik S, Rulkiewicz A, Pruszczyk P, Lisik W, Poboży M, Domienik-Karłowicz J. Reshaping medical education: performance of ChatGPT on a PES medical examination. Cardiol J 2024;31(3):442-450. [doi: 10.5603/cj.97517] [Medline: 37830257]
34. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. Sci Rep 2023 Nov 22;13(1):20512. [doi: 10.1038/s41598-023-46995-z] [Medline: 37993519]
35. Guillen-Grima F, Guillen-Aguinaga S, Guillen-Aguinaga L, et al. Evaluating the efficacy of ChatGPT in navigating the Spanish Medical Residency Entrance Examination (MIR): promising horizons for AI in clinical medicine. Clin Pract 2023 Nov 20;13(6):1460-1487. [doi: 10.3390/clinpract13060130] [Medline: 37987431]
36. Kaneda Y, Takahashi R, Kaneda U, et al. Assessing the performance of GPT-3.5 and GPT-4 on the 2023 Japanese nursing examination. Cureus 2023 Aug;15(8):e42924. [doi: 10.7759/cureus.42924] [Medline: 37667724]
37. Flores-Cohaila JA, García-Vicente A, Vizcarra-Jiménez SF, et al. Performance of ChatGPT on the Peruvian National Licensing Medical Examination: cross-sectional study. JMIR Med Educ 2023 Sep 28;9:e48039. [doi: 10.2196/48039] [Medline: 37768724]
38. Kruger J, Wirtz D, Miller DT. Counterfactual thinking and the first instinct fallacy. J Pers Soc Psychol 2005 May;88(5):725-735. [doi: 10.1037/0022-3514.88.5.725] [Medline: 15898871]
39. Coutinho MVC, Thomas J, Fredricks-Lowman I, Alkaabi S, Couchman JJ. Unskilled and unaware: second-order judgments increase with miscalibration for low performers. Front Psychol 2024;15:1252520. [doi: 10.3389/fpsyg.2024.1252520] [Medline: 38952836]
40. Roos J, Kasapovic A, Jansen T, Kaczmarczyk R. Artificial intelligence in medical education: comparative analysis of ChatGPT, Bing, and medical students in Germany. JMIR Med Educ 2023 Sep 4;9:e46482. [doi: 10.2196/46482] [Medline: 37665620]
41. Fowler T, Pullen S, Birkett L. Performance of ChatGPT and Bard on the official part 1 FRCOphth practice questions. Br J Ophthalmol 2024 Oct;108(10):1379-1383. [doi: 10.1136/bjo-2023-324091]
42. Huang RS, Lu KJQ, Meaney C, Kemppainen J, Punnett A, Leung FH. Assessment of resident and AI chatbot performance on the University of Toronto Family Medicine Residency Progress Test: comparative study. JMIR Med Educ 2023 Sep 19;9:e50514. [doi: 10.2196/50514] [Medline: 37725411]
43. Wong RSY, Ming LC, Raja Ali RA. The intersection of ChatGPT, clinical medicine, and medical education. JMIR Med Educ 2023 Nov 21;9:e47274. [doi: 10.2196/47274] [Medline: 37988149]
44. Xu X, Chen Y, Miao J. Opportunities, challenges, and future directions of large language models, including ChatGPT in medical education: a systematic scoping review. J Educ Eval Health Prof 2024;21(6):6. [doi: 10.3352/jeehp.2024.21.6] [Medline: 38486402]

Abbreviations

AI: artificial intelligence
LLM: large language model
MCQ: multiple-choice question
PNA: Prova Nacional de Acesso à Formação Especializada
USMLE: United States Medical Licensing Examination

Edited by B Lesselroth; submitted 05.08.24; peer-reviewed by B Thies, LA Digiampietri, R Pellegrino; revised version received 30.11.24; accepted 12.12.24; published 05.03.25.

Please cite as:
Prazeres F
ChatGPT's Performance on Portuguese Medical Examination Questions: Comparative Analysis of ChatGPT-3.5 Turbo and ChatGPT-4o Mini
JMIR Med Educ 2025;11:e65108
URL: https://mededu.jmir.org/2025/1/e65108
doi:10.2196/65108

Performance of ChatGPT-4 on Taiwanese Traditional Chinese Medicine Licensing Examinations: Cross-Sectional Study

Liang-Wei Tseng[1], MD; Yi-Chin Lu[2], MD; Liang-Chi Tseng[3], MSc; Yu-Chun Chen[4,5,6], MD, MSc; Hsing-Yu Chen[2,7], MD, PhD

[1]Division of Chinese Acupuncture and Traumatology, Center of Traditional Chinese Medicine, Chang Gung Memorial Hospital, Taoyuan, Taiwan

[2]Division of Chinese Internal Medicine, Center for Traditional Chinese Medicine, Chang Gung Memorial Hospital, No. 123, Dinghu Rd, Gueishan Dist, Taoyuan, Taiwan

[3]Google International LLC Taiwan Branch, Taipei, Taiwan

[4]School of Medicine, Faculty of Medicine, National Yang-Ming Chiao Tung University, Taipei, Taiwan

[5]Taipei Veterans General Hospital, Yuli Branch, Taipei, Taiwan

[6]Institute of Hospital and Health Care Administration, National Yang-Ming Chiao Tung University, Taipei, Taiwan

[7]School of Traditional Chinese Medicine, College of Medicine, Chang Gung University, Taoyuan, Taiwan

Corresponding Author:
Hsing-Yu Chen, MD, PhD
Division of Chinese Internal Medicine, Center for Traditional Chinese Medicine, Chang Gung Memorial Hospital, No. 123, Dinghu Rd, Gueishan Dist, Taoyuan, Taiwan

Abstract

Background: The integration of artificial intelligence (AI), notably ChatGPT, into medical education, has shown promising results in various medical fields. Nevertheless, its efficacy in traditional Chinese medicine (TCM) examinations remains understudied.

Objective: This study aims to (1) assess the performance of ChatGPT on the TCM licensing examination in Taiwan and (2) evaluate the model's explainability in answering TCM-related questions to determine its suitability as a TCM learning tool.

Methods: We used the GPT-4 model to respond to 480 questions from the 2022 TCM licensing examination. This study compared the performance of the model against that of licensed TCM doctors using 2 approaches, namely direct answer selection and provision of explanations before answer selection. The accuracy and consistency of AI-generated responses were analyzed. Moreover, a breakdown of question characteristics was performed based on the cognitive level, depth of knowledge, types of questions, vignette style, and polarity of questions.

Results: ChatGPT achieved an overall accuracy of 43.9%, which was lower than that of 2 human participants (70% and 78.4%). The analysis did not reveal a significant correlation between the accuracy of the model and the characteristics of the questions. An in-depth examination indicated that errors predominantly resulted from a misunderstanding of TCM concepts (55.3%), emphasizing the limitations of the model with regard to its TCM knowledge base and reasoning capability.

Conclusions: Although ChatGPT shows promise as an educational tool, its current performance on TCM licensing examinations is lacking. This highlights the need for enhancing AI models with specialized TCM training and suggests a cautious approach to utilizing AI for TCM education. Future research should focus on model improvement and the development of tailored educational applications to support TCM learning.

(JMIR Med Educ 2025;11:e58897) doi:10.2196/58897

KEYWORDS

artificial intelligence; AI language understanding tools; ChatGPT; natural language processing; machine learning; Chinese medicine license exam; Chinese medical licensing examination; medical education; traditional Chinese medicine; large language model

Introduction

Traditional Chinese medicine (TCM), recognizeartid as one of the most renowned traditional medical systems, boasts a history spanning thousands of years. In the modern era, TCM has evolved to form an integral part of the formal health care system in East Asian countries, particularly in China and Taiwan [1,2]. TCM encompasses a wealth of theoretical knowledge and features unique diagnostic and treatment methods, such as acupuncture and herbal therapy. As a highly practical discipline, TCM learning traditionally relies on the accumulation of experience and the mentorship inherent in the master-apprentice

system; hence, this education model may not be sufficiently reliable or comprehensive. However, with the emerging need for integrative medicine over time, TCM has been integrated into the modern medical education system. This integration has led to prominent changes in educational approaches. The incorporation of TCM into academic institutions resulted in the establishment of formal examination systems. For instance, in Taiwan, TCM practitioners must pass a biannual licensing examination, termed the National Senior Professional and Technical Examinations for Chinese Medicine Practitioners (hereinafter called the "TCM licensing examinations"), to practice as a licensed TCM doctor, similar to their Western medicine counterparts [3].

The advancements in technology and the development of artificial intelligence (AI) have begun to impact and challenge the medical field, with TCM being no exception [4,5]. In the past year, significant progress has been made in AI language models, particularly those based on the generative pretrained transformer (GPT) architecture. ChatGPT, a conversational variant of the GPT model, has demonstrated its potential across various domains [6]. Recognized for its foundational medical knowledge and conversational capabilities, ChatGPT is considered a valuable tool in medical education, aiding in the understanding and application of medical knowledge [7], thereby facilitating student learning [8]. However, its responses are not consistently reliable. Unlike humans who answer questions based on an understanding of the content, it generates replies by drawing from a vast database. Therefore, although it can produce human-like conversations and respond to inquiries, it cannot guarantee the accuracy of its responses [9,10].

Discussions have emerged regarding the sufficiency of AI for clinical decision-making and basic medical consultation [7,11]. In addition, to be a potential mentor for medical students, one benchmark is the ability of AI to pass national licensing examinations (the minimum standard for practicing physicians). Thus, the application of ChatGPT in medical examinations has opened a new research direction. Studies have shown that GPT models, especially GPT-4, can achieve commendable scores on a variety of standardized tests for multiple professions, such as physicians [12-14], pharmacists [15], and nurses [16]. This success in examination settings has sparked interest in the potential of ChatGPT as a self-learning tool, suggesting its use for examination preparation and knowledge enhancement [17].

As previously mentioned, while TCM is a traditional medical system distinct from modern medicine, it has been integrated into modern medical education systems and subjected to formal examinations. The question arises: does ChatGPT possess the requisite knowledge level to assist TCM students in their learning? Only 1 study examined GPT's ability to answer TCM questions, but it focused on questions sourced from online TCM texts rather than formally recognized examination questions and utilized older GPT models (GPT-3 Turbo) [18]. In contrast, a more rigorous study on traditional Korean medicine found that, due to the unique nature of traditional medicine, GPT models require specially optimized prompts, such as language-related adjustments, to pass examinations [19]. However, considering the classical Chinese language barrier and different medical theories in TCM, whether GPT models would face challenges in TCM licensing examinations remains unexplored.

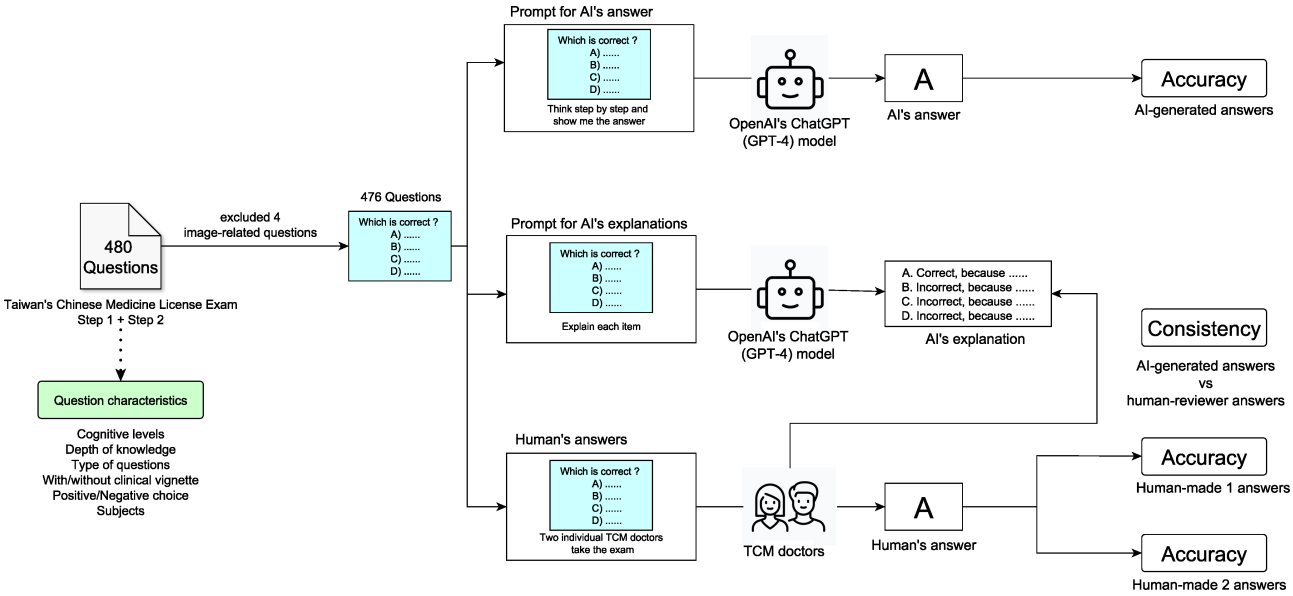
The aim of this study is to evaluate whether ChatGPT can accurately understand and respond to TCM questions by assessing its performance in simulated examination environments. By analyzing the accuracy of AI-generated answers, we sought to identify factors affecting their correctness. This study also aims to understand the consistency between AI-generated answers and their accompanying explanations, offering insights into the depth of understanding of this model. By analyzing the performance of ChatGPT in simulated TCM licensing examinations and comparing it with human performance, this study hopes to provide new insights and recommendations for innovation and development in TCM education.

Methods

Study Design

Figure 1 shows the data processing flowchart of this study. The feasibility of using ChatGPT (GPT-4 model, with a knowledge cutoff date of September 2021), developed by OpenAI, with 2 different prompts on responding to the first National Senior Professional and Technical Examinations for Chinese Medicine Practitioners was assessed by comparing the responses of the model to those of licensed TCM resident doctors. A total of 480 questions from the 2022 examination were inputted into ChatGPT, and 2 different approaches were used to obtain responses from ChatGPT. The first step involved prompting AI to select the correct answer directly from the question options. The second step required ChatGPT to explain why each option was correct or incorrect before selecting the correct answer. For the second step, individual answers and explanations from ChatGPT were manually assessed for accuracy and consistency. Subsequently, accuracy was measured by comparing the AI-selected answers with the correct answers. Additionally, the performance of AI was benchmarked against that of human experts. Two individual TCM resident doctors took the same examination without preparation, and their answers were also evaluated for accuracy. Finally, consistency was evaluated by comparing explanations against a standard set of answers for logical coherence, and the reasons for inconsistency were also verified by the 2 TCM doctors.

Figure 1. Flowchart of this study. GPT: generative pretrained transformer; TCM: traditional Chinese medicine.



The TCM Licensing Examination in Taiwan

In Taiwan, TCM doctors are qualified through 2 stages of licensing examinations after graduation from their TCM course at the university. The contents and answers are freely downloadable after each examination from the following website [20]. The examinations contain 2 stages corresponding to 10 subjects. The first stage consists of basic theory, including 黃帝內經 (Huangdi Neijing), 難經 (Nanjing) (domain I), and basic pharmacology and formulation (domain II). The second stage consists of principles of diagnosis and treatment, including 傷寒論 (Shanghanlun) and 金匱要略 (Jinguiyaolue) (domain III), TCM internal medicine (domain IV), TCM gynecology and obstetrics (domain IV), TCM pediatrics (domain IV), TCM dermatology (domain V), TCM otorhinolaryngology (domain V, including questions regarding the specialty concerning ears, nose, and throat [ENT]) and ophthalmology (domain V), TCM traumatology (domain V), and acupuncture (domain VI). Each domain contains 80 multiple-choice questions with single answers. The full score of each domain is 100. The examination score is calculated by dividing the total score by the number of subjects. Only examinees obtaining average scores ≥60 pass the examination. TCM students are eligible to take the first-stage examination when they have earned the requisite fourth-year university credits. Before the second-stage examination, TCM students must first pass the first-stage examination and graduate from the 7-year university course.

Question Characteristics

A total of 5 factors were used to characterize the examination questions, including the cognitive level, depth of knowledge (DOK), type of questions, vignette style, and polarity of questions (Table S1 in Multimedia Appendix 1). LWT and YCL independently reviewed and classified all questions according to the definitions of these 5 factors. In case of disagreement, HYC was consulted, and the disagreement was resolved by reaching a consensus among all authors. Bloom's taxonomy was modified to classify the questions into lower-order thinking skills (LOTS) and higher-order thinking skills (HOTS). LOTS include remembering, understanding, and applying knowledge to questions, while HOTS include further analyzing, evaluating, and creating after learning [21,22]. For the DOK, 3 levels, ranging from low to high based on Webb's framework on science, were defined as recall, concept, and strategic thinking. Questions with higher levels of DOK indicate the recruitment of sophisticated thinking [23]. Furthermore, the licensing examinations in Taiwan are presented as single-choice questions, adhering to the 1 stem, 4 choices policy. However, 2 types of questions were used to add variety to examination questions, including single-answer multiple-choice (SAMC) and single-answer, multiple-response multiple-choice (SAMRMC) questions. SAMC questions had only 1 most appropriate answer, while SAMRMC questions require the tester to choose the most appropriate answer composed of multiple correct options provided in each question (Table S2 in Multimedia Appendix 2). Moreover, if the content of a question presents clinical scenarios, this question would be categorized as the clinical vignette type. This type of question typically aims to examine the ability of the tester to analyze the clinical conditions and corresponding actions. The polarity of a question depended on whether the question was positively or negatively framed. A "positive-choice question" solicits the correct or affirmative answer, whereas a "negative choice question" demands the identification of the incorrect or negative answer.

Prompt for AI-Generated Answers

To enhance the precision and brevity of responses obtained from ChatGPT (GPT-4 model), we strategically added "think step-by-step" to our queries. This approach aimed to guide the model toward a methodical and sequential problem-solving process. Subsequently, by integrating the command "but show me only the answer, do not explain it," we aimed to extract a more refined and consolidated answer, significantly boosting the response accuracy of the model. An example of a prompt with response is demonstrated in Table S3 in Multimedia Appendix 3. We created a collection of unique prompts derived from an equal number of questions in the question database, submitting them sequentially to the AI model. To solve the issue

of memory retention between submissions, we used a specialized application designed to initiate separate application programming interface requests for each prompt. This approach guaranteed that each application programming interface interaction would be initiated separately. This ensures that the processing of each prompt and the generation of its answer were conducted in isolation, thereby preserving the integrity of the responses without interference from a prior response [24,25].

Prompt for Explanations Provided By AI Through Step-By-Step and Human-Curated Answers

Furthermore, to understand the thinking process of GPT and evaluate the accuracy of its interpretation of our inquiries, we prompted ChatGPT to "explain each item" for each question. This prompt directed the AI to furnish exhaustive explanations for each item [26] (Table S4 in Multimedia Appendix 4). LWT and YCL reviewed all explanations to items and reached decisive responses based on AI-generated explanations. This process was termed "human-curated responses." To authentically represent the logic of AI, we refrained from making any human amendments, even if the explanations provided by AI were incorrect. The answer would be marked as "wrong" if the AI-generated explanations were incorrect.

Outcome Assessment

We evaluated the accuracy of answers generated by the GPT, those made by humans, and explanations provided by the GPT and curated by humans. This was achieved by calculating the ratio of accurate responses to the total number of questions and representing the results as a percentage. This measure of accuracy underwent comparative analysis across different attributes of the questions. The human-curated answers, which encapsulated the interpretation of questions by AI, were evaluated by LWT, YCL, and HYC, who reached a consensus to identify instances of misinterpretation of the question (GPT cannot understand the question and does not provide an answer), misunderstanding of concepts (GPT can understand the question, but lacks knowledge of the topic), and incorrect application of principles (the responses GPT provides are correct in general but fail to answer the question).

Statistical Analysis

Proportions and percentages were used to present categorical data. A logistic regression approach was adopted to assess the effect of various attributes of questions on the correctness of responses generated by GPT-4. The cognitive complexity of the questions, their structural format, the inclusion of clinical vignettes, the overall polarity of questions, and the subjects were used as covariates in the logistic regression with univariable and multivariable models. The influence of each variable on the probability of the AI producing accurate answers was quantified using the adjusted odds ratio, accompanied by 95% CIs. Additionally, the κ statistic was used to evaluate the agreement between responses generated by GPT and curated by humans. This represented the different viewpoints concerning the same explanation between GPT and humans. $P<.05$ was used as the threshold for statistical significance. All statistical evaluation was performed utilizing Stata 17 (StataCorp LLC).

Ethical Considerations

This study did not require ethical approval, as it analyzed data obtained from a publicly available database. The test questions and answers used were originally created and copyrighted by the Taiwan Ministry of Examination and made accessible for academic research purposes. The Ministry retains full copyright over the examination content and confirmed that this research adhered to copyright regulations without any infringement.

Results

Question Characteristics

The examination encompassed a total of 480 questions spanning 10 specialties. Four image-related questions were excluded. Our findings indicated that most questions were HOTS, SAMC, negative-choice, and without a clinical vignette. According to Bloom's taxonomy of cognitive learning, the majority of questions across all subjects required HOTS (263/476, 55.3%; LOTS: 213/476, 44.7%). In particular, principles of diagnosis and treatment, TCM internal medicine, TCM dermatology, and TCM traumatology predominantly featured HOTS (58/80, 72.5%; 37/48, 77.1%; 13/19, 68.4%; and 17/20, 85%, respectively), while TCM pediatrics mainly involved LOTS (11/16, 68.8%). Within the LOTS category, "remembering" was the most common type (121/213, 56.8%), while "analyzing" dominated the HOTS category (255/263, 97%). In terms of Webb's DOK analysis of question types, the basic application of skill/concept represented the largest proportion (248/476, 52.1%), surpassing recall (85/476, 17.9%) and strategic thinking (143/476, 30%). A large portion of the questions were formatted as SAMC (439/476, 92.2%). Negative-choice questions comprised 62.2% (296/476) of the total, while 23.9% (180/476) of the questions included a clinical vignette (Table 1, Figures 2 and 3).

Table . Characteristics of TCM[a] licensing examinations in Taiwan, 2022.

Cognitive level	Total (n=476)	Basic theory (n=80)	Basic pharmacology and formulation (n=80)	Principle of diagnosis and treatment (n=80)	TCM internal medicine (n=48)	TCM GYN/OBS[b] (n=16)	TCM pediatrics (n=16)	TCM dermatology (n=19)	TCM ENT, ophthalmology (n=37)	TCM traumatology (n=20)	TCM acupuncture (n=80)
LOTS [c]											
Remembering	121 (25.4)	28 (35)	19 (23.8)	7 (8.8)	1 (2.1)	6 (37.5)	8 (50)	5 (26.3)	13 (35.1)	3 (15)	31 (38.8)
Understanding	41 (8.6)	10 (12.5)	10 (12.5)	7 (8.8)	2 (4.2)	0 (0)	0 (0)	0 (0)	5 (13.5)	0 (0)	7 (8.8)
Applying	51 (10.7)	7 (8.8)	9 (11.3)	8 (10)	8 (16.7)	3 (18.8)	3 (18.8)	1 (5.3)	4 (10.8)	0 (0)	8 (10)
HOTS [d]											
Analyzing	255 (53.6)	34 (42.5)	40 (50)	54 (67.5)	37 (77.1)	7 (43.8)	5 (31.3)	13 (68.4)	15 (40.5)	17 (85)	33 (41.3)
Evaluating	8 (1.7)	1 (1.3)	2 (2.5)	4 (5.0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (1.3)
Depth of knowledge											
Recall	85 (17.9)	20 (25)	18 (22.5)	6 (7.5)	0 (0)	4 (25)	5 (31.3)	3 (15.8)	3 (8.1)	4 (20)	22 (27.5)
Basic application of skill/concept	248 (52.1)	34 (42.5)	44 (55)	44 (55)	28 (58.3)	7 (43.8)	6 (37.5)	8 (42.1)	25 (67.6)	8 (40)	44 (55)
Strategic thinking	143 (30)	26 (32.5)	18 (22.5)	30 (37.5)	20 (41.7)	5 (31.3)	5 (31.3)	8 (42.1)	9 (24.3)	8 (40)	14 (17.5)
Type of question options and choices											
SAMC[e]	439 (92.2)	78 (97.5)	76 (95)	75 (93.8)	48 (100)	11 (68.8)	13 (81.3)	19 (100)	30 (81.1)	20 (100)	69 (86.3)
SAMRMC[f]	37 (7.8)	2 (2.5)	4 (5)	5 (6.3)	0 (0)	5 (31.3)	3 (18.8)	0 (0)	7 (18.9)	0 (0)	11 (13.8)
Clinical vignette											
Without clinical vignette	362 (76.1)	63 (78.8)	63 (78.8)	61 (76.3)	22 (45.8)	7 (43.8)	14 (87.5)	13 (68.4)	29 (78.4)	16 (80)	74 (92.5)
With clinical vignette	114 (23.9)	17 (21.3)	17 (21.3)	19 (23.8)	26 (54.2)	9 (56.3)	2 (12.5)	6 (31.6)	8 (21.6)	4 (20)	6 (7.5)
Polarity of question options											
Positive	180 (37.8)	22 (27.5)	27 (33.8)	36 (45)	21 (43.8)	3 (18.8)	5 (31.3)	9 (47.4)	8 (21.6)	13 (65)	36 (45)
Negative	296 (62.2)	58 (72.5)	53 (66.3)	44 (55)	27 (56.3)	13 (81.3)	11 (68.8)	10 (52.6)	29 (78.4)	7 (35)	44 (55)

[a]TCM: traditional Chinese medicine.

[b]GYN/OBS: gynecology/obstetrics.

[c]LOTS: lower-order thinking skills.

[d]HOTS: higher-order thinking skills.

[e]SAMC: single-answer multiple-choice.

[f]SAMRMC: single-answer, multiple-response multiple-choice.

Figure 2. Distribution of subjects in TCM licensing examinations. The detailed numbers and proportion of each subject's question types can be seen in Table 1. ENT: ears, nose, and throat; GYN/OBS: gynecology/obstetrics; HOTS: higher-order thinking skills; LOTS: lower-order thinking skills; SAMC: single-answer multiple-choice; SAMRMC: single-answer, multiple-response multiple-choice; TCM: traditional Chinese medicine.

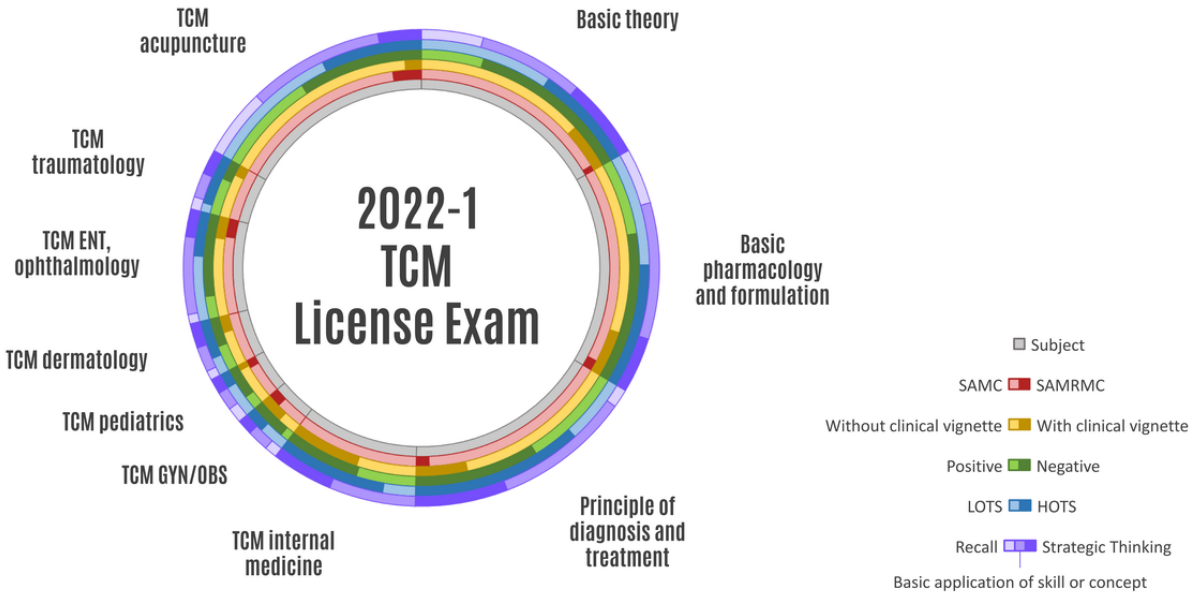
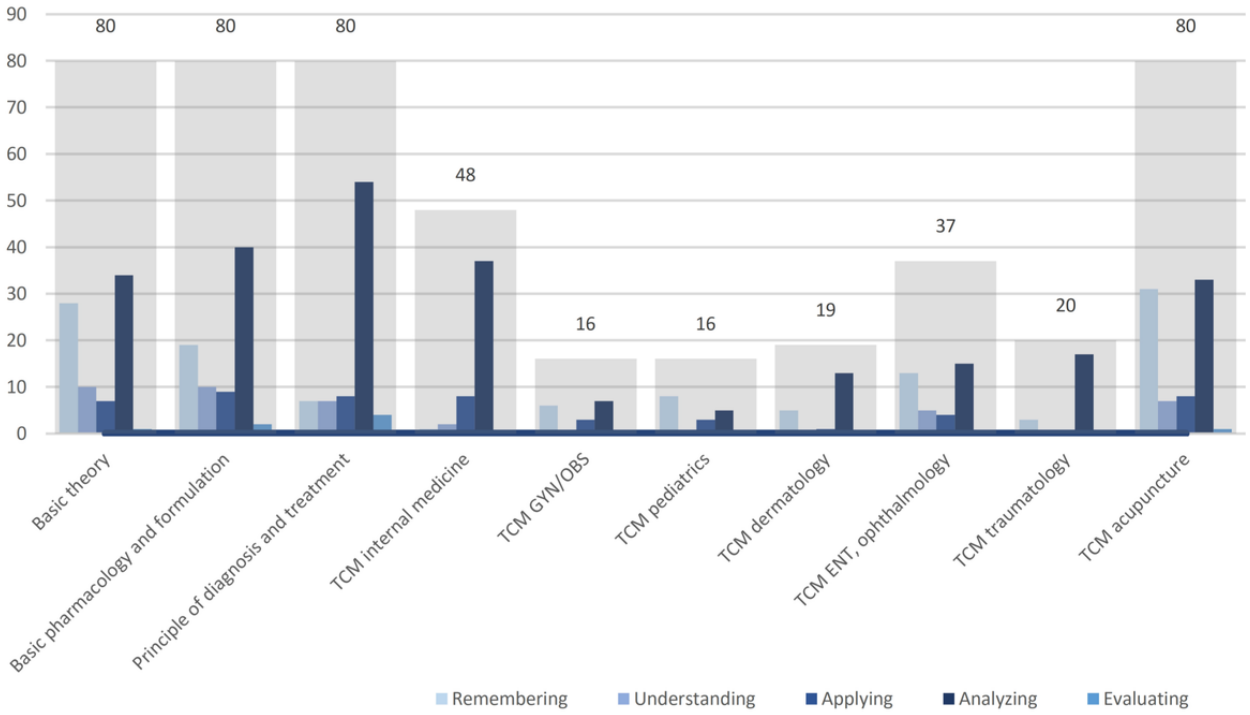


Figure 3. Analysis of question types according to Bloom's cognitive level in TCM licensing examinations. ENT: ears, nose, and throat; TCM: traditional Chinese medicine.



GPT-4 Model Performance and Accuracy Across Different Question Characteristics

We observed that the performance of the GPT-4 model was inferior to that of humans and did not demonstrate significant variation across different categories of examination questions. The GPT-4 model demonstrated an overall accuracy of only 43.9% (209/476). In comparison, 2 human evaluators achieved accuracy rates of 70% (333/476) and 78.4% (373/476), respectively (Table 2). The performance of ChatGPT across various variables is shown in Table 3. The accuracy of AI-generated answers did not show a significant correlation

with the characteristics of the questions, regardless of the classification method used (Figure 4). The GPT-4 model demonstrated a performance close to that of humans in TCM dermatology and TCM traumatology. The accuracy of AI-generated answers varied among the test subjects, ranging from 31.3% in TCM pediatrics to 73.7% in TCM dermatology. Notably, only TCM internal medicine (adjusted odds ratio [aOR] 3.07, 95% CI 1.41 - 6.68; *P*=.005), TCM dermatology (aOR 5.11, 95% CI 1.65 - 15.85; *P*=.005), and TCM acupuncture (aOR 2.14, 95% CI 1.12 - 4.11; *P*=.02) showed statistically significant better performance (Figure 4). On the other hand, GPT had a higher, but not statistically significant, accuracy rate for questions categorized as LOTS (96/213, 45.1%), SAMC (197/439, 44.9%), strategic thinking (66/143, 46.2%), with clinical vignette (52/114, 45.6%), and positive-choice (85/180, 47.2%).

Table . Accuracy rates of testers and ChatGPT-4 for TCM[a] licensing examinations.

	Number of questions	Number of correct responses	Accuracy, %
Human-made 1	476	333	70
Human-made 2	476	373	78.4
ChatGPT-4[b]	476	209	43.9
Human-curated answer 1	476	192	40.3
Human-curated answer 2	476	186	39.1

[a]TCM: traditional Chinese medicine.

[b]ChatGPT did not show answers to 7 questions although an explanation was provided.

Table . Accuracy rates of testers and ChatGPT-4 across different types and subjects of questions.

	Accuracy, %				
	Human-made 1	Human-made 2	ChatGPT-4	Human-curated 1	Human-curated 2
Bloom's cognitive level					
LOTS[a]	150 (70.4)	164 (77)	96 (45.1)	78 (36.6)	75 (35.2)
HOTS[b]	183 (69.6)	209 (79.5)	113 (43)	114 (43.3)	111 (42.2)
Depth of knowledge					
Recall	57 (67.1)	65 (76.5)	34 (40)	27 (31.8)	22 (25.9)
Basic application of skill/concept	172 (69.4)	193 (77.8)	109 (44)	103 (41.5)	102 (41.1)
Strategic thinking	104 (72.7)	115 (80.4)	66 (46.2)	62 (43.4)	62 (43.4)
Type of questions					
SAMC[c]	312 (71.1)	346 (78.8)	197 (44.9)	180 (41)	176 (40.1)
SAMRMC[d]	21 (56.8)	27 (73)	12 (32.4)	12 (32.4)	10 (27)
Vignette style question					
Without clinical vignette	248 (68.5)	283 (78.2)	157 (43.4)	143 (39.5)	137 (37.8)
With clinical vignette	85 (74.6)	90 (78.9)	52 (45.6)	49 (43)	49 (43)
Polarity of question					
Positive	129 (71.7)	142 (78.9)	85 (47.2)	78 (43.3)	76 (42.2)
Negative	204 (68.9)	231 (78)	124 (41.9)	114 (38.5)	110 (37.2)
Subjects					
Basic theory	51 (63.7)	63 (78.8)	29 (36.3)	29 (36.3)	28 (35)
Basic pharmacology and formulation	63 (78.8)	66 (82.5)	30 (37.5)	32 (40)	28 (35)
Principle of diagnosis and treatment	57 (71.3)	58 (72.5)	29 (36.3)	29 (36.3)	29 (36.3)
TCM[e] internal medicine	41 (85.4)	44 (91.7)	30 (62.5)	24 (50)	24 (50)
TCM gynecology and obstetrics	10 (62.5)	12 (75)	8 (50)	4 (25)	4 (25)
TCM pediatrics	11 (68.8)	13 (81.3)	5 (31.3)	7 (43.8)	7 (43.8)
TCM dermatology	14 (73.7)	17 (89.5)	14 (73.7)	12 (63.2)	12 (63.2)
TCM ENT[f], ophthalmology	21 (56.8)	26 (70.3)	12 (32.4)	13 (35.1)	13 (35.1)
TCM traumatology	9 (45)	14 (70)	9 (45)	8 (40)	8 (40)
TCM acupuncture	56 (70)	60 (75)	43 (53.8)	34 (42.5)	33 (41.3)

[a]LOTS: lower-order thinking skills.

[b]HOTS: higher-order thinking skills.

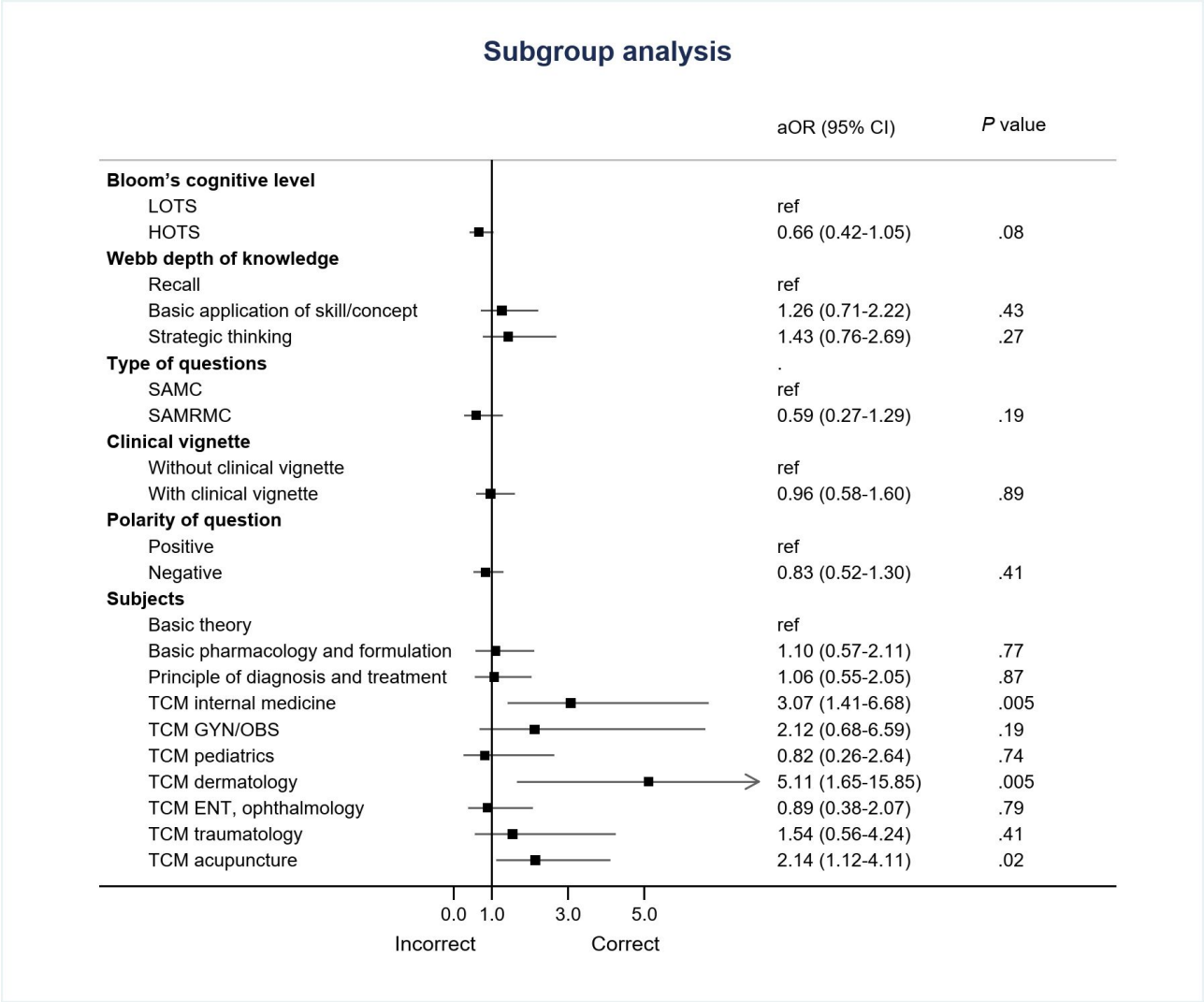
[c]SAMC: single-answer multiple-choice.

[d]SAMRMC: single-answer, multiple-response multiple-choice.

[e]TCM: traditional Chinese medicine.

[f]ENT: ears, nose, and throat.

Figure 4. Factors associated with correct answers provided by ChatGPT-4. aOR: adjusted odds ratio; ENT: ears, nose, and throat; GYN/OBS: gynecology/obstetrics; HOTS: higher-order thinking skills; LOTS: lower-order thinking skills; SAMC: single-answer multiple-choice; SAMRMC: single-answer, multiple-response multiple-choice; TCM: traditional Chinese medicine.



Consistency Between AI-Generated Answers and Human-Curated Answers and Analysis of Incorrect Responses Provided by the GPT-4 Model

The consistency between AI-generated and human-curated results was low (κ=0.504; Figure 5). After human review, the accuracy of the human-curated answers showed an overall trend of slight decrease, except for some minor increases in basic pharmacology and formulation, TCM pediatrics, and TCM otorhinolaryngology and ophthalmology. The accuracies for the remaining specialties were slightly lower, ranging from 43.9% to 40.3% (Table 2, Figures 5 and 6). For human reviewer 1, discrepancies were observed between AI-generated responses and those reviewed by humans, with 23.96% (115 of 480 questions) of the answers provided by AI conflicting with its own explanations. For 33% of correctly answered questions (69 of 209 questions), the AI provided an incorrect explanation, indicating a scenario of "correct answer, incorrect explanation." Conversely, for 17% of incorrectly answered questions (46 of 267 questions), the AI provided a correct explanation, suggesting a case of "incorrect answer, correct explanation." This reduced the overall accuracy of the AI model to 43.9%.

Figure 5. Accuracy rates of humans and ChatGPT-4 for TCM licensing examinations. The passing standard is an average score of 60. With 476 questions, the threshold is at least 286 correct answers (red dashed line). AI: artificial intelligence; TCM: traditional Chinese medicine.

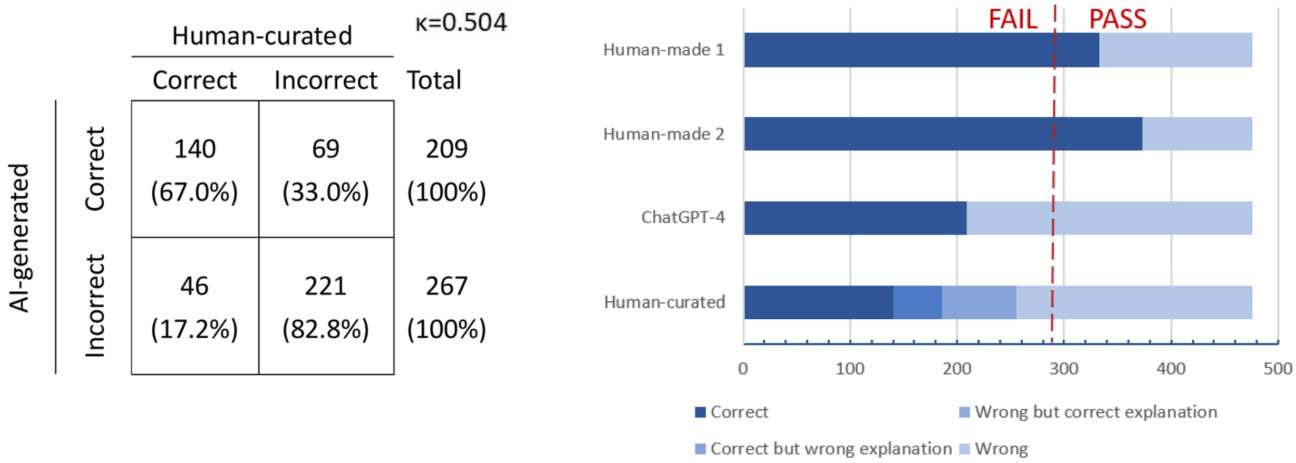
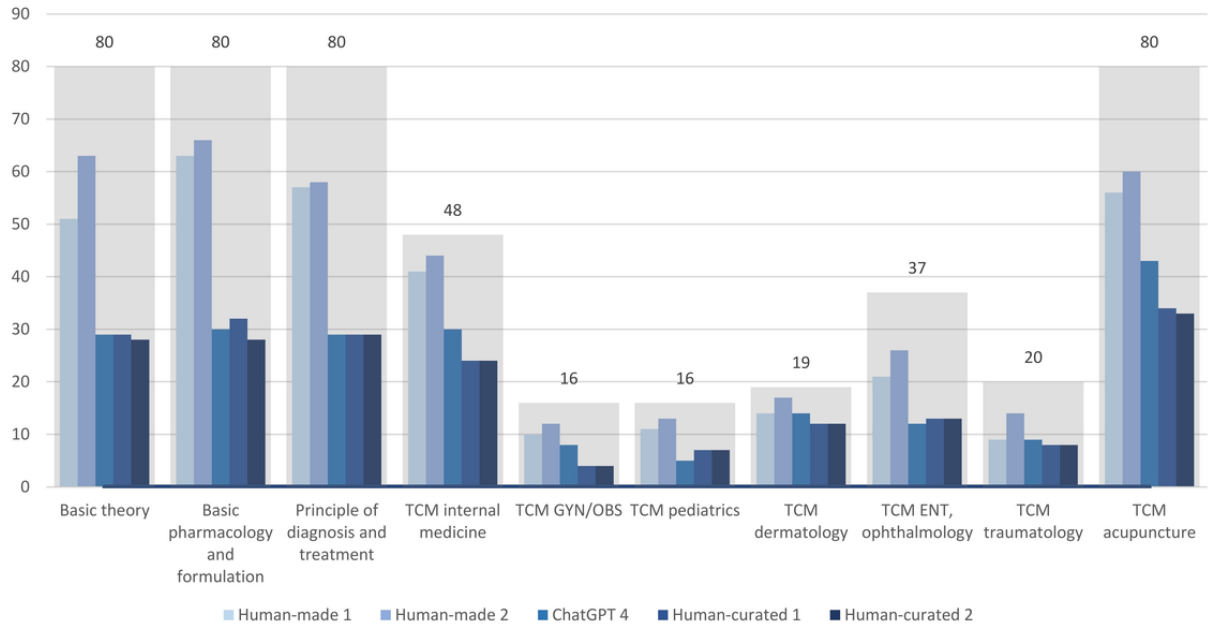


Figure 6. Performance of humans and ChatGPT-4 across various subjects. ENT: ears, nose, and throat; GYN/OBS: gynecology/obstetrics; TCM: traditional Chinese medicine.



We further analyzed the reasons responsible for the incorrect answers provided by the GPT. For this purpose, we categorized the potential reasons for these errors into 3 types: misinterpretation of the question (failing to understand the question), misunderstanding of concepts (lacking knowledge of the topic), and incorrect application of principles (the content is correct, but it does not answer the question). The results revealed that most of the errors (263/476, 55.3%) were attributed to the misunderstanding of concepts (Table 4, Figure 7). However, a closer examination of the different characteristics of the questions indicated that misunderstanding of concepts was more common in LOTS, recall, and SAMRMC compared to their counterparts. The second most common cause of error was incorrect application of principles (20/476, 4.2%), followed by misinterpretation of questions (7/476, 1.5%).

Table . Reasons responsible for incorrect artificial intelligence–generated responses (human-curated).

	Correct (n=186)	Misinterpretation of the question (n=7)	Misunderstanding of concepts (n=263)	Incorrect application of principles (n=20)	*P* value
Bloom's cognitive level					.25
LOTS[a]	75 (40.3)	5 (71.4)	124 (47.1)	9 (45)	
HOTS[b]	111 (59.7)	2 (28.6)	139 (52.9)	11 (55)	
Depth of knowledge					.06
Recall	22 (11.8)	0 (0)	60 (22.8)	3 (15)	
Basic application of skill/concept	102 (54.8)	5 (71.4)	132 (50.2)	9 (45)	
Strategic thinking	62 (33.3)	2 (28.6)	71 (27)	8 (40)	
Type of questions					.16
SAMC[c]	176 (94.6)	6 (85.7)	237 (90.1)	20 (100)	
SAMRMC[d]	10 (5.4)	1 (14.3)	26 (9.9)	0 (0)	
Vignette style question					.39
Without clinical vignette	137 (73.7)	6 (85.7)	206 (78.3)	13 (65)	
With clinical vignette	49 (26.3)	1 (14.3)	57 (21.7)	7 (35)	
Polarity of question					.28
Positive	76 (40.9)	1 (14.3)	98 (37.3)	5 (25)	
Negative	110 (59.1)	6 (85.7)	165 (62.7)	15 (75)	
Subjects					<.001
Basic theory	28 (15.1)	0 (0)	35 (13.3)	17 (85)	
Basic pharmacology and formulation	28 (15.1)	1 (14.3)	49 (18.6)	2 (10)	
Principle of diagnosis and treatment	29 (15.6)	3 (42.9)	48 (18.3)	0 (0)	
TCM[e] internal medicine	24 (12.9)	0 (0)	24 (9.1)	0 (0)	
TCM gynecology and obstetrics	4 (2.2)	0 (0)	12 (4.6)	0 (0)	
TCM pediatrics	7 (3.8)	1 (14.3)	8 (3.0)	0 (0)	
TCM dermatology	12 (6.5)	0 (0)	7 (2.7)	0 (0)	
TCM ENT[f], ophthalmology	13 (7)	1 (14.3)	22 (8.4)	1 (5)	
TCM traumatology	8 (4.3)	0 (0)	12 (4.6)	0 (0)	
TCM acupuncture	33 (17.7)	1 (14.3)	46 (17.5)	0 (0)	

[a]LOTS: lower-order thinking skills.

[b]HOTS: higher-order thinking skills.

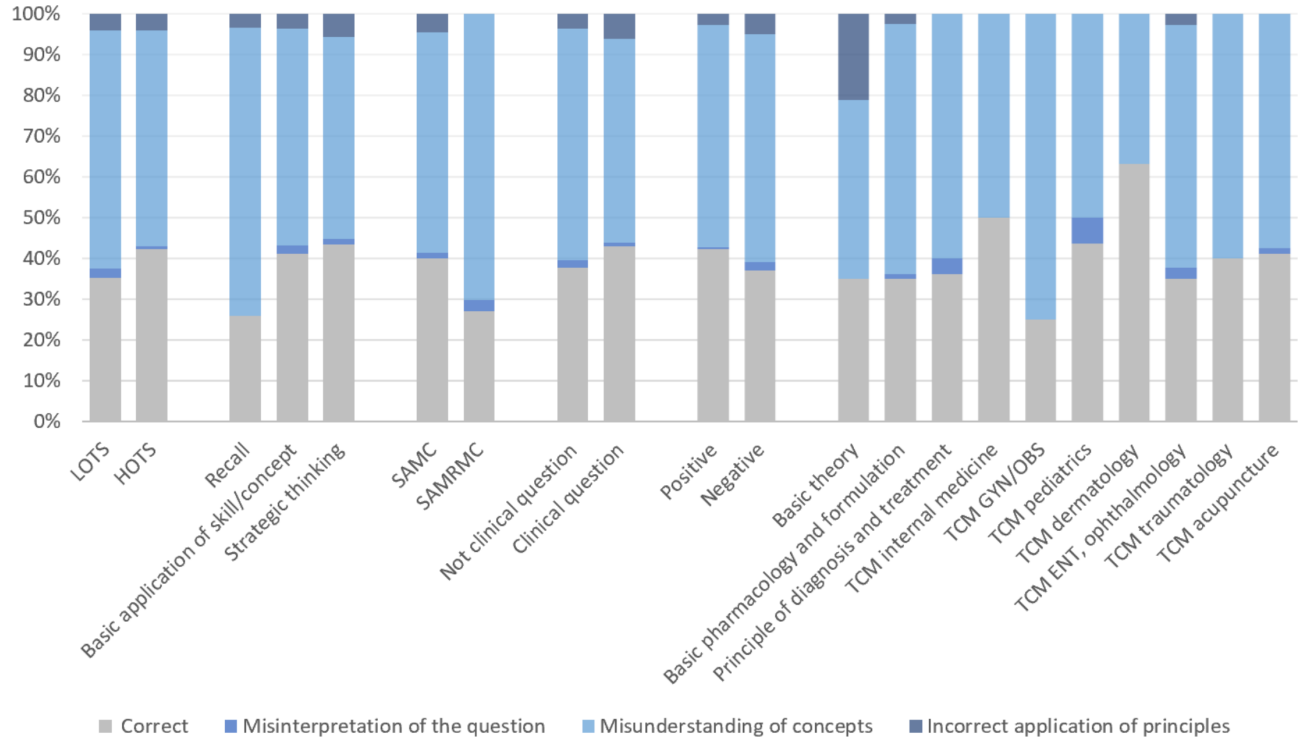
[c]SAMC: single-answer multiple-choice.

[d]SAMRMC: single-answer, multiple-response multiple-choice.

[e]TCM: traditional Chinese medicine.

[f]ENT: ears, nose, and throat.

Figure 7. Distribution of reasons for incorrect answers provided by ChatGPT-4. ENT: ears, nose, and throat; GYN/OBS: gynecology/obstetrics; HOTS: higher-order thinking skills; LOTS: lower-order thinking skills; SAMC: single-answer multiple-choice; SAMRMC: single-answer, multiple-response multiple-choice; TCM: traditional Chinese medicine.



Discussion

Performance of ChatGPT in Medical Examinations

This is the first study to test the capabilities of ChatGPT in TCM examinations. ChatGPT has undergone rigorous testing for its proficiency in medical examinations. Nonetheless, its effectiveness in TCM licensing examinations remains unexplored. Hence, this study fills a research void by examining the capability of an advanced language model like ChatGPT in the context of TCM. Generally, most studies indicate ChatGPT can meet the medical examination pass standards. For example, ChatGPT 3.5 scored around the pass mark on the United States Medical Licensing Examination [14] and exhibited strong performance in specialties such as radiation oncology and neurosurgery [27,28]. GPT-4 surpassed 70% in its score for UK medical licensing examinations [12], and its competency extends to examinations in different languages. For example, GPT 3.5 typically scored around the passing mark on the Japanese nursing examinations [16] and Korean medical student parasitology examinations [29]. Although GPT-3.5 Turbo is not yet capable, GPT-4 passed the medical licensing examinations of China [30,31] and achieved 88.6% accuracy in the equivalent examinations of Saudi Arabia [32]. Interestingly, it even outperformed human residents in the residency training examinations of Japan [33].

Published research has identified 2 trends in this setting. First, GPT-4 surpasses GPT-3.5 in identical medical examinations, as demonstrated in medical student finals in Poland [34] and the medical licensing examinations of Peru [35]. A systematic review and meta-analysis of ChatGPT use in medical licensing examinations worldwide observed similar results [36]. Second, ChatGPT models showed higher accuracy when answering questions translated into English compared with the original language [34,37]. In Taiwan, traditional Chinese is the language used for medical licensing examinations. Despite this disadvantage, ChatGPT performed near the pass threshold for the nursing [38] and pharmacy licensing examinations in Taiwan [15]; translating pharmacy examination questions into English indeed improved scores across all subjects [15]. Thus, it was hypothesized that GPT-4 would perform similarly in TCM licensing examinations. However, the results were surprising. The study used the first 2022 TCM licensing examinations in Taiwan as a case study to assess the performance of the model. GPT-4 failed the exam with an overall accuracy of 43.9%; following human revision of AI-provided explanations, the accuracy further decreased to 40.3% (human 1) and 39.1% (human 2). These results underscore the need for further research and development on the application of AI models to TCM examination preparation and highlight the existing knowledge gap. The reasons behind these outcomes merit further investigation.

Challenges Encountered by ChatGPT When Answering Medical Questions

Previous literature has discussed the shortcomings and challenges of ChatGPT in answering examination questions, including a decreased proficiency in languages other than English [34,37], AI “hallucinations” originating from erroneous data [10,38], and proficiency limited to certain types of questions [13,39]. The tendency for ChatGPT to be less proficient in answering questions posed in languages other than English stems from the fact that ChatGPT is an LLM trained primarily on English language data, which includes a wide variety of sources such as books, websites, and news articles [6]. The questions for TCM licensing examinations are not presented in

English. Although ChatGPT can fluently interact in traditional Chinese, its responses to medical examination questions, which require specific expertise and have standard answers, may reveal its inadequacies. AI "hallucinations" indicate a tendency to produce "hallucinations" or factually incorrect content due to incorrect data. This poses the risk of generating misleading or fabricated information, which complicates the use of AI as a reliable self-learning tool [7,10]. We also encountered seemingly plausible but incorrect content in AI-generated responses in our research. We even found that verifying the authenticity of these answers is more time-consuming and requires deeper professional knowledge than the questions themselves. Our study also showed that ChatGPT had higher, albeit not statistically significant, accuracy rates for questions posed such as SAMC (n=197, 44.9%) and presented with clinical vignettes (n=52, 45.6%). This trend aligns with findings of previous studies, such as a lower proficiency in multiple-choice questions [13] and a poorer aptitude for conceptual questions compared with clinical scenarios [39]. Despite these limitations, which we have also encountered, other research has shown that ChatGPT can pass examinations. Therefore, the use of ChatGPT in the context of TCM may pose its own unique set of challenges and necessitates further investigation.

Challenges Encountered by ChatGPT When Answering TCM Examination Questions

We identified 3 main reasons for incorrect answers according to AI-generated responses, namely misinterpretation of the question, misunderstanding of concepts, and incorrect application of principles. Misunderstanding of concepts was the most prevalent, especially in questions with lower cognitive demand such as recall and LOTS, as well as in questions where a single item encompasses multiple questions (eg, SAMRMC), indicating either a lack of knowledge or incorrect knowledge. We believe that this primarily stems from 2 factors. First, the database for TCM is currently incomplete. Second, compared with Western medicine, TCM is often considered alternative medicine. If an LLM such as ChatGPT answers questions based solely on the Western medical knowledge system, then TCM content may be ignored. Additionally, TCM focuses on personalized treatment without a golden standard, leading to the absence of definitive answers for the same disease.

The incomplete TCM database is due to challenges such as insufficient data, lack of standardization, and unrepresentative data sources. Although the specific TCM data that ChatGPT uses for training are unclear, it is evident that the current online data for TCM are significantly less comprehensive than those for Western medicine. For instance, a bibliometric analysis over the past 20 years did not show a significant presence of TCM-related keywords in the context of pediatric allergic rhinitis [40]. However, the usage rate of TCM for allergic diseases in Taiwan is approximately 30% - 50% [41]. Therefore, a model constructed based on such a database is likely to exhibit discrepancies with reality. Furthermore, online data often contain inaccuracies or incomplete information. Previous research has shown that uncleaned training texts can affect performance and could underpin the subpar performance of the trained model [42].

It is important to note that, due to challenges in translation and cultural appropriation, certain medical terms have different connotations in the TCM and Western medical systems. However, ChatGPT tends to interpret these terms with a preference for their meanings within Western medicine. For instance, in some AI-generated responses, the TCM term for "肝" was mistakenly translated and described as the physical organ "liver" in Western medicine. Similarly, the term for "瘧" in TCM was translated and described as "malaria" in some AI-generated responses. The understanding of "肝" in TCM is not entirely the same as in modern medicine, and "瘧" in TCM refers to a broad category of symptoms similar to malaria but not restricted to infections caused by *Plasmodium*.

The crux of TCM is personalized treatment, which is antithetical to gold-standard treatments. Hence, multiple therapeutic approaches may exist for the same disease. If the examination questions do not specify a particular scope or clear criteria, there may be no standard answer or multiple possible solutions. This study revealed that the decrease in the overall accuracy rate after human review was primarily driven by a reduction in accuracy for LOTS questions, whereas the accuracy rate for HOTS remained stable or even increased. Regarding DOK, the decrease in accuracy following human review was primarily in recall, with less of a decrease noted in more advanced DOK (eg, basic application of skill/concept, strategic thinking). This suggests that GPT-4 is more adept at providing detailed explanations for complex logical reasoning questions, as opposed to simple memorization, which might be influenced by incorrect information. In addition, if users intend to use GPT to answer TCM questions, they should be particularly cautious of potential hallucinations in lower cognitive demand questions.

Our study revealed that the GPT-4 model is currently unable to pass the TCM licensing examinations. This research underscores the limitations of the performance of AI in TCM licensing examinations, as well as illuminates broader challenges within the realm of integrating TCM knowledge into AI development.

Limitations

Although this study provides valuable insights into the use of the GPT-4 model for TCM licensing examination preparation, several limitations have been identified. The focus solely on the GPT-4 model of ChatGPT might neglect the complexities and potential capabilities of other recently developed AI-driven language models, such as Claude 3 by Anthropic, Bard (Gemini Pro) by Google, or LLaMa2 by Facebook. Notably, we did not use expert-level AI, such as Med-PaLM by Google [43]. Moreover, we did not use other traditional Chinese-language LLMs, such as Taiwan-LLM [44,45]. Nevertheless, GPT models are the most widely used and studied models, and it is necessary to use the same tool to facilitate comparisons with other research studies [36].

Considering the cultural context specific to the TCM licensing examination of Taiwan, the generalizability of our findings to different regions or educational systems may be limited. Notably, model performance may change over time, indicating that our results may not be replicated in the future. This study also did not account for potential inconsistencies in responses provided by ChatGPT to identical queries during different

sessions. However, this issue could be minimized by explicitly setting the parameters of ChatGPT.

Additionally, the difficulty of each exam can vary, which might affect ChatGPT's performance. However, the difficulty is generally controlled and, as a national exam, the pass rates have been stable over the years [46]. Previous exam questions could potentially be part of the GPT model's training data (with a knowledge cutoff date of September 2021), introducing bias. Therefore, we only used the first exam of 2022 to mitigate this issue.

Implications for Practice and Future Research

This study investigated the use of the GPT-4 model for TCM licensing examination preparation. The findings revealed that AI-driven tools are not yet valuable assets for TCM educators and students. The observed limitations (ie, often providing responses based on incorrect facts) highlight the need for further development before this model can be effectively used as a self-learning tool. As the AI field continues to advance with the introduction of new models, educators must stay informed and utilize the most effective tools while being cognizant of their limitations. This study sets the stage for 2 potential research directions. In terms of TCM, considering the suboptimal examination results, we speculate that the primary drawback lies in the quality of the front-end data. Future improvements may include incorporating ancient TCM texts and customizing training for LLMs.

We must deliberately incorporate relevant resources into our training database materials, such as textbooks on TCM in Chinese and ancient TCM texts. Currently, the majority of descriptions and knowledge regarding TCM are in Chinese. When these data are published in journals or translated into English, they often adopt the framework and language of modern medicine as a medium for knowledge transmission. This approach tends to underemphasize the original content of TCM, which is mostly documented in Chinese literature. Therefore, the inclusion of TCM materials in LLM training and the standardization of TCM should be targeted for improvement.

Tailoring training data for LLMs presents another promising avenue for improvement. TCM comprises different schools, suggesting that narrowing the knowledge domain could be more advantageous. Hence, to excel in TCM, developing specialized ChatGPT models or custom LLMs might be a beneficial strategy. Considering the current limitations in enhancing the database, integrating specific prompts offers an alternative solution. For example, the chain-of-thoughts method, used in LLMs for complex problem-solving, articulates intermediate steps in reasoning. This approach is particularly effective for models with extensive parameters, enhancing their ability to manage multistep tasks [26]. It has been confirmed that this method can also improve the performance of ChatGPT in medical examinations [47]. Hence, the adoption of chain-of-thoughts may be a viable strategy to address the complexity of TCM examinations. Additionally, previous research indicated that restricting ChatGPT to a single response in a Basic Life Support examination may introduce bias. When ChatGPT generates 3 responses per question, it successfully passes the examination. Moreover, rephrasing incorrectly answered questions as open-ended questions significantly boosts the accuracy of ChatGPT. This implies that open-ended questioning or multiple inquiries might be more effective than single-choice formats [48].

Conclusion

Our study represents the first comprehensive assessment of the performance of ChatGPT in TCM licensing examinations. Despite advances in AI and its success in various medical licensing tests, ChatGPT demonstrated a limited ability to accurately respond to TCM examination questions, achieving an overall accuracy rate significantly lower than that of its human counterparts. This shortfall underscores the challenges posed by the unique concepts and terminologies of TCM, highlighting a significant knowledge gap in the understanding of TCM principles by AI. Our findings call for further advancements in AI training, specifically tailored toward the intricate domain of TCM, to enhance its utility in this specialized field of medicine.

Acknowledgments

This study was partially supported by Chang Gung Medical Foundation (grant CGRPG1Q0011), the Ministry of Health and Welfare (grants MOHW112-CMAP-M-113-000006-D, MOHW113-CMAP-M-113-000002-D, and MOHW113-CMAP-M-113-000003-B), and the National Science and Technology Council in Taiwan (grant MOST111-2320-B-182-035-MY3).

Authors' Contributions

LWT contributed to manuscript writing. HYC and YCC were responsible for the statistical analysis, project administration, funding acquisition, manuscript revision, and study design. Results were interpreted by LCT and YCL. HYC and YCC contributed equally as co-corresponding authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
List of the 5 factors, data definitions, and source citations.

[DOCX File, 18 KB - mededu_v11i1e58897_app1.docx]

Multimedia Appendix 2
Examples of single-answer multiple-choice and single-answer, multiple-response multiple-choice questions.
[DOCX File, 16 KB - mededu_v11i1e58897_app2.docx]

Multimedia Appendix 3
Examples of the prompt used to generate responses from questions.
[DOCX File, 57 KB - mededu_v11i1e58897_app3.docx]

Multimedia Appendix 4
Examples of the prompts used to generate responses from questions with explanations for each item.
[DOCX File, 179 KB - mededu_v11i1e58897_app4.docx]

References

1. Chi C. Integrating traditional medicine into modern health care systems: examining the role of Chinese medicine in Taiwan. Soc Sci Med 1994 Aug;39(3):307-321. [doi: 10.1016/0277-9536(94)90127-9] [Medline: 7939847]
2. Chi C, Lee JL, Lai JS, Chen CY, Chang SK, Chen SC. The practice of Chinese medicine in Taiwan. Soc Sci Med 1996 Nov;43(9):1329-1348. [doi: 10.1016/0277-9536(95)00429-7] [Medline: 8913003]
3. Park YL, Huang CW, Sasaki Y, Ko Y, Park S, Ko SG. Comparative study on the education system of traditional medicine in China, Japan, Korea, and Taiwan. Expl NY 2016;12(5):375-383. [doi: 10.1016/j.explore.2016.06.004] [Medline: 27546589]
4. Wang Y, Shi X, Li L, Efferth T, Shang D. The impact of artificial intelligence on traditional Chinese medicine. Am J Chin Med 2021;49(6):1297-1314. [doi: 10.1142/S0192415X21500622] [Medline: 34247564]
5. Li W, Ge X, Liu S, Xu L, Zhai X, Yu L. Opportunities and challenges of traditional Chinese medicine doctors in the era of artificial intelligence. Front Med (Lausanne) 2023;10:1336175. [doi: 10.3389/fmed.2023.1336175] [Medline: 38274445]
6. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. 2020 Presented at: Proceedings of the 34th International Conference on Neural Information Processing Systems; Dec 6-12, 2020; Vancouver, BC, Canada.
7. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. N Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
8. Han JW, Park J, Lee H. Analysis of the effect of an artificial intelligence chatbot educational program on non-face-to-face classes: a quasi-experimental study. BMC Med Educ 2022 Dec 1;22(1):830. [doi: 10.1186/s12909-022-03898-3] [Medline: 36457086]
9. Branum C, Schiavenato M. Can ChatGPT accurately answer a PICOT question? Assessing AI response to a clinical question. Nurse Educ 2023;48(5):231-233. [doi: 10.1097/NNE.0000000000001436] [Medline: 37130197]
10. Borji A. A categorical archive of ChatGPT failures. arXiv. Preprint posted online on Feb 6, 2023. [doi: 10.21203/rs.3.rs-2895792/v1]
11. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023;6:1169595. [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
12. Lai UH, Wu KS, Hsu TY, Kan JKC. Evaluating the performance of ChatGPT-4 on the United Kingdom Medical Licensing Assessment. Front Med (Lausanne) 2023;10:1240915. [doi: 10.3389/fmed.2023.1240915] [Medline: 37795422]
13. Alfertshofer M, Hoch CC, Funk PF, et al. Sailing the seven seas: a multinational comparison of ChatGPT's performance on medical licensing examinations. Ann Biomed Eng 2024 Jun;52(6):1542-1545. [doi: 10.1007/s10439-023-03338-3] [Medline: 37553555]
14. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Dig Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
15. Wang YM, Shen HW, Chen TJ. Performance of ChatGPT on the pharmacist licensing examination in Taiwan. J Chin Med Assoc 2023 Jul 1;86(7):653-658. [doi: 10.1097/JCMA.0000000000000942] [Medline: 37227901]
16. Taira K, Itaya T, Hanada A. Performance of the large language model ChatGPT on the National Nurse Examinations in Japan: evaluation study. JMIR Nurs 2023 Jun 27;6:e47305. [doi: 10.2196/47305] [Medline: 37368470]
17. Wang H, Wu W, Dou Z, He L, Yang L. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: pave the way for medical AI. Int J Med Inform 2023 Sep;177:105173. [doi: 10.1016/j.ijmedinf.2023.105173] [Medline: 37549499]
18. Yizhen L, Shaohan H, Jiaxing Q, Lei Q, Dongran H, Zhongzhi L. Exploring the comprehension of ChatGPT in Traditional Chinese Medicine knowledge. arXiv. Preprint posted online on Mar 14, 2024. [doi: 10.48550/arXiv.2403.09164]
19. Jang D, Yun TR, Lee CY, Kwon YK, Kim CE. GPT-4 can pass the Korean National Licensing Examination for Korean Medicine Doctors. PLOS Dig Health 2023 Dec;2(12):e0000416. [doi: 10.1371/journal.pdig.0000416] [Medline: 38100393]

20. Taiwan Ministry of Examination - Examination Question Inquiry Platform [Website in Mandarin]. URL: https://wwwq.moex.gov.tw/exam/wFrmExamQandASearch.aspx [accessed 2025-01-04]
21. Zaidi NB, Hwang C, Scott S, Stallard S, Purkiss J, Hortsch M. Climbing Bloom's taxonomy pyramid: lessons from a graduate histology course. Anat Sci Ed 2017 Sep;10(5):456-464. [doi: 10.1002/ase.1685]
22. Krathwohl DR. A revision of Bloom's taxonomy: an overview. Theor Pract 2002 Nov 1;41(4):212-218. [doi: 10.1207/s15430421tip4104_2]
23. Webb NL. Depth-of-knowledge levels for four content areas. LA 2002;28(March):1-9.
24. Giannos P, Delardas O. Performance of ChatGPT on UK standardized admission tests: insights from the BMAT, TMUA, LNAT, and TSA examinations. JMIR Med Educ 2023 Apr 26;9:e47737. [doi: 10.2196/47737] [Medline: 37099373]
25. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
26. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. 2022 Presented at: 36th Conference on Neural Information Processing Systems; Nov 28 to Dec 9, 2022; New Orleans, LA, USA p. 24824-24837.
27. Huang Y, Gomaa A, Semrau S, et al. Benchmarking ChatGPT-4 on a radiation oncology in-training exam and Red Journal Gray Zone cases: potentials and challenges for ai-assisted medical education and decision making in radiation oncology. Front Oncol 2023;13:1265024. [doi: 10.3389/fonc.2023.1265024] [Medline: 37790756]
28. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. Neurosurgery 2023 Dec 1;93(6):1353-1365. [doi: 10.1227/neu.0000000000002632] [Medline: 37581444]
29. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. J Educ Eval Health Prof 2023;20:1. [doi: 10.3352/jeehp.2023.20.1] [Medline: 36627845]
30. Fang C, Wu Y, Fu W, et al. How does ChatGPT-4 preform on non-English national medical licensing examination? An evaluation in Chinese language. PLOS Dig Health 2023 Dec;2(12):e0000397. [doi: 10.1371/journal.pdig.0000397] [Medline: 38039286]
31. Zong H, Li J, Wu E, Wu R, Lu J, Shen B. Performance of ChatGPT on Chinese national medical licensing examinations: a five-year examination evaluation study for physicians, pharmacists and nurses. BMC Med Educ 2024 Feb 14;24(1):143. [doi: 10.1186/s12909-024-05125-7] [Medline: 38355517]
32. Aljindan FK, Al Qurashi AA, Albalawi IAS, et al. ChatGPT conquers the Saudi Medical Licensing Exam: exploring the accuracy of artificial intelligence in medical knowledge assessment and implications for modern medical education. Cureus 2023 Sep;15(9):e45043. [doi: 10.7759/cureus.45043] [Medline: 37829968]
33. Watari T, Takagi S, Sakaguchi K, et al. Performance comparison of ChatGPT-4 and Japanese medical residents in the General Medicine In-Training Examination: comparison study. JMIR Med Educ 2023 Dec 6;9:e52202. [doi: 10.2196/52202] [Medline: 38055323]
34. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. Sci Rep 2023 Nov 22;13(1):20512. [doi: 10.1038/s41598-023-46995-z] [Medline: 37993519]
35. Flores-Cohaila JA, García-Vicente A, Vizcarra-Jiménez SF, et al. Performance of ChatGPT on the Peruvian National Licensing Medical Examination: cross-sectional study. JMIR Med Educ 2023 Sep 28;9:e48039. [doi: 10.2196/48039] [Medline: 37768724]
36. Liu M, Okuhara T, Chang X, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. J Med Internet Res 2024 Jul 25;26:e60807. [doi: 10.2196/60807] [Medline: 39052324]
37. Wang X, Gong Z, Wang G, et al. ChatGPT performs on the Chinese National Medical Licensing Examination. J Med Syst 2023 Aug 15;47(1):86. [doi: 10.1007/s10916-023-01961-0] [Medline: 37581690]
38. Huang H. Performance of ChatGPT on Registered Nurse License Exam in Taiwan: a descriptive study. Healthcare (Basel) 2023 Oct 30;11(21):2855. [doi: 10.3390/healthcare11212855] [Medline: 37958000]
39. Scaioli G, Lo Moro G, Conrado F, Rosset L, Bert F, Siliquini R. Exploring the potential of ChatGPT for clinical reasoning and decision-making: a cross-sectional study on the Italian Medical Residency Exam. Ann Ist Super Sanita 2023;59(4):267-270. [doi: 10.4415/ANN_23_04_05] [Medline: 38088393]
40. Liu F, Chen N, Wang R, Zhang L, Li Y. Visual analysis of allergic rhinitis in children based on web of science and CiteSpace software. Front Pediatr 2022;10:911293. [doi: 10.3389/fped.2022.911293] [Medline: 36245734]
41. Lin PY, Chu CH, Chang FY, Huang YW, Tsai HJ, Yao TC. Trends and prescription patterns of traditional Chinese medicine use among subjects with allergic diseases: a nationwide population-based study. World Allergy Organ J 2019;12(2):100001. [doi: 10.1016/j.waojou.2018.11.001] [Medline: 30937136]
42. Rejeleene R, Xu X, Talburt J. Towards trustable language models: investigating information quality of large language models. arXiv. Preprint posted online on 2024.

43. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nat New Biol 2023 Aug;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
44. Chen PH, Cheng S, Chen WL, Lin YT, Chen YN. Measuring Taiwanese Mandarin language understanding. arXiv. Preprint posted online on Mar 29, 2024arXiv:2403.20180.
45. Lin YT, Chen YN. Taiwan llm: bridging the linguistic divide with a culturally aligned language model. arXiv. Preprint posted online on Nov 29, 2023arXiv:2311.17487.
46. Taiwan Ministry of Examination - Examination Statistics [Website in Mandarin]. URL: https://wwwc.moex.gov.tw/main/examreport/wfrmexamstatistics.aspx?menu_id=158 [accessed 2025-01-04]
47. Ting YT, Hsieh TC, Wang YF, et al. Performance of ChatGPT incorporated chain-of-thought method in bilingual nuclear medicine physician board examinations. Dig Health 2024;10:20552076231224074. [doi: 10.1177/20552076231224074] [Medline: 38188855]
48. Zhu L, Mou W, Yang T, Chen R. ChatGPT can pass the AHA exams: open-ended questions outperform multiple-choice format. Resuscitation 2023 Jul;188:109783. [doi: 10.1016/j.resuscitation.2023.109783] [Medline: 37349064]

Abbreviations

AI: artificial intelligence
aOR: adjusted odds ratio
DOK: depth of knowledge
ENT: ears, nose, and throat
GPT: generative pretrained transformer
GYN/OBS: gynecology/obstetrics
HOTS: higher-order thinking skills
LLM: large language model
LOTS: lower-order thinking skills
SAMC: single-answer multiple-choice
SAMRMC: single-answer, multiple-response multiple-choice
TCM: traditional Chinese medicine

Edited by B Lesselroth; submitted 27.03.24; peer-reviewed by B Shen, T Kikuchi, Z Hou; revised version received 27.07.24; accepted 09.11.24; published 19.03.25.

Please cite as:
Tseng LW, Lu YC, Tseng LC, Chen YC, Chen HY
Performance of ChatGPT-4 on Taiwanese Traditional Chinese Medicine Licensing Examinations: Cross-Sectional Study
JMIR Med Educ 2025;11:e58897
URL: https://mededu.jmir.org/2025/1/e58897
doi:10.2196/58897

Generative Artificial Intelligence in Medical Education—Policies and Training at US Osteopathic Medical Schools: Descriptive Cross-Sectional Survey

Tsunagu Ichikawa[1*], BS; Elizabeth Olsen[2*], BS, MS; Arathi Vinod[3*], BA; Noah Glenn[4*]; Karim Hanna[5*], MD; Gregg C Lund[*], MS, DO; Stacey Pierce-Talsma[1*], MS, DO

[1]College of Osteopathic Medicine, University of New England, 11 Hills Beach Road, Biddeford, ME, United States

[2]College of Osteopathic Medicine, Rocky Vista University, Parker, CO, United States

[3]College of Osteopathic Medicine, Touro University California, Vallejo, CA, United States

[4]McCombs School of Business, University of Texas at Austin, Austin, TX, United States

[5]Morsani College of Medicine, University of South Florida, Tampa, FL, United States

[*]all authors contributed equally

Corresponding Author:
Stacey Pierce-Talsma, MS, DO
College of Osteopathic Medicine, University of New England, 11 Hills Beach Road, Biddeford, ME, United States

Abstract

Background: Interest has recently increased in generative artificial intelligence (GenAI), a subset of artificial intelligence that can create new content. Although the publicly available GenAI tools are not specifically trained in the medical domain, they have demonstrated proficiency in a wide range of medical assessments. The future integration of GenAI in medicine remains unknown. However, the rapid availability of GenAI with a chat interface and the potential risks and benefits are the focus of great interest. As with any significant medical advancement or change, medical schools must adapt their curricula to equip students with the skills necessary to become successful physicians. Furthermore, medical schools must ensure that faculty members have the skills to harness these new opportunities to increase their effectiveness as educators. How medical schools currently fulfill their responsibilities is unclear. Colleges of Osteopathic Medicine (COMs) in the United States currently train a significant proportion of the total number of medical students. These COMs are in academic settings ranging from large public research universities to small private institutions. Therefore, studying COMs will offer a representative sample of the current GenAI integration in medical education.

Objective: This study aims to describe the policies and training regarding the specific aspect of GenAI in US COMs, targeting students, faculty, and administrators.

Methods: Web-based surveys were sent to deans and Student Government Association (SGA) presidents of the main campuses of fully accredited US COMs. The dean survey included questions regarding current and planned policies and training related to GenAI for students, faculty, and administrators. The SGA president survey included only those questions related to current student policies and training.

Results: Responses were received from 81% (26/32) of COMs surveyed. This included 47% (15/32) of the deans and 50% (16/32) of the SGA presidents (with 5 COMs represented by both the deans and the SGA presidents). Most COMs did not have a policy on the student use of GenAI, as reported by the dean (14/15, 93%) and the SGA president (14/16, 88%). Of the COMs with no policy, 79% (11/14) had no formal plans for policy development. Only 1 COM had training for students, which focused entirely on the ethics of using GenAI. Most COMs had no formal plans to provide mandatory (11/14, 79%) or elective (11/15, 73%) training. No COM had GenAI policies for faculty or administrators. Eighty percent had no formal plans for policy development. Furthermore, 33.3% (5/15) of COMs had faculty or administrator GenAI training. Except for examination question development, there was no training to increase faculty or administrator capabilities and efficiency or to decrease their workload.

Conclusions: The survey revealed that most COMs lack GenAI policies and training for students, faculty, and administrators. The few institutions with policies or training were extremely limited in scope. Most institutions without current training or policies had no formal plans for development. The lack of current policies and training initiatives suggests inadequate preparedness for integrating GenAI into the medical school environment, therefore, relegating the responsibility for ethical guidance and training to the individual COM member.

(JMIR Med Educ 2025;11:e58766) doi:10.2196/58766

KEYWORDS

artificial intelligence; medical education; faculty development; policy; AI; training; United States; school; university; college; institution; osteopathic; osteopathy; curriculum; student; faculty; administrator; survey; cross-sectional

Introduction

Artificial intelligence (AI) is a technology capable of performing tasks traditionally requiring human intelligence [1]. AI has a long-standing presence in medicine across clinical, educational, and administrative domains [2-4]. Generative artificial intelligence (GenAI) technologies are a subset of AI that can create new content.

In the clinical domain, GenAI has demonstrated proficiency in performing tasks ranging from passing the United States Medical Licensing Examination to providing empathetic patient communication [5,6]. At a more advanced level, these tools have answered real-world medical questions with more factual accuracy and more empathy than human physicians [7,8]. Such capabilities highlight GenAI's potential as a pivotal tool in both the learning environment of medical students and the broader context of patient care. However, the integration of GenAI into medical education raises important questions regarding the ethical, legal, and practical implications of its use.

Increased computing power, the development of a user-friendly conversational interface that lowers the technical barriers to use, and the availability to the public at little or no direct cost have made this technology nearly as available as web-based search engines or document spell-checking for medical educators and students. This has stimulated a great deal of interest by all constituencies in medicine and medical education. GenAI is only 1 component of the general field of AI. However, with the recent nearly ubiquitous availability to the general population in the United States, the yet clearly defined risks and benefits have significant implications for the short term in all aspects of medicine and the need for training and policies for medical trainees.

The rapid evolution of GenAI highlights the responsibility of medical schools to take a proactive approach to adapt their curricula and policies to harness the benefits of these technologies while mitigating potential risks. How medical schools currently fulfill their responsibilities is unclear. There are published reports highlighting individual AI-related training programs, as well as recommendations for AI curriculum, content, delivery, and challenges in medical schools [9-11]. While insightful, they do not describe the full educational landscape of US medical schools that grant either DO or MD degrees. This is particularly crucial in Colleges of Osteopathic Medicine (COMs) in the United States, which account for a significant and growing proportion of the country's medical student population. Understanding the current landscape of GenAI policies and training in COMs is essential for identifying gaps, setting benchmarks, and guiding future initiatives aimed at effectively integrating GenAI into medical education.

GenAI has rapidly become nearly ubiquitous in the United States and has the potential for significant benefits and risks. It is unclear whether COMs have included training or policy guidance in this domain. This study aimed to describe the status of policy and training, specifically in one aspect of AI, GenAI, for medical students, faculty, and administrators, as well as near-term plans for policy and training development at COMs. This analysis will provide an overview of the current state of GenAI integration in osteopathic medical education, which will demonstrate opportunities for future development.

Methods

Study Design and Population

This descriptive cross-sectional study targeted US COMs that held full accreditation by the Commission on Osteopathic College Accreditation as of the end of the 2022 - 2023 academic year. These COMs have at least 1 graduating class, ensuring that they possess a comprehensive experience with the full spectrum of undergraduate medical education. Approximately 28% of all US medical students are enrolled in COMs [12,13] in academic settings ranging from large public research universities to small private institutions. Therefore, we believe that studying COMs will offer a representative sample of the current GenAI integration in US medical education.

Ethical Considerations

Before initiating contact with potential participants, the institutional review board (number 0723-10) of the University of New England, Biddeford, Maine, granted this project an exemption status. Participation in the study was voluntary, and informed consent was provided in both the email invitation and beginning of the survey. Data collection procedures were designed for privacy and confidentiality with deidentification of respondents. There was no compensation for survey participation.

Survey Development and Data Collection

Due to the novel and rapidly developing field of GenAI, a survey was developed using an iterative process to obtain the availability, content, and development plans for training and policies for students, faculty, and administrators. The survey was designed to prioritize the general details of these domains. This strategy was to maximize the survey participation and to provide direction for potential future projects. The design was led by team members with experience in the user interface (GCL), survey development (GCL and SPT), COM medical curriculum development (GCL and SPT), and COM administrative management and operations (GCL and SPT). The survey was tested before implementation with a convenience sample of administrators and students to ensure that the questions were straightforward and the web-based survey system was usable. The order of survey items was the same for all participants in each group, with each question being presented on an individual screen. However, the surveys used an adaptive methodology to expose participants only to pertinent questions. For example, only those participants who answered that they currently provided training would be asked about the

content of the training. If a COM stated that they do not have a GenAI policy, they would be asked about future development.

Data were collected using a survey distributed via a web-based tool (Qualtrics XM). The recruitment for participation was sent by an email directly to the potential participant. The recruitment email described the project purpose and survey details, including that the survey was on the web, anonymous, and no incentives were provided for their participation. No personal data were collected, including the respondent's IP address. Two separate surveys were developed: one for the deans of the COMs and another for the presidents of the Student Government Association (SGA). The dean's survey included questions about current and planned GenAI policies and training for students, faculty, and administrators, as well as questions about the content of existing policies and training (Multimedia Appendix 1). Recognizing that students are unlikely to have knowledge of policy, curriculum planning, or those related to faculty or administrators, the SGA president's survey exclusively encompassed questions about current student policies and training (Multimedia Appendix 2). In both the dean and SGA president recruitment email, the recipient was informed that if there was a more appropriate survey responder, they may forward the email to that person, such as, the dean to an appropriate administrator, and the SGA president to a student.

Each dean and SGA president recruitment email included a unique survey URL to ensure that only 1 response represented each COM for each category. Qualtrics provides distribution data that are separate from the survey results. This allowed follow-up emails to nonresponders while maintaining the anonymity of the data. Data were collected from July 28, 2023, to September 14, 2023.

Data Analysis

Descriptive statistics were used to analyze the survey results. Response rates for both surveys were calculated as the number of completed surveys as a percentage of total COMs surveyed. The number of started but not completed surveys was calculated as a percentage of total COMs surveyed. For each COM not providing training or having a policy, the status of development was reported as the percentage of COMs surveyed without that characteristic. Due to the anonymity of the respondents and the institutional overlap of the dean and SGA presidents, no statistical comparison between the 2 groups was made.

Results

Response Rates

Of the 32 COMs surveyed, 47% (15/32) deans and 50% (16/32) SGA presidents completed the survey. Five surveys overlapped deans and SGA presidents. The dean or SGA president responded from 81% (26/32) of the COMs surveyed, providing a comprehensive understanding of the COMs. All surveys started were completed (100%).

GenAI Policies for Students

A vast majority of COMs reported a lack of established policies regarding the use of GenAI by students. Specifically, 93% (14/15) of deans and 88% (14/16) of SGA presidents indicated that their institutions had no student-focused GenAI policies. Among the few COMs with existing policies, the scope was primarily limited to GenAI use in graded assignments. Of the COMs with no policy, 79% (11/14) had no formal plans for policy development. The stages of planning for student policy are shown in Table 1.

Table . Status of student generative artificial intelligence policy and training development (Colleges of Osteopathic Medicine without policy or training).

	Student GenAI[a] policy	Student mandatory education	Student elective education
Total surveys, n	14	14	15
Status, n (%)			
Not working on a policy or education	3 (21.4)	3 (21.4)	8 (53.3)
Informal conversations	8 (57.1)	8 (57.1)	3 (20)
Workgroup in place	1 (7.1)	3 (21.4)	2 (13.3)
Being drafted and under review	2 (14.3)	0 (0)	1 (6.7)
Approved to take effect after July 1, 2023	0 (0)	0 (0)	1 (6.7)

[a]GenAI: generative artificial intelligence.

GenAI Training for Students

Only 1 COM was identified as having mandatory student training, which focused entirely on the ethics of using GenAI. None of the COMs offered any elective training. Most COMs had no formal plans to provide mandatory (11/14, 79%) or elective (11/15, 73%) training. The stages of planning for student training are shown in Table 1.

GenAI Policies for Faculty or Administrators

None of the COMs studied had a GenAI policy for faculty or administrators. Similar to the students, 80% (12/15) had no formal plans to develop one. The stages of planning for faculty or administrator policy are shown in Table 2.

Table . Status of faculty or administrator generative artificial intelligence policy and training development for Colleges of Osteopathic Medicine (COMs) with no policy or training.

		Faculty/administrator policy	Faculty/administrator training
Total surveys, n		15	10
Status, n (%)			
	We are not working on a policy or training	6 (40)	2 (20)
	Informal conversations	6 (40)	3 (30)
	Workgroup in place	2 (13.3)	2 (20)
	Being drafted and under review	1 (6.7)	3 (30)
	Approved and will take effect after July 1, 2023	0 (0)	0 (0)

GenAI Training for Faculty or Administrators

Only 33.3% (5/15) of COMs had initiated faculty or administrator-focused GenAI training. These predominantly covered basic use and ethical considerations. Except for examination question development, there was no specific focus on skills to enhance educational efficiency or reduce workload (Table 3). Fifty percent (5/10) of the COMs without faculty or administrator training had no formal plans to develop training (Table 2).

Table . Content of current faculty or administrator generative artificial intelligence training.

	Deans, n (%)
Total surveys	5 (100)
How to use the technology	4 (80)
Benefits/limitations of the technology	4 (80)
Ethics of using it	3 (60)
Legal perspective on using it	2 (40)
Development of examination questions	2 (40)

Discussion

Principal Findings

Our survey uncovers a pronounced gap in GenAI policies and training across US COMs, with the vast majority of institutions surveyed lacking formal policy guidelines (93% dean responses and 88% SGA president responses), and of the COMs with no current student policies, 79% (11/14) had no formal plans for future development. Furthermore, no COMs described any student GenAI elective training, with 73% (11/15) reporting no plans for mandatory educational programs. This underscores an urgent GenAI training imperative for medical schools to prepare future physicians for the imminent AI-enhanced health care landscape. Little has been done to support COM faculty to address these needs as no COMs surveyed had a formal policy regarding Gen AI for faculty or administration, 80% (12/15) did not have a plan to develop one, and only 33% (5/15) had focused training mainly in the realm of utilization and ethical considerations.

Comparison With Prior Work

In a recent national survey of US postsecondary schools, 8% had GenAI policies in place [14]. In that report, the focus of the policies was not described. If these were related to students, it is comparable with the data of this project, where 7% (1/15) of the deans or 12% (2/16) of the SGA presidents responded that they had student GenAI policies. In our sample of student GenAI policies, the focus was on using GenAI in graded assignments. While there were few COMs with student-focused policies, none of the COMs had faculty or administrator policies.

The survey results indicated that the status of COM AI policies is unlikely to change significantly in the near future, with few COMs having formal plans to evaluate and develop GenAI policies. The 21% (3/14) of COMs with formal plans for student policies and 20% (3/15) with plans for faculty or administrator policies demonstrate that they are far less engaged than the postsecondary programs, in which 57% are evaluating and developing policies [14].

As with policy, training for COM students, faculty, and administrators is minimal and does not focus on enabling students, faculty, or administrators to increase productivity, improve effectiveness, or decrease workload. Because the majority do not have formal plans to develop training, this situation is unlikely to change significantly in the near future.

Implications for Future Practice

The rapid advancement of AI technologies, including GenAI, necessitates a proactive stance from medical education institutions to integrate these tools effectively and ethically into teaching, learning, and clinical practice. COMs must move more quickly to develop AI policies and training. However, we do not propose indiscriminately replicating the nascent policies or

training approaches of other institutions, which may not be appropriate for their institution. Furthermore, we caution against a hasty and thoughtless development process merely for the sake of establishing provisional measures. Instead, we propose that medical educators and administrators use the growing body of resources to strategically and methodically create policies and training resources using interdisciplinary teams and continually improve them as future GenAI innovations progressively transform the paradigm of technology-assisted human labor.

One example of resources to be reviewed is the study by Chan [15] that presented an AI policy framework integrating their local data and the UNESCO (United Nations Educational, Scientific and Cultural Organization) AI policy guidance [16]. This policy framework is divided into 3 dimensions, governance, operational, and pedagogical, and can also be used as a competency framework, as shown in Table 4.

Table . Artificial intelligence (AI) education policy framework [15].

Domain	Explanation	Content	Leadership
Pedagogical	Teaching and learning aspects of AI integration.	• Rethinking assessments and examinations. Developing student holistic competencies/generic skills • Preparing students for the AI-driven workplace • Encouraging a balanced approach to AI adoption	Teachers
Operational	Practical implementation of AI in university settings	• Monitoring and evaluating AI implementation • Providing training and support for teachers, staff, and students in AI literacy	Teaching and learning and IT staff
Governance	Governance considerations surrounding AI usage in education	• Understanding, identifying, and preventing academic misconduct and ethical dilemmas • Addressing governance of AI: data privacy, transparency, accountability, and security • Attributing AI technologies • Ensuring equity in access to AI	Senior management

Further frameworks for describing AI literacy and learner competencies have emerged [9,10,17-20] and can form a starting point for COMs when developing a curriculum consistent with their institution's educational mission and existing pedagogical architecture. Building upon this framework, in addition to work done internally, the growing body of published content resources can be accessed and, where appropriate, integrated into their development process. Some resources may be adapted from general educational domains, including skills such as writing [21] or faculty development of course content [22]. Other resources are specific to clinical care [20,23], education [24], or ethical use [25,26]. By adopting and evolving these frameworks with growing evidence-based resources, medical schools can ensure that their curricula not only cover the operational aspects of GenAI but also address the ethical, social, and professional implications.

This general framework is appropriate for learners at any developmental stage. However, as in other areas of medical education, the learners' level of training [11,27] must be considered. For faculty or administrators, responsibilities in developing, integrating, and operationalizing the curriculum must also be considered [28].

In addition to the trainee level, medical school policy makers and educators must consider the systems in which future physicians will work. Physicians should be part of a team with diverse backgrounds and professional training to be most effective. With further AI development, these teams will include AI-powered computer assistants. The team must know how to interact effectively and appropriately with this new "team member," including how it affects the patients and families they care for. This awareness is similar to the early assessments of the effects of electronic health records during clinical encounters [29,30].

Implementing GenAI competencies or any new content is a challenge with an already crowded curriculum. We propose that GenAI be integrated into the current system, where other tools are used to minimize the negative effect. When trainees learn to search and evaluate background scientific publications, GenAI can be incorporated where appropriate as one of the tools they are trained with. Furthermore, when practicing for clinical encounters, whether an actual clinical encounter or their objective structured clinical exams, using GenAI as a tutor may potentially reinforce their preparation. There are many similar uses that will integrate GenAI as a tool and not necessitate a significant increase in curriculum time and may additionally make other aspects of their curriculum more effective. However, these efforts will need further evaluation.

By developing clear policies and offering robust training, medical schools can ensure that future physicians are adept at leveraging GenAI to improve health care outcomes while navigating the ethical and professional complexities it presents.

Limitations

This study's findings must be interpreted in light of several limitations. The availability of data limits this project. Ongoing assessment is needed that includes a larger group of medical schools, including those that grant either doctor of osteopathic medicine or doctor of medicine degrees. In addition, other aspects of the physician's life cycle (graduate medical education, clinical practice, and continuing education) must be studied.

The rapidly evolving nature of GenAI requires institutional policies and training initiatives that can quickly adapt, necessitating ongoing research to capture these developments accurately.

Conclusions and Future Directions

Most COMs do not provide AI policy guidance or training for medical students, faculty, or administrators. There also does not seem to be an appropriate prioritization by COMs to remedy this deficiency. While many philosophers, including the great baseball legend Yogi Berra, have opined that "It is difficult to make predictions, especially about the future" [31], this difficulty does not negate medical schools' responsibility while waiting for the future to become clear. They must assess future physicians' needs and implement appropriate training and guidance in their programs. If the COMs do not lead, their trainees will be unprepared for the future. This risks inappropriate use of AI and the medical equivalent to the lawyer who used GenAI to submit a brief in court that included fabricated references or "hallucinations" [32].

Future research should explore effective strategies for implementing GenAI education and policy development, including interdisciplinary approaches and stakeholder engagement.

Acknowledgments

The authors wish to thank Christopher Callaway, PhD, for his assistance with survey design and data collection.

Conflicts of Interest

None declared.

Multimedia Appendix 1
COM (College of Osteopathic Medicine) dean survey.
[DOCX File, 23 KB - mededu_v11i1e58766_app1.docx]

Multimedia Appendix 2
SGA (Student Government Association) president survey.
[DOCX File, 19 KB - mededu_v11i1e58766_app2.docx]

References

1. McCarthy J, Hayes P. Some philosophical problems from the standpoint of artificial intelligence. In: Meltzer B, Michie D, editors. Machine Intelligence 4: Edinburgh University Press.
2. Vyborny CJ, Giger ML. Computer vision and artificial intelligence in mammography. AJR Am J Roentgenol 1994 Mar;162(3):699-708. [doi: 10.2214/ajr.162.3.8109525] [Medline: 8109525]
3. Keel S, Lee PY, Scheetz J, et al. Feasibility and patient acceptability of a novel artificial intelligence-based screening model for diabetic retinopathy at endocrinology outpatient services: a pilot study. Sci Rep 2018 Mar 12;8(1):4330. [doi: 10.1038/s41598-018-22612-2] [Medline: 29531299]
4. Takiddin A, Schneider J, Yang Y, Abd-Alrazaq A, Househ M. Artificial intelligence for skin cancer detection: scoping review. J Med Internet Res 2021 Nov 24;23(11):e22934. [doi: 10.2196/22934] [Medline: 34821566]
5. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
6. Brin D, Sorin V, Vaid A, Soroush A, Glicksberg BS, Charney AW, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep 2023 Oct 1;13(1):16492. [doi: 10.1038/s41598-023-43436-9] [Medline: 37779171]
7. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 1;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
8. Goodman RS, Patrinely JR, Stone CA Jr, Zimmerman E, Donald RR, Chang SS, et al. Accuracy and reliability of chatbot responses to physician questions. JAMA Netw Open 2023 Oct 2;6(10):e2336483. [doi: 10.1001/jamanetworkopen.2023.36483]

9. Lee J, Wu AS, Li D, Kulasegaram KM. Artificial intelligence in undergraduate medical education: a scoping review. Acad Med 2021 Nov 1;96(11S):S62-S70. [doi: 10.1097/ACM.0000000000004291] [Medline: 34348374]
10. Charow R, Jeyakumar T, Younus S, et al. Artificial intelligence education programs for health care professionals: scoping review. JMIR Med Educ 2021 Dec 13;7(4):e31043. [doi: 10.2196/31043] [Medline: 34898458]
11. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019 Dec 3;5(2):e16048. [doi: 10.2196/16048] [Medline: 31793895]
12. American Association of Colleges of Osteopathic Medicine (AACOM). Osteopathic medical college total enrollment by race/ethnicity 2000-2023. 2023 Oct 30. URL: https://www.aacom.org/docs/default-source/research-reports/2000-23-tebycom-re.xlsx?sfvrsn=7ee8361a_24 [accessed 2025-01-30]
13. Association of American Medical Colleges (AAMC). Applicants, matriculants, enrollment, and graduates of US MD-granting medical schools, 2013-2014 through 2022-202. 2013. URL: https://www.aamc.org/media/37816/download?attachment [accessed 2025-01-30]
14. Sebesta J, Davis VL. Supporting instruction and learning through artificial Intelligence: a survey of institutional practices & policies.: WICHE Cooperative for Educational Technologies; 2023. URL: https://wcet.wiche.edu/wp-content/uploads/sites/11/2023/07/AI-Survey-In-Depth-Analysis-Report-Summer-2023.pdf [accessed 2023-09-23]
15. Chan CKY. A comprehensive AI policy education framework for university teaching and learning. Int J Educ Technol High Educ 2023;20(1):1-25. [doi: 10.1186/s41239-023-00408-3]
16. Miao F, Holmes W. Artificial intelligence and education guidance for policymakers United Nations Educational, Scientific and Cultural Organization (UNESCO): Paris, France. 2021. URL: https://unesdoc.unesco.org/ark:/48223/pf0000376709 [accessed 2023-10-17]
17. Kong SC, Man-Yin Cheung W, Zhang G. Evaluation of an artificial intelligence literacy course for university students with diverse study backgrounds. Comput Educ Artif Intell 2021;2:100026. [doi: 10.1016/j.caeai.2021.100026]
18. Ng DTK, Leung JKL, Chu SKW, Qiao MS. Conceptualizing AI literacy: an exploratory review. Comput Educ Artif Intell 2021;2:100041. [doi: 10.1016/j.caeai.2021.100041]
19. McCoy LG, Nagaraj S, Morgado F, Harish V, Das S, Celi LA. What do medical students actually need to know about artificial intelligence? NPJ Digit Med 2020;3:86. [doi: 10.1038/s41746-020-0294-7] [Medline: 32577533]
20. Russell RG, Lovett Novak L, Patel M, et al. Competencies for the use of artificial intelligence-based tools by health care professionals. Acad Med 2023 Mar 1;98(3):348-356. [doi: 10.1097/ACM.0000000000004963] [Medline: 36731054]
21. Byrd A, Flores L, Green D, et al. MLA-CCCC joint task force on writing and AI working paper: overview of the issues, statement of principles, and recommendations. 2023 Jul. URL: https://hcommons.org/app/uploads/sites/1003160/2023/07/MLA-CCCC-Joint-Task-Force-on-Writing-and-AI-Working-Paper-1.pdf [accessed 2023-09-07]
22. Dickey E, Bejarano A. A model for integrating generative AI into course content development. . 2023 23. [doi: 10.48550/arXiv.2308.12276]
23. Ethical application of artificial intelligence in family medicine. American Academy of Family Physicians (AAFP). URL: https://www.aafp.org/about/policies/all/ethical-ai.html [accessed 2023-09-06]
24. Global Forum on Innovation in Health Professional Education, Board on Global Health, Health and Medicine Division, National Academies of Sciences, Engineering, and Medicine. Artificial intelligence in health professions education. In: Proceedings of a Workshop: National Academies Press; 2023:38166138 URL: https://www.nap.edu/catalog/27174 [doi: 10.17226/27174]
25. Nguyen A, Ngo HN, Hong Y, Dang B, Nguyen BPT. Ethical principles for artificial intelligence in education. Educ Inf Technol (Dordr) 2023;28(4):4221-4241. [doi: 10.1007/s10639-022-11316-w] [Medline: 36254344]
26. Weidener L, Fischer M. Teaching AI ethics in medical education: a scoping review of current literature and practices. Perspect Med Educ 2023;12(1):399-410. [doi: 10.5334/pme.954] [Medline: 37868075]
27. Sapci AH, Sapci HA. Artificial intelligence education and tools for medical and health informatics students: systematic review. JMIR Med Educ 2020 Jun 30;6(1):e19285. [doi: 10.2196/19285] [Medline: 32602844]
28. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. Acad Med 2024 Jan 1;99(1):22-27. [doi: 10.1097/ACM.0000000000005439] [Medline: 37651677]
29. Margalit RS, Roter D, Dunevant MA, Larson S, Reis S. Electronic medical record use and physician-patient communication: an observational study of Israeli primary care encounters. Patient Educ Couns 2006 Apr;61(1):134-141. [doi: 10.1016/j.pec.2005.03.004] [Medline: 16533682]
30. Ventres W, Kooienga S, Vuckovic N, Marlin R, Nygren P, Stewart V. Physicians, patients, and the electronic health record: an ethnographic analysis. Ann Fam Med 2006;4(2):124-131. [doi: 10.1370/afm.425] [Medline: 16569715]
31. Dickstein DP. Editorial: it's difficult to make predictions, especially about the future: risk calculators come of age in child psychiatry. J Am Acad Child Adolesc Psychiatry 2021 Aug;60(8):950-951. [doi: 10.1016/j.jaac.2020.12.029] [Medline: 33383160]
32. New York Times. The ChatGPT lawyer explains himself. 2023. URL: https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html [accessed 2023-09-06]

Abbreviations

AI: artificial intelligence
COMs: Colleges of Osteopathic Medicine
GenAI: generative artificial intelligence
SGA: Student Government Association
UNESCO: United Nations Educational, Scientific and Cultural Organization

Edited by B Lesselroth; submitted 24.03.24; peer-reviewed by C Zanetti, M Lin; revised version received 09.10.24; accepted 02.01.25; published 11.02.25.

Please cite as:
Ichikawa T, Olsen E, Vinod A, Glenn N, Hanna K, Lund GC, Pierce-Talsma S
Generative Artificial Intelligence in Medical Education—Policies and Training at US Osteopathic Medical Schools: Descriptive Cross-Sectional Survey
JMIR Med Educ 2025;11:e58766
URL: https://mededu.jmir.org/2025/1/e58766
doi:10.2196/58766

Assessing Familiarity, Usage Patterns, and Attitudes of Medical Students Toward ChatGPT and Other Chat-Based AI Apps in Medical Education: Cross-Sectional Questionnaire Study

Safia Elwaleed Elhassan, MBBS; Muhammad Raihan Sajid, MBBS, MMed; Amina Mariam Syed, MBBS; Sidrah Afreen Fathima, MBBS; Bushra Shehroz Khan, MBBS; Hala Tamim, PhD

College of Medicine, Alfaisal University, Takhasussi street, Riyadh, Saudi Arabia

Corresponding Author:
Muhammad Raihan Sajid, MBBS, MMed
College of Medicine, Alfaisal University, Takhasussi street, Riyadh, Saudi Arabia

Abstract

Background: There has been a rise in the popularity of ChatGPT and other chat-based artificial intelligence (AI) apps in medical education. Despite data being available from other parts of the world, there is a significant lack of information on this topic in medical education and research, particularly in Saudi Arabia.

Objective: The primary objective of the study was to examine the familiarity, usage patterns, and attitudes of Alfaisal University medical students toward ChatGPT and other chat-based AI apps in medical education.

Methods: This was a cross-sectional study conducted from October 8, 2023, through November 22, 2023. A questionnaire was distributed through social media channels to medical students at Alfaisal University who were 18 years or older. Current Alfaisal University medical students in years 1 through 6, of both genders, were exclusively targeted by the questionnaire. The study was approved by Alfaisal University Institutional Review Board. A χ^2 test was conducted to assess the relationships between gender, year of study, familiarity, and reasons for usage.

Results: A total of 293 responses were received, of which 95 (32.4%) were from men and 198 (67.6%) were from women. There were 236 (80.5%) responses from preclinical students and 57 (19.5%) from clinical students, respectively. Overall, males (n=93, 97.9%) showed more familiarity with ChatGPT compared to females (n=180, 90.09%; *P*=.03). Additionally, males also used Google Bard and Microsoft Bing ChatGPT more than females (*P*<.001). Clinical-year students used ChatGPT significantly more for general writing purposes compared to preclinical students (*P*=.005). Additionally, 136 (46.4%) students believed that using ChatGPT and other chat-based AI apps for coursework was ethical, 86 (29.4%) were neutral, and 71 (24.2%) considered it unethical (all *P*s>.05).

Conclusions: Familiarity with and usage of ChatGPT and other chat-based AI apps were common among the students of Alfaisal University. The usage patterns of these apps differ between males and females and between preclinical and clinical-year students.

(JMIR Med Educ 2025;11:e63065) doi:10.2196/63065

KEYWORDS

ChatGPT; artificial intelligence; large language model; medical students; ethics; chat-based; AI apps; medical education; social media; attitude; AI

Introduction

ChatGPT is a sophisticated large language model of artificial intelligence (AI) that was created by OpenAI and released to the public in November 2022 [1]. It generates human-like responses to natural language inputs. The users can hold a conversation with the model where they input a prompt and receive a response [2]. It has many applications including email writing, solving math problems, grammar checking, generating answers to complex questions, and more [3]. Other similar chat-based AI apps include Google Bard, Microsoft Bing ChatGPT, Socrative by Google, Hugging Chat, Snapchat AI, Perplexity AI, and YouChat, among others. All these apps are similar to ChatGPT in terms of generating natural responses to prompts [4].

There has been a rise in new literature pertaining to the use of ChatGPT and other AI tools among medical students. Many published articles show that medical students have a positive attitude toward using ChatGPT in education [5-8]. Many students are eager to use AI tools as they believe it can revolutionize medicine and dentistry [9,10]. Additionally, ChatGPT and other chat-based AIs are continuing to evolve to expand their scope of usage, for example, making virtual histology slides for interactive learning [11-13]. Moreover,

ChatGPT can be used in medical education and medical specialties [14]. Its implications in the cardiovascular, cerebrovascular, and radiology fields are being extensively studied, as it can interpret medical imaging and potentially provide a diagnosis [15-17].

Regarding medical research, ChatGPT and similar AI apps can expedite the writing processes by enabling authors to allocate their time and resources more efficiently, by reducing the time spent on the laborious process of searching for relevant literature [18].

A few studies have been conducted to determine students' willingness to integrate AI tools such as ChatGPT into education. One study demonstrated that both undergraduate and postgraduate students in Hong Kong had a positive attitude toward integrating AI tools into higher education due to its ability to provide immediate solutions, help generate ideas, and handle tedious tasks, allowing students to focus on more important work [6]. Similarly, another study performed on students and faculty at Texas University showed a favorable perception of ChatGPT usage. The responses highlighted the benefits of having access to an AI instructor, which can assist in simplifying concepts by providing examples, offering study advice, and working with students on individual projects [5]. However, the studies were relatively recent and recommend further research, targeting different majors to understand the specialized use of AI in different fields.

Within the Middle East, limited recent studies have assessed medical students' knowledge and attitudes toward AI. A recent study assessed the awareness, perceptions, and opinions of pharmacy undergraduate students toward AI at King Saud University in Riyadh. The findings indicated a generally positive attitude, with demographic factors such as gender and year of study influencing their perceptions [19]. Another qualitative study investigated the knowledge, benefits, concerns, and limitations associated with the use of ChatGPT among medical college faculty and students in Saudi Arabia; the results highlighted both positive aspects such as enhanced communication and learning, and concerns regarding reliability and privacy [20]. Another study conducted at the University of Jordan involving 623 randomly selected medical students demonstrated a strong positive inclination toward using ChatGPT for learning. The findings recommended integrating ChatGPT into the university curricula, emphasizing benefits for students and the potential for misuse [21].

Due to the rise in popularity of ChatGPT and other chat-based AI in medical education, further research must be conducted to understand students' familiarity, usage habits, and attitudes toward these technologies. Despite data from other parts of the world and colleges, there is a significant lack of information on this topic in medical education and research, especially in Saudi Arabia. Therefore, this study was designed to study the familiarity, usage, and attitudes of medical students at Alfaisal University toward ChatGPT and other chat-based AI apps for medical education and research. Furthermore, it explores the perceived limitations, advantages, and ethical concerns that arise from their use. This paper addresses the research question "What are the familiarity, usage patterns, and attitudes of Alfaisal University medical students toward ChatGPT and other chat-based AI apps in medical education?" Based on existing literature, we hypothesize that Alfaisal University medical students are familiar with chat-based AI apps and hold positive attitudes toward their use.

Methods

Study Design and Enrollment

This study was a closed cross-sectional survey that was conducted among medical students at Alfaisal University. Alfaisal University is a private university in Riyadh, Saudi Arabia, that has around 1500 enrolled medical students.

Only current Alfaisal University medical students in years 1 through 6, of both genders, aged 18 years and above were targeted by the questionnaire; students who did not meet the eligibility criteria were not included in the study. The target sample size was calculated to be in the range of 290 - 310 students to achieve a 95% confidence level with a 5% CI, using a sample size calculator.

The online questionnaire was made using Google Forms, a web-based tool used to distribute surveys. The survey was open for responses over 6 weeks, from October 8, 2023, to November 22, 2023. The current survey was modified based on earlier published research [5,6]; the published surveys were chosen in accordance with the IDEE (Identify, Discern, Ethics, Engage) framework, which evaluates how students utilize chat-based AI to achieve specific educational goals, assesses the perceived level of AI integration, examines the effectiveness of AI tools, and explores the ethical considerations involved. The survey was revised to align with our requirements and complement the goals of the study, as previous articles targeted different populations. The answer choices were adapted to reflect the context specific to medical students. The survey was sent to students via email and through Whatsapp groups and other social media outlets including Instagram and Twitter. The survey was designed in accordance with the Checklist for Reporting Results of Internet E-Surveys (CHERRIES) [22,23].

The questionnaire (Multimedia Appendix 1) consisted of 21 questions distributed over 5 pages to assess familiarity, usage, and attitude of medical students toward ChatGPT and other AI apps. The survey consisted of four sections. The first section addressed demographic aspects, including gender (males and females) and academic year (preclinical: years 1 - 3; clinical: years 4 - 6). The second section had questions regarding the knowledge and use of ChatGPT and other chat-based AI apps, including familiarity, frequency of use, and purposes of usage. Participants were asked to rank their familiarity and frequency through Likert-scale questions. For the purpose of usage, the questions were divided according to uses in medical education and medical research. Participants were asked to select all the relevant choices. In the third section, participants were asked to rate their attitudes toward using ChatGPT or other chat-based AI apps in medical training using Likert-scale questions. They rated their beliefs about the enhancement of medical education through such tools, and their intentions to incorporate them into their future learning practices. The final section investigated

ethical considerations of students, including concerns about academic dishonesty.

Descriptive statistics were performed to describe the level of familiarity, reasons for usage, and attitudes of students toward ChatGPT and other chat-based AI apps. A χ^2 test was conducted to assess the relationships between gender and year of studies in terms of familiarity and reasons for usage of ChatGPT and other AI apps. Data analysis was carried out using SPSS (version 29.0; IBM Corp). Statistical significance was set at $P<.05$.

Ethical Considerations

This study received ethical approval from the Institutional Review Board at Alfaisal University (approval number: IRB-20247). The participants were informed of the purpose of the study; the survey was 4 - 5 minutes long and the principal investigator's email was provided for inquiries. The students provided written informed consent to participate in the research. Participation was voluntary and the students were not given any compensation. To maintain confidentiality, no personally identifiable information such as names or college identity numbers were gathered. The responses were only available to the primary investigators and coinvestigators, and data were anonymized.

Results

In total, 293 responses fit the inclusion criteria, 95 (32.4 %) of which were from men and 198 (67.6%) from women. There were 236 (80.5%) responses from preclinical and 57 (19.5%) from clinical students, respectively. Participant familiarity with various AI apps is summarized in Table 1. Most students were familiar with ChatGPT and other chat-based AI apps. However, men used ChatGPT, Google Bard, and Microsoft Bing ChatGPT significantly more than women (all *P*s<.05). Additionally, Socrative by Google was used more by students in the preclinical years when compared to students in clinical years (*P*=0.11) (Table 1).

Table . Familiarity of students toward various artificial intelligence (AI) apps.

AI apps	Total number of responses (N=293), n (%)	Gender			Academic year		
		Men (N=95), n (%)	Women (N=198), n (%)	*P* value	Preclinical (N=236), n (%)	Clinical (N=57), n (%)	*P* value
ChatGPT	273 (93)	93 (97.9)	180 (90.9)	.03	218 (92.4)	55 (96.5)	.27
Google Bard	63 (21.5)	32 (33.7)	31 (15.7)	<.001	50 (21.2)	13 (22.8)	.79
Microsoft Bing ChatGPT	89 (30.4)	41 (43.2)	48 (24.2)	<.001	73 (30.9)	16 (28.1)	.67
Socrative by Google	41 (14)	13 (13.7)	28 (14.1)	.92	39 (16.5)	2 (3.5)	.01
Snapchat AI	181 (61.8)	59 (62.8)	122 (61.6)	.85	145 (61.7)	36 (63.2)	.84
Perplexity AI	7 (2.4)	3 (3.2%)	4 (2)	.55	5 (2.1)	2 (3.5)	.54
YouChat	9 (3.1)	1 (1.1)	8 (4)	.17	9 (3.8)	0 (0.0)	.13
Poe-Telegram-Chatsonic-Replika-Huggingchat	14 (4.8)	7 (7.4)	7 (3.5)	.15	10 (4.2)	4 (7)	.38

Reasons for using various AI apps are summarized in Table 2. Men used ChatGPT for technical questions and solving practice questions significantly more than women (both *P*s<.05). Additionally, clinical students used ChatGPT significantly more for general writing compared to preclinical students (*P*=.005) (Table 2).

Table . Reasons for using various AI apps.

	Total (N=293), n (%)	Gender			Academic years		
		Men (N=95), n (%)	Women (N=198), n (%)	*P* value	Preclinical (N=236), n (%)	Clinical (N=57), n (%)	*P* value
Usage of Chat-GPT/other chat-based apps for medical education							
Asking technical questions	104 (35.5)	44 (46.3)	60 (30.3)	.007	82 (34.7)	22 (38.6)	.59
Asking general knowledge questions/advice on medical issues	113 (38.6)	39 (41.1)	74 (37.4)	.55	86 (36.4)	27 (47.4)	.13
Solving practice questions	84 (28.7)	34 (35.8)	50 (25.3)	.06	73 (30.9)	11 (19.3)	.08
Generating flashcards	34 (11.6)	13 (13.7)	21 (10.6)	.44	29 (12.3)	5 (8.8)	.46
Asking quick questions when stuck on a problem	115 (39.2)	39 (41.1)	76 (38.4)	.66	95 (40.3)	20 (35.1)	.47
Explaining concepts	101 (34.5)	39 (41.1)	62 (31.3)	.10	78 (33.1)	23 (40.4)	.30
Summarizing text	110 (37.5)	32 (33.7)	78 (39.8)	.31	87 (37)	23 (41.1)	.57
Usage of Chat-GPT/other chat-based apps for medical research							
Helping with assignments, making notes, drafting emails	9 (3.1)	3 (3.2)	6 (3)	.95	8 (3.4)	1 (1.8)	.52
General writing	4 (1.4)	2 (2.1)	2 (1)	.45	1 (0.4)	3 (5.3)	.005
Summarizing texts	97 (33.1)	35 (37.2)	62 (31.3)	.32	79 (33.6)	18 (31.6)	.77
Proofreading	59 (20.1)	23 (24.2)	36 (18.2)	.23	48 (20.3)	11 (19.3)	.86
Grammar checking	82 (28)	29 (30.5)	53 (26.8)	.50	67 (28.4)	15 (26.3)	.75
Paraphrasing	121 (41.3)	41 (43.2)	80 (40.4)	.65	91 (38.6)	30 (52.6)	.053
Writing sections of research	46 (15.7)	16 (17)	30 (15.2)	.68	31 (13.2)	15 (26.3)	.02
Generating citations	39 (13.3)	10 (10.5)	29 (14.6)	.33	35 (14.8)	4 (7)	.12
Searching for relevant articles	63 (21.5)	20 (21.1)	43 (21.7)	.90	45 (19.1)	18 (31.6)	.04
Analyzing literature	39 (13.3)	17 (17.9)	22 (11.1)	.11	33 (14)	6 (10.5)	.49

Attitudes and ethical knowledge toward AI apps are reported in Table 3. Notably, the findings showed that 136 (46.4%) of the participants believed using ChatGPT and other chat-based AI apps for coursework was ethical, 86 (29.4%) were neutral, and 71 (24.2%) considered it unethical (all *P*s>.05) (Table 3).

Table . Attitude and ethical knowledge toward AI apps.

Aspect	Agree/ethical, n (%)	Neutral, n (%)	Disagree/nonethical, n (%)
ChatGPT/other chat-based AI apps can enhance my medical education	171 (58.4)	96 (32.8)	26 (8.9)
In the future, I plan to incorporate ChatGPT/other chat-based AI apps into my learning procedures	149 (50.9)	96 (32.8)	48 (16.4)
ChatGPT/other chat-based AI apps can help me save time in medical research	188 (64.2)	79 (27)	26 (8.9)
ChatGPT/other chat-based AI apps can provide me with unique perspectives that I may not have thought of myself	193 (65.9)	80 (27.3)	20 (6.8)
ChatGPT/other chat-based AI apps can provide me with personalized and immediate feedback for my assignments	180 (61.4)	81 (27.6)	32 (10.9)
I can become overreliant on ChatGPT/other chat-based AI apps	104 (35.5)	78 (26.6)	111 (37.9)
ChatGPT/other chat-based AI apps will enable academic dishonest behaviors	226 (77.1)	54 (18.4)	13 (4.4)
I understand ChatGPT/other chat-based AI apps can generate output that is factually inaccurate	206 (70.3)	64 (21.8)	23 (7.8)
To what extent do you think using ChatGPT/other chat-based AI apps is ethical for coursework?	136 (46.4)	86 (29.4)	71 (24.2)

Discussion

Principal Findings

This study investigated the familiarity, usage patterns, and attitudes toward chat-based AI apps among medical students at Alfaisal University, Riyadh, Saudi Arabia. The findings reveal interesting insights into how this technology is integrated into medical education and research.

When evaluating familiarity, it was found that a significant majority of students (>90%) were familiar with ChatGPT, the most popular application. Additionally, male students exhibited a statistically greater familiarity with, and use of certain apps compared to female students. Furthermore, preclinical students were more familiar with Socrative by Google than other AI apps.

For usage, the primary reasons for using chat-based AI were related to medical education, including asking questions, solving practice problems, generating flash cards, and summarizing texts. Nearly 40% of the students reported using AI to ask quick questions when stuck on a problem and explain concepts. While less prevalent, AI was also used for tasks such as summarizing research texts, proofreading, and paraphrasing.

When questioned about attitudes, most students agreed that chat-based AI could enhance learning, save time, and provide unique perspectives. A vast majority of medical students were willing to incorporate ChatGPT and similar AI apps in their learning strategies and believed that it enabled them to save time. It also provided them with unique perspectives and personalized and immediate feedback on their assignments. Despite the positive outlook, a significant portion of students (37.9%) expressed concerns about overreliance on AI; they also had varying opinions regarding the ethical use of AI for coursework. Despite the positive views on chat-based AI for learning, a significant concern emerged among students. Nearly 77% students feared that these AI apps could contribute to academic dishonesty.

Implications of Findings

The findings have significant implications for medical education. The high awareness of chat-based AI, particularly among male students suggests that integrating technology into early medical education could enhance learning outcomes. The varied app usage between preclinical and clinical students highlights the importance of tailored educational tools at different training stages. Furthermore, students' comfort in using AI for daily problem-solving underscores its potential to streamline research workflows and enhance study efficiency, emphasizing the importance of incorporating AI literacy and ethical considerations into curricula.

The findings of this study reinforce the idea that the conventional memory-based medical curriculum, which is primarily memory based, must be followed by advancements in AI. This model has been effective for centuries but demonstrates limitations in the context of the AI age, where

technology is evolving to assist with information retrieval, data processing, and clinical decision-making. While memory and foundational knowledge remain important, there is an increasing need for critical thinking, problem-solving, and technological literacy.

Competence in the efficient integration and using knowledge from an expanding range of sources, including the ethical use of AI must be taught to aspiring doctors[16,24]. These findings, unique to this study, reinforce the importance of using AI in medical education [9,16,21,25]. However, students also acknowledged the potential for misuse, highlighting the importance of clear guidelines and fostering a culture of academic integrity alongside the integration of AI in medical education.

Comparison of Literature

Regarding awareness, our findings are similar to a previously published study from Saudi Arabia that assessed the awareness, perceptions, and opinions toward AI among pharmacy undergraduates. Several students had a positive awareness toward AI and its implications in health care [19].

A cross-sectional study on medical and dental students' perceptions of AI noted a lack of basic AI education in medical and dental schools. Furthermore, raised concerns about AI-competent doctors may replace doctors those less knowledgeable in using AI. This suggests that educational resources are crucial during earlier stages of medical training to keep up with advancements in AI [9]. Additionally, another study showed that pharmacy students deemed it essential to incorporate AI into college curriculum to effectively educate students on apps in the health care field [19].

A study in Canada reported similar results in terms of attitudes toward AI in research. Conducted on Canadian entry-to–health care students, it found that students who were interested in research generally had a more favorable outlook toward AI [26]. This suggests a potential role of AI in enhancing research efficiency.

However, concerns about the overreliance on AI were similarly found in other studies. For instance, a cross-sectional study conducted among pharmacy students in Saudi Arabia found that 46% of students believed that the use of AI reduced the humanistic aspect of health care, while 7.6% believed that AI devalued the medical profession [12]. Similarly, another study conducted on Canadian health care students expressed those concerns that AI could eventually take over their jobs [26].

There were also varying opinions about the ethical use of AI for coursework. A study at the University of Jordan encouraged educators to integrate ChatGPT into medical curricula and teaching practices, while also addressing student concerns and the potential for misuse [21]. Similarly, a cross-sectional study by Weidner and Fischer [27] in German-speaking European countries highlighted the necessity of incorporating teaching AI ethics into the undergraduate medical curricula. This highlights the need for discussion around responsible AI integration in medical education [9,16,25,28].

A previously published study emphasized the potential for misuse, raising concerns that students might rely on ChatGPT to outsource their assessment tasks [29]. Additionally, in a qualitative study conducted at the Faculty of Medicine, Jazan University in Saudi Arabia, respondents expressed ethical concerns related to threats to academic integrity, plagiarism, privacy, and confidentiality issues [7]. Our findings are similar to these studies, highlighting the importance of clear guidelines and fostering a culture of academic integrity [8,30].

Limitations of the Study

This study focuses on self-reported data, which may not always reflect actual practices and can cause information bias. There may be a chance of selection bias due to convenient sampling. Additionally, the results may not be generalizable to other countries, as cultural differences could lead to varying attitudes and responses in different contexts.

Conclusion

Overall, this study provides valuable insights into the growing integration of chat-based AI apps within medical education. As technology evolves, it will be crucial to address ethical concerns and ensure responsible use while maximizing the potential benefits for student learning and research.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey.
[DOCX File, 22 KB - mededu_v11i1e63065_app1.docx]

References

1. Marr B. A short history of ChatGPT: how we got to where we are today. Forbes. URL: https://www.forbes.com/sites/bernardmarr/2023/05/19/a-short-history-of-chatgpt-how-we-got-to-where-we-are-today/ [accessed 2024-01-30]
2. How does ChatGPT work? Zapier. URL: https://zapier.com/blog/how-does-chatgpt-work/ [accessed 2024-01-30]
3. 50 ChatGPT use cases with real-life examples in 2024. AI Multiple Research. URL: https://research.aimultiple.com/chatgpt-use-cases/ [accessed 2024-04-29]
4. The best AI productivity tools in 2024. Zapier. URL: https://zapier.com/blog/best-ai-productivity-tools/ [accessed 2024-04-29]

5. Amani S, White L, Balart T, et al. Generative AI perceptions: a survey to measure the perceptions of faculty, staff, and students on generative AI tools in academia. arXiv. Preprint posted online on Apr 21, 2023. [doi: 10.48550/arXiv.2304.14415]
6. Chan CKY, Hu W. Students' voices on generative AI: perceptions, benefits, and challenges in higher education. Int J Educ Technol High Educ 2023;20(1). [doi: 10.1186/s41239-023-00411-8]
7. Salih SM. Perceptions of faculty and students about use of artificial intelligence in medical education: a qualitative study. Cureus. URL: https://www.cureus.com/articles/237686-perceptions-of-faculty-and-students-about-use-of-artificial-intelligence-in-medical-education-a-qualitative-study [accessed 2024-10-21]
8. Magalhães Araujo S, Cruz-Correia R. Incorporating ChatGPT in medical informatics education: mixed methods study on student perceptions and experiential integration proposals. JMIR Med Educ 2024 Mar 20;10:e51151 [FREE Full text] [doi: 10.2196/51151] [Medline: 38506920]
9. Bisdas S, Topriceanu CC, Zakrzewska Z, et al. Artificial intelligence in medicine: a multinational multi-center survey on the medical and dental students' perception. Front Public Health 2021;9:795284. [doi: 10.3389/fpubh.2021.795284] [Medline: 35004598]
10. Sit C, Srinivasan R, Amlani A, et al. Attitudes and perceptions of UK medical students towards artificial intelligence and radiology: a multicentre survey. Insights Imaging 2020 Feb 5;11(1):14. [doi: 10.1186/s13244-019-0830-7] [Medline: 32025951]
11. Lee H. The rise of ChatGPT: exploring its potential in medical education. Anat Sci Educ 2024;17(5):926-931. [doi: 10.1002/ase.2270] [Medline: 36916887]
12. Alkhaaldi SMI, Kassab CH, Dimassi Z, et al. Medical student experiences and perceptions of ChatGPT and artificial intelligence: cross-sectional study. JMIR Med Educ 2023 Dec 22;9:e51302. [doi: 10.2196/51302] [Medline: 38133911]
13. Veras M, Dyer JO, Rooney M, Barros Silva PG, Rutherford D, Kairy D. Usability and efficacy of artificial intelligence chatbots (ChatGPT) for health sciences students: protocol for a crossover randomized controlled trial. JMIR Res Protoc 2023 Nov 24;12:e51873. [doi: 10.2196/51873] [Medline: 37999958]
14. Mohammad B, Supti T, Alzubaidi M, et al. The pros and cons of using ChatGPT in medical education: a scoping review. Stud Health Technol Inform 2023 Jun 29;305:644-647. [doi: 10.3233/SHTI230580] [Medline: 37387114]
15. Srivastav S, Chandrakar R, Gupta S, et al. ChatGPT in radiology: the advantages and limitations of artificial intelligence for medical imaging diagnosis. Cureus 2023 Jul;15(7):e41435. [doi: 10.7759/cureus.41435] [Medline: 37546142]
16. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019 Dec 3;5(2):e16048. [doi: 10.2196/16048] [Medline: 31793895]
17. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887. [doi: 10.3390/healthcare11060887] [Medline: 36981544]
18. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023;6:1169595 [FREE Full text] [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
19. Syed W, Basil A Al-Rawi M. Assessment of awareness, perceptions, and opinions towards artificial intelligence among healthcare students in Riyadh, Saudi Arabia. Medicina (Kaunas) 2023 Apr 24;59(5):828. [doi: 10.3390/medicina59050828] [Medline: 37241062]
20. Abouammoh N, Alhasan K, Raina R, et al. Exploring perceptions and experiences of ChatGPT in medical education: a qualitative study among medical college faculty and students in Saudi Arabia. medRxiv. Preprint posted online on Jul 16, 2023. [doi: 10.1101/2023.07.13.23292624]
21. Ajlouni AO, Wahba FAA, Almahaireh AS. Students' attitudes towards using ChatGPT as a learning tool: the case of the University of Jordan. Int J Interact Mob Technol 2023;17(18):99-117. [doi: 10.3991/ijim.v17i18.41753]
22. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
23. Su J, Yang W. Unlocking the power of ChatGPT: a framework for applying generative AI in education. ECNU Review of Education 2023;6(3):355-366. [doi: 10.1177/20965311231168423]
24. Wartman SA, Combs CD. Medical education must move from the information age to the age of artificial intelligence. Acad Med 2018 Aug;93(8):1107-1109. [doi: 10.1097/ACM.0000000000002044] [Medline: 29095704]
25. Gong B, Nugent JP, Guest W, et al. Influence of artificial intelligence on Canadian medical students' preference for radiology specialty: a national survey study. Acad Radiol 2019 Apr;26(4):566-577. [doi: 10.1016/j.acra.2018.10.007] [Medline: 30424998]
26. Teng M, Singla R, Yau O, et al. Health care students' perspectives on artificial intelligence: countrywide survey in Canada. JMIR Med Educ 2022 Jan 31;8(1):e33390. [doi: 10.2196/33390] [Medline: 35099397]
27. Weidener L, Fischer M. Artificial intelligence in medicine: cross-sectional study among medical students on application, education, and ethical aspects. JMIR Med Educ 2024 Jan 5;10:e51247. [doi: 10.2196/51247] [Medline: 38180787]
28. Kapsali MZ, Livanis E, Tsalikidis C, Oikonomou P, Voultsos P, Tsaroucha A. Ethical concerns about ChatGPT in healthcare: a useful tool or the tombstone of original and reflective thinking? Cureus 2024 Feb;16(2):e54759. [doi: 10.7759/cureus.54759] [Medline: 38523987]

29. Zhai X. ChatGPT user experience: implications for education. SSRN Journal 2022 Dec 27. [doi: 10.2139/ssrn.4312418]
30. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT - reshaping medical education and clinical management. Pak J Med Sci 2023;39(2):605-607. [doi: 10.12669/pjms.39.2.7653] [Medline: 36950398]

Abbreviations

AI: artificial intelligence
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
IDEE : Identify, Discern, Ethics, Engage

Edited by B Lesselroth; submitted 09.06.24; peer-reviewed by E Bai, I Zaletel; revised version received 21.10.24; accepted 02.01.25; published 30.01.25.

Please cite as:
Elhassan SE, Sajid MR, Syed AM, Fathima SA, Khan BS, Tamim H
Assessing Familiarity, Usage Patterns, and Attitudes of Medical Students Toward ChatGPT and Other Chat-Based AI Apps in Medical Education: Cross-Sectional Questionnaire Study
JMIR Med Educ 2025;11:e63065
URL: https://mededu.jmir.org/2025/1/e63065
doi:10.2196/63065

Performance Evaluation and Implications of Large Language Models in Radiology Board Exams: Prospective Comparative Analysis

Boxiong Wei, MD

Department of Ultrasound, Peking University First Hospital, 8 Xishiku Rd, Xicheng District, Beijing, China

Corresponding Author:
Boxiong Wei, MD
Department of Ultrasound, Peking University First Hospital, 8 Xishiku Rd, Xicheng District, Beijing, China

Abstract

Background: Artificial intelligence advancements have enabled large language models to significantly impact radiology education and diagnostic accuracy.

Objective: This study evaluates the performance of mainstream large language models, including GPT-4, Claude, Bard, Tongyi Qianwen, and Gemini Pro, in radiology board exams.

Methods: A comparative analysis of 150 multiple-choice questions from radiology board exams without images was conducted. Models were assessed on their accuracy for text-based questions and were categorized by cognitive levels and medical specialties using χ^2 tests and ANOVA.

Results: GPT-4 achieved the highest accuracy (83.3%, 125/150), significantly outperforming all other models. Specifically, Claude achieved an accuracy of 62% (93/150; *P*<.001), Bard 54.7% (82/150; *P*<.001), Tongyi Qianwen 70.7% (106/150; *P*=.009), and Gemini Pro 55.3% (83/150; *P*<.001). The odds ratios compared to GPT-4 were 0.33 (95% CI 0.18 - 0.60) for Claude, 0.24 (95% CI 0.13 - 0.44) for Bard, and 0.25 (95% CI 0.14 - 0.45) for Gemini Pro. Tongyi Qianwen performed relatively well with an accuracy of 70.7% (106/150; *P*=0.02) and had an odds ratio of 0.48 (95% CI 0.27 - 0.87) compared to GPT-4. Performance varied across question types and specialties, with GPT-4 excelling in both lower-order and higher-order questions, while Claude and Bard struggled with complex diagnostic questions.

Conclusions: GPT-4 and Tongyi Qianwen show promise in medical education and training. The study emphasizes the need for domain-specific training datasets to enhance large language models' effectiveness in specialized fields like radiology.

(JMIR Med Educ 2025;11:e64284) doi:10.2196/64284

KEYWORDS

large language models; LLM; artificial intelligence; AI; GPT-4; radiology exams; medical education; diagnostics; medical training; radiology; ultrasound

Introduction

Artificial intelligence (AI) in radiology has significantly improved diagnostic accuracy and educational methods for radiologists. By using advanced machine learning and deep learning techniques, AI applications have evolved from enhancing image interpretation to supporting complex diagnostic decisions [1]. These advancements not only increase the efficiency of diagnostic processes but also provide radiologists with interactive training simulations, crucial for their professional growth and certification readiness [2-9].

Recent advancements have also emerged with the development of large language models (LLMs) like GPT-4, Claude, Bard, Tongyi Qianwen and Gemini Pro. These models have added a new aspect to medical education by producing medically accurate content and supporting advanced diagnostic reasoning exercises [10,11]. These features are crucial for establishing safe learning spaces where future radiologists can practice detailed diagnostic reasoning and decision-making without real-world clinical risks [12,13]. Moreover, these LLMs are crucial in developing and clarifying complex medical scenarios and test questions, improving the educational experience and boosting the diagnostic abilities of students [14-16].

Despite these advancements, recent research has pinpointed limitations in the use of LLMs in medical exams, particularly in specialties like radiology that demand extensive clinical insight. Studies have shown that while LLMs such as GPT-4 can manage simple diagnostic questions effectively, they encounter difficulties with more complex cases that require a deeper clinical understanding and the integration of diverse medical information [17,18]. These findings highlight a significant gap in the existing literature; there is a lack of

comprehensive comparative studies that evaluate the performance of various LLMs across different diagnostic scenarios in radiology [19].

This study addresses this gap by comparing several mainstream LLMs in text-based radiology board exams, without imaging components, evaluating their overall performance. While a secondary objective is to analyze performance by question type and topic. This study hypothesizes that GPT-4 will outperform other models, particularly in handling complex diagnostic questions.

Methods

Study Design

This research was structured as a prospective, comparative analysis that aimed to test the effectiveness of various notable LLMs within a controlled environment resembling radiology board examinations without images. The radiology exams comprehensively evaluated a candidate's radiology knowledge, reasoning, and clinical skills. China does not currently have a unified national licensing exam specifically for radiologists. Given that the Canadian Royal College and American Board of Radiology exams are viewed as authoritative and widely recognized, test questions were selected according to the standards of these two exams for model testing [20]. Both of the exams assess candidates on a broad spectrum of radiology topics using multiple-choice questions.

Ethical Considerations

Despite the reliance on nonpersonal, pre-existing data and the lack of direct involvement of human or animal subjects, ethical approval and the need for informed consent were waived by the Institutional Review Board of Peking University First Hospital, Beijing, China. The radiologists who participated in question validation and categorization were compensated at a rate of 300 Chinese Yuan (US $40.91) per hour for their professional expertise. All data used in the study were anonymized exam questions, with no personal identifiable information involved. The research strictly adhered to ethical standards, with data integrity meticulously upheld throughout the study.

Models Selection

The models chosen for this investigation included GPT-4 (OpenAI), Claude 2.1 (Anthropic), Bard (Google, PaLM 2), Tongyi Qianwen (Alibaba, Qwen-72B), and Gemini Pro 1.0 (Google). All models were tested from late November to early December 2023. These models represent significant advancements in AI, particularly in natural language processing. They were selected based on their demonstrated success in academic and professional settings, indicating their potential effectiveness in educational applications.

Dataset Composition

The dataset for this study consisted of 150 multiple-choice questions drawn from historical radiology board exams similar to those given by the Canadian Royal College and the American Board of Radiology. These questions were sourced from the websites of Board Vitals [21] and CanadaQBank [22], which are widely recognized for providing questions that closely reflect the content and format of North American radiology board exams. Each question was individually reviewed and validated by two academic radiologists—one specializing in ultrasound with 20 years of experience and the other in abdominal radiology with 4 years of experience. Questions were only included if both reviewers concurred on their relevance and appropriateness for this study. Questions that involved images were excluded.

Question Categorization

All questions were classified according to their primary assessment objectives using Bloom's Taxonomy, including two main categories: lower-order thinking (remembering and understanding) and higher-order thinking (applying, analyzing, and evaluating) [23]. Higher-order thinking questions were further divided into specific groups such as description and analysis of image findings, application of concepts, clinical management, and calculation and classification. Additionally, questions were also classified based on the specific area of disease focus, including digestive, genitourinary, musculoskeletal, respiratory, cardiovascular (including angiography and intervention), nervous, breast and thyroid, pediatrics, and imaging basics and physics. Each question was reviewed and categorized independently by the two board-certified radiologists mentioned above. Any disagreements were then discussed collectively to arrive at a consensus.

Scoring Criteria

The Canadian Royal College examination uses a pass-fail system based on achieving at least 70% on all written components of the examination. The American Board of Radiology uses a criterion-referenced scoring system. This means that candidates are evaluated against a predefined standard, not in comparison to other test-takers. The passing standard is typically set by a group of experts, including residency program directors and experienced clinicians, who determine the difficulty level of each question to ensure it aligns with the required competency for independent practice. To pass, candidates must meet or exceed the passing standard for all categories scored together. For both exams, the questions undergo psychometric validation, and questions that are not effective in discriminating between candidates or are found too difficult may be removed. The threshold for passing in this study was set at 70% to align with the standards of the Royal College examinations in Canada. This study did not use the criterion-referenced scoring system used by the American Board of Radiology because its standards were difficult to ascertain. Each multiple-choice question was inputted into different LLMs, and the first response from each model was recorded as the subject of analysis.

Statistical Analysis

To evaluate the association between model type and accuracy for categorical variables, χ^2 tests were used. For categories with small sample sizes, the Fisher exact test was used to ensure the validity of the statistical results. Odds ratios and their corresponding 95% CIs were calculated using GPT-4 as the benchmark. ANOVA was used to compare the mean accuracy rates across different models. Following the results from the ANOVA, Tukey's honestly significant difference test was

applied to identify specific pairs of models that demonstrated significant differences in performance. Cohen *d* was calculated to quantify the magnitude of differences between the models, providing a clearer understanding of the practical significance of the findings. Split-half reliability testing was used to assess the consistency of each model's performance across different subsets of data, ensuring the reliability of the models over varied test conditions. Statistical significance was set at an α level of .05.

Results

Overall Model Performance

GPT-4 emerged as the leading model with an accuracy rate of 83.3% (125/150), significantly outperforming its peers. Tongyi Qianwen also displayed strong performance, recording a 70.7% (106/150) accuracy. Moderate effectiveness was observed in models like Claude and Gemini Pro, with accuracy rates of 62.0% (93/150) and 55.3% (83/150), respectively. Bard trailed with a 54.7% (82/150) accuracy rate, highlighting its challenges in handling complex medical data under exam conditions (Table 1).

Table . Performance of different large language models on radiology board–styled multiple-choice questions without images.

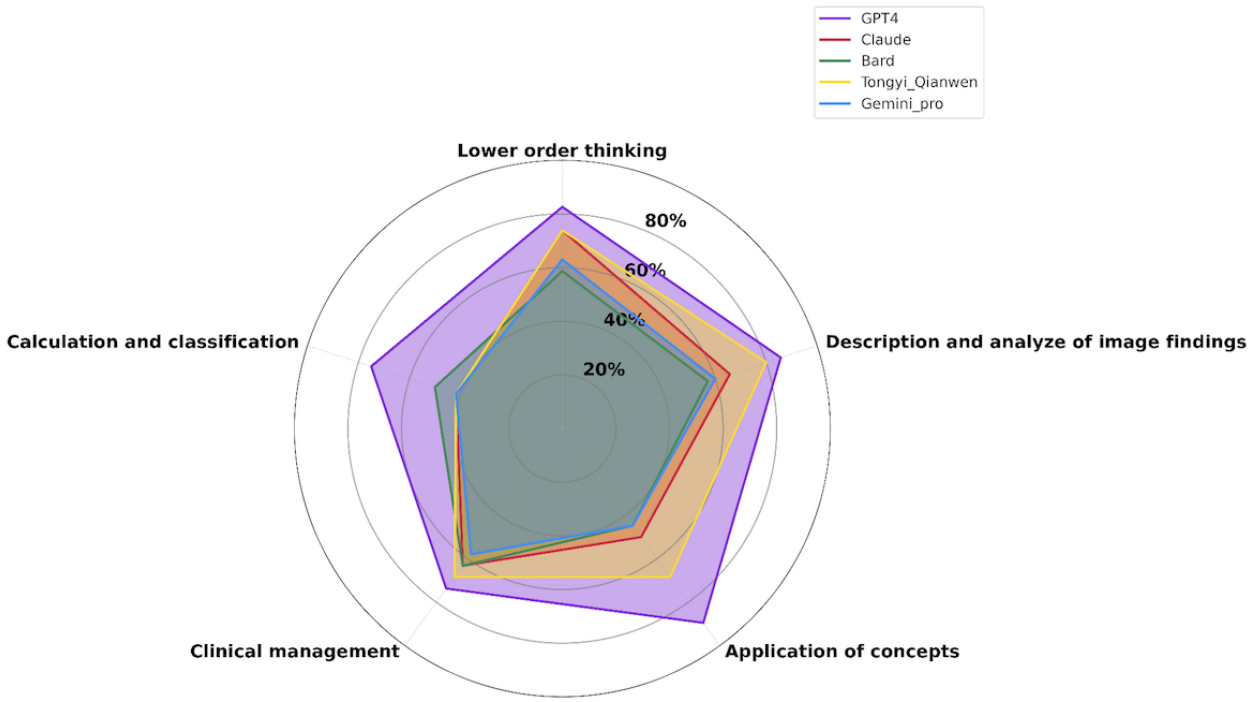
Parameter	Test score, n (%)				
	GPT4	Claude	Bard	Tongyi Qianwen	Gemini Pro
All questions (n=150)	125 (83.3)	93 (62.0)	82 (54.7)	106 (70.7)	83 (55.3)
Question type					
Lower order thinking (n=46)	38 (82.6)	34 (73.9)	27 (58.7)	34 (73.9)	29 (63)
Higher order thinking (n=104)	87 (83.7)	59 (56.7)	55 (52.9)	72 (69.2)	54 (51.9)
Higher order thinking question categories					
Description and analyze of image findings (n=35)	30 (85.7)	23 (65.7)	20 (57.1)	28 (80)	21 (60)
Application of concepts (n=38)	34 (89.5)	19 (50)	17 (44.7)	26 (68.4)	17 (44.7)
Clinical management (n=19)	14 (73.7)	12 (63.2)	12 (63.2)	13 (68.4)	11 (57.9)
Calculation and classification (n=12)	9 (75)	5 (41.7)	6 (50)	5 (41.7)	5 (41.7)
Question topic					
Digestive (n=15)	10 (66.7)	7 (46.7)	5 (33.3)	10 (66.7)	9 (60)
Genitourinary (n=21)	19 (90.5)	15 (71.4)	14 (66.7)	15 (71.4)	11 (52.4)
Musculoskeletal (n=11)	8 (72.7)	6 (54.5)	7 (63.6)	9 (81.8)	7 (63.6)
Respiratory (n=15)	12 (80)	9 (60)	8 (53.3)	8 (53.3)	8 (53.3)
Cardiovascular (n=22)	19 (86.4)	14 (63.6)	8 (36.4)	18 (81.8)	11 (50)
Nervous (n=11)	11 (100)	9 (81.8)	7 (63.6)	8 (72.7)	9 (81.8)
Breast and thyroid (n=14)	11 (78.6)	9 (64.3)	9 (64.3)	9 (64.3)	7 (50)
Pediatrics (n=19)	15 (78.9)	11 (57.9)	11 (57.9)	13 (68.4)	9 (47.4)
Imaging Basics and physics (n=22)	19 (86.4)	11 (50)	12 (54.5)	15 (68.2)	12 (54.5)

Detailed Performance Analysis by Question Type

The breakdown by question type revealed that GPT-4 consistently excelled in both lower-order and higher-order thinking questions, scoring 82.6% (38/46) and 83.7% (87/104), respectively. This indicated GPT-4's capability to manage both basic recall and more complex analytical tasks effectively. In contrast, models such as Claude and Bard demonstrated a drop in performance with higher-order thinking questions, achieving only 56.7% (59/104) and 52.9% (55/104) accuracy in this

category, respectively. This gradient in performance highlighted the difficulties faced by current LLMs in simulating the complex cognitive processes involved in clinical reasoning (Figure 1).

Figure 1. Model accuracy by question type, illustrating the differentiation in model performance between lower-order and higher-order thinking questions.



Performance Across Medical Specialties

Performance analysis segmented by medical specialty showed marked variances. GPT-4 demonstrated exceptional proficiency in neurology with a perfect score of 100% (11/11), and also performed well in genitourinary and cardiovascular categories, with accuracies of 90.5% (19/21) and 86.4% (19/22), respectively. However, challenges were apparent in areas like musculoskeletal and digestive categories, where high-performing models like GPT-4 experienced reduced accuracy rates of 72.7% (8/11) and 66.7% (10/15), respectively. These results indicated that some specialties may need more tailored domain-specific training for models to enhance their effectiveness (Table 1).

Detailed odds ratios and CIs for each model are presented in Multimedia Appendix 1. The odds ratio results show that GPT-4 had the highest performance. All the other models had significantly lower odds ratios compared to GPT-4. Tongyi Qianwen had the highest odds ratio among the other models. As shown in Multimedia Appendix 2, the pairwise comparisons showed that GPT-4 significantly outperformed all other models, with statistically significant differences observed in its comparison with Claude (P<.001), Bard (P<.001), Tongyi Qianwen (P=.009), and Gemini Pro (P<.001). Additionally, Tongyi Qianwen exhibited a significantly higher accuracy compared to Bard (P=.004) and Gemini Pro (P=.006). In contrast, no statistically significant differences were found between Claude and Bard (P=.20), Claude and Gemini Pro (P=.24), or Bard and Gemini Pro (P=.90). These results suggest that the performance of these models was relatively similar in this dataset.

Discussion

Principal Findings

The exceptional performance of GPT-4 in this study aligns with recent findings that highlight its advanced reasoning capabilities and improvements over previous versions, such as GPT-3.5, in various professional contexts, including various kinds of medical exams [24]. GPT-4's extensive training on diverse datasets and its refined architecture enable it to adeptly handle complex questions, which are typical in the specialized language and scenario-based queries found in medical board examinations [25]. Nevertheless, the performance differences observed among models like Bard and Claude can be attributed to the nature of their training and inherent limitations in processing complex cognitive tasks, which are crucial in radiology examinations. This is largely due to the absence of specialized medical training data during their development phases. These findings are in line with the research, which indicated that while GPT-4's textual reasoning is strong, its integration and analysis of image-based information remains inadequate [26].

Models such as GPT-4 and Tongyi Qianwen, which displayed superior performance, likely benefited from training datasets that included medical scenarios. The significance of domain-specific training is well-documented, emphasizing that for LLMs to excel in specialized fields like radiology, they require training with pertinent medical data. Both GPT-4 and Tongyi Qianwen exceeded the 70% passing threshold for the simulated radiology board exams. This marks a significant achievement and shows the potential of these models in

academic and professional environments. The threshold mirrors real medical licensing exam criteria, offering a realistic measure of AI's potential performance in actual educational assessments. The robust performance of Tongyi Qianwen, particularly in an English-based setup, is notable. Despite generally not being ranked as highly as Western models in AI benchmarks, its performance indicates significant progress in China's AI development [27]. This supports calls for more inclusive and diverse training datasets to reduce biases and improve the global applicability of AI technologies.

GPT-4 has demonstrated the capability to pass simulated UK Radiology Fellowship Examinations, especially in sections focused on physics and single best answers [28]. However, challenges remain when these models are tested with image-based questions, highlighting a persisting gap between current AI capabilities and the complex demands of radiological diagnostics [26]. While integrating LLMs into medical education and assessments promises transformative changes in how content is delivered and evaluated, there is a risk of excessive reliance on AI. This overdependence could potentially undermine the development of critical thinking and diagnostic skills vital for medical practice [25].

Limitations

This study's limitations include its sole focus on text-based questions and the exclusion of visual components, which are integral to radiology. Future research should incorporate multimodal assessments and also aim to integrate image recognition capabilities with textual analysis to improve the applicability of LLMs in radiology. These models will need to be fine-tuned with domain-specific datasets to enhance their practical utility in medical education and clinical diagnostics. Another notable limitation is the delay between the submission and publication of peer-reviewed articles, which can result in outdated assessments of rapidly evolving LLMs. The models evaluated in this paper were based on their versions from late November to early December 2023, and significant advancements have occurred since then, particularly with models like Claude, which has been regularly updated, with multiple new versions released by Anthropic. In future work, we intend to continue discussing the accuracy comparisons among new models as they are released. Additionally, if sufficient technical resources are available, we aim to create a platform to maintain an up-to-date database of LLM performance on this benchmark.

Conclusion

This article underscores the evolving capabilities and limitations of LLMs in medical education. While models like GPT-4 show promise, the path to their effective integration in clinical practice requires ongoing refinement and a deeper understanding of their operational dynamics in complex medical settings.

Data Availability

The data sets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The odds ratios and CIs of each model using GPT-4 as the benchmark.
[DOCX File, 17 KB - mededu_v11i1e64284_app1.docx]

Multimedia Appendix 2
Hypothetical pairwise comparison table.
[DOCX File, 17 KB - mededu_v11i1e64284_app2.docx]

References

1. Aggarwal R, Sounderajah V, Martin G, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digit Med 2021 Apr 7;4(1):65. [doi: 10.1038/s41746-021-00438-z] [Medline: 33828217]
2. Cabitza F, Campagner A, Balsano C. Bridging the "last mile" gap between AI implementation and operation: "data awareness" that matters. Ann Transl Med 2020 Apr;8(7):501. [doi: 10.21037/atm.2020.03.63] [Medline: 32395545]
3. Thrall JH, Li X, Li Q, et al. Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success. J Am Coll Radiol 2018 Mar;15(3 Pt B):504-508. [doi: 10.1016/j.jacr.2017.12.026] [Medline: 29402533]
4. Akinci D'Antonoli T, Stanzione A, Bluethgen C, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. Diagn Interv Radiol 2024 Mar 6;30(2):80-90. [doi: 10.4274/dir.2023.232417] [Medline: 37789676]
5. Nassiri K, Akhloufi MA. Recent advances in large language models for healthcare. BioMed Inform 2024;4(2):1097-1143. [doi: 10.3390/biomedinformatics4020062]

6. Duong MT, Rauschecker AM, Rudie JD, et al. Artificial intelligence for precision education in radiology. Br J Radiol 2019 Nov;92(1103):20190389. [doi: 10.1259/bjr.20190389] [Medline: 31322909]
7. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health 2019 Oct;1(6):e271-e297. [doi: 10.1016/S2589-7500(19)30123-2] [Medline: 33323251]
8. Papadimitroulas P, Brocki L, Christopher Chung N, et al. Artificial intelligence: deep learning in oncological radiomics and challenges of interpretability and data harmonization. Phys Med 2021 Mar;83:108-121. [doi: 10.1016/j.ejmp.2021.03.009] [Medline: 33765601]
9. Seah JCY, Tang CHM, Buchlak QD, et al. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. Lancet Digit Health 2021 Aug;3(8):e496-e506. [doi: 10.1016/S2589-7500(21)00106-0] [Medline: 34219054]
10. Nichols JA, Herbert Chan HW, Baker MAB. Machine learning: applications of artificial intelligence to imaging and diagnosis. Biophys Rev 2019 Feb;11(1):111-118. [doi: 10.1007/s12551-018-0449-9] [Medline: 30182201]
11. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 2020 Mar 25;368:m689. [doi: 10.1136/bmj.m689] [Medline: 32213531]
12. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. 2020 Presented at: 34th Conference on Neural Information Processing Systems; Dec 6-12, 2020; Vancouver, Canada p. 1877-1901 URL: https://dl.acm.org/doi/abs/10.5555/3495724.3495883 [accessed 2025-01-06]
13. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Presented at: 31st International Conference on Neural Information Processing Systems; Dec 4-9, 2017; Long Beach, CA p. 6000-6010 URL: https://dl.acm.org/doi/10.5555/3295222.3295349 [accessed 2025-01-06]
14. Yu AC, Mohajer B, Eng J. External validation of deep learning algorithms for radiologic diagnosis: a systematic review. Radiol Artif Intell 2022 May;4(3):e210064. [doi: 10.1148/ryai.210064] [Medline: 35652114]
15. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiology 2018 Mar;286(3):800-809. [doi: 10.1148/radiol.2017171920] [Medline: 29309734]
16. Fischer AM, Eid M, De Cecco CN, et al. Accuracy of an artificial intelligence deep learning algorithm implementing a recurrent neural network with long short-term memory for the automated detection of calcified plaques from coronary computed tomography angiography. J Thorac Imaging 2020 May;35 Suppl 1:S49-S57. [doi: 10.1097/RTI.0000000000000491] [Medline: 32168163]
17. McBee MP, Awan OA, Colucci AT, et al. Deep learning in radiology. Acad Radiol 2018 Nov;25(11):1472-1480. [doi: 10.1016/j.acra.2018.02.018] [Medline: 29606338]
18. Li L, Qin L, Xu Z, et al. Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: evaluation of the diagnostic accuracy. Radiology 2020 Aug;296(2):E65-E71. [doi: 10.1148/radiol.2020200905] [Medline: 32191588]
19. Rauschecker AM, Rudie JD, Xie L, et al. Artificial intelligence system approaching neuroradiologist-level differential diagnosis accuracy at brain MRI. Radiology 2020 Jun;295(3):626-637. [doi: 10.1148/radiol.2020190283] [Medline: 32255417]
20. Wang YE, Liu M, Jin L, et al. Radiology education in China. J Am Coll Radiol 2013 Mar;10(3):213-219. [doi: 10.1016/j.jacr.2012.11.006] [Medline: 23571062]
21. The standard in healthcare board exam prep & CME. Board Vitals. URL: https://www.boardvitals.com/ [accessed 2025-01-06]
22. CanadaQBank. URL: https://www.canadaqbank.com/ [accessed 2025-01-06]
23. Krathwohl DR. A revision of Bloom's taxonomy: an overview. Theor Pract 2002 Nov 1;41(4):212-218. [doi: 10.1207/s15430421tip4104_2]
24. Bhayana R, Bleakney RR, Krishna S. GPT-4 in radiology: improvements in advanced reasoning. Radiology 2023 Jun;307(5):e230987. [doi: 10.1148/radiol.230987] [Medline: 37191491]
25. Lourenco AP, Slanetz PJ, Baird GL. Rise of ChatGPT: it may be time to reassess how we teach and test radiology residents. Radiology 2023 Jun;307(5):e231053. [doi: 10.1148/radiol.231053] [Medline: 37191490]
26. Kim H, Kim P, Joo I, Kim JH, Park CM, Yoon SH. ChatGPT vision for radiological interpretation: an investigation using medical school radiology examinations. Korean J Radiol 2024 Apr;25(4):403-406. [doi: 10.3348/kjr.2024.0017] [Medline: 38528699]
27. Jiang L, Wu Z, Xu X, et al. Opportunities and challenges of artificial intelligence in the medical field: current application, emerging problems, and problem-solving strategies. J Int Med Res 2021 Mar;49(3):3000605211000157. [doi: 10.1177/03000605211000157] [Medline: 33771068]
28. Ariyaratne S, Jenko N, Mark Davies A, Iyengar KP, Botchu R. Could ChatGPT pass the UK radiology fellowship examinations? Acad Radiol 2024 May;31(5):2178-2182. [doi: 10.1016/j.acra.2023.11.026] [Medline: 38160089]

Abbreviations

AI: artificial intelligence
LLM: large language model

Edited by B Lesselroth; submitted 14.07.24; peer-reviewed by B Thies, R Yin; revised version received 10.10.24; accepted 03.12.24; published 16.01.25.

Please cite as:
Wei B
Performance Evaluation and Implications of Large Language Models in Radiology Board Exams: Prospective Comparative Analysis
JMIR Med Educ 2025;11:e64284
URL: https://mededu.jmir.org/2025/1/e64284
doi:10.2196/64284

Factors Associated With the Accuracy of Large Language Models in Basic Medical Science Examinations: Cross-Sectional Study

Naritsaret Kaewboonlert, MD; Jiraphon Poontananggul, MD; Natthipong Pongsuwan, MD; Gun Bhakdisongkhram, PhD, MD

Institute of Medicine, Suranaree University of Technology, 111 University Avenue, Nakhon Ratchasima, Thailand

Corresponding Author:
Naritsaret Kaewboonlert, MD
Institute of Medicine, Suranaree University of Technology, 111 University Avenue, Nakhon Ratchasima, Thailand

Abstract

Background: Artificial intelligence (AI) has become widely applied across many fields, including medical education. Content validation and its answers are based on training datasets and the optimization of each model. The accuracy of large language model (LLMs) in basic medical examinations and factors related to their accuracy have also been explored.

Objective: We evaluated factors associated with the accuracy of LLMs (GPT-3.5, GPT-4, Google Bard, and Microsoft Bing) in answering multiple-choice questions from basic medical science examinations.

Methods: We used questions that were closely aligned with the content and topic distribution of Thailand's Step 1 National Medical Licensing Examination. Variables such as the difficulty index, discrimination index, and question characteristics were collected. These questions were then simultaneously input into ChatGPT (with GPT-3.5 and GPT-4), Microsoft Bing, and Google Bard, and their responses were recorded. The accuracy of these LLMs and the associated factors were analyzed using multivariable logistic regression. This analysis aimed to assess the effect of various factors on model accuracy, with results reported as odds ratios (ORs).

Results: The study revealed that GPT-4 was the top-performing model, with an overall accuracy of 89.07% (95% CI 84.76% - 92.41%), significantly outperforming the others (P<.001). Microsoft Bing followed with an accuracy of 83.69% (95% CI 78.85% - 87.80%), GPT-3.5 at 67.02% (95% CI 61.20% - 72.48%), and Google Bard at 63.83% (95% CI 57.92% - 69.44%). The multivariable logistic regression analysis showed a correlation between question difficulty and model performance, with GPT-4 demonstrating the strongest association. Interestingly, no significant correlation was found between model accuracy and question length, negative wording, clinical scenarios, or the discrimination index for most models, except for Google Bard, which showed varying correlations.

Conclusions: The GPT-4 and Microsoft Bing models demonstrated equal and superior accuracy compared to GPT-3.5 and Google Bard in the domain of basic medical science. The accuracy of these models was significantly influenced by the item's difficulty index, indicating that the LLMs are more accurate when answering easier questions. This suggests that the more accurate models, such as GPT-4 and Bing, can be valuable tools for understanding and learning basic medical science concepts.

(JMIR Med Educ 2025;11:e58898) doi:10.2196/58898

KEYWORDS

accuracy; performance; artificial intelligence; AI; ChatGPT; large language model; LLM; difficulty index; basic medical science examination; cross-sectional study; medical education; datasets; assessment; medical science; tool; Google

Introduction

Advances in artificial intelligence (AI), machine learning, and large language models (LLMs) have made these tools widely used across a variety of industries. Education and other fields are increasingly using these technologies for decision-making and predictive analysis, using machine learning fed by large databases [1]. Their utility has expanded to a wide range of applications, including speech recognition, image categorization, and language translation [2].

The application of computer technologies to study and create models for decision-making, prediction, and simulation is known as machine learning. Model performance is based on training datasets. The incorporation of AI into traditional health care and medical education has had a substantial impact on medical practices [3]. It has accelerated diagnostic processes in radiography [4], pathology, endoscopy, and ultrasonography, has improved clinical decision-making, and has decreased the workloads of health care personnel. AI has had an impact on pharmaceutical development and management and medical education, resulting in a new paradigm [5].

A study on the accuracy of ChatGPT in answering questions that were contextually similar to those in the United States Medical Licensing Examination (USMLE) reported accuracy rates of 44% - 64% for step 1 and 42% - 57.8% for step 2, depending on the dataset [6]. This research indicated that the model's accuracy in answering questions matched the passing score for third-year medical students, suggesting that further development is required for ChatGPT to meet or exceed the USMLE passing criteria [7]. Additionally, the model has the potential to generate insightful content that could aid human learners in studying medical sciences [8].

Evaluations of ChatGPT's accuracy in answering university-level physiology examination questions have shown it can correctly answer more than 75% of them. Furthermore, it can provide explanations that align with expert assessments [9]. For specialized surgical studies, ChatGPT's GPT-4 model, an evolution of GPT-3.5, has been used to assess surgical question accuracy, revealing an overall accuracy of 76.4%, compared to 46.8% with GPT-3.5, a statistically significant difference ($P<.05$). GPT-4 showed an accuracy range of 63.6% - 88.3% across different topics, outperforming GPT-3.5 in every subtopic [10].

In terms of answering questions for family medicine experts in Taiwan, ChatGPT demonstrated an accuracy of 41.6% in a study that also found that the length of the questions did not affect the model's accuracy. However, the authors noted that the AI's accuracy might depend on the difficulty of the test, the local language, and medical practices, which differ by region and could reduce the model's accuracy [11].

This study investigated the accuracy of responses from widely used LLM AIs, including ChatGPT (with GPT-3.5 and GPT-4), Bing, and Google Bard. Also, we compared their accuracy and determined relationships with the difficulty index for multiple-choice questions closely related to the content of the Thailand Center for Medical Competency Assessment step 1, as well as other factors that may affect the AI's accuracy, such as the length of the question, the presence of negatively worded questions, and the variety of topics across various systems. This research was undertaken to explore these dimensions.

Methods

Study Design and Setting

This study was carried out at the Institute of Medicine, Suranaree University of Technology, Thailand. The curriculum has been accredited by the World Federation for Medical Education since 2021, and the program enrolls 92 medical students annually. Preclinical medical students receive instruction through a collaboration between the School of Preclinic, the Institute of Science, and the Institute of Medicine.

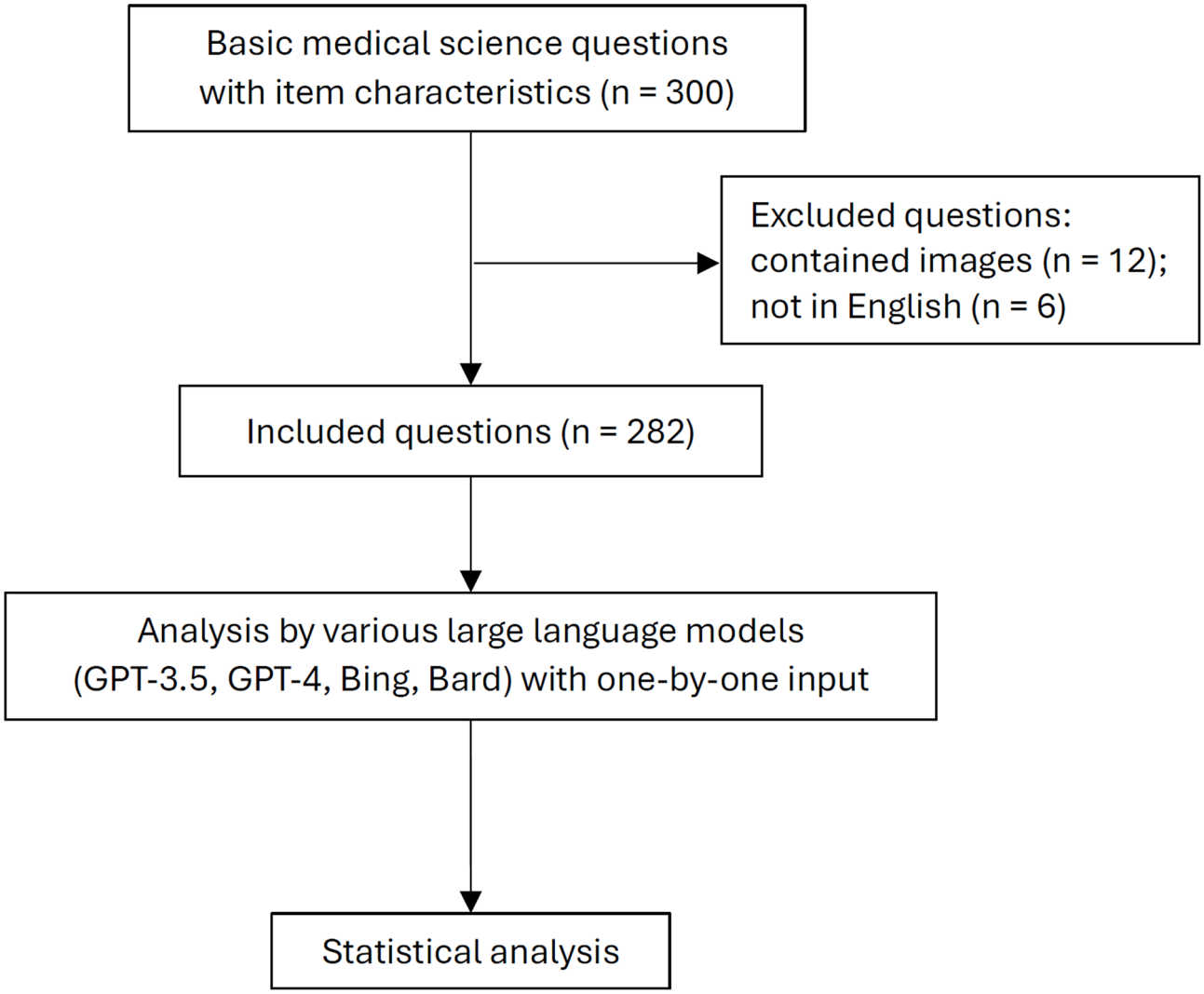
Ethical Considerations

The Human Research Ethics Committee at Suranaree University of Technology approved an exemption (certificate of exemption 117/2566) for this study, which was conducted in accordance with international guidelines for human research.

Data Source

This study used a set of 300 multiple-choice questions that closely matched the content and topic distribution of Thailand's step 1 National Medical Licensing Examination. These questions were voluntarily administered to third-year medical students in February 2021 and 2022. This timing was chosen because the students had already completed courses relevant to the examination. The difficulty index and discrimination index of each question were assessed from the test. The same set of questions was used for both years without any modifications to the content of the exam. The study excluded questions that contained pictures or were not written in English. These exclusion criteria were applied to ensure consistency in the type of questions assessed and to maintain a focus on the textual comprehension and response accuracy of the LLMs (Figure 1).

Figure 1. Study flow chart.



Question Characteristics

Question length was defined as the number of words contained within a question. Negative word questions were identified as those containing the terms "not," "no," "exclude," or "neither." Case scenario questions were characterized by the inclusion of a clinical case scenario, providing a contextual background to the question being asked.

We also used item analyses [12-14], such as the difficulty index, discrimination index, and internal consistency reliability, as independent factors associated with the LLMs' accuracy.

Difficulty index (represented by the letter *p*) is the proportion of examinees who answered a specific question correctly. If a question is easy and every examinee answers it correctly, *p* will be 1. Conversely, if no examinees answer the question correctly, *p* will be 0. This index helps in evaluating the relative difficulty of each question in an examination [12].

Discrimination index (represented by the letter *r*) refers to a question's ability to differentiate between examinees who have high scores and those who do not. Questions with a high discrimination ability are characterized by high scorers typically answering them correctly, while low scorers tend to answer them incorrectly [13]. The most widely used metric for assessing a question's discrimination ability is the point-biserial correlation. The point-biserial correlation coefficient ranges from −1 to 1. A higher point-biserial correlation indicates a question with better discriminatory power.

Internal consistency reliability was measured with Cronbach α. It ranges from 0 to 1, with higher values indicating greater internal consistency. A Cronbach α value above 0.7 is generally considered acceptable, values above 0.8 are considered good, and values above 0.9 are considered excellent.

Prompt Input for LLMs

We used the prompt "Choose the best one answer." Each question was asked to each LLM after inputting the prompt during the same period, from January 18 to 24, 2024. We individually inputted the selected questions into various LLMs, including ChatGPT (with GPT-3.5 and GPT-4), Microsoft Bing, and Google Bard (one session contained one prompt and individual question). The responses from these models were then categorized as either correct or incorrect.

Statistical Analysis

In this study, discrete variables are represented as percentages, while continuous variables are represented as either the mean (SD) or median (IQR). The association between categorical

variables was analyzed using the χ^2 test or Fisher exact test. The relationships between variables and the ability of the LLMs to provide correct answers was examined using multivariable logistic regression, with results reported as odds ratios (ORs) and 95% CIs. Statistical significance was determined at a *P* value of <.05 for all tests. The analysis was facilitated by Stata (version 17; StataCorp), which was used for data analysis and chart creation.

Results

We evaluated the LLMs by using a set of 300 multiple-choice questions that were closely aligned with the content and topic distribution of Thailand's Step 1 National Medical Licensing Examination. According to the exclusion criteria, 12 picture-containing questions and 6 non-English questions were excluded; therefore, 282 eligible questions were included. All eligible questions were concurrently input into various LLMs (Figure 1). The responses were then recorded, categorizing the outcomes as either correct or incorrect.

The questions were categorized according to the block system (Table 1), with distributions as follows: 32.3% on general principles, 5.7% on the hematopoietic system, 8.2% on the nervous system, 3.9% on skin and connective tissues, 4.3% on the musculoskeletal system, 7.8% on the respiratory system, 8.9% on the cardiovascular system, 7.5% on the gastrointestinal system, 6.7% on the urinary system, 7.1% on the reproductive system, and 7.8% on the endocrine system. The average question length was 49.10 (SD 18.94) words, with 24 questions (8.2%) containing negative wording. More than half of the questions, specifically 53.2%, were based on clinical case scenarios (more descriptive statistics for the item analysis for each block are provided in Multimedia Appendix 1). The mean difficulty index was 0.35, indicating moderately difficult to difficult questions. The discrimination index was 0.16, suggesting a poor ability to distinguish between higher and lower performers. Otherwise, the internal consistency reliability, at 0.84, highlighted an acceptable level of consistency across the examination.

Table . Question characteristics (n=282).

Characteristics	Values
Number of questions by block, n (%)	
General principles[a]	91 (32.3)
Hematopoietic system	16 (5.7)
Nervous system	23 (8.2)
Skin and connective tissue	11 (3.9)
Musculoskeletal system	12 (4.3)
Respiratory system	22 (7.8)
Cardiovascular system	25 (8.9)
Gastrointestinal system	21 (7.5)
Urinary system	19 (6.7)
Reproductive system	20 (7.1)
Endocrine system	22 (7.8)
Question length (words), mean (SD)	49.10 (18.94)
Negative-word questions, n (%)	24 (8.5)
Case scenario questions, n (%)	150 (53.2)
Average difficulty index (*p*)	0.35
Average discrimination index (*r*)	0.16
Internal consistency reliability (α)	0.84

[a]"General principle" questions refer to fundamental principles in biochemistry, molecular biology, human development, genetics, normal immune responses, basic pathological processes, laboratory investigations, general pharmacology, epidemiology, and biostatistics.

The overall accuracy of the LLMs in the basic medical science examination was as follows (Table 2): GPT-4 achieved the highest accuracy at 89.07% (95% CI 84.76% - 92.41%), Microsoft Bing had an accuracy of 83.69% (95% CI 78.85% - 87.80%), GPT-3.5 recorded an accuracy of 67.02% (95% CI 61.20% - 72.48%), and Google Bard demonstrated an accuracy of 63.83% (95% CI 57.92% - 69.44%). The Fisher exact test showed that GPT-4 performed more accurately than Microsoft Bing, and that the difference was statistically significant (*P*<.001)

Table . Accuracy of large language models with 95% CIs, compared based on category (n=282).

	GPT-3.5	GPT-4	Microsoft Bing	Google Bard
Number of correct answers	189	251	236	180
Overall accuracy, % (95% CI)	67.02 (61.20 - 72.48)	89.07 (84.76 - 92.41)	83.69 (78.85 - 87.80)	63.83 (57.92 - 69.44)
General principles, % (95% CI)	84.62 (75.54 - 91.33)	90.11 (82.05 - 95.38)	84.62 (75.54 - 91.33)	72.53 (62.17 - 81.37)
Block system, % (95% CI)	61.78 (54.49 - 68.70)	88.48 (83.08 - 92.64)	83.25 (77.18 - 88.25)	59.69 (52.36 - 66.70)

The GPT-4 model demonstrated the highest accuracy among the LLMs in the general principles section for basic science, achieving 90.11% (95% CI 82.05% - 95.38%), as shown in Table 2. GPT-3.5 and Bing exhibited equal accuracy in this section, with the lowest accuracy being 72.53% (95% CI 62.17% - 81.37%) for Bard. Additionally, GPT-4 maintained its position as the top performer in the block system with an accuracy of 88.48% (95% CI 83.08% - 92.64%), whereas Bard again displayed the lowest performance in this segment (Figure 2). Overall, GPT-4 stood out for its superior performance in overall accuracy, general principles, and the block system.

Figure 2. Comparative accuracy with 95% CIs for artificial intelligence models across different question categories.

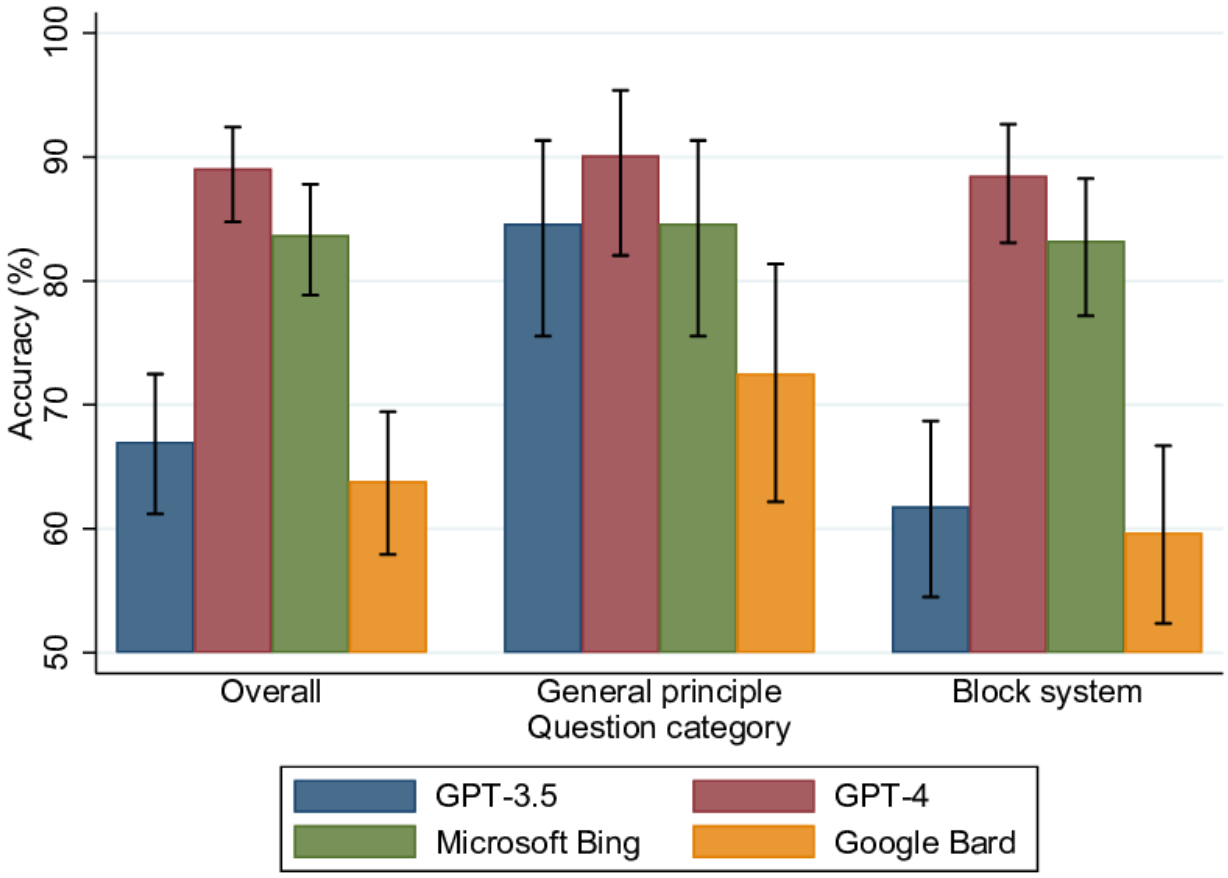


Table 3 presents the number of correct answers stratified by the block system alongside the proportion of correct answers relative to the total number of questions. The GPT-4 model exhibited the best performance, with its accuracy ranging from 84% to 95%. Following GPT-4, the Microsoft Bing model demonstrated block system accuracies between 68% and 91%. The accuracy of GPT-3.5 and Google Bard was comparable in this study, with GPT-3.5 achieving between 53% and 85%, and Google Bard ranging from 53% to 72%.

Table . Number of correct answers stratified by block system (n=282)

Topic	Correct answers, n (%)			
	GPT-3.5	GPT-4	Microsoft Bing	Google Bard
General principles (n=91)	77 (85)	82 (90)	77 (85)	66 (73)
Hematopoietic system (n=16)	8 (50)	14 (88)	13 (81)	11 (69)
Nervous system (n=23)	13 (57)	21 (91)	19 (83)	12 (52)
Skin and connective tissue (n=11)	8 (73)	10 (91)	9 (82)	5 (46)
Musculoskeletal system (n=12)	9 (75)	11 (92)	10 (83)	8 (67)
Respiratory system (n=22)	12 (55)	17 (77)	19 (86)	13 (59)
Cardiovascular system (n=25)	14 (56)	21 (84)	20 (80)	16 (64)
Gastrointestinal system (n=21)	15 (71)	20 (95)	18 (86)	14 (67)
Urinary system (n=19)	10 (53)	17 (90)	13 (68)	10 (53)
Reproductive system (n=20)	13 (65)	18 (90)	18 (90)	12 (60)
Endocrine system (n=22)	16 (73)	20 (91)	20 (91)	13 (59)

Table 4 illustrates the question characteristics associated with correct answers. There was a correlation between the difficulty index and the accuracy in all 4 models, with the strongest association observed in the GPT-4 model (OR 90.13, 95% CI 4.30 - 1887.54; *P*=.004). This was followed by GPT-3.5, which had an OR of 28.03 (95% CI 4.68 - 167.98; *P*<.001). Microsoft Bing and Google Bard demonstrated similar correlations with correct answers, with ORs of 18.9 (95% CI 1.84 - 195.42; *P*=.01) and 18.73 (95% CI 3.12 - 112.45; *P*=.001), respectively, as shown in Table 4. There was no statistically significant correlation between the accuracy of GPT-3.5, GPT-4, and Bing and question length, negative word questions, clinical case scenario questions, or the discrimination index.

Table . Multivariable logistic regression analysis showing question characteristics associated with correct answer of large language model artificial intelligence (n=282).

Variable	GPT-3.5		GPT-4		Microsoft Bing		Google Bard	
	OR[a] (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value
Question length (word)	0.99 (0.97 - 1.00)	.07	1.00 (0.98 - 1.02)	.96	1.00 (0.98 - 1.02)	.94	0.98 (0.97 - 1.00)	.02
Negative word question	0.55 (0.22 - 1.35)	.19	0.44 (0.15 - 1.30)	.14	0.46 (0.18 - 1.22)	.12	0.26 (0.10 - 0.69)	.007
Case scenario question	0.94 (0.50 - 1.77)	.85	1.57 (0.63 - 3.93)	.34	0.94 (0.43 - 2.04)	.87	0.56 (0.30 - 1.07)	.08
Difficulty index (*p*)	28.03 (4.68 - 167.98)	<.001	90.13 (4.30 - 1887.54)	.004	18.9 (1.84 - 195.42)	.01	18.73 (3.12 - 112.45)	.001
Discrimination index (*r*)	2.80 (0.34 - 23.32)	.34	4.85 (0.20 - 116.54)	.33	9.66 (0.67 - 140.06)	.10	9.31 (1.02 - 84.68)	.048

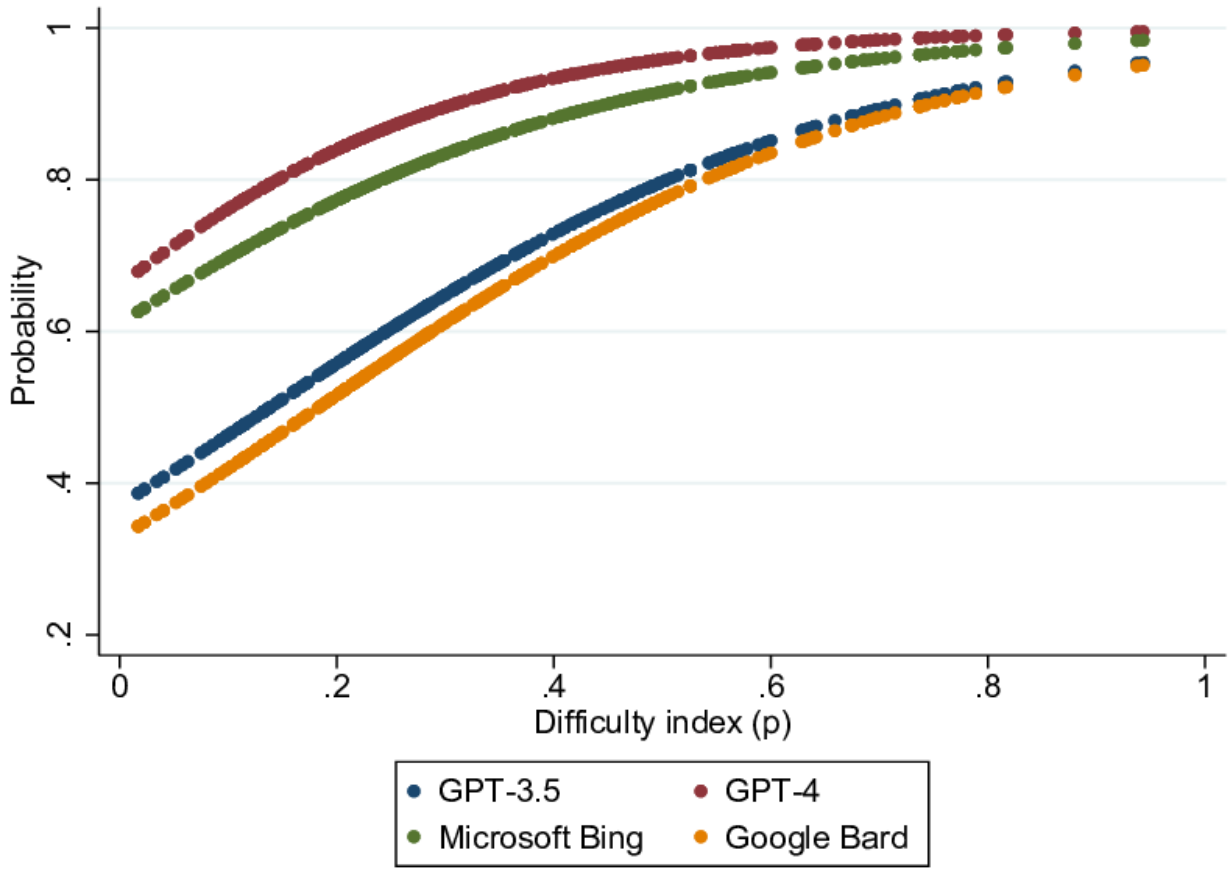
[a]OR: odds ratio.

On the other hand, for Google Bard, longer questions had a higher OR, of 0.98 (95% CI 0.97 - 1.00; *P*=.02), for the model to provide the correct answer than shorter questions. The negative-word questions were less likely to be answered correctly by the model, with an OR of 0.26 (95% CI 0.10 - 0.69; *P*=.007), compared to those without negative words. Furthermore, questions with a higher discrimination index were more likely to be correctly answered with statistical significance by the model, with an OR of 9.31 (95% CI 1.02 - 84.68, *P*=.048), as compared to those with a lower discrimination index. No statistically significant correlation was observed between the accuracy of the AIs in answering clinical case scenario questions, as presented in Table 4.

The correlation between the difficulty index and the estimated accuracy of the various AI models, analyzed with binary logistic regression, is shown in Figure 3. The GPT-4 model consistently demonstrated the highest accuracy across all levels of question difficulty index (Figure 3). Google Bard, on the other hand, had the lowest estimated accuracy. The accuracy of the various

LLMs improved as the difficulty index increased, indicating that these models performed better on easier questions.

Figure 3. Accuracy of various artificial intelligence models estimated based on difficulty index.



Discussion

Accuracy of the LLMs on Basic Medical Science Examinations

This study compared the accuracy of LLMs in answering questions from a basic medical science examination related to the National Medical Licensing Examination, finding that GPT-4 had the highest accuracy, at 89.07%, and Google Bard had the lowest accuracy, at 63.83%, when tasked with answering questions in this context. The most frequently studied AI models were GPT-3.5 and GPT-4.

These results align with the 2023 findings of Yanagita et al [15], who used questions from the National Medical Licensing Examination in Japan, administered by the Japanese Ministry of Health, Labour and Welfare. When inputting Japanese questions into the prompt, they reported an accuracy for GPT-4 of 81.5%, significantly higher than GPT-3.5's accuracy of 42.8%, with GPT-4 surpassing the National Medical Licensing Examination passing standard of 72%.

Our results are similar to those of the study conducted by Gilson et al [6] in 2023, which found that the performance of GPT-3.5 on AMBOSS-Step1 and NBME-Free-Step1 was 44% and 64.4%, respectively. Flores-Cohaila et al [16] conducted a study on the accuracy of LLMs on the Peruvian National Licensing Medical Examination and discovered that GPT-4 had 86% accuracy, following by GPT-3.5 at 77%, with moderately difficult to difficult questions being associated with incorrect answers (the OR for GPT-3.5 was 6.6, 95% CI 2.73 - 15.95; for GPT-4, the OR was 33.23, 95% CI 4.3 - 257.12).

A literature review from China (Wang et al [17]) evaluated the performance of GPT-3.5 and GPT-4 on the China National Medical Licensing Examination and reported 56% and 84% accuracy for GPT-3.5 and GPT-4, respectively, demonstrating GPT-4's superiority over GPT-3.5 in terms of accuracy on basic medical science examinations.

The accuracy of GPT-4 and GPT-3.5 is influenced by the variety within the question dataset. This results in diverse outcomes across different countries, changing according to the environmental context, difficulty level of the examination, and the proportion of subcomponents within the examination question sets, which may vary from one country to another. Consequently, the estimated accuracy of AI models for each dataset is not constant.

Difficulty Index and the LLMs' Accuracy

In this study, we identified factors correlated with the accuracy of AI models in answering questions. We found that for every model, the difficulty index was associated with correctly answering questions. Moreover, across all models, there was a tendency to answer questions correctly as the difficulty index increased (indicating easier questions). Specifically, GPT-4

demonstrated the highest OR at 90.13 (95% CI 4.30 - 1887.54; *P*=.004), followed by GPT-3.5, with an OR of 28.03 (95% CI 4.68 - 167.98; *P*<.001).

This result aligns with findings from Antaki et al [18] showing that question difficulty was the most predictive factor of GPT-3.5's answer accuracy (likelihood ratio 24.05; *P*<.001) and that GPT-4 was more accurate than GPT-3.5. The current research reveals the accuracy of AI models in answering questions across various disciplines, particularly studies focusing on the renowned GPT-3.5 model. However, this study focused on the relationship between the difficulty index, derived from human examination observations, and the accuracy of every simple-to-access LLM that is widely used. There was also a variation in accuracy among all models, with GPT-4 being the most accurate, and there was an obvious correlation with the difficulty index for each model, indicating that easier questions had higher accuracy.

The Implication of LLMs for Medical Education

This study's findings hold significant implications for medical education, particularly regarding the use of LLMs such as GPT-4, Microsoft Bing, GPT-3.5, and Google Bard as educational tools [19]. There are 3 major ways that this study's findings can be applied to augment traditional study methods.

First, enhancing study efficiency: the high accuracy rates of LLMs, especially GPT-4, in answering medical examination questions suggest their utility as effective study aids. By providing immediate and accurate answers with explanations, these models can help students identify areas of weakness and reinforce their learning more efficiently than traditional study methods alone.

Second, supplementing traditional education methods: LLMs can act as supplementary tools in medical education, alongside lectures, textbooks, and clinical scenarios. Integrating LLMs into the curriculum provides students with an additional resource for study and review to enhance the overall educational experience.

Last, preparing for licensing examinations: given the study's focus on medical licensing examinations, LLMs could play a crucial role in preparing students for these critical assessments. The ability of LLMs to accurately answer examination questions, such as those tackled by GPT-4, and explain reasoning processes can assist students in better preparing for the format and content of licensing exams.

LLMs may have a negative impact on medical education. Excessive dependence on LLMs might impede the development of independent critical thinking skills. Students may become reliant on the model's suggestions instead of developing their own reasoning processes. LLMs can sometimes provide incorrect, incomplete, or biased information [20,21]. This can interfere with the development of critical appraisal skills, leading students to accept inaccurate information, which may hinder their critical thinking and medical reasoning abilities [22]. Additionally, reduced peer and mentor interaction can hinder the development of professional judgment, depriving students of diverse perspectives and collaborative problem-solving experiences.

To maximize the benefits while minimizing the negative impact of incorporating LLMs into medical education [23], 4 strategies can be considered. First, structured use: LLMs can be incorporated as supplementary tools in a structured curriculum rather than as primary sources of information. Second, critical appraisal training: the importance of critically appraising information provided by LLMs should be emphasized, and students should be taught how to cross-reference and validate information. Third, independent thought should be encouraged: environments should be fostered that encourage independent thinking and problem-solving, using LLMs to support (not replace) these processes. Fourth, monitoring and evaluation: the impact of LLMs on students' learning and reasoning skills should be assessed, and educational approaches should be adjusted based on these assessments.

Limitations

One significant limitation of this study is the LLMs' ability to accurately respond to complex medical examination questions. Moreover, despite GPT-4's high performance, the study's focus on a single culturally and geographically specific medical licensing examination (Thailand Step 1 National Medical Licensing Examination) may limit the generalizability of the findings to other medical examinations and educational contexts. The exclusion of questions containing images and those not in English restricted the comprehensiveness of the assessment, considering the importance of questions on visual diagnostics. Updates to LLMs can significantly affect their accuracy, leading to a potential increase in the capabilities of the models over time. Furthermore, different LLMs can respond differently to different prompts. They can generate different answers across independent sessions, even with identical prompts. Therefore, a sensitivity analysis of the accuracy of the LLMs' responses should be conducted with a variety of prompt and session settings.

Conclusion

Our results show a significant variation in performance among different LLMs, with the most accurate model being GPT-4. This study has shed light on the role of LLMs as supplementary tools in medical education, as well as the need for more research to increase the generalizability of the findings to different educational settings. We advocate for the ongoing development and modification of LLMs to match the unique demands of medical education internationally, which has important implications for the future integration of AI in medical training and test preparation.

Acknowledgments

This research received funding support from the Unit of Teaching and Learning Development and Research, Suranaree University of Technology.

Authors' Contributions

NK was responsible for the entire project and prepared the initial draft of the manuscript. JP conceptualized the project. GB and NP contributed by proofreading and editing the manuscript. All authors participated in interpreting the results and in preparing the final version of the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Descriptive statistics of item analysis for each block system.
[DOCX File, 18 KB - mededu_v11i1e58898_app1.docx]

References

1. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. Science 2015 Jul 17;349(6245):255-260. [doi: 10.1126/science.aaa8415] [Medline: 26185243]
2. Chen L, Chen P, Lin Z. Artificial intelligence in education: a review. IEEE Access 2020;8:75264-75278. [doi: 10.1109/ACCESS.2020.2988510]
3. Novak LL, Russell RG, Garvey K, et al. Clinical use of artificial intelligence requires AI-capable organizations. JAMIA Open 2023 Jul;6(2):ooad028. [doi: 10.1093/jamiaopen/ooad028] [Medline: 37152469]
4. Sedaghat S. Early applications of ChatGPT in medical practice, education and research. Clin Med (Lond) 2023 May;23(3):278-279. [doi: 10.7861/clinmed.2023-0078] [Medline: 37085182]
5. Liu PR, Lu L, Zhang JY, Huo TT, Liu SX, Ye ZW. Application of artificial intelligence in medicine: an overview. Curr Med Sci 2021 Dec;41(6):1105-1115. [doi: 10.1007/s11596-021-2474-3] [Medline: 34874486]
6. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
7. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Dig Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
8. Tolsgaard MG, Pusic MV, Sebok-Syer SS, et al. The fundamentals of artificial intelligence in medical education research: AMEE Guide No. 156. Med Teach 2023 Jun;45(6):565-573. [doi: 10.1080/0142159X.2023.2180340] [Medline: 36862064]
9. Subramani M, Jaleel I, Krishna Mohan S. Evaluating the performance of ChatGPT in medical physiology university examination of phase I MBBS. Adv Physiol Educ 2023 Jun 1;47(2):270-271. [doi: 10.1152/advan.00036.2023] [Medline: 36971685]
10. Oh N, Choi GS, Lee WY. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. Ann Surg Treat Res 2023 May;104(5):269-273. [doi: 10.4174/astr.2023.104.5.269] [Medline: 37179699]
11. Weng TL, Wang YM, Chang S, Chen TJ, Hwang SJ. ChatGPT failed Taiwan's Family Medicine Board Exam. J Chin Med Assoc 2023 Aug 1;86(8):762-766. [doi: 10.1097/JCMA.0000000000000946] [Medline: 37294147]
12. Kumar D, Jaipurkar R, Shekhar A, Sikri G, Srinivas V. Item analysis of multiple choice questions: a quality assurance test for an assessment tool. Med J Armed Forces India 2021 Feb;77(Suppl 1):S85-S89. [doi: 10.1016/j.mjafi.2020.11.007] [Medline: 33612937]
13. Dhanvijay AKD, Dhokane N, Balgote S, et al. The effect of a one-day workshop on the quality of framing multiple choice questions in physiology in a medical college in India. Cureus 2023 Aug;15(8):e44049. [doi: 10.7759/cureus.44049] [Medline: 37746478]
14. Bhattacherjee S, Mukherjee A, Bhandari K, Rout AJ. Evaluation of multiple-choice questions by item analysis, from an online internal assessment of 6th semester medical students in a rural medical college, West Bengal. Ind J Community Med 2022;47(1):92-95. [doi: 10.4103/ijcm.ijcm_1156_21] [Medline: 35368481]
15. Yanagita Y, Yokokawa D, Uchida S, Tawara J, Ikusaka M. Accuracy of ChatGPT on medical questions in the National Medical Licensing Examination in Japan: evaluation study. JMIR Form Res 2023 Oct 13;7:e48023. [doi: 10.2196/48023] [Medline: 37831496]
16. Flores-Cohaila JA, García-Vicente A, Vizcarra-Jiménez SF, et al. Performance of ChatGPT on the Peruvian National Licensing Medical Examination: cross-sectional study. JMIR Med Educ 2023 Sep 28;9:e48039. [doi: 10.2196/48039] [Medline: 37768724]
17. Wang H, Wu W, Dou Z, He L, Yang L. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: pave the way for medical AI. Int J Med Inform 2023 Sep;177:105173. [doi: 10.1016/j.ijmedinf.2023.105173] [Medline: 37549499]

18. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. Ophthalmol Sci 2023 Dec;3(4):100324. [doi: 10.1016/j.xops.2023.100324] [Medline: 37334036]
19. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023 Jun 1;9:e48291. [doi: 10.2196/48291] [Medline: 37261894]
20. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ 2024 Nov;58(11):1276-1285. [doi: 10.1111/medu.15402] [Medline: 38639098]
21. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023 Sep 22;23(1):689. [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
22. Safranek CW, Sidamon-Eristoff AE, Gilson A, Chartash D. The role of large language models in medical education: applications and implications. JMIR Med Educ 2023 Aug 14;9:e50945. [doi: 10.2196/50945] [Medline: 37578830]
23. Shimizu I, Kasai H, Shikino K, et al. Developing medical education curriculum reform strategies to address the impact of generative AI: qualitative study. JMIR Med Educ 2023 Nov 30;9:e53466. [doi: 10.2196/53466] [Medline: 38032695]

Abbreviations

AI: artificial intelligence
LLM: large language model
USMLE: United States Medical Licensing Examination

Edited by B Lesselroth; submitted 29.03.24; peer-reviewed by AKD Dhanvijay, C Rose, GP Privitera, P Shingru, W Xu; revised version received 22.05.24; accepted 04.12.24; published 13.01.25.

Please cite as:
Kaewboonlert N, Poontananggul J, Pongsuwan N, Bhakdisongkhram G
Factors Associated With the Accuracy of Large Language Models in Basic Medical Science Examinations: Cross-Sectional Study
JMIR Med Educ 2025;11:e58898
URL: https://mededu.jmir.org/2025/1/e58898
doi:10.2196/58898

Enhancing Medical Student Engagement Through Cinematic Clinical Narratives: Multimodal Generative AI–Based Mixed Methods Study

Tyler Bland, PhD

Department of Medical Education, University of Idaho, 875 Perimeter Drive MS 4061, WWAMI Medical Education, Moscow, ID, United States

Corresponding Author:
Tyler Bland, PhD
Department of Medical Education, University of Idaho, 875 Perimeter Drive MS 4061, WWAMI Medical Education, Moscow, ID, United States

Abstract

Background: Medical students often struggle to engage with and retain complex pharmacology topics during their preclinical education. Traditional teaching methods can lead to passive learning and poor long-term retention of critical concepts.

Objective: This study aims to enhance the teaching of clinical pharmacology in medical school by using a multimodal generative artificial intelligence (genAI) approach to create compelling, cinematic clinical narratives (CCNs).

Methods: We transformed a standard clinical case into an engaging, interactive multimedia experience called "Shattered Slippers." This CCN used various genAI tools for content creation: GPT-4 for developing the storyline, Leonardo.ai and Stable Diffusion for generating images, Eleven Labs for creating audio narrations, and Suno for composing a theme song. The CCN integrated narrative styles and pop culture references to enhance student engagement. It was applied in teaching first-year medical students about immune system pharmacology. Student responses were assessed through the Situational Interest Survey for Multimedia and examination performance. The target audience comprised first-year medical students (n=40), with 18 responding to the Situational Interest Survey for Multimedia survey (n=18).

Results: The study revealed a marked preference for the genAI-enhanced CCNs over traditional teaching methods. Key findings include the majority of surveyed students preferring the CCN over traditional clinical cases (14/18), as well as high average scores for triggered situational interest (mean 4.58, SD 0.53), maintained interest (mean 4.40, SD 0.53), maintained-feeling interest (mean 4.38, SD 0.51), and maintained-value interest (mean 4.42, SD 0.54). Students achieved an average score of 88% on examination questions related to the CCN material, indicating successful learning and retention. Qualitative feedback highlighted increased engagement, improved recall, and appreciation for the narrative style and pop culture references.

Conclusions: This study demonstrates the potential of using a multimodal genAI-driven approach to create CCNs in medical education. The "Shattered Slippers" case effectively enhanced student engagement and promoted knowledge retention in complex pharmacological topics. This innovative method suggests a novel direction for curriculum development that could improve learning outcomes and student satisfaction in medical education. Future research should explore the long-term retention of knowledge and the applicability of learned material in clinical settings, as well as the potential for broader implementation of this approach across various medical education contexts.

(JMIR Med Educ 2025;11:e63865) doi:10.2196/63865

KEYWORDS

artificial intelligence; cinematic clinical narratives; cinemeducation; medical education; narrative learning; AI; medical student; pharmacology; preclinical education; long-term retention; AI tools; GPT-4; image; applicability

Introduction

Background

Student and trainee engagement is a critical factor in medical education, influencing outcomes such as academic achievement, overall well-being, satisfaction, and reduced burnout [1,2]. High levels of engagement have been linked to increased motivation and better learning experiences, as active participation encourages deeper understanding and application of complex material [3]. In contrast, traditional lecture-based learning often results in passive absorption of information, limiting student engagement and negatively affecting the ability to interact meaningfully with content [4]. To address this, we developed a cinematic clinical narrative (CCN), an interactive multimedia learning experience designed to enhance student engagement by integrating cinematic storytelling and narrative-based learning techniques. This method builds upon the principles of

cinemeducation, a teaching approach that uses film to create emotional connections and foster active learning [5]. By using generative artificial intelligence (genAI) tools, we have further enhanced the learning experience and decreased the barrier to entry for instructors, making it more immersive and adaptable to current educational needs. GenAI has been recognized as a transformative tool in reshaping medical education, offering new opportunities for interactive, technology-driven learning environments that promote active student engagement [6,7].

The target audience for our CCN comprises first-year medical students learning pharmacology related to the immune system. Medical students often face a knowledge gap in understanding complex pharmacological interactions and the intricacies of immune responses largely due to the difficulty of the material [8,9]. Furthermore, there is speculated to be a skill gap in medical and other professional health science students in applying theoretical knowledge to clinical scenarios [10] and the real problem of burnout due to many factors, one of which is the large amount of knowledge required to retain in a short amount of time [11]. The CCN aims to address these issues by enhancing comprehension, clinical application skills, and empathy toward patients with autoimmune diseases.

The CCN used a unique instructional approach by merging cinemeducation [5] with multiple genAI platforms, tailored for first-year medical students in pharmacology. This method addresses the challenge of enhancing engagement and knowledge retention in complex subjects such as immune system pharmacology. Unlike traditional didactic teaching, our approach, supported by others advocating for innovative teaching strategies, uses storytelling to deepen understanding and empathy [12-14]. Use of genAI in medical training, particularly in personalizing learning experiences and competencies for genAI-based tools, is also a current area of active research [15,16]. This aligns with other researchers who highlight the importance of interactive and engaging content in medical education [17]. Our project also leverages the effectiveness of narrative-based learning, which offers an experiential learning environment over conventional teaching methods and is more accurate to real-world situations [18].

Medical students often struggle to engage with and retain complex pharmacological concepts, especially in preclinical education, where traditional teaching methods can lead to passive learning and poor knowledge retention. To address this challenge, we developed and implemented a novel instructional approach, CCNs, which leverages multimodal genAI tools to create immersive, engaging learning experiences. The aim of this study is to evaluate the effectiveness of these genAI-enhanced CCNs in increasing student engagement, interest, and knowledge retention in medical pharmacology concepts. We tested this intervention by assessing student interest using the Situational Interest Survey for Multimedia (SIS-M) and measuring examination performance on content covered by the CCNs. We hypothesize that students exposed to CCNs will report higher levels of engagement compared with traditional case-based learning and have passing examination grades on questions related to the CCN.

Theoretical Framework

The instructional method in the CCN uses contemporary educational theories emphasizing active, learner-centered approaches. Drawing inspiration from the Constructivist Learning Theory, which advocates for knowledge construction through experience [19], our approach uses an adaptation of cinemeducation to create an immersive learning environment [5]. This also aligns with Mayer's Cognitive Theory of Multimedia Learning, which suggests that learning is enhanced through multimodal presentations [20]. Furthermore, our multimodal use of various genAI platforms for content development is informed by the Technological Pedagogical Content Knowledge (TPACK) framework [21], ensuring an effective integration of technology in teaching. This methodology responds to identified needs in medical education for more engaging and effective teaching strategies, bridging theory and practice in a novel and impactful way.

Methods

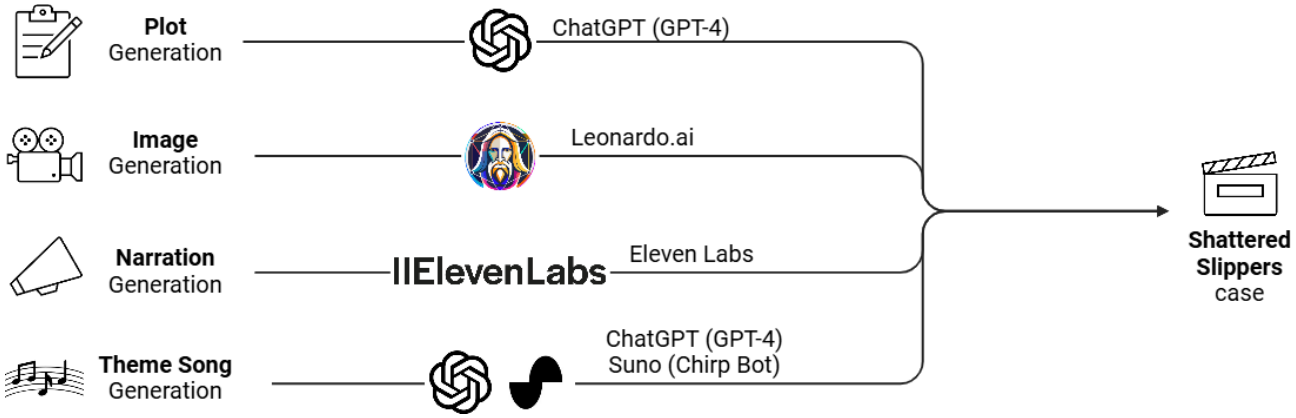
Participants and CCN Design Overview

This study was conducted at the University of Idaho WWAMI Medical Education Program, which is part of a collaborative University of Washington School of Medicine program serving Washington, Wyoming, Alaska, Montana, and Idaho. The WWAMI program provides medical education to students across these states, offering them the opportunity to complete their first 2 preclinical years of medical school in their home states before transitioning to clinical training. The target learners for this study were first-year medical students in the WWAMI program enrolled in a 6-week foundational infections and immunity course, which included topics covering immune system pharmacology. Students in this course attend pharmacology lectures that culminate in clinical cases, allowing them to apply their newly acquired knowledge of medications to real-world patient scenarios.

We decided to reimagine one of these cases into "Shattered Slippers," a CCN that was presented as a fictional sequel to the movie "Another Cinderella Story" (Multimedia Appendices 1 and 2). This fictional sequel features the star from the original movie, Selena Gomez, which was purposeful, given her real-life battle with lupus and her experience receiving a kidney transplant. This choice not only provides a strong thematic link connecting the CCN to the source material but also serves to humanize and demystify the conditions under study.

The development of "Shattered Slippers" used a suite of genAI platforms to create an immersive and engaging learning experience (Figure 1). The plot was crafted using GPT-4, known for its language understanding and generation capabilities. For visual imagery, Leonardo.ai and Stable Diffusion were used to generate high-quality, contextually relevant images. Narration was produced using Eleven Labs, ensuring a coherent and captivating storytelling experience. Furthermore, the theme song, integral to setting the tone of the educational module, was composed using the combined efforts of GPT-4 and Suno.

Figure 1. Multimodal generative artificial intelligence (genAI) case generation approach. Each portion of the case used a different genAI platform for material generation. These included ChatGPT (GPT-4), Leonardo.ai, Eleven Labs, and Suno.



These artificial intelligence (AI)–generated materials were all integrated into 2 PowerPoint presentations. Part I of the CCN was presented at the end of a 1-hour pharmacology lecture on immunomodulatory drugs with specific focus on nonsteroidal anti-inflammatory drugs, glucocorticoids, and innate immune system inhibitors. Part II of the CCN was presented 4 weeks later at the end of a 1-hour pharmacology lecture on immunomodulatory and transplant drugs with specific focus on cytokine inhibitors, cytotoxic drugs, and antimetabolites. Both lectures were presented in-person with >90% of students attending both lectures. The combined CCN is provided as a supplemental file (Multimedia Appendix 2).

At the conclusion of the course, students were informed about Selena Gomez's actual medical journey. This revelation effectively bridged the gap between the fictional narrative of "Shattered Slippers" and real-world medical scenarios, thereby enhancing the educational impact and relevance of the clinical cases discussed.

Plot Development

The process of developing the plot for "Shattered Slippers" began with a reimagining of a clinical case initially presented in the first-year medical school curriculum. This original case centered around a ballerina struggling with rheumatoid arthritis, where students were tasked with diagnosing the sources of her pain and inflammation and selecting suitable immunomodulatory medications.

Using ChatGPT (GPT-4) [22], a large language model (LLM), we transformed this clinical scenario into a compelling narrative for "Shattered Slippers." The sequential steps of the medical case were input into GPT-4, with instructions to adapt these into a fictional storyline (Figure 2 and Multimedia Appendix 3). To enhance thematic resonance and real-world connection, the ballerina's diagnosis in the plot was altered from rheumatoid arthritis to lupus, mirroring the real-life medical condition of Selena Gomez, who stars in the CCN.

Further expanding the scope of the narrative, the plot incorporated a kidney transplant storyline. This addition served a dual purpose. First, it aligned with the second lecture on immunoregulatory pharmacology focusing on organ transplant pharmacology. Second, it resonated with Selena Gomez's personal medical history, as she has undergone a kidney transplant. This incorporation not only ensured continuity with the educational objectives of the course but also added depth and authenticity to the fictional narrative, making it more engaging and relatable for the students.

Figure 2. Excerpt of plot generation. The initial prompt in the conversation covered the development of a separate CCN. Prompt engineering techniques in this initial prompt included Persona Prompting [23,24] and a modified version of Zero-Shot CoT [25]. Excerpts of the first prompt and output related to the Shattered Slippers CCN are provided. CCN: cinematic clinical narrative; CoT: Chain of Thought; LLM: large language model.

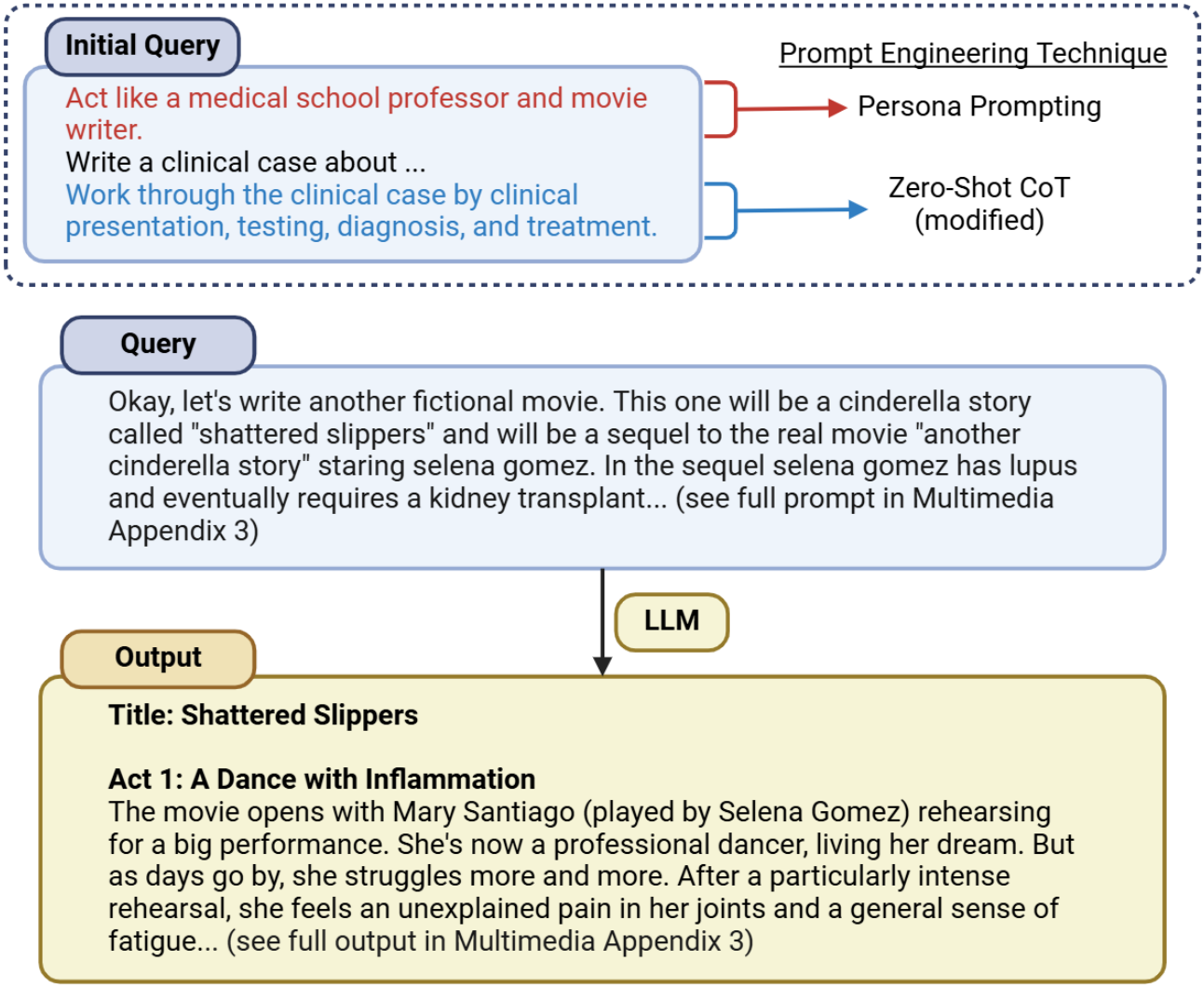


Image Generation

In order to create a more immersive educational experience, fictional images were integrated into the “Shattered Slippers” case study. These images were generated using the Leonardo.ai platform [26], which harnesses the capabilities of the Stable Diffusion XL image–generating technology (Figure 3 and Multimedia Appendix 4).

In an effort to maintain transparency and distinguish between real and AI-generated content, all images depicting real people were marked with an “AI-generated image” icon. This icon, chosen for its symbolic significance, is the spinning top from the movie “Inception.” The selection of this particular icon was purposeful; it serves as a metaphor for the increasingly blurred lines between reality and artificial constructs, mirroring the movie’s thematic exploration of distinguishing reality from illusion. This concept was explained to the students prior to their engagement with the case, setting the stage for a thoughtful consideration of the role and impact of genAI in content creation. This iconography not only helped in identifying AI-generated images but also subtly underscored the advanced capabilities of genAI in creating hyperrealistic images.

Figure 3. Artificial intelligence (AI)-generated image of Selena Gomez singing with Justin Bieber. The prompt used was "adult Selena Gomez and Justin Bieber singing together." The spinning top in the bottom right corner was added as a watermark to denote an AI-generated image. Generated with Leonardo.ai.

Narration Generation

Enhancing the immersive aspect of the CCN, an audio narration was incorporated to accompany the text on the PowerPoint slides. This element was designed to emulate the experience of listening to a movie narrator, thereby bringing the story of "Shattered Slippers" to life in an auditory format. To achieve this, the finalized script of the plot was submitted to the Eleven Labs platform [27], which specializes in converting text into lifelike audio narration (Multimedia Appendix 5).

Each of these audio narrations were incorporated into their corresponding PowerPoint slides. As each slide was presented during the course, the audio narration played automatically, further synchronizing the visual and auditory elements of the learning experience. This integration of audio narration with the visual content not only enriched the storytelling aspect of the module but also supported diverse learning styles, facilitating a more engaging and multisensory educational experience for the students.

Theme Song Generation

Although not directly educational, a theme song for "Shattered Slippers" was created to complete the immersive experience. The inclusion of a theme song aimed to add an additional layer of engagement and context to the fictional movie, contributing to a more comprehensive and cinematic learning environment.

The lyrics for the theme song were generated using GPT-4 [22]. Following the lyric generation, Suno Chirp Bot, a genAI tool for music composition [28], was used to create the melody and vocals for the theme song. This genAI-driven process allowed for a harmonious blend of lyrics and music, resulting in a fully rendered theme song (Multimedia Appendix 6).

Once completed, the theme song was embedded into the PowerPoint presentation. This musical addition served as a capstone to the multisensory educational module, further enriching the student's experience by providing a unique auditory element that complemented the visual and textual components of "Shattered Slippers."

Data Collection

The "Shattered Slippers" CCN was integrated into 2 distinct pharmacology lectures, both of which focused on medications used in immune system modulation. The target audience for this CCN was a class of 40 first-year medical students (n=40). This approach aimed not only to enrich their understanding of immunomodulatory pharmacology but also to engage them in a unique and memorable learning experience.

To evaluate student interest in the CCN as an educational tool, at the conclusion of the course, students were invited to participate in a feedback process using the SIS-M [29-31] (Table 1) of which 18 students responded (n=18). The SIS-M was developed by Dr Tonia Dousay, a professor in instructional design and educational technology, to assess various constructs of situational interest in multimedia-based learning environments. Originally created for the educational field, the SIS-M focuses on adult learners and measures constructs such

as triggered situational interest (initial engagement with multimedia), maintained interest, and value interest (perceived usefulness of the content). The survey was originally used to evaluate the effectiveness of multimedia in promoting engagement and motivation in higher education and adult learning settings [29,30] and has recently been used in medical education research [31], making it an appropriate tool for assessing learner engagement in this study. This survey was used to capture their views and opinions on the "Shattered Slippers" case, providing insights into student engagement, interest, and the overall impact of the CCN on their learning experience. The survey includes items to rank on a 1 - 5 scale (1=strongly disagree, 5=strongly agree), a question asking for preference of clinical case format, and an open-ended question asking, "Why do you think this is your preference." The CHERRIES report for this survey is supplied (Multimedia Appendix 7).

Table . SIS items.

SIS[a] type	Survey item
SI-triggered	The multimedia presentation was interesting.
SI-triggered	The multimedia presentation grabbed my attention.
SI-triggered	The multimedia presentation was often entertaining.
SI-triggered	The multimedia presentation was so exciting, it was easy to pay attention.
SI-maintained-feeling	What I learned in the multimedia presentation is fascinating to me.
SI-maintained-feeling	I am excited about what I learned in the multimedia presentation.
SI-maintained-feeling	I like what I learned in the multimedia presentation.
SI-maintained-feeling	I found the information in the multimedia presentation interesting.
SI-maintained-value	What I studied in the multimedia presentation is useful for me to know.
SI-maintained-value	The things I studied in the multimedia presentation are important to me.
SI-maintained-value	What I learned in the multimedia presentation can be applied to my job.
SI-maintained-value	I learned valuable things in the multimedia presentation.

[a]SIS: Situational Interest Survey.

Data Analysis

The research team used Microsoft Excel for the analysis of the SIS-M survey results. The average class pharmacology examination grades (n=40) from questions covered by the "Shattered Slippers" case study (n=2) were analyzed for achievement data. These included a multiple-choice question, selected by the course lead (not the study author) from a pool of questions that tested pharmacology content covered in each pharmacology lecture. The questions were administered during the students' weekly examinations, scheduled for the week immediately following the presentation of the material. Importantly, these questions were modeled after USMLE-style step 1 board questions, which assess students' ability to apply their pharmacological knowledge in a clinical context. Using this format provides a rigorous and standardized measure of student understanding of the material, ensuring that the assessment reflects the type of knowledge and critical thinking required for success on future board examinations.

The SIS-M survey's analysis focused on various dimensions of situational interest: triggered interest, maintained-value (MV), maintained interest, and maintained-feeling (MF). Thematic analysis was conducted using ChatGPT (GPT4o and o1-preview) and Claude 3.5 Sonnet. This involved generating initial codes and identifying themes, followed by the researcher combining and refining these themes for overlap and relevancy between the 3 LLMs [31]. Prompt engineering techniques used included Persona Prompting [23,24], Zero-Shot Chain of Thought (CoT) [25], and Self-Criticism [32]. The Zero-Shot Chain of Thought prompting was not used with the ChatGPT o1-preview model, as it has built-in Tree-of-Thought functionality in every output. The initial prompt was the following:

> *Act like a brilliant medical education researcher. I am doing a study on a Cinematic Clinical Narrative (CCN) which is an educational tool that combines clinical case studies with storytelling techniques typically seen in movies or TV shows. By embedding medical information within a compelling fictional storyline, CCNs help medical students retain complex medical concepts in an engaging, memorable way. The CCN in the study was called "Shattered Slippers," was a fictional sequel to the movie "Another Cinderella Story," and stars Selena Gomez. It covered the topics of immunomodulatory medications for treating lupus, and kidney transplants. I surveyed the participants on their preference of the CCN over traditional clinical cases and asked them to explain their preference. Please perform a thematic analysis on the below participant responses marked between <response> </response>. Let's work this out in a step by step way to be sure we have the right answer.*
>
> *<response>*
>
> *Participant responses here*
>
> *</response>*

This was then followed by the following Self-Criticism prompt: "Please reflect on your previous answer for any errors."

Ethical Considerations

This educational research was approved as exempt by the institutional review board of the University of Idaho (21-223). As the CCN incorporated references to real celebrities and included AI-generated images of actual people, we consulted legal counsel to ensure compliance. The counsel advised that, given the educational context and the clear labeling of images as AI-generated rather than real, the usage was permissible. Furthermore, we end the CCN with a brief description of the real-life health struggles of the celebrities, which is all public information. However, since this remains a legally gray area, we recommend exercising caution in future projects that use similar techniques. The SIS-M was conducted anonymously to ensure the confidentiality of participants' responses. No identifying information was collected, allowing students to provide honest feedback without concern for personal attribution.

Results

The quantitative assessment of the "Shattered Slippers" CCN using the SIS-M is summarized in Table 2. The results indicated high levels in participants' interest with the "Shattered Slippers" CCN, with the majority of students (14/18) indicating a preference for the CCN over traditionally presented clinical cases, only 1 student preferring the traditional approach, and 3 expressing no preference (Table 3).

Table . Situational Interest Survey for Multimedia results (N=18): scores.

Question	Minimum[a]	Maximum[a]	Mean[a]	SD	Variance
The Shattered Slippers case was interesting.	4.00	5.00	4.61	0.49	0.24
The Shattered Slippers case grabbed my attention.	4.00	5.00	4.72	0.45	0.20
The Shattered Slippers case was often entertaining.	3.00	5.00	4.67	0.58	0.33
The Shattered Slippers case was so exciting, it was easy to pay attention.	3.00	5.00	4.33	0.58	0.33
What I learned from the Shattered Slippers case is fascinating to me.	4.00	5.00	4.39	0.49	0.24
I am excited about what I learned from the Shattered Slippers case.	4.00	5.00	4.39	0.49	0.24
I like what I learned from the Shattered Slippers case.	3.00	5.00	4.39	0.59	0.35
I found the information from the Shattered Slippers case interesting.	4.00	5.00	4.33	0.47	0.22
What I studied in the Shattered Slippers case is useful for me to.	4.00	5.00	4.50	0.50	0.25
The things I studied in the Shattered Slippers case are important to me.	3.00	5.00	4.28	0.56	0.31
What I learned from the Shattered Slippers case can be applied to my major/career.	3.00	5.00	4.44	0.60	0.36
I learned valuable things from the Shattered Slippers case.	4.00	5.00	4.44	0.50	0.25

[a]Rated on a 5-point scale (1=Strongly disagree, 5=Strongly agree).

Table . Situational Interest Survey for Multimedia results (N=18): preferences for case type.

Which case type do you prefer?	Count
Traditional case studies	1
Shattered Slippers case study	14
No preference	3

Participants indicated a high average triggered situational interest in the CCN (mean 4.58, SD 0.53), as well as high maintained interest scores indicated by the students (mean 4.40, SD 0.53).

The results for MF interest indicated high MF in students receiving the CCN (mean 4.38, SD 0.51). A feeling of educational value by the participants was supported by high scores for MV interest (mean 4.42, SD 0.54).

Bridging quantitative data with qualitative insights, the survey conducted among participants also provided an open-ended question for students to reflect on their opinion of the CCN. Thematic analysis of the responses revealed the following:

- *Enhanced engagement through storytelling and entertainment*: The combination of storytelling and entertainment in the CCN heightened student engagement, making the learning process more enjoyable and effective compared with traditional methods.
- *Improved memorability and recall of medical concepts*: The CCN's engaging narrative and multimedia elements enhanced memory retention, making complex medical information more accessible and memorable.
- *Relatability through pop culture and personal connection*: Leveraging familiar pop culture icons such as Selena Gomez helped students form a personal connection with the material, enhancing engagement and motivation to learn.
- *Preference for interactive and detailed learning*: Some students value interactive learning environments and detailed information, suggesting that while the CCN is engaging, it could be further enhanced by incorporating active learning elements and comprehensive content.
- *Suggestions for improvement*: Attention to technical elements, such as the use of genAI voice narration, could improve the overall effectiveness and reception of the CCN.

The thematic analysis reveals that the CCN "Shattered Slippers" was preferred over traditional case studies due to its engaging storytelling, enhanced memorability, and relatability through pop culture references. While students appreciated the innovative approach, some expressed a desire for more interactive learning methods and provided suggestions for technical improvements. Incorporating these insights can further refine the CCN as a valuable tool in medical education.

In addition to the survey feedback from the SIS-M, the success of the "Shattered Slippers" CCN was further demonstrated academically. Students displayed strong comprehension and knowledge of the material covered, achieving an average score of 88% on examination questions pertaining to the case study content. This high performance underscores the effectiveness of the CCN as a teaching tool, suggesting that it may also be useful in promoting academic performance as well as student preference and interest.

Discussion

Principal Findings

The "Shattered Slippers" CCN supports the pedagogical value of integrating innovative genAI-driven methods and culturally resonant themes into medical education. Our study shows the capacity of this approach to not only enhance student interest but also promote their understanding and retention of complex subject matter. Furthermore, it adds very little to no extra time to the lecture material, as it basically reskins the existing material into a more cinematic experience. This is particularly important, as many new active learning teaching methodologies either extend the amount of time students spend with the material or cause instructors to remove large amounts of material in order to incorporate novel active learning activities. We considered it ethical to clearly mark AI-generated images of real individuals to avoid confusion but did not deem it necessary to label AI-generated material such as text or audio that was not mimicking a real-world person. As genAI models continue to improve in generating realistic images and cloned voices, it will become increasingly important to label AI-generated materials that mimic real-world individuals to prevent confusion with reality and avoid potential legal issues.

This study shows the importance of engaging students beyond conventional didactic methods, suggesting that the inclusion of elements such as plot development, multimedia, and popular culture can make learning more relatable and impactful. The feedback from the SIS-M supports that this approach can effectively address the initial problem of student disengagement and the need for more effective educational strategies as identified in the introduction.

The process of creating CCNs with genAI tools is highly efficient and cost-effective. Designing the case outline took about a day, while plot and narration generation were completed in seconds using GPT-4 and Eleven Labs. Image and theme song generation took under an hour each, with slight delays due to iterative refinement. Overall, the time investment was minimal compared with traditional methods. The required technical skills are basic, involving familiarity with genAI platforms for text, image, and audio generation and standard project management skills to integrate these elements into a PowerPoint slide deck. In terms of cost, the only expense was a US $20 per month subscription to ChatGPT; other platforms were used on free tiers. This low cost, combined with fast production times, makes migrating to this format highly accessible and efficient for educators, offering significant time and cost savings compared with traditional content creation methods of this caliber.

Future directions of this work will explore how similar immersive educational experiences can be scaled and adapted for diverse student populations and learning environments. The versatility of genAI-enhanced CCNs extends beyond pharmacology, offering potential applications in other areas such as anatomy, pathology, and clinical skills. This pedagogical strategy can be adapted to various medical disciplines, making abstract topics more engaging and accessible to diverse learners. It also asks questions on how educational policies might evolve to integrate this type of AI-generated material into curricula systematically. As genAI becomes more integral to education, policies must address both the ethical use of genAI and the need for genAI literacy among educators and students. Personalized, genAI-driven learning experiences could revolutionize how content is delivered, providing flexibility and tailored learning opportunities. There is an opportunity to explore interdisciplinary collaborations, merging medical education with fields such as AI, storytelling, and multimedia design. These collaborations could further refine educational tools and help bridge the gap between traditional learning and modern health care technologies, fostering genAI literacy in future medical professionals. This promising pilot study shows potential for scalability and broad applicability of genAI-enhanced CCNs. The strategy offers a model for

transforming how complex medical topics are taught, providing a scalable, engaging solution that can be adapted across different medical content areas to meet evolving educational needs.

Limitations

Our project has limitations in terms of cultural adaptability due to its reliance on specific cultural references and celebrity figures, which may not resonate with all audiences. Furthermore, the use of genAI technologies presents challenges in environments with varying levels of technological resources and differing instructor familiarity with these platforms. While the skills required to effectively use genAI can vary depending on the model, these challenges are mitigated by the increasing availability of more user-friendly genAI platforms. These platforms are simplifying AI integration in educational contexts, expanding the potential for their broader application. For instance, prompt engineering, which is crucial for optimizing output from LLMs, is becoming less essential with newer versions such as ChatGPT's o1-preview model, which incorporates many of these strategies into the system itself. This reduces the need for advanced user expertise and lowers the barrier to efficient LLM use.

Another limitation of our study is the process of validity checking for AI-generated content. Although the materials were reviewed by medical professionals, including physicians, PhDs, and PharmDs, to ensure accuracy, the use of genAI introduces potential risks in content reliability, especially as AI-generated content may produce subtle inaccuracies or lack the nuanced context that a human expert might provide. Future implementations of this approach would benefit from a formalized validation process to ensure that the clinical and educational integrity of AI-generated materials is maintained.

The evaluation methodology, focusing on immediate reactions via the SIS-M, provides a single time point of the resource's impact but does not capture the longevity of knowledge retention or the applicability of the learned material in clinical settings. Furthermore, the study included a limited sample size, with only 18 respondents to the SIS-M survey, which may not provide a comprehensive view of the broader student population. Future research could explore longitudinal studies to measure the lasting educational benefits of such methodologies with a larger participant population.

Furthermore, our study lacked a control or comparison group, a common challenge in medical education research. All students in the study were exposed only to the CCN case, and without a traditional case-based learning comparison, it is difficult to isolate the exact impact of the CCN on student performance. While we acknowledge that a control group could provide valuable insights, the integration of such comparisons is often logistically difficult in medical school settings. Future studies could address this by designing more controlled experimental conditions or through the use of quasi-experimental designs to better understand the differential effects of various educational interventions on learning.

Conclusions

The "Shattered Slippers" CCN demonstrates the effectiveness of combining cinemeducation with genAI in medical education. This approach enhanced student engagement, promoted knowledge retention, and offered a novel perspective on complex pharmacological clinical cases. The application and positive student feedback suggest that this multimodal genAI approach to educational content creation has potential for broader application in medical education. Our project also highlights the need for continuous innovation and adaptation in teaching methodologies to meet the evolving demands of health care education. Future research and development in this area could further transform medical education, making it more engaging, effective, and aligned with modern technological advancements.

Acknowledgments

The author would like to extend his heartfelt gratitude to his students for their participation and invaluable contributions to the "Shattered Slippers" project. Their engagement and feedback were essential in shaping this educational endeavor and its success.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Shattered Slippers: cinematic clinical narrative.
[PDF File, 878 KB - mededu_v11i1e63865_app1.pdf]

Multimedia Appendix 2
Shattered Slippers full presentation.
[MP4 File, 134196 KB - mededu_v11i1e63865_app2.mp4]

Multimedia Appendix 3
ChatGPT plot generation.
[DOCX File, 20 KB - mededu_v11i1e63865_app3.docx]

Multimedia Appendix 4
Leonardo.ai image generation.
[DOCX File, 3053 KB - mededu_v11i1e63865_app4.docx]

Multimedia Appendix 5
Eleven Labs narration generation and audio clips.
[DOCX File, 13 KB - mededu_v11i1e63865_app5.docx]

Multimedia Appendix 6
ChatGPT and Suno Chirp Bot theme song generation and audio clip.
[DOCX File, 15 KB - mededu_v11i1e63865_app6.docx]

Multimedia Appendix 7
Situations Interest Survey of Multimedia CHERRIES (Checklist for Reporting Results of Internet E-Surveys) report.
[DOCX File, 15 KB - mededu_v11i1e63865_app7.docx]

References

1. Kassab SE, Taylor D, Hamdy H. Student engagement in health professions education: AMEE Guide No. 152. Med Teach 2023 Sep;45(9):949-965. [doi: 10.1080/0142159X.2022.2137018] [Medline: 36306374]
2. Kassab SE, Al-Eraky M, El-Sayed W, Hamdy H, Schmidt H. Measurement of student engagement in health professions education: a review of literature. BMC Med Educ 2023 May 20;23(1):354. [doi: 10.1186/s12909-023-04344-8] [Medline: 37210491]
3. Hodges LC. Student engagement in active learning classes. In: Active Learning in College Science: The Case for Evidence-Based Practice: Springer Cham; 2020:27-41. [doi: 10.1007/978-3-030-33600-4_3]
4. Miller CJ, McNear J, Metz MJ. A comparison of traditional and engaging lecture methods in a large, professional-level course. Adv Physiol Educ 2013 Dec;37(4):347-355. [doi: 10.1152/advan.00050.2013] [Medline: 24292912]
5. Rueb M, Siebeck M, Rehfuess EA, Pfadenhauer LM. Cinemeducation in medicine: a mixed methods study on students' motivations and benefits. BMC Med Educ 2022 Mar 12;22(1):172. [doi: 10.1186/s12909-022-03240-x] [Medline: 35279156]
6. Narayanan S, Ramakrishnan R, Durairaj E, Das A. Artificial intelligence revolutionizing the field of medical education. Cureus 2023 Nov;15(11):e49604. [doi: 10.7759/cureus.49604] [Medline: 38161821]
7. Arango-Ibanez JP, Posso-Nuñez JA, Díaz-Solórzano JP, Cruz-Suárez G. Evidence-based learning strategies in medicine using AI. JMIR Med Educ 2024 May 24;10(1):e54507. [doi: 10.2196/54507] [Medline: 38801706]
8. Karim MY. Using clinical cases to restore basic science immunology knowledge in physicians and senior medical students. Front Immunol 2020;11:1756. [doi: 10.3389/fimmu.2020.01756] [Medline: 32973743]
9. Haidaris CG, Frelinger JG. Inoculating a new generation: immunology in medical education. Front Immunol 2019;10:2548. [doi: 10.3389/fimmu.2019.02548] [Medline: 31749807]
10. Ousey K, Gallagher P. The theory-practice relationship in nursing: a debate. Nurse Educ Pract 2007 Jul;7(4):199-205. [doi: 10.1016/j.nepr.2007.02.001] [Medline: 17689445]
11. Bhugra D, Molodynski A, Ventriglio A. Well-being and burnout in medical students. Ind Psychiatry J 2021;30(2):193-197. [doi: 10.4103/ipj.ipj_224_21] [Medline: 35017800]
12. Bhargava M, Naik PR, Hegde P, Navya N, Sachith M, Vineetha S. Integrating narrative medicine through story-telling: a feasibility study in a community medicine curriculum for undergraduate and postgraduate students. Cureus 2023 Jul;15(7):e41851. [doi: 10.7759/cureus.41851] [Medline: 37581154]
13. Kagawa Y, Ishikawa H, Son D, et al. Using patient storytelling to improve medical students' empathy in Japan: a pre-post study. BMC Med Educ 2023 Jan 27;23(1):67. [doi: 10.1186/s12909-023-04054-1] [Medline: 36707818]
14. Liao HC, Wang YH. Storytelling in medical education: narrative medicine as a resource for interdisciplinary collaboration. Int J Environ Res Public Health 2020 Feb 11;17(4):1135. [doi: 10.3390/ijerph17041135] [Medline: 32053911]
15. Cooper A, Rodman A. AI and medical education—a 21st-century Pandora's box. N Engl J Med 2023 Aug 3;389(5):385-387. [doi: 10.1056/NEJMp2304993] [Medline: 37522417]
16. Russell RG, Lovett Novak L, Patel M, et al. Competencies for the use of artificial intelligence-based tools by health care professionals. Acad Med 2023 Mar 1;98(3):348-356. [doi: 10.1097/ACM.0000000000004963] [Medline: 36731054]
17. Tuma F. The use of educational technology for interactive teaching in lectures. Ann Med Surg (Lond) 2021 Feb;62:231-235. [doi: 10.1016/j.amsu.2021.01.051] [Medline: 33537136]
18. Charon R. Narrative medicine: a model for empathy, reflection, profession, and trust. JAMA 2001 Oct;286(15):1897-1902. [doi: 10.1001/JAMA.286.15.1897] [Medline: 11597295]
19. Dennick R. Constructivism: reflections on twenty five years teaching the constructivist approach in medical education. Int J Med Educ 2016 Jun 25;7:200-205. [doi: 10.5116/ijme.5763.de11] [Medline: 27344115]

20. Mayer RE. Cognitive theory of multimedia learning. In: The Cambridge Handbook of Multimedia Learning: Cambridge University Press; 2012:31-48. [doi: 10.1017/CBO9780511816819.004]
21. Niess ML. Technological Pedagogical Content Knowledge (TPACK) Framework for K-12 Teacher Preparation: Emerging Research and Opportunities: Information Science Reference; 2016:1-173. [doi: 10.4018/978-1-5225-1621-7]
22. ChatGPT. OpenAI. URL: https://openai.com/chatgpt [accessed 2023-12-19]
23. Wang J, Liu Z, Zhao L, et al. Review of large vision models and visual prompt engineering. Meta Rad 2023 Nov;1(3):100047. [doi: 10.1016/j.metrad.2023.100047]
24. White J, Fu O, Hays S, et al. A prompt pattern catalog to enhance prompt engineering with ChatGPT. arXiv. Preprint posted online on Feb 21, 2023 URL: https://arxiv.org/abs/2302.11382v1 [accessed 2024-09-23] [doi: 10.48550/arXiv.2302.11382]
25. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. Presented at: 36th Conference on Neural Information Processing Systems (NeurIPS 2022); Nov 28 to Dec 9, 2022; New Orleans, Louisiana URL: https://arxiv.org/abs/2205.11916v4 [accessed 2023-09-23]
26. Leonardo.ai. URL: https://leonardo.ai/ [accessed 2023-12-19]
27. Text to speech & AI voice generator. ElevenLabs. URL: https://elevenlabs.io/ [accessed 2023-12-19]
28. SUNO AI. URL: https://www.suno.ai/ [accessed 2023-12-19]
29. Dousay TA. Effects of redundancy and modality on the situational interest of adult learners in multimedia learning. Education Tech Research Dev 2016 Dec;64(6):1251-1271. [doi: 10.1007/s11423-016-9456-3]
30. Dousay TA, Trujillo NP. An examination of gender and situational interest in multimedia learning environments. Brit J Educ Tech 2019 Mar;50(2):876-887. [doi: 10.1111/bjet.12610]
31. Bland T, Guo M, Dousay TA. Multimedia design for learner interest and achievement: a visual guide to pharmacology. BMC Med Educ 2024 Feb 5;24(1):113. [doi: 10.1186/s12909-024-05077-y] [Medline: 38317141]
32. Huang J, Gu S, Hou L, et al. Large language models can self-improve. Presented at: EMNLP 2023—Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; Dec 6-10, 2023; Singapore p. 1051-1068. [doi: 10.18653/v1/2023.emnlp-main.67]

Abbreviations

AI: artificial intelligence
CCN: cinematic clinical narrative
genAI: generative artificial intelligence
LLM: large language model
MF: maintained-feeling
MV: maintained-value
SIS-M: Situational Interest Survey for Multimedia
TPACK: Technological Pedagogical Content Knowledge

Edited by B Lesselroth, G Eysenbach; submitted 01.07.24; peer-reviewed by M Guo, MI Knopp; revised version received 26.09.24; accepted 07.11.24; published 06.01.25.

Please cite as:
Bland T
Enhancing Medical Student Engagement Through Cinematic Clinical Narratives: Multimodal Generative AI–Based Mixed Methods Study
JMIR Med Educ 2025;11:e63865
URL: https://mededu.jmir.org/2025/1/e63865
doi:10.2196/63865

Original Paper

Perceptions and Earliest Experiences of Medical Students and Faculty With ChatGPT in Medical Education: Qualitative Study

Noura Abouammoh[1,2], MBBS, PhD; Khalid Alhasan[1,3,4], MBBS; Fadi Aljamaan[1,5], MD; Rupesh Raina[6], MD; Khalid H Malki[1,7], PhD; Ibraheem Altamimi[1], MBBS; Ruaim Muaygil[1,8], MBBS; Hayfaa Wahabi[2,9], MD, PhD; Amr Jamal[1,2,9], MBBS; Ali Alhaboob[1,3], MBBS; Rasha Assad Assiri[10], MBBS; Jaffar A Al-Tawfiq[11,12,13], MBBS; Ayman Al-Eyadhy[1,3], MD; Mona Soliman[1,8], MBBS, PhD; Mohamad-Hani Temsah[1,3,9], MD

[1]College of Medicine, King Saud University, Riyadh, Saudi Arabia

[2]Department of Family and Community Medicine, King Saud University Medical City, King Saud University, Riyadh, Saudi Arabia

[3]Pediatric Department, King Saud University Medical City, King Saud University, Riyadh, Saudi Arabia

[4]Department of Kidney and Pancreas Transplant, Organ Transplant Center of Excellence, King Faisal Specialist Hospital & Research Centre, Riyadh, Saudi Arabia

[5]Critical Care Department, King Saud University Medical City, King Saud University, Riyadh, Saudi Arabia

[6]Department of Nephrology, Cleveland Clinic Akron General and Akron Children Hospital, Akron, OH, United States

[7]Research Chair of Voice, Swallowing, and Communication Disorders, Department of Otolaryngology, College of Medicine, King Saud University, Riyadh, Saudi Arabia

[8]Medical Education Department, King Saud University Medical City, King Saud University, Riyadh, Saudi Arabia

[9]Evidence-Based Health Care & Knowledge Translation Research Chair, Family & Community Medicine Department, College of Medicine, King Saud University, Riyadh, Saudi Arabia

[10]Department of Basic Medical Sciences, College of Medicine, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia

[11]Specialty Internal Medicine and Quality Department, Johns Hopkins Aramco Healthcare, Dhahran, Saudi Arabia

[12]Infectious Disease Division, Department of Medicine, Indiana University School of Medicine, Indianapolis, IN, United States

[13]Infectious Disease Division, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, United States

Corresponding Author:
Mohamad-Hani Temsah, MD
Pediatric Department
King Saud University Medical City
King Saud University
King Abdullah Road
Riyadh, 11424
Saudi Arabia
Phone: 966 114692002
Email: mtemsah@ksu.edu.sa

Abstract

Background: With the rapid development of artificial intelligence technologies, there is a growing interest in the potential use of artificial intelligence–based tools like ChatGPT in medical education. However, there is limited research on the initial perceptions and experiences of faculty and students with ChatGPT, particularly in Saudi Arabia.

Objective: This study aimed to explore the earliest knowledge, perceived benefits, concerns, and limitations of using ChatGPT in medical education among faculty and students at a leading Saudi Arabian university.

Methods: A qualitative exploratory study was conducted in April 2023, involving focused meetings with medical faculty and students with varying levels of ChatGPT experience. A thematic analysis was used to identify key themes and subthemes emerging from the discussions.

Results: Participants demonstrated good knowledge of ChatGPT and its functions. The main themes were perceptions of ChatGPT use, potential benefits, and concerns about ChatGPT in research and medical education. The perceived benefits included collecting and summarizing information and saving time and effort. However, concerns and limitations centered around the potential lack of critical thinking in the information provided, the ambiguity of references, limitations of access, trust in the output of ChatGPT, and ethical concerns.

Conclusions: This study provides valuable insights into the perceptions and experiences of medical faculty and students regarding the use of newly introduced large language models like ChatGPT in medical education. While the benefits of ChatGPT were recognized, participants also expressed concerns and limitations requiring further studies for effective integration into medical education, exploring the impact of ChatGPT on learning outcomes, student and faculty satisfaction, and the development of critical thinking skills.

(JMIR Med Educ 2025;11:e63400) doi:10.2196/63400

KEYWORDS

ChatGPT; medical education; Saudi Arabia; perceptions; knowledge; medical students; faculty; chatbot; qualitative study; artificial intelligence; AI; AI-based tools; universities; thematic analysis; learning; satisfaction

Introduction

Artificial intelligence (AI) is a computer-based technology invented as a digital system to imitate and aid human intellect and skills. The wide use of AI technology is changing the medical field considerably, aiming for more efficient patient management. Medical education is one of the vital domains of health care practice, in which AI has a promising contribution by providing an alternative and efficient means of information access, achieving teaching goals and skills development. As an example, the integration of AI in simulated surgical skills learning showed comparable results compared to remote expert instructions [1], but it led to unintended outcomes in another study, which affected trainees' efficiency metrics on the cost of safer skills development [2]. Case-based learning is another potential field harnessing AI technology in medical education, which has shown promising results [3]. AI technology has also been used in teaching clinical examination skills, such as breast self-examination, yielding mixed results: high levels of student satisfaction paired with increased anxiety [4]. Such AI-driven interventions will be leading health care practice in the future, such as the introduction of machine-based surgical treatment with robotic surgery, which has effectively promoted diagnostic accuracy, achieving treatment goals and saving health care professionals' workload [5-7]. AI technology integration in medical education and medical research will not only contribute to patients' care but also improve if not revolutionize the medical education system [8,9]. All these changes of AI integration into the medical practice need to be accompanied by evolution in the medical teaching and training curricula [8,9], facing significant interest among educators and researchers recently on AI's rapid involvement in medical education [10-13].

One of the pioneer and popular generative AI-based tools is ChatGPT, a language model developed by OpenAI that uses natural language processing to generate humanlike responses to queries, with many potential applications in health care [14,15]. ChatGPT was perceived by health care workers to positively impact the future of health care systems by 76.7% in a recent study [16]. However, little is known specifically about the perceptions and experiences of faculty and students or trainees against the use of ChatGPT in the context of medical education within Saudi Arabia.

The health care sector in Saudi Arabia is experiencing dramatic growth and reformatting, with a strong emphasis on prioritizing medical education and digitizing the health systems. Therefore, using AI technology in the health care system is a promising strategy for substantial investments in medical, nursing, and other specialized educational disciplines [17]. As medical education evolves, the use of AI-based tools like ChatGPT could potentially transform the way medical education is delivered [18]. Literature has a gap in assessing the perceptions and attitudes of medical education stakeholders regarding integrating AI technology in curricula, clinical teaching, and simulation skills development. Most literature addressed specific AI technology adoption in medical practice or certain educational domains but did not assess it collectively in multiple domains related to medical education. Therefore, it is crucial to explore the medical faculty staff and students' knowledge, perceived benefits, concerns, and limitations of ChatGPT application in medical education.

This qualitative study seeks to explore the perception on the use of newly introduced AI chatbots, like ChatGPT3.5, in medical education from the perspective of faculty and medical students. By deepening our understanding of faculty and students' knowledge about ChatGPT and its applications in medical education, this study identifies both the facilitators and barriers to its use. The research offers valuable preliminary insights into the acceptance of AI-based tools in medical education and informs the development of effective strategies for integrating such tools within medical education systems in Saudi Arabia and similar contexts, as more AI models evolve.

Methods

Study Design

This study was conducted using a focus group technique at the College of Medicine, King Saud University, a leading university in Saudi Arabia [19]. The study included faculty and students from different levels.

The study aims to preliminarily explore and understand participants' perceptions of ChatGPT, a newly introduced large language model. A qualitative methodology was chosen, as it is well suited to exploring experiences, meanings, and perspectives from participants' viewpoints [20-22]. Examining the perceptions of both faculty and students enables a comparative head-to-head analysis of their viewpoints. Qualitative methodology provides a deep explanation of different viewpoints participants may have about ChatGPT use in medical education. It can also allow the authors to propose probing questions to understand and explore users' perceptions. Although individual interviews would elicit a more detailed picture of an issue, focus group discussion was used, as the aim

of the study is to explore different viewpoints using participants' dynamics and thought sharing to enrich the discussion [23]. Data source triangulation was applied to support the trustworthiness of the findings and allow prelude comparison.

Participants were recruited from the College of Medicine through purposive sampling. As a small number of faculty and students used ChatGPT at the time of data collection, a purposive sample was applied. A student was asked to announce the need to interview students who have ever used ChatGPT. Another announcement to faculty was made, and an invitation was sent to random faculties from 3 departments who use or want to share their ideas about ChatGPT in medical education. The sample included 6 medical faculty members (2 associate professors and 4 professors) and 6 medical students (2 second year, 2 third year, 1 fourth year, and 1 fifth year). Two focus group discussions were conducted in April 2023 on the Zoom platform (Zoom Video Communications), one with faculty members and the other with students, and each group consisted of 6 participants. The discussions were conducted in English language as preferred by the participants. Two of the authors (NA and MHT) served as moderators, and each discussion lasted for approximately 1 hour.

Using the Zoom platform in data collection facilitated gathering participants at the same time after working hours. As the team acknowledged that nonverbal cues may not be detected as participants refrained from opening their cameras, follow-up questions and probing were used to minimize subjectivity in understanding participants' responses. Not all participants knew each other; hence, the setting was more private to freely share opposing views.

A topic guide was prepared by the author (NA) to cover aspects such as participants' familiarity with ChatGPT, its uses, facilitators, and limiting factors of its incorporation in medical education. Probing and follow-up questions were allowed depending on participants' responses. The themes were saturated at that time after the second interview possibly due to the limited experience of participants in the early stages of ChatGPT launch. Thematic analysis was used to analyze the data using a priori themes and allowing new themes to emerge from the data [24]. The discussions were transcribed using Zoom's automatic transcription feature. This feature had the advantage of identifying the name that the participants chose for themselves in the discussion and linking it with the speaker.

The transcripts were revised and read multiple times to identify patterns and themes that emerged from the data. A coding framework was developed from the data by each coder (NA and MHT) and applied using NVivo software (version 12; QSR International) [25]. Themes were identified and refined through an iterative process of coding, reviewing, and discussing the data among the research team until a consensus was reached [24]. Initial codes were developed by 2 different authors (NA and MHT) and then, comparison and discussion were made to agree on the coding framework. Coding themes were similar, and no major changes were made in the thematic framework.

Ethical Considerations

This study received ethics approval from the Institutional Review Board at King Saud University (approval 23/0155/IRB). All participants provided verbal informed consent prior to their inclusion in the study, including consent for the audio recording of interviews. Participants were fully informed about the purpose of the study, the voluntary nature of their participation, and their right to withdraw at any time without any consequences. To ensure participant privacy and confidentiality, pseudonyms were assigned, and no identifying information was included in the transcripts or final report. The data were securely stored and accessible only to authorized members of the research team. No compensation was provided to participants for their participation in the study.

Results

Overview

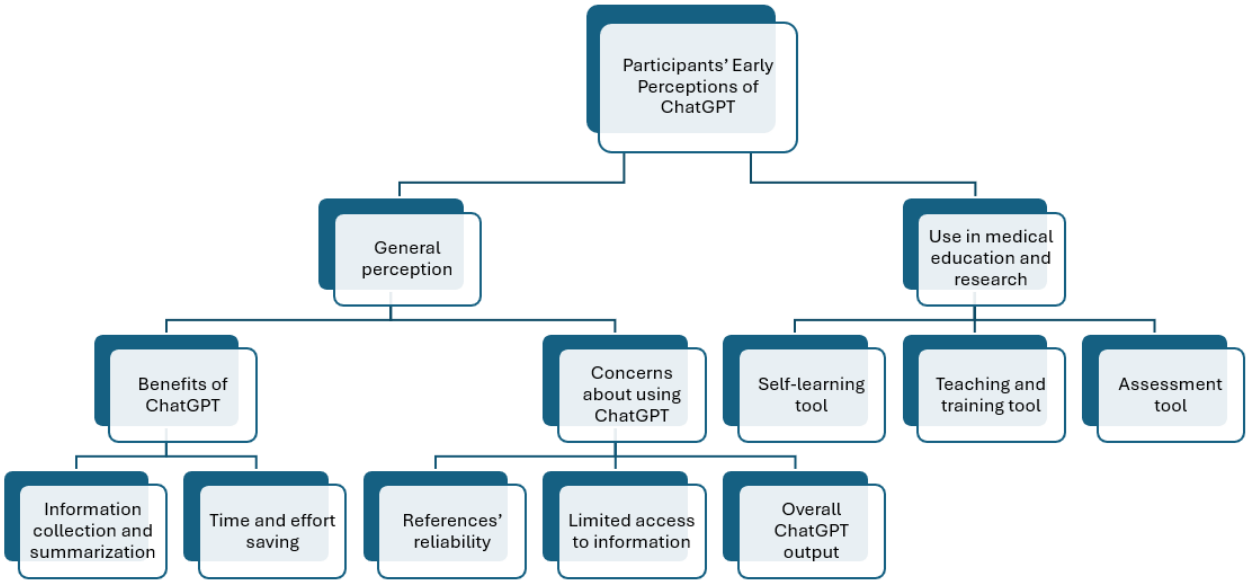
In total, 6 medical faculty staff and 6 medical students with different experiences with ChatGPT participated in the study. Table 1 shows their demographic data. Figure 1 displays the thematic framework used to assess participants' perception of ChatGPT in general and in medical education.

Analysis of the data from the discussion generated two main themes: (1) participants' general perception of ChatGPT and (2) ChatGPT use in medical education and research.

Table 1. Participants position, department, and frequency of ChatGPT use.

Participant code	Position	Department	Age (years) and sex	Using ChatGPT
Participant 1	Faculty	Critical Care Department, College of Medicine	44 and male	Regular user in medical education
Participant 2	Faculty	Ear, Nose, and Throat department, College of Medicine	56 and male	Regular user in medical education
Participant 3	Faculty	Family medicine, College of Medicine	60 and female	Not a user
Participant 4	Faculty	Pediatrics department, College of Medicine	38 and male	Not a user
Participant 5	Faculty	Pediatrics department, College of Medicine	41 and male	Regular user in medical education
Participant 6	Faculty	Medical Education Department, College of Medicine	58 and female	Not a user
Student 1	Student	College of Medicine	20 and male	Regular user for general search
Student 2	Student	College of Medicine	22 and male	Regular user for general search
Student 3	Student	College of Medicine	21 and male	Regular user for general search
Student 4	Student	College of Medicine	21 and male	Regular user for general search
Student 5	Student	College of Medicine	19 and male	Regular user for general search
Student 6	Student	College of Medicine	20 and male	Regular user for general search

Figure 1. Thematic framework of participants' perception on using ChatGPT.



Participants' General Perception of ChatGPT

Overview

All participants expressed good knowledge of ChatGPT's main goal and functions. One participant noted:

> *The idea from this software is that it will chat with you regarding any topic you will ask about...it chats with me in a human like manner, and collect for me the answers from all over resources, and display them.* [Participant 1]

One student described ChatGPT as an "assistant," and others elaborated:

> *Artificial intelligence helps me execute the command that I'm asked to execute.* [Student 2]

> *It's another way of searching for highly accurate information, depending on what I search for and how I search for it.* [Student 5]

Participants were challenged about ChatGPT compared to other traditional search engines: "It is not at similar to Google, even Google started invention of AI application to enrich its platform" (Participant 2). Most participants supported the use of ChatGPT but were not concerned about its information sources.

Benefits of ChatGPT

Two main subthemes emerged from the focus group discussions about the benefits of using ChatGPT.

Collects and Summarizes Information

The majority of the participants believed that searching for information through ChatGPT is more efficient compared to

standard search engines, as the former saves time by summarizing and textualizing the raw information output from the search: "ChatGPT is beautiful in collecting information and presenting it to me in a simplified text that I can easily comprehend" (Participant 2).

Few participants used ChatGPT to review scientific papers or provide ideas for new papers: "I used it to study limitations of studies and the future recommendations for studies I was asked to review...it gives me ideas" (Participant 2).

Students also noted:

> *I find ChatGPT more directive towards what I ask, and to the point, mostly because when I look for something on classical search engine, such as Google...I have to go into some sub web pages which has an answer and look between all the thousands of answers to find one. While ChatGPT will give it to me concisely like this is option A, option B, option C.* [Student 3]

Another faculty added, "It will do the search for me; then even with critically appraise it and give me the final result" (Participant 1). However, one faculty participant was more conservative in her comments about using AI in collecting data and did not perceive the information displayed by ChatGPT as reliable because it lacks the "critical thinking" skill to enable it to reach a final scientific plausible conclusion, "The problem of collecting all the information in one place is that collecting the information and giving it in a nutshell, in one place. This machine is not critically thinking" (Participant 3).

Saves Time and Efforts

Opinions varied in terms of whether using ChatGPT saves time and effort, considering the perceived benefits. One faculty mentioned: "It saves time when I'm stuck in generating exam question" (Participant 2). A student added: "It's not accurate, but at least it saves me time. This is the most important point" (Student 4).

On the other hand, another faculty participant subtly disclosed her denunciation about the functionality of ChatGPT. She believed that ChatGPT helps partly in performing tasks, but that advantage is contradicted by paying time to verifying and authenticating the ChatGPT output.

> *Me as a researcher. When I search for information I'm putting it together, ChatGPT tries to put it for me. So far, I can't see it superior to the human mind...It does some of the work for me, but I have to take it with a bunch of salt.* [Participant 3]

Concerns About Using ChatGPT

When the participants discussed the drawbacks of using ChatGPT, they mentioned expressions such as "hallucination" and "blinding euphorically." The following subthemes emerged as perceived drawbacks of ChatGPT.

References Reliability

While few participants were not sure about the source of ChatGPT information, most participants believed that it is the internet: "It's the same data retriever as Google" (Participant 3). Other participants had a deeper view: "ChatGPT generated references, and citations have to be taken with caution" (Participant 2).

Faculty experienced situations where they doubted the reference of the information provided by ChatGPT. For example, one faculty noted: "I'm not sure what are the sources used to extract information, even if I ask for references, it might not mention them...or at least it will not volunteer in mentioning them" (Participant 1). While others defended that: "If it doesn't have access to the reference, it will tell that it doesn't have access, but if the reference is online, it can refer to that" (Participant 2).

Similarly, a student commented: "It is multitasking, rather than searching for the source of information, it presents the answer and references" (Student 5). One participant pointed that the unreliability of ChatGPT sources supports her view of not relying on ChatGPT.

Limited Access to Information

Several participants acknowledged ChatGPT's limitation in accessing all available information, driving caution while using ChatGPT:

> *One of the restrictions regarding medical search, it's restricted to certain resources like PubMed...there are some other medical websites that it cannot access yet.* [Participant 4]

> *We don't know the algorithm behind the search nor exactly how it looks for information.* [Participant 6]

One of the participants elaborated that ChatGPT is invented by humans; therefore, they may manipulate or restrict its search and output.

> *It's not free of bias. If I am asking for something morally wrong or illegal. It will not answer because it is constrained. So, it is not fully free from human constraints.* [Student 3]

Some faculty participants raised an ethical concern that may affect the trust in ChatGPT information. One participant explained:

> *Can drug company pay ChatGPT to display answers that are in favour of certain medication? Could ChatGPT be manipulated? ChatGPT inventors are for sure looking for money somehow by anyway!* [Participant 1]

Overall ChatGPT Output

All participants believed that ChatGPT users should not fully trust the information presented and practice caution, while others elaborated that it is ideal for new topics as a jumpstart:

> *I should not take it (information from ChatGPT) for granted; I have to review what's there, but it gives me a nice idea, very excellent ideas...It sheds the light on some certain angles that I was not looking for.* [Participant 2]

Some participants pointed that trusting ChatGPT output depends on your previous background about the topic:

> *I should have the ability to differentiate between what is reliable and what is not reliable...Myself, I am not well-versed in medical education. For example, I am highly qualified in research, but regarding education I take for granted whatever output from ChatGPT in that regard, while I can filter information regarding research and judge it well.* [Participant 3]

One student agreed:

> *It depends on what I am looking for. Sometimes it's very accurate. Sometimes it's not...But as a human mind I have an idea about what I am looking for, therefore, I can judge if its accurate or doubt the answer.* [Student 4]

All participants agreed that the unfamiliarity of ChatGPT users with its search algorithm enforced the participants' trust issue.

A faculty explained:

> *Do we know the ChatGPT searching methodology? is it scientific methodology? How it extracts the information from the paper, how it appraises it? What are the sources that this engine has access to? All this will augment the reliability of my experience.* [Participant 1]

One participant mentioned that ChatGPT cannot be used for critical thinking in certain contexts; thus, it cannot be fully trusted:

> *It cannot give me what is relevant to me, my community and population and my students...It might be dangerous to put ChatGPT superior to human intellect!* [Participant 3]

Another faculty participant defended the ChatGPT's reliability, noting that it declares its level of expertise and specialty ahead of each information presented:

> *If I ask ChatGPT about something in geology, it will start with "I am not a geologist" and then move on with the dialogue...and it finishes the response by "it is very important to refer to those sources."* [Participant 2]

ChatGPT Use in Medical Education and Research

Participants discussed ChatGPT use in medical education from 3 aspects as discussed below, but in general, they raised concerns about using it without appropriate and dedicated training.

Self-Learning Tool

The majority of faculty participants supported using ChatGPT in the teaching process. A faculty participant commented:

> *Students are no longer enjoying the usual long lectures, or didactic lectures but they enjoy more challenging aspects exploring a new experience, and living it...I think the ChatGPT could be used as a very good trigger for the students to go and read and find out more, discuss among themselves and go explore this with their seniors, with their educators.* [Participant 5]

Another faculty added that it should be used to get an idea about a topic, but further reading is important for students:

> *ChatGPT is like a short fast access to a topic, it helps to get the most important information...they (students) need to read the references.* [Participant 1]

However, another faculty participant raised concerns using ChatGPT for concluding opinions and summing debates:

> *If they (students) use ChatGPT just for recalling information then no problem...But if they want to make inferences, they should not use it.* [Participant 3]

ChatGPT methodology was raised by another faculty participant who did not support using it in learning at all because of its unclear methodology and unverified information sources.

On the other side, the majority of student participants did not support using ChatGPT to obtain information and felt the traditional search engines are more reliable and easier to use:

> *I do not perceive it as a search engine. I don't look up medical information on it, or anything, because I find the classic search engines easier.* [Student 2]
>
> *I know exactly where the reliable sources are. Then I can take the information from other sources with confidence, and more simple steps.* [Student 1]

Some participants, while supportive of ChatGPT's use in medical practice, emphasized its role in clinical medicine education. They raised concerns about its impact on decision-making, particularly due to ChatGPT's inadequate or unclear strategy for disclosing information sources:

> *If I look at the other search engines for which support medical information, they present like up-to-date information...ChatGPT is very complex, and the methodology and the algorithm it uses is not clear so, it is not a reliable source of information for decision making and for serious information.* [Participant 1]

The issue of updated sources in ChatGPT was also raised:

> *We need to be cautious about using the information...the medical field information is changing very quickly, so we have to be careful about this point.* [Participant 4]

Other participants debated that the information accuracy depends on user searching and prompt engineering skills:

> *Prompt questions will make the difference in getting the response, and I recommend digging into the prompts technology to get more accurate answers, and doing this is important to acquire the right answer.* [Participant 2]
>
> *You get the response according to the precision of the search.* [Student 1]

Interestingly, a faculty participant raised the concern of students and faculty losing their critical thinking skills if they depend on ChatGPT:

> *It is dangerous...because we are replacing critical thinking. We are prioritizing this thing over human intellect.* [Participant 3]

A student participant who expressed poor research skills was concerned about such skills being affected or even weakened by dependence on ChatGPT in research. In general, students did not support the use of ChatGPT as the primary source of information, especially for new topics, but as a collateral resource.

Teaching and Training Tool

Some participants believed that teaching modalities should change after the introduction of AI technology. They expressed optimism of more teaching methodology shifting from memorization to critical thinking; however, this aim was not perceived achievable through ChatGPT so far:

> *We must invest more in the skills of our medical students and problem-solving critical thinking analysis. These are the areas that is lacking in the ChatGPT, and that we need to focus more on.* [Participant 4]

Faculty participants raised concerns about students' replacement of traditional lectures with AI applications like ChatGPT, which might be risky in general, especially in the current stage of unverified and undedicated AI applications for medical education.

Another concern raised from one faculty regarding the lecturers and trainers:

> *Do our faculty have enough knowledge to use and recommend ChatGPT for their students and instruct them how to use it and get maximum benefit from it?* [Participant 3]

However, all students did not see themselves relying on ChatGPT for learning: "We just need to be familiar on how to use ChatGPT and use it as a tool that supports our search rather than completely relying on it" (Student 2).

Faculty participants differentiated between the needs of postgraduate and undergraduate students and their use of ChatGPT. One faculty (Participant 3) felt that using AI in training postgraduate trainees would be difficult because postgraduate training depends on building skills, while undergraduate depends on memorization as per him.

ChatGPT might be a tool to generate clinical scenarios and draw a framework for discussions with the students:

> *One problem would take weeks from our team and long hours of sitting together and creating the medical problems that we teach in the problem-based learning sessions. So, it would be interesting to see how ChatGPT deals with this.* [Participant 4]

An interesting point mentioned by some faculty is the inability of ChatGPT to teach students human, emotional, and social skills: "Using AI is not designed to help in teaching some skills such as Humanity and the communication, the teamwork" (Participant 4).

Assessment Tool

Most faculty participants mentioned using ChatGPT for academic assessment like examination questions generation:

> *I asked ChatGPT to generate questions for me with scenario and without scenario...it was good to Very good. It's not reaching to excellent level. I have to review and modify.* [Participant 2]

In addition, most faculty participants mentioned using ChatGPT for medical problems, clinical scenarios, and bedside teaching. Some faculty participants raised the idea of using AI applications like ChatGPT to assess the quality and objectives of examinations in order to guide certain questions to assess critical thinking rather than recall knowledge only. Cheating and plagiarism were one of the raised concerns by the faculty during the discussion: "We have to be very careful about cheating and misuse of ChatGPT by our medical students in medical assignments" (Participant 4).

In line with the former comment, one student defended his use of ChatGPT, raising a debatable point of using AI applications for academic assignments is ethical or not:

> *I mainly use it for writing, and then I just review it and edit it...mainly for research or some essays...for example I'd give it some data, and I ask it to write a paragraph that summarizes this data, or an introduction to something for example (Disease X).* [Student 2]

Another student mentioned that his use of ChatGPT in assignments is mainly for summarization. Others use it to collect information resources: "It can make my job way easier. For example, if I have a research assignment to just collect the resources about a topic" (Student 3).

Overall, all participants reached a conclusion of being open-minded and accepting for the ChatGPT intrusion into our lives: "I think it's coming in the near future, and we need to live in the reality to adjust and take the best out of it" (Participant 4).

Discussion

Principal Findings

This paper presents a general snapshot of the faculty and undergraduate medical students' perceptions of ChatGPT and its use in medical education. All participants demonstrated a good understanding of ChatGPT and its functionalities; some described its role as assistive, while others found it as a mere information search tool. Almost all participants were impressed by ChatGPT's ability to provide a concise summary of search results compared to traditional search engines, which is in line with the literature [26]. On the other hand, few students in our study perceived Google as a better tool for learning.

In line with other publications [27], our participants believe that ChatGPT provides a more user-appealing and faster solution for busy users by delivering a summarized, high-caliber textual output. One of the major challenges they mentioned regarding ChatGPT use is its sources of information, which is in line with previously published similar studies showing that students and

faculty are aware of the limitations of ChatGPT that influence its accuracy [26,27]. In a comparative study between platforms, ChatGPT-generated responses were considered to be reliable and beneficial, while others deemed them potentially risky [28]. For example, a study showed that there were concerns about ChatGPT advice regarding antimicrobial stewardship, general course lengths were accurate but the duration varied, and source control was either incorrectly cited as justification for prolonging therapy or ignored entirely [29]. Therefore, ChatGPT output should be dealt with skeptically and selectively, as poor users' baseline knowledge might lead to risky, dangerous, or suboptimal conclusions. Previous literature has shown that in comparison with Google, the majority of the participants tend to doubtfully trust ChatGPT output for reasons related to the novelty of AI and users, lack of understanding of its algorithm, and information sources as studied previously [14,30]. Notably, only 40% of these experts concluded that the perceived value of ChatGPT's responses outperformed those from Google [31].

Therefore, participants tend to trust ChatGPT responses if they have a previous background about the search topic. Participants suggested that while ChatGPT might be helpful in certain aspects of medical education, users should approach the information with caution and apply their medical judgment.

The participants' concerns about ChatGPT output also stemmed partly from the observed phenomenon of references' hallucinations, which raised serious concerns about its reliability and validity [32-34]. In addition, they stressed on the point of ChatGPT's limited access to updated medical literature. A previous study had cautioned authors regarding references generated by ChatGPT [35]. To overcome these limitations, developers should work on expanding the access of ChatGPT's resources, improving its search methodology, and ensuring a more comprehensive and reliable source of information.

Faculty participants explored the potential of ChatGPT in generating examination questions and clinical scenarios, enhancing bedside teaching, and reviewing assessments. Still, they emphasized the need for reviewing and modifying AI-generated content as well as the importance of developing policies and strategies to tackle potential academic misconduct related to ChatGPT use. Previous studies showed ChatGPT's excellent performance as it passed the American Heart Association examination with 84% accuracy, but it failed Taiwan's family medicine examination and fared poorly on the urology self-assessment examination [36-38]. A study concluded that ChatGPT responses were frequently incomplete and sometimes misleading [26]. However, a recent expletory review showed that ChatGPT has a potential impact on medical education, scientific research, and medical writing [14]. Thus, the ChatGPT's generated questions need to be carefully examined and revised especially regarding scientific content. Other research highlighted that generated output in that regard is not highly different among different AI platforms, as the multiple-choice question–based examination performance of ChatGPT was marginally better than that of Google's Bard [39].

Both faculty and students appreciated the time-saving advantage of ChatGPT and its fast access to information. Therefore, faculty used it in preparing lecture materials and examination questions. While students used it in their academic assignments, this mirrors a previous study about ChatGPT perception among students who used it for generating academic content, brainstorming ideas, and writing texts [40,41].

Faculty in our study and previous research raised concerns about students' ChatGPT overuse [13,27,42]. According to our participants, using it by students may interfere with their critical thinking, writing, and information retrieval skills. Faculty highlighted the students' need to critically review and modify the AI-generated content, ensuring it aligns with academic standards and expectations. Banerjee et al [11] reported that postgraduate trainee doctors have an overall positive perception of the impact of AI on clinical training; however, they found that AI will eventually reduce the trainees' clinical judgment and practical skills. In line with that, the faculty participants were concerned about students' self-reliance on AI applications on the cost of traditional teaching methods, which might deprive them from skills best learned in person or group teaching. One study listed the following as disadvantages: lack of originality, inaccurate content, or unknown data sources [14]. It is also uncertain how ChatGPT handles offensive material, false information, or plagiarism [34].

Ethical concerns, such as potential manipulation by pharmaceutical companies, were raised by participants. Maintaining transparency and integrity in AI-generated information is vital to address these concerns. Implementing measures such as third-party audits, strict guidelines, data transparency, and continuous monitoring of ChatGPT's information sources can help ensure the unmanipulated ethical use of ChatGPT in medical education [43-45].

We recommend creating guidelines for students on the appropriate use of AI applications, specifying tasks they should complete independently and the extent to which AI tools can assist. Additionally, we propose incorporating teaching sessions to help students critically evaluate AI-generated outputs. At this early stage of AI adoption [46], group teaching sessions comparing the critical appraisal of medical topics using AI tools versus traditional search methods would be beneficial. We also emphasize leveraging AI applications primarily as advanced search engines and using their summarization capabilities rather than relying entirely on their final outputs.

Participants emphasized the importance of being open-minded and adopting new technologies like AI chatbots including ChatGPT. As AI chatbots could have cultural bias, addressing cultural differences in learning styles is vital [46,47].

The potential implications of using ChatGPT in medical education include improved efficiency, streamlined information gathering, and time-saving benefits. However, future research is needed to explore the impact of AI-based tools on medical education in terms of quality, student and faculty satisfaction, and the development of critical thinking skills. Ongoing research and evaluation are essential to ensure the effective integration of AI-based tools like ChatGPT into medical education while addressing potential concerns and limitations.

In preparation for the future of medical education, educational institutions should be proactive in integrating AI technologies

like ChatGPT into their curricula and teaching methodologies [48,49]. Educators and policy makers need to remain vigilant about reliability concerns and actively take steps to be ready to address the ethical challenges and possibilities arising from the use of AI in health professions education [37,45]. This process should involve regular evaluations, ongoing improvements, and a strong emphasis on maintaining the essential human aspects of medical education, such as critical thinking, communication, and empathy.

Strengths

One of the strengths of this study is the qualitative design, which allowed for an in-depth exploration of participants' experiences, perceptions, and concerns related to the use of ChatGPT in medical education, revealing diverse viewpoints and generating valuable insights into the potential benefits and challenges of integrating ChatGPT into medical education [50]. Moreover, the study involved participants with varying levels of experience with ChatGPT, ensuring a comprehensive understanding of the perspectives of both novices and experienced users. The identification of themes and subthemes has laid a solid foundation for further research and exploration of AI-based tools like ChatGPT in medical education.

Limitations

There are some limitations to our study. The sample size was relatively small, and the participants were primarily drawn from a single institution, which may limit the generalizability of some findings to other medical education settings. The study did not quantitatively assess the impact of ChatGPT on learning outcomes, satisfaction, or other measurable aspects of medical education, which could in the future provide valuable data to supplement the qualitative findings. Additionally, since the study's focus was on understanding the perception of faculty and students, the perspectives of other stakeholders, such as administrators and policy makers, were not captured, and this could be explored in future research [51-53]. Furthermore, the study, which was conducted in the early phase of ChatGPT launching, did not explore the long-term implications and potential changes in perception and use of ChatGPT over time, as participants' experience with the tool may evolve, altering their views on its benefits and limitations [54].

Therefore, future research should incorporate larger and more diverse samples from multiple institutions as well as conduct quantitative studies to measure the impact of ChatGPT on various aspects of medical education in Saudi Arabia specifically and globally. Longitudinal studies could be conducted to assess the changes in perception and use of ChatGPT over time and evaluate the long-term effects of its integration into medical education.

Conclusions

Participants praised the advantages of ChatGPT, such as time-saving and excellent summarizing skills. However, concerns were raised regarding the accuracy and critical appraisal of information provided by ChatGPT and the need to approach the information with caution. ChatGPT-delivered information and cited references' hallucination were concerns seriously raised by participants, which needs urgent assessment and solution in addition to limited access to certain medical databases. This study highlights the need for ongoing research and evaluation to ensure that AI-based tools like ChatGPT are effectively integrated into medical education while addressing potential concerns and limitations. Educators and students must also maintain a strong foundation in critical thinking and judgment. As medical education continues to evolve, the integration of AI technologies like ChatGPT has the potential to transform the way medical education is delivered but must be done with a thoughtful and ethical approach.

Acknowledgments

We have used ChatGPT [55], an artificial intelligence chatbot developed by OpenAI, to improve some readability and language of this work, without replacing researchers' tasks. This was done with human oversight, and the authors then carefully reviewed and edited the generated text. We assure that the authors are ultimately responsible and accountable for the originality, accuracy, and integrity of their work. We would like to acknowledge the efforts in data curation in the focus groups, namely (listed with their permission): Abdulaziz Alomar, Faisal Alomri, Hadi Alhemsi, Homoud Algadhib, and Ibrahim Alhezam. The authors are grateful to Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R148), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Data Availability

The preprint of this research is available [56]. The datasets generated and analyzed during this study are not publicly available due to institutional review board privacy regulations but are available from the corresponding author on reasonable request after obtaining institutional review board approval.

Conflicts of Interest

None declared.

References

1. Fazlollahi AM, Bakhaidar M, Alsayegh A, Yilmaz R, Winkler-Schwartz A, Mirchi N, et al. Effect of artificial intelligence tutoring vs expert instruction on learning simulated surgical skills among medical students: a randomized clinical trial. JAMA Netw Open 2022;5(2):e2149008 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.49008] [Medline: 35191972]
2. Fazlollahi AM, Yilmaz R, Winkler-Schwartz A, Mirchi N, Ledwos N, Bakhaidar M, et al. AI in surgical curriculum design and unintended outcomes for technical competencies in simulation training. JAMA Netw Open 2023;6(9):e2334658 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.34658] [Medline: 37725373]
3. Zhao C, Xu T, Yao Y, Song Q, Xu B. Comparison of case-based learning using Watson for oncology and traditional method in teaching undergraduate medical students. Int J Med Inform 2023;177:105117. [doi: 10.1016/j.ijmedinf.2023.105117] [Medline: 37301132]
4. Simsek-Cetinkaya S, Cakir SK. Evaluation of the effectiveness of artificial intelligence assisted interactive screen-based simulation in breast self-examination: an innovative approach in nursing students. Nurse Educ Today 2023;127:105857. [doi: 10.1016/j.nedt.2023.105857] [Medline: 37253303]
5. Sheng B, Chen X, Li T, Ma T, Yang Y, Bi L, et al. An overview of artificial intelligence in diabetic retinopathy and other ocular diseases. Front Public Health 2022;10:971943 [FREE Full text] [doi: 10.3389/fpubh.2022.971943] [Medline: 36388304]
6. Tosaki T, Yamakawa M, Shiina T. A study on the optimal condition of ground truth area for liver tumor detection in ultrasound images using deep learning. J Med Ultrason (2001) 2023;50(2):167-176 [FREE Full text] [doi: 10.1007/s10396-023-01301-2] [Medline: 37014524]
7. Alip SL, Kim J, Rha KH, Han WK. Future platforms of robotic surgery. Urol Clin North Am 2022;49(1):23-38. [doi: 10.1016/j.ucl.2021.07.008] [Medline: 34776052]
8. Han ER, Yeo S, Kim MJ, Lee YH, Park KH, Roh H. Medical education trends for future physicians in the era of advanced technology and artificial intelligence: an integrative review. BMC Med Educ 2019;19(1):460 [FREE Full text] [doi: 10.1186/s12909-019-1891-5] [Medline: 31829208]
9. Oh N, Choi G, Lee WY. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. Ann Surg Treat Res 2023;104(5):269-273 [FREE Full text] [doi: 10.4174/astr.2023.104.5.269] [Medline: 37179699]
10. Malhotra A, Molloy EJ, Bearer CF, Mulkey SB. Emerging role of artificial intelligence, big data analysis and precision medicine in pediatrics. Pediatr Res 2023;93(2):281-283. [doi: 10.1038/s41390-022-02422-z] [Medline: 36807652]
11. Banerjee M, Chiew D, Patel KT, Johns I, Chappell D, Linton N, et al. The impact of artificial intelligence on clinical education: perceptions of postgraduate trainee doctors in London (UK) and recommendations for trainers. BMC Med Educ 2021;21(1):429 [FREE Full text] [doi: 10.1186/s12909-021-02870-x] [Medline: 34391424]
12. Lee J, Wu AS, Li D, Kulasegaram KM. Artificial intelligence in undergraduate medical education: a scoping review. Acad Med 2021;96(11S):S62-S70. [doi: 10.1097/ACM.0000000000004291] [Medline: 34348374]
13. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019;5(2):e16048 [FREE Full text] [doi: 10.2196/16048] [Medline: 31793895]
14. Temsah O, Khan SA, Chaiah Y, Senjab A, Alhasan K, Jamal A, et al. Overview of early ChatGPT's presence in medical literature: insights from a hybrid literature review by ChatGPT and human experts. Cureus 2023;15(4):e37281 [FREE Full text] [doi: 10.7759/cureus.37281] [Medline: 37038381]
15. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst 2023;47(1):33 [FREE Full text] [doi: 10.1007/s10916-023-01925-4] [Medline: 36869927]
16. Temsah MH, Aljamaan F, Malki KH, Alhasan K, Altamimi I, Aljarbou R, et al. ChatGPT and the future of digital health: a study on healthcare workers' perceptions and expectations. Healthcare (Basel) 2023;11(13):1812 [FREE Full text] [doi: 10.3390/healthcare11131812] [Medline: 37444647]
17. Albejaidi F, Nair KS. Building the health workforce: Saudi Arabia's challenges in achieving Vision 2030. Int J Health Plann Manage 2019;34(4):e1405-e1416. [doi: 10.1002/hpm.2861] [Medline: 31402508]
18. Wartman SA, Combs CD. Medical education must move from the information age to the age of artificial intelligence. Acad Med 2018;93(8):1107-1109. [doi: 10.1097/ACM.0000000000002044] [Medline: 29095704]
19. List of 38 best universities in Saudi Arabia. EduRank. 2023. URL: https://edurank.org/geo/sar/ [accessed 2025-01-16]
20. Busetto L, Wick W, Gumbinger C. How to use and assess qualitative research methods. Neurol Res Pract 2020;2:14 [FREE Full text] [doi: 10.1186/s42466-020-00059-z] [Medline: 33324920]
21. Mukred M, Mokhtar UA, Hawash B. Exploring the acceptance of ChatGPT as a learning tool among academicians: a qualitative study. J Komunikasi Malays J Commun 2023;39(4):306-323. [doi: 10.17576/jkmjc-2023-3904-16]
22. Iqbal N, Ahmed H, Azhar KA. Exploring teachers' attitudes towards using ChatGPT. GJMAS 2022;3(4):97-111. [doi: 10.46568/gjmas.v3i4.163]
23. Bryman A. Social Research Methods. Oxford, United Kingdom: Oxford University Press; 2016:0199689458.
24. Ritchie J, Spencer L. Qualitative data analysis for applied policy research. In: Analyzing Qualitative Data. London, United Kingdom: Routledge; 2002:173-194.

25. Dhakal K. NVivo. J Med Libr Assoc 2022;110(2):270-272 [FREE Full text] [doi: 10.5195/jmla.2022.1271] [Medline: 35440911]
26. Shoufan A. Exploring students' perceptions of chatGPT: thematic analysis and follow-up survey. IEEE Access 2023;11:38805-38818. [doi: 10.1109/access.2023.3268224]
27. Gülhan Güner S, Yiğit S, Berşe S, Dirgar E. Perspectives and experiences of health sciences academics regarding ChatGPT: a qualitative study. Med Teach 2024:1-10. [doi: 10.1080/0142159X.2024.2413425] [Medline: 39392461]
28. van Bulck L, Moons P. What if your patient switches from Dr. Google to Dr. ChatGPT? A vignette-based survey of the trustworthiness, value, and danger of ChatGPT-generated responses to health questions. Eur J Cardiovasc Nurs 2024;23(1):95-98. [doi: 10.1093/eurjcn/zvad038] [Medline: 37094282]
29. Howard A, Hope W, Gerada A. ChatGPT and antimicrobial advice: the end of the consulting infection doctor? Lancet Infect Dis 2023;23(4):405-406. [doi: 10.1016/S1473-3099(23)00113-5] [Medline: 36822213]
30. Kamoun F, El Ayeb W, Jabri I, Sifi S, Iqbal F. Exploring students' and faculty's knowledge, attitudes, and perceptions towards ChatGPT: a cross-sectional empirical study. J Inf Technol Educ Res 2024;23(4):1-33. [doi: 10.28945/5239]
31. Deladem Kumordzie C. All you need to know about ChatGPT and why its a threat to Google. LinkedIn. URL: https://www.linkedin.com/posts/opheliamtsenuokpor_all-you-need-to-know-about-chatgpt-why-activity-7018610251717332993-zcHV/ [accessed 2025-01-19]
32. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus 2023;15(2):e35179 [FREE Full text] [doi: 10.7759/cureus.35179] [Medline: 36811129]
33. Masters K. Medical teacher's first ChatGPT's referencing hallucinations: lessons for editors, reviewers, and teachers. Med Teach 2023;45(7):673-675. [doi: 10.1080/0142159X.2023.2208731] [Medline: 37183932]
34. Aljamaan F, Temsah MH, Altamimi I, Al-Eyadhy A, Jamal A, Alhasan K, et al. AI chatbots' medical hallucination: innovation of references hallucination score and comparison of six large language models. JMIR Preprints Preprint posted online on November 6, 2023 [FREE Full text] [doi: 10.2196/preprints.54345]
35. Sanchez-Ramos L, Lin L, Romero R. Beware of references when using ChatGPT as a source of information to write scientific articles. Am J Obstet Gynecol 2023;229(3):356-357. [doi: 10.1016/j.ajog.2023.04.004] [Medline: 37031761]
36. Zhu L, Mou W, Yang T, Chen R. ChatGPT can pass the AHA exams: open-ended questions outperform multiple-choice format. Resuscitation 2023;188:109783. [doi: 10.1016/j.resuscitation.2023.109783] [Medline: 37349064]
37. Huynh LM, Bonebrake BT, Schultis K, Quach A, Deibert CM. New artificial intelligence ChatGPT performs poorly on the 2022 self-assessment study program for urology. Urol Pract 2023;10(4):409-415. [doi: 10.1097/UPJ.0000000000000406] [Medline: 37276372]
38. Weng TL, Wang YM, Chang S, Chen TJ, Hwang SJ. ChatGPT failed Taiwan's Family Medicine Board Exam. J Chin Med Assoc 2023;86(8):762-766. [doi: 10.1097/JCMA.0000000000000946] [Medline: 37294147]
39. Meo SA, Al-Khlaiwi T, AbuKhalaf AA, Meo AS, Klonoff DC. The scientific knowledge of Bard and ChatGPT in endocrinology, diabetes, and diabetes technology: multiple-choice questions examination-based performance. J Diabetes Sci Technol 2023:19322968231203987. [doi: 10.1177/19322968231203987] [Medline: 37798960]
40. Magalhães Araujo S, Cruz-Correia R. Incorporating ChatGPT in medical informatics education: mixed methods study on student perceptions and experiential integration proposals. JMIR Med Educ 2024;10:e51151 [FREE Full text] [doi: 10.2196/51151] [Medline: 38506920]
41. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023;6:1169595 [FREE Full text] [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
42. Sedaghat S. Early applications of ChatGPT in medical practice, education and research. Clin Med (Lond) 2023;23(3):278-279 [FREE Full text] [doi: 10.7861/clinmed.2023-0078] [Medline: 37085182]
43. Mohammad B, Supti T, Alzubaidi M, Shah H, Alam T, Shah Z, et al. The pros and cons of using ChatGPT in medical education: a scoping review. Stud Health Technol Inform 2023;305:644-647. [doi: 10.3233/SHTI230580] [Medline: 37387114]
44. Aljamaan F, Temsah MH, Altamimi I, Al-Eyadhy A, Jamal A, Alhasan K, et al. Reference hallucination score for medical artificial intelligence chatbots: development and usability study. JMIR Med Inform 2024;12:e54345 [FREE Full text] [doi: 10.2196/54345] [Medline: 39083799]
45. Masters K. Ethical use of artificial intelligence in health professions education: AMEE Guide No. 158. Med Teach 2023;45(6):574-584 [FREE Full text] [doi: 10.1080/0142159X.2023.2186203] [Medline: 36912253]
46. Aljamaan F, Malki KH, Alhasan K, Jamal A, Altamimi I, Khayat A, et al. ChatGPT-3.5 System Usability Scale early assessment among healthcare workers: horizons of adoption in medical practice. Heliyon 2024 Apr 15;10(7):e28962 [FREE Full text] [doi: 10.1016/j.heliyon.2024.e28962] [Medline: 38623218]
47. Rodgers CM, Ellingson SR, Chatterjee P. Open data and transparency in artificial intelligence and machine learning: a new era of research. F1000Res 2023;12:387 [FREE Full text] [doi: 10.12688/f1000research.133019.1] [Medline: 37065505]
48. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT—reshaping medical education and clinical management. Pak J Med Sci 2023;39(2):605-607 [FREE Full text] [doi: 10.12669/pjms.39.2.7653] [Medline: 36950398]

49. Jamal A, Solaiman M, Alhasan K, Temsah M, Sayed G. Integrating ChatGPT in medical education: adapting curricula to cultivate competent physicians for the AI era. Cureus 2023;15(8):e43036 [FREE Full text] [doi: 10.7759/cureus.43036] [Medline: 37674966]
50. Gupta R, Pande P, Herzog I, Weisberger J, Chao J, Chaiyasate K, et al. Application of ChatGPT in cosmetic plastic surgery: ally or antagonist? Aesthet Surg J 2023;43(7):NP587-NP590. [doi: 10.1093/asj/sjad042] [Medline: 36840479]
51. Shahsavar Y, Choudhury A. User intentions to use ChatGPT for self-diagnosis and health-related purposes: cross-sectional survey study. JMIR Hum Factors 2023;10:e47564 [FREE Full text] [doi: 10.2196/47564] [Medline: 37195756]
52. Das D, Kumar N, Longjam LA, Sinha R, Deb Roy A, Mondal H, et al. Assessing the capability of ChatGPT in answering first- and second-order knowledge questions on microbiology as per competency-based medical education curriculum. Cureus 2023;15(3):e36034 [FREE Full text] [doi: 10.7759/cureus.36034] [Medline: 37056538]
53. Hasanein AM, Sobaih AEE. Drivers and consequences of ChatGPT use in higher education: key stakeholder perspectives. Eur J Investig Health Psychol Educ 2023;13(11):2599-2614 [FREE Full text] [doi: 10.3390/ejihpe13110181] [Medline: 37998071]
54. Temsah MH, Altamimi I, Jamal A, Alhasan K, Al-Eyadhy AA. ChatGPT surpasses 1000 publications on PubMed: envisioning the road ahead. Cureus 2023;15(9):e44769 [FREE Full text] [doi: 10.7759/cureus.44769] [Medline: 37809155]
55. ChatGPT. OpenAI. URL: https://chatgpt.com/ [accessed 2025-01-23]
56. Abouammoh N, Alhasan K, Raina R, Malki KA, Aljamaan F, Tamimi I, et al. Exploring perceptions and experiences of ChatGPT in medical education: a qualitative study among medical college faculty and students in Saudi Arabia. medRxiv Preprint posted online on July 16, 2023. [doi: 10.1101/2023.07.13.23292624]

Abbreviations

AI: artificial intelligence

Edited by B Lesselroth; submitted 18.06.24; peer-reviewed by F Kamoun, SM Araujo, KA Morbitzer; comments to author 02.09.24; revised version received 03.11.24; accepted 02.01.25; published 20.02.25.

Please cite as:
Abouammoh N, Alhasan K, Aljamaan F, Raina R, Malki KH, Altamimi I, Muaygil R, Wahabi H, Jamal A, Alhaboob A, Assiri RA, Al-Tawfiq JA, Al-Eyadhy A, Soliman M, Temsah MH
Perceptions and Earliest Experiences of Medical Students and Faculty With ChatGPT in Medical Education: Qualitative Study
JMIR Med Educ 2025;11:e63400
URL: https://mededu.jmir.org/2025/1/e63400
doi:10.2196/63400
PMID:39977012

Original Paper

Detecting Artificial Intelligence–Generated Versus Human-Written Medical Student Essays: Semirandomized Controlled Study

Berin Doru[1]; Christoph Maier[1]; Johanna Sophie Busse[1]; Thomas Lücke[1]; Judith Schönhoff[2]; Elena Enax- Krumova[3]; Steffen Hessler[4]; Maria Berger[5*]; Marianne Tokic[6*]

[1]University Hospital of Paediatrics and Adolescent Medicine, St. Josef-Hospital, Ruhr University Bochum, Bochum, Germany

[2]Departement of German Philology, General and Comparative Literary Studies, Ruhr University Bochum, Bochum, Germany

[3]Department of Neurology, BG University Hospital Bergmannsheil gGmbH Bochum, Ruhr University Bochum, Bochum, Germany

[4]German Department, German Linguistics, Ruhr University Bochum, Bochum, Germany

[5]German Department, Digital Forensic Linguistics, Ruhr University Bochum, Bochum, Germany

[6]Department for Medical Informatics, Biometry and Epidemiology, Ruhr University Bochum, Bochum, Germany

[*]these authors contributed equally

Corresponding Author:
Berin Doru
University Hospital of Paediatrics and Adolescent Medicine, St. Josef-Hospital
Ruhr University Bochum
Alexandrinenstraße 5
Bochum, 44791
Germany
Phone: 49 234 509 2611
Email: Berin.Doru@rub.de

Abstract

Background: Large language models, exemplified by ChatGPT, have reached a level of sophistication that makes distinguishing between human- and artificial intelligence (AI)–generated texts increasingly challenging. This has raised concerns in academia, particularly in medicine, where the accuracy and authenticity of written work are paramount.

Objective: This semirandomized controlled study aims to examine the ability of 2 blinded expert groups with different levels of content familiarity—medical professionals and humanities scholars with expertise in textual analysis—to distinguish between longer scientific texts in German written by medical students and those generated by ChatGPT. Additionally, the study sought to analyze the reasoning behind their identification choices, particularly the role of content familiarity and linguistic features.

Methods: Between May and August 2023, a total of 35 experts (medical: n=22; humanities: n=13) were each presented with 2 pairs of texts on different medical topics. Each pair had similar content and structure: 1 text was written by a medical student, and the other was generated by ChatGPT (version 3.5, March 2023). Experts were asked to identify the AI-generated text and justify their choice. These justifications were analyzed through a multistage, interdisciplinary qualitative analysis to identify relevant textual features. Before unblinding, experts rated each text on 6 characteristics: linguistic fluency and spelling/grammatical accuracy, scientific quality, logical coherence, expression of knowledge limitations, formulation of future research questions, and citation quality. Univariate tests and multivariate logistic regression analyses were used to examine associations between participants' characteristics, their stated reasons for author identification, and the likelihood of correctly determining a text's authorship.

Results: Overall, in 48 out of 69 (70%) decision rounds, participants accurately identified the AI-generated texts, with minimal difference between groups (medical: 31/43, 72%; humanities: 17/26, 65%; odds ratio [OR] 1.37, 95% CI 0.5-3.9). While content errors had little impact on identification accuracy, stylistic features—particularly redundancy (OR 6.90, 95% CI 1.01-47.1), repetition (OR 8.05, 95% CI 1.25-51.7), and thread/coherence (OR 6.62, 95% CI 1.25-35.2)—played a crucial role in participants' decisions to identify a text as AI-generated.

Conclusions: The findings suggest that both medical and humanities experts were able to identify ChatGPT-generated texts in medical contexts, with their decisions largely based on linguistic attributes. The accuracy of identification appears to be independent of experts' familiarity with the text content. As the decision-making process primarily relies on linguistic attributes—such as stylistic features and text coherence—further quasi-experimental studies using texts from other academic disciplines should be

conducted to determine whether instructions based on these features can enhance lecturers' ability to distinguish between student-authored and AI-generated work.

(JMIR Med Educ 2025;11:e62779) doi:10.2196/62779

KEYWORDS

artificial intelligence; ChatGPT; large language models; textual analysis; writing style; AI; chatbot; LLMs; detection; authorship; medical student; textual analysis; linguistic quality; decision-making; logical coherence

Introduction

The rapid development of artificial intelligence (AI) and the emergence of large language models (LLMs), such as ChatGPT, have increasingly blurred the lines between human-written and AI-generated text. This has created a significant challenge in identifying the authorship of written work, especially as the use of AI has become ubiquitous since chatbots have become freely available [1,2]. Consequently, critical concerns have arisen in the educational and academic sectors, where the reliability and authenticity of written work are fundamental.

According to a recent nationwide survey, nearly two-thirds of the German students reported using AI-based tools for their studies, with ChatGPT being the most commonly used chatbot [3]. ChatGPT, developed by OpenAI, an American AI research laboratory, is a state-of-the-art AI chatbot capable of assisting users with a wide range of tasks, from text generation to problem-solving [4,5]. Its capabilities have opened up significant opportunities for educational and academic contexts. For example, AI-based tools like ChatGPT can support tasks such as text analysis, translation, and proofreading for research purposes [6]. They also can provide support to students by enhancing their understanding of scientific methods, improving and refining written work, and assisting with examination preparation [3,7].

However, the widespread use of such tools raises concerns about their impact on students' development of critical and independent thinking skills [8,9]. In addition, it is possible that ChatGPT could provide incomplete or inaccurate information, potentially leading to misunderstandings of academic concepts and topics [10,11]. Further concerns arise from the potential for academic dishonesty and plagiarism, particularly in the context of written assignments and academic essays [9,12].

The lack of clarity on how to handle such cases is putting universities in a quandary, leading to the first court cases and, more recently, to the University of Munich being vindicated in its decision to reject such written work [13]. In this case, an essay submitted as part of a Master's application was rejected because it was "too well written," raising suspicions that the text was likely generated by an AI tool such as ChatGPT [13]. This case highlights that AI-generated texts are often characterized by their seemingly perfected style of formulation that refers to the linguistic level [13,14], a characteristic that is known to be particularly pronounced and even more nuanced in English output texts compared with other languages such as German [15-18].

The relevance of the problem for medical studies is not obvious at first glance because medical students generally do not have to write long scientific texts on a regular basis during their studies, but usually only for their doctoral thesis. However, the increasing use of AI tools such as ChatGPT also poses challenges in the medical field, where assessments rely not only on linguistic quality but also on content accuracy. The potential misattribution of authorship in medical texts—such as research articles, patient information, or promotional materials—has particularly serious implications, as errors or inaccuracies in these contexts can have grave consequences. It can be assumed that medical texts do not fall as much within the scope of ChatGPT and are therefore more difficult to reproduce accurately, especially because AI authors have no "moral scruples" about concealing ignorance and replacing verified sources with falsified ones [19-22]. However, a few studies have addressed the problem of AI-generated content on medical texts [21-26], reporting that ChatGPT has, at times, managed to mislead medical professionals [26], which may suggest that familiarity with content plays a minor role in authorship identification.

Existing research on LLM-based text generators such as ChatGPT frequently focuses on their role in assisting the writing process rather than evaluating the quality and detectability of longer scientific texts [27-29]. In addition, studies often investigate the detection of texts written by chatbots using automatic tools or even detectors specifically designed for this purpose [18,25,27-30]. The detection rate of these detectors is often higher than that of human reviewers, but the accuracy can vary greatly depending on the text genre and the classifier used [14,31]. Moreover, linguistic features appear to be the most important subset of features influencing the performance of feature-based classifiers [2,5]. In the academic domain, educators still face the challenge of qualitatively assessing the authenticity of student texts, often without the aid of automated detection tools.

It is therefore of particular interest to determine how well human readers from the academic field can detect differences when directly comparing 2 texts on the same topic—an original student text and an AI-generated text—and which features stand out as particularly conspicuous and decisive for them. To better distinguish between the relevance of content-related and text-analytical attributes, we assembled a group of language experts from the humanities field alongside a group of medical experts specializing in pediatrics and neurology. Another key novelty and prerequisite of our study is the use of fully reproduced, longer scientific texts on medical topics written in German by medical students.

Therefore, we conducted a study to determine whether medical experts and humanities lecturers could distinguish between texts written by medical students and those generated by ChatGPT

(specifically ChatGPT version 3.5, March 23). Unlike the interactive "Turing Test" [32], this task was not performed through a dialogue with a machine but rather through an internal, personal evaluation of 2 texts. We hypothesize that, in line with the Turing prophecy [32], the correct identification rate for AI-generated texts within a German medical sample is approximately 70%, with content familiarity playing a secondary role, while formal and linguistic features exert a greater influence.

Through a prospective analysis of longer German-language scientific texts written by students in the specialized health-related field of medicine, this study aims to provide new insights into AI-generated texts and the influence of content familiarity and linguistic expertise. The findings are intended to inform the development of guidelines to help lecturers (and others) recognize AI-generated texts, even in the absence of a comparable "original" text, and to contribute to future projects addressing the challenges posed by AI tools in academia.

Methods

Recruitment Process

This semirandomized controlled trial was conducted between May and August 2023 at the University Hospital of Pediatrics and Adolescent Medicine, St. Josef-Hospital, Ruhr University Bochum (RUB), Germany. To recruit participants, an open call was issued to the Department of Pediatrics and Adolescent Medicine and the Department of Neurology at the University Hospital Bergmannsheil (both RUB). Senior physicians and members of scientific working groups were invited to participate, ensuring the involvement of clinical experts familiar with the content of the texts. Participation was voluntary, with clinical employment or medical expertise serving as key inclusion criteria, along with an interest in scientific texts. In the next phase of recruitment, a call was made to the Faculty of Humanities at Ruhr University to include participants with experience in text reception and analysis, with teaching experience as an additional inclusion criterion.

Design

Each participant received 2 pairs of texts, totaling 4 printed texts. Each pair consisted of 1 of 18 available term papers written by a medical student and a corresponding text generated by ChatGPT (version 3.5, March 2023), with the order of presentation randomized. The medical experts received 1 pair of texts on a topic closely related to their specialty and another pair on a less familiar topic. For example, a pediatrician received the text "Autoantibodies in Diabetes Mellitus," while a neurologist received "Measurement of Aδ and C Fibers in Electrophysiology." The second pair of texts covered a topic less directly related to their field of expertise.

Given the exploratory nature of the study, it became increasingly evident—only after completing the experimental phase with the medical experts—that the extent to which content familiarity influences the identification process needed further examination, particularly in comparison to formal and linguistic aspects.

As a result, a second phase of the study was initiated, involving a new group of experts with greater expertise in formal and linguistic analyses. This allowed for a comparison of results and evaluations across groups. In this group of humanities experts, each participant received the same 2 pairs of texts to ensure better comparability and verification of subject unfamiliarity. The first pair addressed a more general topic also familiar to nonmedical fields: "Iodine Deficiency." The second pair analyzed a more specialized medical topic: "Autoantibodies in Diabetes Mellitus."

Participants were asked to read a pair of texts and, based on their personal experience with student-written texts, decide within a week—without extensive research—which of the 2 they believed was generated by ChatGPT. To ensure the blinding of both interviewers and participants, the headers of the texts contained only a randomly generated 3-digit number and a "chatbot or student" checkbox. Before the subsequent interview, participants documented their decision by ticking the corresponding box for their chosen text version. They were also instructed not to discuss the task or the texts with one another.

About 1-2 weeks after the texts were distributed, participants were invited to a semistructured interview—conducted in person, by telephone, or via Zoom (Zoom Communications/Qumu Corporation)—to discuss their decisions, reasoning, and evaluations of the texts. Unblinding occurred after the interviews.

Creation of the ChatGPT-Generated Versions

Eighteen German-language medical essays served as templates for the ChatGPT-generated texts. These essays were written by doctoral students from the University Pediatric Clinic and the Clinic for Neurology at Bergmannsheil University Hospital, RUB, Germany. They originated from the Doctoral Colloquium pool at the University Hospital of Pediatrics and Adolescent Medicine in Bochum. As part of the colloquium, each doctoral student is encouraged to write a scientific essay thematically related to their announced dissertation topic. To provide an initial experience with scientific research and writing—and to give the reviewing study coordinator a first impression of their academic level and skills—students do not receive specific instructions. For this study, all available German texts from this pool that were written before the general introduction of ChatGPT were considered, provided their authors consented to their use.

We used ChatGPT version 3.5, March 14 to replicate the texts. To generate a version with the same title and outline as the original papers while avoiding text breaks, 2 separate prompts were required to produce a continuous text from the introduction to the conclusion (Table 1).

For the main part, depending on the type of original paper, several commands were necessary. For example, see Table 2.

We then merged the individual sections to create a complete term paper, supplementing it with a bibliography that listed the sources provided by ChatGPT in sequential order. To ensure consistency, we harmonized the formatting of both ChatGPT-generated and student-written texts as much as possible, using the Arial font (size 11 for body text and size 12 for headings) with justified alignment. Sections or sentences specific to a student's individual dissertation project were

removed to maintain general applicability. However, we did not alter the choice of words, sentence structure, punctuation, spelling, or citation style.

Table 1. Prompts used to create ChatGPT text.

German (original prompt)	English (translation)
• "Schreibe bitte einen Abschnitt über das Thema [TITEL DES TEILTHEMAS] der [wissenschaftlichen/medizinischen] Hausarbeit [TITEL DER HAUSARBEIT]." • "Belege Deine Aussagen mit Quellen, die bei Pubmed auffindbar sind."	• "Please write a section on the topic [NAME OF SUBTOPIC] of the [scientific/medical] term paper [NAME OF TERM PAPER]." • "Support your statements with sources that can be found on PubMed."

Table 2. Additional instructions used to create ChatGPT text.

German (original prompt)	English (translation)
"Schreibe einen Abschnitt zum Thema 'Potenziell reversible Pathomechanismen als mögliche Ursachen von Hyposmie oder Anosmie bei Kindern' der wissenschaftlichen Hausarbeit 'Ursachen und Diagnostik von Riechstörungen bei Kindern und Jugendlichen', der an den vorherigen Abschnitt anknüpft."	"Write a section on the topic 'Potentially reversible pathomechanism as possible causes of hyposmia or anosmia in children' of the scientific paper 'Causes and diagnostics of olfactory dysfunction in children and adolescents', which ties in with the previous section."

Data Assessment During the Interview

The medical expert group was interviewed by 2 blinded interviewers (BD and JSB), while the humanities expert group was interviewed by a partially blinded interviewer (CM). Initially, participants provided demographic information, including age, experience in academic and student teaching, academic qualifications, publication history, and prior experience with ChatGPT. They were then asked to assess how well the following questions were addressed in each text, using the German grading system from 1 (very good) to 6 (unsatisfactory) (see Multimedia Appendix 1 for details). How would you rate (1) linguistic fluency, (2) scientific quality (eg, are the re-definitions scientifically derived and are studies cited that lead to certain conclusions?), (3) internal logic, (4) description of the limitations of current knowledge, (5) future research questions, and (6) citations and references of the text? Participants were then asked to identify which text version, using the corresponding 3-digit number, they had categorized as being generated by ChatGPT and to list the key reasons for their decision, which the interviewer recorded using keywords. Next, they rated their confidence in their decision on a scale from 1 (very confident) to 6 (not confident at all). After this initial assessment, participants were unblinded and informed which text had been written by a student and which by ChatGPT. In cases of misidentification, they were asked about their suspected reasons, which the interviewer also documented using keywords.

Construction of Categories

Beyond the identification rate and the text evaluations by each group, a qualitative analysis of participants' statements regarding their reasoning for assigning authorship proved essential. This deeper analysis aimed to examine the influence of content-related versus formal-linguistic aspects and to better attribute global features to either student or chatbot authorship. For this purpose, the free-text responses (recorded by interviewers using keywords) were first thematically clustered based on the terms mentioned (see sample statements in the *Free-Text Analysis* section). Subsequently, through multiturn discussions between medical and linguistic experts, these thematic clusters were refined into distinct, nonoverlapping categories that encompassed all interviewee statements while reducing redundancy and multiple classifications.

Many of these categories align with standard text-analytical frameworks, which typically cover a broad range of attributes, including morphology, syntax, style, structure, coherence and cohesion, content quality, form, and even sociolinguistic aspects [33-35]. However, the categories derived in this study are directly based on the text types used in the experiments. As a result, they provide a more precise representation of the emerging and still undefined text type "AI-generated" and are therefore preferable (see Multimedia Appendix 2 for an overview of the categories).

Statistical Analysis

Data were analyzed using Microsoft Excel, SPSS version 29.3 (IBM Corp.), and R-4.1.2 (R Foundation). Descriptive statistics are presented as numbers (n) and percentages or as means (SD), where appropriate. Univariate odds ratios and 95% CIs from the Fisher exact test were used to examine the association between demographic markers, participants' field (medicine or humanities), and their expertise with the likelihood of correctly identifying a text's source. The relationship between interview scores and response accuracy was assessed using the 2-sided Wilcoxon signed rank test for paired values. Additionally, correlations among all 5 responses were tested using a Friedman 2-way analysis of variance for ranks with Bonferroni correction.

To analyze how participants attempted to identify the machine-generated text, we modeled the association of the derived categories (items) from the interviews based on their likelihood of being mentioned in the context of a chatbot-generated text. For each participant and interview, we recorded whether an item was cited in reference to a perceived chatbot text or a perceived student text. This association was analyzed using repeated-measures logistic regression, incorporating a random participant and sequence effect. The

model was further adjusted for age group, the expert group (medical vs humanities), and prior experience with ChatGPT (binary).

Ethical Considerations

An application for the study project was submitted to the Ethics Committee of the Medical Faculty at RUB (reference number 23-7837; April 2023). As the study did not involve direct research on human participants or patient data, the committee informed us that ethical approval was not required.

Results

Interviewee Sample

The biographical data of the 22 participating physicians (14 pediatricians, 3 nutritionists, 4 neurologists, and 1 neuroscientist) and 13 humanities scholars (8 literary scholars, 3 Germanists or linguists, 1 classical philologist, and 1 Romance philologist) are presented in Table 3.

As there were more participating experts than available term papers, 3 pairs of texts were each assessed by 3 or 4 medical experts.

At the time of the survey, only one-fifth of the participants reported having prior experience with ChatGPT. As the number of participating experts exceeded the number of available term papers, 3 pairs of texts were each assessed by 3 or 4 medical experts.

Table 3. Interviewee sample.

Characteristics	All participants (N=35)	Medical experts (n=22)	Humanities experts (n=13)
Age (years), n (%)			
<40	17 (49)	13 (59)	4 (31)
≥40	18 (51)	9 (41)	9 (69)
Experience in academic teaching[a] (years), n (%)			
None	2 (6)	2 (9)	N/A[b]
<5	8 (23)	8 (36)	N/A
≥5	25 (71)	12 (55)	13 (100)
PhD/professorship, n (%)			
Yes	28 (80)	18 (82)	10 (77)
Authorship in a publication, n (%)			
Yes	32 (91)	20 (91)	12 (92)
Experience with ChatGPT[a], n (%)			
Yes	7 (20)	2 (9)	5 (38)
Only a little	7 (20)	4 (18)	3 (23)
No	21 (60)	16 (73)	5 (38)

[a]Self-assessed.

[b]N/A: not applicable.

Detection Rate

With 35 participants evaluating 2 text pairs each—excluding 1 misaligned and omitted case—a total of 69 decision rounds were conducted. In 48 out of 69 (70%) decision rounds, participants correctly identified the authorship of the texts. Medical and humanities experts showed a slight but nonsignificant difference in detection rates, with medical experts correctly identifying 31 out of 43 (72%) decision rounds compared with 17 out of 26 (65%) decision rounds by humanities experts (odds ratio 1.37, 95% CI 0.5-3.9). Among the 35 participants, 21 (60%) misidentified the authorship of at least one text pair, including 12 medical experts. Additionally, 5 (14%) participants, including 3 physicians, misidentified both text pairs.

Notably, familiarity with the topic did not significantly impact identification accuracy (Table 4), nor did personal characteristics such as age, academic qualifications, or years of teaching experience. However, younger participants without advanced academic titles showed a slight tendency to better identify ChatGPT-generated texts. Confidence in participants' own judgments did not differ significantly across groups (odds ratio 0.6, 95% CI 0.2-1.79).

Table 4. Characteristics of the participants with correct and incorrect decisions about the authorship of the respective text.

Characteristics	All participants			Medical experts			Humanities experts		
	Decision			Decision			Decision		
	Correct	False	OR[a] (95% CI)	Correct	False	OR (95% CI)	Correct	False	OR (95% CI)
Tests, n (%)	48 (70)	21 (30)	N/A[b]	31 (72)	12 (28)	N/A	17 (65)	9 (35)	
Age (years), n (%)									
<40	24 (73)	9 (27)	N/A	17 (68)	8 (32)	N/A	7 (88)	1 (12)	N/A
≥40	24 (67)	12 (33)	0.75 (0.27-2.11)	14 (78)	4 (22)	1.65 (0.41-6.63)	10 (56)	8 (44)	0.18 (0.02-1.77)
PhD/professorship, n (%)									
Yes	37 (67)	18 (33)	N/A	25 (71)	10 (29)	N/A	12 (60)	8 (40)	N/A
No	11 (78.6)	3 (21.4)	1.78 (0.44-7.2)	6 (75)	2 (25)	1.2 (0.21-6.98)	5 (83.3)	1 (16.7)	3.33 (0.33-34.12)
Experience in academic teaching[c] (years), n (%)									
≥5	30 (67)	15 (33)	N/A	16 (70)	7 (30)	N/A	14 (64)	8 (36)	N/A
<5 or none	18 (75)	6 (25)	1.5 (0.49-4.56)	15 (75)	5 (25)	1.31 (0.34-5.05)	3 (75)	1 (25)	1.71 (0.15-19.36)
Authorship in a publication, n (%)									
Yes	44 (70)	19 (30)	N/A	28 (72)	11 (28)	N/A	16 (67)	8 (33)	N/A
No	4 (67)	2 (33)	0.86 (0.15-5.12)	3 (75)	1 (25)	1.18 (0.11-12.59)	1 (50)	1 (50)	0.5 (0.03-9.08)
Experience with ChatGPT[c], n (%)									
Yes	20 (74)	7 (26)	N/A	9 (82)	2 (18)	N/A	11 (69)	5 (31)	N/A
No	28 (67)	14 (33)	0.7 (0.24-2.05)	22 (69)	10 (31)	0.49 (0.09-2.69)	6 (60)	4 (40)	0.68 (0.13-3.55)
Text pair (sequence), n (%)									
First	23 (66)	12 (34)	N/A	16 (73)	6 (27)	N/A	7 (54)	6 (46)	N/A
Second	25 (74)	9 (26)	0.75 (0.27-2.09)	15 (71)	6 (29)	0.94 (0.25-3.56)	10 (77)	3 (23)	2.86 (0.53-15.47)
Familiar with the topic, n (%)									
More	25 (68)	12 (32)	N/A	18 (75)	6 (25)	N/A	7 (54)	6 (46)	N/A
Less	23 (72)	9 (28)	1.01 (0.36-2.8)	13 (68)	6 (32)	0.72 (0.19-2.75)	10 (77)	3 (23)	2.86 (0.53-15.47)
Self-confidence in the decision[c], n (%)									
Rather sure	35 (73)	13 (27)	N/A	22 (76)	7 (24)	N/A	13 (68)	6 (32)	N/A
Unsure	13 (62)	8 (38)	0.6 (0.2-1.79)	9 (64)	5 (36)	0.57 (0.14-2.29)	4 (57)	3 (43)	0.62 (0.1-3.66)

[a]OR: odds ratio for the correct decision.

[b]N/A: not applicable.

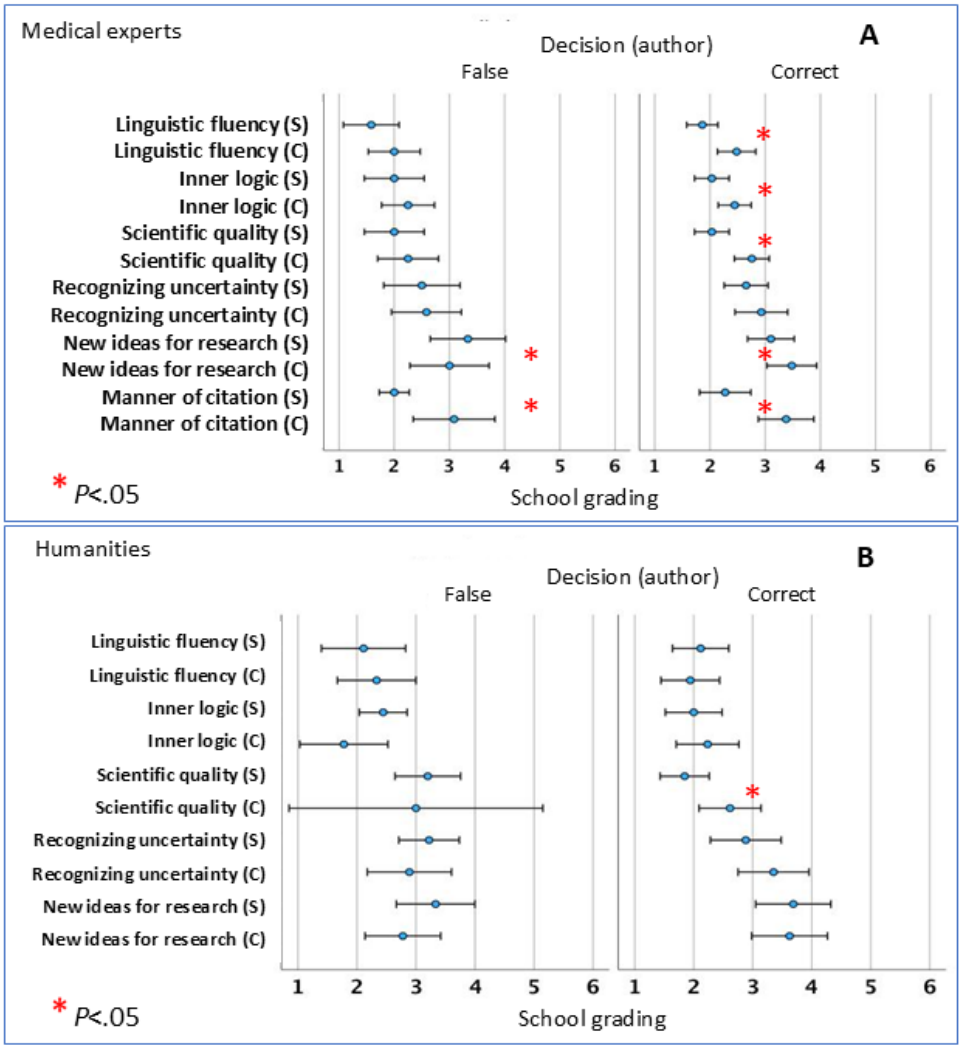
[c]Self-assessed.

Interview Analysis

When authorship was correctly identified, texts written by medical students received notably higher ratings for stylistic fluency, the internal logic of argumentation, and scientific quality (see Figure 1A, right-hand side). These differences were less pronounced in the assessment of knowledge limitations and future research directions. Regardless of correct identification, the way sources were cited was consistently rated higher in student-authored texts (Figure 1). Among humanities experts, score differences between correctly and incorrectly classified texts were minimal, though the academic quality of student-written texts was still rated significantly (r=0.336, P=.009) higher overall (Figure 1B, right-hand side). Many participants who misidentified the authorship attributed their errors to either underestimating the quality of student work or overestimating ChatGPT's capabilities—particularly in terms

of logical coherence, the presentation of scientific knowledge limitations, and the formulation of new research ideas.

Figure 1. Association of mean school grade (using German school grades 1=very good to 6=unsatisfactory) and correctness of authoring identification in (A) medical experts and (B) humanities experts. Participants were not yet unblinded at the time of assessment (see Multimedia Appendix 1 for details). Left side: incorrect attribution, right side: correct attribution of authorship. *P<.05 (2-sided Wilcoxon signed rank test). C: chatbot-generated text; S: student text.



Free-Text Analysis

We categorized the 187 freely formulated reasons participants provided for their decisions into 1 of 12 derived categories (see Multimedia Appendix 2). Three categories were excluded from statistical analysis due to their low frequency: inconsistency of writing style (n=4), other issues (n=4), and errors in content (n=3). Notably, all 3 content-error attributions came from the medical expert group. Of the remaining 176 statements, 88 (50%) were contributed by medical experts and 88 (50%) by humanities scholars (Table 5).

The experiment revealed that significantly more statements were made about (suspected) ChatGPT-generated texts (130/176, 73.9%) than about student-written texts, regardless of whether the suspicion was correct (Table 5). Sample statements from both groups are provided in Tables 6 and 7. Medical experts' explanations were often concise, frequently critiquing a "superficial" style with "unnecessary additional information." By contrast, humanities experts tended to provide more detailed justifications, describing characteristics such as "smooth style" and "redundancies."

We analyzed the likelihood of specific categories being mentioned in reference to texts suspected to be generated by ChatGPT. The results indicate that "redundancy" (12/14, 86%, associated with GPT vs 2/14, 14%, with student texts), "repetition" (20/22, 91% vs 2/22, 9%), and "common thread and coherence" (21/24, 88% vs 3/24, 13%) were the most frequently cited characteristics (Figure 2).

Table 5. The remaining 9 categories and item frequency by presumed nature of the text (for a detailed explanation in German and English, see Multimedia Appendix 2).

Item	Category	How often mentioned				
		Overall, n (%)	Humanities		Medical experts	
			Chatbot, n (%)	Students, n (%)	Chatbot, n (%)	Students, n (%)
		176 (100)	59 (33.5)	29 (16.5)	71 (40.3)	17 (9.7)
1	Differentiated content	16 (9.1)	4 (6.8)	4 (13.8)	4 (5.6)	4 (23.5)
2	Superficial content	13 (7.4)	2 (3.4)	2 (6.9)	8 (11.3)	1 (5.9)
3	Redundancy	14 (8.0)	7 (11.9)	2 (6.9)	5 (7.0)	N/A[a]
4	Monotonous structure of sentences	14 (8.0)	6 (10.2)	4 (13.8)	4 (5.6)	N/A
5	Repetition	22 (12.5)	8 (13.6)	1 (3.4)	12 (16.9)	1 (5.9)
6	Common thread coherency	24 (13.6)	10 (16.9)	2 (6.9)	11 (15.5)	1 (5.9)
7	Distinctive literature style	16 (9.1)	3 (5.1)	2 (6.9)	8 (11.3)	3 (17.6)
8	Form	25 (14.2)	6 (10.2)	6 (20.7)	9 (12.7)	4 (23.5)
9	Distinctive Wording	32 (18.2)	13 (22.0)	6 (20.7)	10 (14.1)	3 (17.6)

[a]N/A: not applicable.

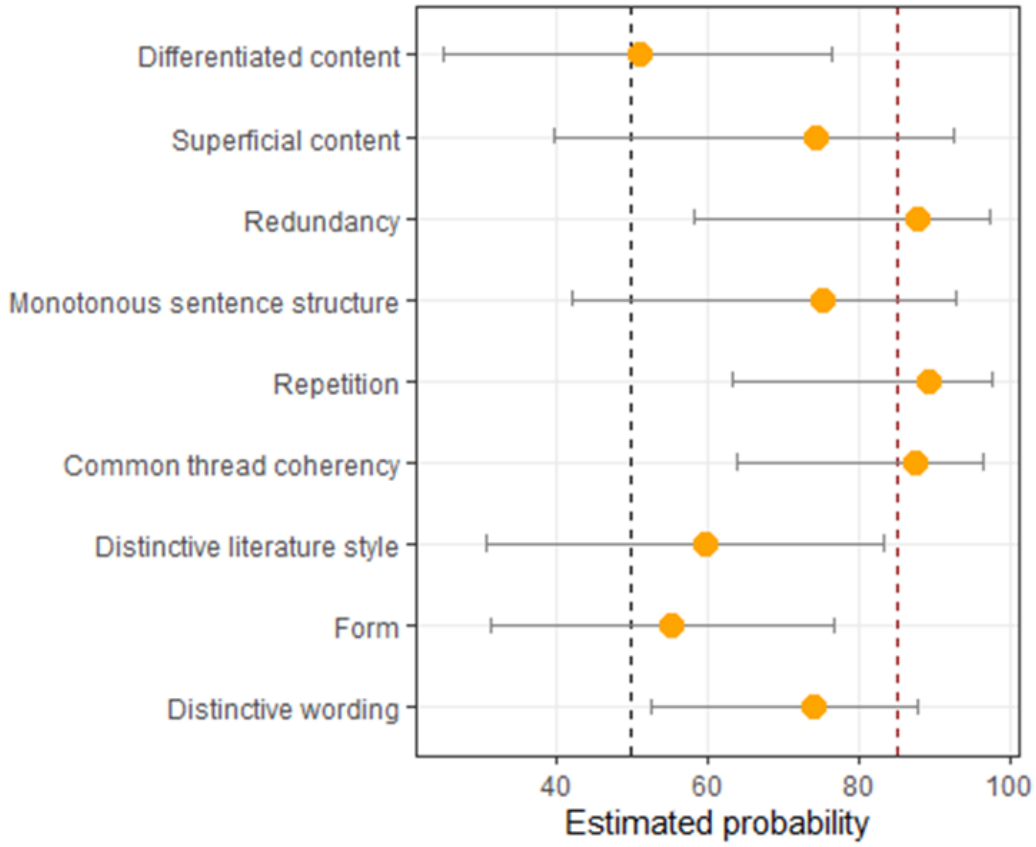
Table 6. Excerpt from the statements of the medical expert group on the main reasons for choosing ChatGPT as the author.

German (original statement)	English (translation)
“‚erschien zu perfekt geschrieben, oberflächlich‘”	“‚seemed too perfectly written, superficial’”
“‚unnützes Wissen, Zusatzinfos, die nicht notwendig für die Arbeit wären‘”	“‚useless knowledge, additional information that is not necessary for the work‘”
“‚fehlender roter Faden, fehlende Kontinuität der Logik’”	“‚lack of common thread, lack of continuity of logic’”

Table 7. Excerpt from the statements of the humanities expert group on the main reasons for choosing ChatGPT as the author.

German (original statement)	English (translation)
“‚fehlende Kohärenz, Redundanz, Monotonie […] ist sehr redundant, wiederholt Formulierungen teils mehrfach in Variationen. Der Definitionsteil wirkt reihend, stilistisch homogen, […] teils erfährt man, was man sich hätte denken können, […]. Der letzte Absatz wiederholt in etwa, was vorher da stand - was so wirkt, als hätte er diesen schon 'vergessen'.”	“‚lack of coherence, redundancy, monotony [...] is very redundant, repeats formulations sometimes several times in variations. The definition section appears to be sequential, stylistically homogeneous, [...] partly one learns what one could have imagined, [...]. The last paragraph roughly repeats what was there before - which makes it seem as if he had already 'forgotten' it.”
“‚Smooth, fließende Übergänge, aber übertextlich schlechter, d.h. gesamt Kohärenz schlechter (Wiederholung), habe nichts gelernt‘.”	“‚Smooth, fluent transitions, but overtextually worse, i.e. overall coherence worse (repetition), 'haven't learned anything'.”
“‚Wiederholung vieler Sätze und Inhalte; ausgeprägte Tendenz zu bestimmten schablonenartigen Formulierungen im Sinne einer ‘Anmoderation’, z.T. phrasenhaft ohne Inhalt. Optisch gute Gliederung (vorgegeben durch die Überschriften), jedoch roter Faden nicht gut erkennbar, vieles wirkt lediglich wie aufgelistete Einzelformationen. […] Nennung vieler Quellen, deren Zuordnung zu einzelnen Aussagen ist oft nicht konkret.”	“‚Repetition of many sentences and contents; pronounced tendency towards certain template-like formulations in the sense of a ‘presentation’, sometimes phrase-like without content. Visually well structured (given by the headings), but a common thread is not easily recognizable, many things appear to be merely a list of individual formations. [...] Mention of many sources, their assignment to individual statements is often not concrete.”

Figure 2. Estimated probability of thinking of ChatGPT as authorship.



Discussion

Overview

This analysis offers insights into the current identification rate of AI-generated texts and their evaluation compared with medical student texts by 2 different expert groups. It also provides an initial overview of the decision-making processes of medical and humanities experts during these assessments. Our findings suggest that both medical and humanities experts can effectively identify ChatGPT-generated texts in medical contexts and that linguistic and stylistic features play a significant role in distinguishing AI-generated from human-written texts, regardless of content familiarity. This supports the broader notion that linguistic analysis is crucial in identifying AI-generated text, aligning with foundational theories in human-robot interaction, such as Turing's predictions [32].

Identification Rate

In the 1950s, Alan Turing [32] predicted that within 50 years, AI would advance to the point where the likelihood of identifying a machine as nonhuman in a dialogue or an "imitation game" would be no more than 70% [32]. With a slight delay of about 20 years, his prediction was almost precisely fulfilled in an online game inspired by the Turing Test [5]. Unlike Turing's method and the large-scale Israeli study, our research did not involve direct dialogue between humans and machines [5,32]. However, when participants were presented with 2 texts of different authorship, an internal dialogue was essential for making an authorship determination. Ultimately, our study's main finding aligns almost exactly with Turing's prediction: only in 48 out of 69 (70%) decision rounds, participants correctly identified the ChatGPT-generated text. This accuracy rate remained consistent regardless of whether participants were experts in the content of the text or in linguistic analysis, and irrespective of their prior experience with ChatGPT. Notably, familiarity with the subject matter did not appear to be a decisive factor, as humanities experts performed similarly to medical experts who specialized in the respective topics. Moreover, at the individual participant level, no significant differences were found between the 2 expert groups in terms of their proximity to the text's subject matter.

A Chinese study by Ma et al [2], which also examined the identification rate of chatbot-generated texts, reported similar findings, with approximately 66% of texts correctly identified. This study analyzed around 40 scientific texts, including 20 scientific paper abstracts and 20 wiki item descriptions, assessed by 2 PhD students with a background in computer science [2]. Ma et al [2] also highlighted notable differences in writing style between AI-generated and human-written scientific texts, a conclusion that aligns with our findings. In our study, participants primarily based their decisions on text-analytical features, while content errors influenced their judgment in only 3 instances.

The study by Waltzer et al [36], which closely resembles our research in design, reported similar findings. In their study, 140 college instructors were presented with pairs of essays and correctly identified the ChatGPT-generated text 70% of the time. Like our results, Waltzer et al [36] found that neither prior

experience with ChatGPT nor subject-specific expertise—measured by self-reported familiarity with the topic—significantly improved identification accuracy [36]. However, a key difference is that their study analyzed English-language essays written for a psychology program, whereas our research focused on German-language texts authored by medical students [36].

Performance

The evaluation of a text can focus on different levels and aspects, often emphasizing either content or linguistic features. Currently, AI programs such as ChatGPT are recognized for their seemingly perfected linguistic style [13,14]. A notable case at a German university (TU Munich) illustrates this: an essay submitted as part of a Master's application was rejected—and this decision was upheld by a court—on the grounds that it was "too well written," strongly suggesting AI authorship [13]. However, it is important to note that this essay was written in English [13]. While ChatGPT is also proficient in translating languages such as German and Chinese [15], its performance in German differs from English. Research suggests that AI-generated texts tend to be more nuanced and varied in English than in German [16,17]. This discrepancy is likely due to the greater availability of digital data in English, which results in more refined and contextually accurate outputs. Nevertheless, AI language models continuously improve as they interact with users, enhancing their capabilities in non–English languages over time.

Interestingly, the humanities group, despite their focus on linguistic features, identified ChatGPT-generated texts less accurately than the medical expert group—though this difference was not statistically significant. Notably, humanities experts rated the linguistic quality of ChatGPT texts higher than those written by medical students, a contrast that was significant compared with the evaluations of the medical experts. The decision-making process behind text identification revealed key patterns: participants were more likely to suspect a human author when encountering spelling and grammatical errors, greater variation in sentence structure, medical-specific terminology, a writing style aimed at a professional readership, or shifts in citation style.

An "AI author," by contrast, was suspected if there was a monotonous sentence structure, partly "English" grammar, a "smooth" wording style, that is, good readability/understandability, but overall more superficial, an intended less professional readership, better overall formal structure of the text (derivation, outline, weft), frequent repetitions, and a lack of supra-textual coherence of the argumentation in contrast to the coherent and easily comprehensible sequence of arguments within individual paragraphs.

Many studies explored the identifiability of chatbot-generated text using machine learning–based detectors, a subset of AI technologies [27-29]. These detectors often achieve higher identification rates than human evaluators. However, direct human comparison is rarely included, and accuracy and F_1-scores vary significantly depending on the text genre and the specific machine learning classifier used. For instance, when various LLM-based classifiers are applied to different data sets, their accuracy ranges from 70% (DetectGPT classifier on Wikipedia articles) to 97% (GPT-Pat classifier on COVID-19–related question-answer data sets). Similarly, perplexity-based classifiers achieve around 70% accuracy on ACL paper abstracts, whereas RoBERTa (Robustly Optimized BERT Pretraining Approach)–based classifiers reach up to 97% on COVID-19–related data sets [37].

Another challenge is that while these tools are generally reliable in detecting AI-generated text, they are not always sufficiently accurate in identifying human-authored text. This suggests that the tools may struggle with the complexity of human writing while also highlighting a key limitation—especially in cases where a lecturer, for example, must evaluate a single piece of writing without comparison [14,31].

Our study also compares the performance of medical students and ChatGPT. Notably, the texts written by medical students received higher professional evaluations. However, the humanities experts specifically rated the linguistic quality of ChatGPT-generated texts more favorably. Additionally, when comparing ChatGPT's performance with that of medical students in an examination setting—such as in the study by Huh [38]—ChatGPT performed worse than medical students [31]. In a parasitology examination, ChatGPT correctly answered 60.8% of the questions, whereas the average score among 77 medical students was significantly higher at 90.8% [38]. In comparison, a German study by Friederichs et al [39] found that ChatGPT correctly answered two-thirds of all multiple-choice questions at the level of the German state licensing examination in the Progress Test Medicine. It even outperformed most medical students in their first 3 years and performed comparably to students in the later stages of their studies [39]. Our study also revealed that participants who overestimated ChatGPT's writing capabilities and underestimated those of the students were more likely to misidentify the author. This misconception was particularly evident in cases where participants misclassified texts in both sessions, suggesting that their biased perception significantly influenced their decisions.

Interpretation

Our study demonstrates that the identification rate predicted by Turing holds within a group primarily engaged in student teaching and academic writing. Our findings confirm the expectation that linguistic features play a more significant role in identifying AI-generated texts than content familiarity or specialized expertise. In both expert groups, text-analytical features were the primary factors influencing their decisions. This aligns with the emerging field of stylometric analysis, which is increasingly being applied to the detection of AI-generated content [40]. ChatGPT-generated text, especially in comparison to authors from the (fictional) literature domain, exhibits limited stylistic variety [41]. Notably, there was no significant difference in the identification rate between the 2 expert groups, despite 1 being more familiar with the subject matter. Higher proximity to the topic was also not a predictive factor at the individual participant level. Instead, certain linguistic characteristics played a key role in the

decision-making process and were consistently associated with AI-generated texts. In particular, redundancy, repetition, and a lack of coherence were distinctive features attributed to ChatGPT-generated texts. While these traits influenced the perception of AI authorship, they ultimately did not prove to be reliable predictors for correct identification. The linguistic features of ChatGPT-generated texts are often perceived as superior due to their smoother wording and better structural organization. This aligns with findings from [42], which indicate that AI-generated texts tend to exhibit relatively low lexical density, high reading ease, and frequent use of the simple present tense. Whether these linguistic characteristics, if systematically outlined in a manual and provided to participants beforehand, could enhance identification accuracy remains an open question. However, this presents an intriguing avenue for future research.

Limitations

While numerous studies are currently investigating the performance of LLM-based text generators such as ChatGPT, many focus primarily on their assistive role in the writing process rather than assessing the quality of fully generated long-form scientific texts. A key contribution of our study is that it examines complete, AI-generated scientific texts rather than partial outputs. Additionally, instead of relying on specialized AI detection tools, we analyze how individuals working in academia recognize such texts without assistance and how they evaluate their performance while identifying distinct linguistic features. This study enables the compilation of categorical features that could aid in identifying AI-generated text in both academic and everyday reading. However, limitations in generalizability arise due to the relatively small sample size and the exclusive use of a single AI model, ChatGPT version 3.5. Nevertheless, for an exploratory study, we do not consider this a critical issue. Additionally, participants were aware that 1 of the texts had to be AI-generated, raising the question of whether they would have identified an AI-authored text without this prior knowledge. A further limitation arises from the use of different interviewers for the 2 expert groups, who also differed methodologically in terms of blinding. However, it should be noted that the decision to identify the authors was always made before the interview process. Additionally, the interview was transcribed in bullet points, so some information may have been lost in this process. Finally, the dynamic nature of development should also be acknowledged, as ChatGPT, like other AI programs, is continuously being developed and improved.

Conclusion

Our study shows that linguistic and text-analytical features, in particular, play a role in the decision-making process for correctly identifying a chatbot author. In our sample, both nonspecialists and specialists identified AI-generated texts with an accuracy rate of approximately 70% (48/69). Further quasi-experimental studies using texts from other academic disciplines should be conducted to determine whether instructions based on these features can enhance lecturers' ability to distinguish between student-authored and AI-generated work.

A follow-up study could be conducted in a few years to track the evolution of AI-generated text identification and examine whether identification success changes as AI technology and tools advance.

Acknowledgments

The authors thank the students for providing their work. We also extend our gratitude to our colleagues at the Paediatric and Neurological University Clinics of the Ruhr-University of Bochum, as well as the teachers at the Department of Philology, for their voluntary participation, contributions, and suggestions. Additionally, we thank the Medical Dean of Studies, Prof. T. Schäfer, for the initial exchange, and Dr. A. Lucke for her critical endorsement of our work.

Authors' Contributions

BD and JSB contributed equally to conceptualization and data acquisition, with BD leading the qualitative analysis and the original draft writing. CM played a leading role in conceptualization, supervision, and writing—review and editing—while contributing equally to formal analysis and data acquisition. TL and MB shared equal responsibilities in supervision and writing—review and editing. JS contributed equally to conceptualization, qualitative analysis, and writing, whereas EEK was involved in writing—review and editing on an equal basis. SH took the lead in proofreading and contributed equally to review and editing. MT led the formal analysis while sharing equal responsibilities in supervision and writing—review and editing. All authors have reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questionnaire with English translation and abbreviation label.
[DOCX File , 15 KB - mededu_v11i1e62779_app1.docx]

Multimedia Appendix 2

Categories for the qualitative analysis, created from the various reasons given in free form for the decision on the authorship of a text.
[DOCX File , 18 KB - mededu_v11i1e62779_app2.docx]

References

1. Jannai D, Meron A, Lenz B, Levine Y, Shoham Y. Human or not? A gamified approach to the Turing Test. arXiv Preprint posted online on January 30, 2023 [FREE Full text]
2. Ma Y, Liu J, Yi F, Cheng Q, Huang Y, Lu W, et al. AI vs. human -- differentiation analysis of scientific content generation. arXiv. Preprint posted online on January 21, 2023 URL: https://doi.org/10.48550/arXiv.2301.10416 [accessed 2023-05-01]
3. von Garrel J, Mayer J, Mühlfeld M. Hochschule Darmstadt. 2023. URL: https://doi.org/10.48444/h_docs-pub-395 [accessed 2024-11-01]
4. ChatGPT: optimizing language models for dialogue. OpenAI. URL: https://openai.com/blog/chatgpt/ [accessed 2023-07-01]
5. Taecharungroj V. "What can ChatGPT do?" Analyzing early reactions to the innovative AI chatbot on Twitter. BDCC 2023 Feb 16;7(1):35 [FREE Full text] [doi: 10.3390/bdcc7010035]
6. Berdejo-Espinola V, Amano T. AI tools can improve equity in science. Science 2023 Mar 10;379(6636):991 [FREE Full text] [doi: 10.1126/science.adg9714] [Medline: 36893248]
7. Marx JPS. ChatGPT im studium: die top 10 befehle für effektives Lernen. Shribe. 2023. URL: https://shribe.de/chatgpt-studium/ [accessed 2024-11-01]
8. Arif TB, Munaf U, Ul-Haque I. The future of medical education and research: is ChatGPT a blessing or blight in disguise? Med Educ Online 2023 Dec 21;28(1):2181052 [FREE Full text] [doi: 10.1080/10872981.2023.2181052] [Medline: 36809073]
9. Choi EPH, Lee JJ, Ho MH, Kwok JYY, Lok KYW. Chatting or cheating? The impacts of ChatGPT and other artificial intelligence language models on nurse education. Nurse Educ Today 2023 Jun;125:105796. [doi: 10.1016/j.nedt.2023.105796] [Medline: 36934624]
10. Flanagin A, Bibbins-Domingo K, Berkwits M, Christiansen SL. Nonhuman "authors" and implications for the integrity of scientific publication and medical knowledge. JAMA 2023 Feb 28;329(8):637-639 [FREE Full text] [doi: 10.1001/jama.2023.1344] [Medline: 36719674]
11. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, et al. ChatGPT and other large language models are double-edged swords. Radiology 2023 Apr;307(2):e230163 [FREE Full text] [doi: 10.1148/radiol.230163] [Medline: 36700838]
12. O'Connor S. Open artificial intelligence platforms in nursing education: tools for academic progress or abuse? Nurse Educ Pract 2023 Jan;66:103537 [FREE Full text] [doi: 10.1016/j.nepr.2022.103537] [Medline: 36549229]
13. Zenthöfer J. Erstes urteil zu ChatGPT an hochschulen. FAZ. 2024. URL: https://www.faz.net/aktuell/karriere-hochschule/student-nutzt-chatgpt-fuer-bewerbung-erstes-urteil-zu-ki-an-hochschulen-19564795.html [accessed 2024-05-01]
14. Abburi H, Roy K, Suesserman M, Pudota N, Veeramani B, Bowen E, et al. A simple yet efficient ensemble approach for AI-generated text detection. arXiv Preprint posted online on January 23, 2023 [FREE Full text]
15. Jiao W, Wang W, Huang J, Wang X, Tu Z. Is ChatGPT a good translator? A preliminary study. arXiv Preprint posted online on January 14, 2023 [FREE Full text]
16. Richtscheid W. Some hints for the use of "chatGPT" when using with German input language. Medium. 2023. URL: https://medium.com/@walter.richtscheid_93860/notes-on-the-use-of-chatgpt-in-german-language-c4941a14da54 [accessed 2023-07-01]
17. De Vries A. Computergenerierter Zufall als kreatives Moment in Malerei und Literatur. Potentiale und Grenzen von Machine-Learning-Modellen am Beispiel von GPT. In: Lucke A, Alexa H, editors. Literaturwissenschaft und Informatik. Transdisziplinäre Perspektiven, digitale Methoden und selbstlernende Algorithmen. Bielefeld, Germany: Universität Bielefeld; 2024:93-121.
18. Corizzo R, Leal-Arenas S. One-GPT: a one-class deep fusion model for machine-generated text detection. 2023 Presented at: IEEE International Conference on Big Data (BigData); December 15-18, 2023; Sorrento, Italy p. 5743-5752.
19. Parviainen J, Rantala J. Chatbot breakthrough in the 2020s? An ethical reflection on the trend of automated consultations in health care. Med Health Care Philos 2022 Mar 04;25(1):61-71 [FREE Full text] [doi: 10.1007/s11019-021-10049-w] [Medline: 34480711]
20. Alturaiki AM, Banjar HR, Barefah AS, Alnajjar SA, Hindawi S. A smart chatbot for interactive management in beta thalassemia patients. Int J Telemed Appl 2022;2022:9734518 [FREE Full text] [doi: 10.1155/2022/9734518] [Medline: 35601050]
21. Bin Sawad A, Narayan B, Alnefaie A, Maqbool A, Mckie I, Smith J, et al. A systematic review on healthcare artificial intelligent conversational agents for chronic conditions. Sensors (Basel) 2022 Mar 29;22(7):1-20 [FREE Full text] [doi: 10.3390/s22072625] [Medline: 35408238]
22. Cao X, Liu X. Artificial intelligence-assisted psychosis risk screening in adolescents: practices and challenges. World J Psychiatry 2022 Oct 19;12(10):1287-1297 [FREE Full text] [doi: 10.5498/wjp.v12.i10.1287] [Medline: 36389087]
23. Gaffney H, Mansell W, Tai S. Conversational agents in the treatment of mental health problems: mixed-method systematic review. JMIR Ment Health 2019 Oct 18;6(10):e14166 [FREE Full text] [doi: 10.2196/14166] [Medline: 31628789]

24. Schick A, Feine J, Morana S, Maedche A, Reininghaus U. Validity of chatbot use for mental health assessment: experimental study. JMIR Mhealth Uhealth 2022 Oct 31;10(10):e28082 [FREE Full text] [doi: 10.2196/28082] [Medline: 36315228]
25. Liao W, Liu Z, Dai H, Xu S, Wu Z, Zhang Y, et al. Differentiating ChatGPT-generated and human-written medical texts: quantitative study. JMIR Med Educ 2023 Dec 28;9:e48904 [FREE Full text] [doi: 10.2196/48904] [Medline: 38153785]
26. Khairatun Hisan U, Miftahul Amri M. ChatGPT and medical education: a double-edged sword. Journal of Pedagogy and Education Science 2023 Mar 11;2(01):71-89 [FREE Full text] [doi: 10.56741/jpes.v2i01.302]
27. Hayawi K, Shahriar S, Mathew S. The imitation game: detecting human and AI-generated texts in the era of ChatGPT and BARD. Journal of Information Science 2024 Feb 14;The Imitation Game:1-36 [FREE Full text] [doi: 10.1177/01655515241227531]
28. Orenstrakh MS, Karnalim O, Suarez CA, Liut M. Detecting LLM-generated text in computing education: a comparative study for ChatGPT cases. arXiv Preprint posted online on January 17, 2023 [FREE Full text] [doi: 10.1109/compsac61105.2024.00027]
29. Mitrovic S, Andreoletti D, Ayoub O. ChatGPT or human? Detect and explain. Explaining decisions of machine learning model for detecting short ChatGPT-generated text. arXiv. Preprint posted online on January 11, 2023 URL: https://doi.org/10.48550/arXiv.2301.13852 [accessed 2023-05-01]
30. Fröhling L, Zubiaga A. Feature-based detection of automated language models: tackling GPT-2, GPT-3 and Grover. PeerJ Comput Sci 2021;7:e443 [FREE Full text] [doi: 10.7717/peerj-cs.443] [Medline: 33954234]
31. Elkhatat A, Elsaid K, Almeer S. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. Int J Educ Integr 2023 Sep 01;19(1):17 [FREE Full text] [doi: 10.1007/s40979-023-00140-5]
32. Turing AM. Computing machinery and intelligence. Mind 1950 Oct 01;LIX(236):433-460 [FREE Full text] [doi: 10.1093/mind/lix.236.433]
33. Schäer R. Einführung in Die Grammatische Beschreibung des Deutschen: Dritte. Berlin, Germany: Language Science Press; 2018.
34. Fobbe E. Forensische Linguistik: Eine Einführung. Tübingen, Germany: Narr Francke Attempto Verlag; 2011.
35. Hessler S. Autorschaftserkennung und Verstellungsstrategien: Textanalysen und-vergleiche im Spektrum forensischer Linguistik, Informationssicherheit und Machine-Learning (Volume 585). Tübingen, Germany: Narr Francke Attempto Verlag; 2023.
36. Waltzer T, Pilegard C, Heyman GD. Can you spot the bot? Identifying AI-generated writing in college essays. Int J Educ Integr 2024 Jul 08;20(1):11. [doi: 10.1007/s40979-024-00158-3]
37. Yu X, Qi Y, Chen K, Chen G, Yang X, Zhu P, et al. LLM paternity test: generated text detection with LLM genetic inheritance. arXiv Preprint posted online on January 30, 2024 [FREE Full text]
38. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. J Educ Eval Health Prof 2023 Jan 11;20:1 [FREE Full text] [doi: 10.3352/jeehp.2023.20.01]
39. Friederichs H, Friederichs WJ, März M. ChatGPT in medical school: how successful is AI in progress testing? Med Educ Online 2023 Dec;28(1):2220920 [FREE Full text] [doi: 10.1080/10872981.2023.2220920] [Medline: 37307503]
40. Uzun L. ChatGPT and academic integrity concerns: detecting artificial intelligence generated content. Language Education & Technology. 2023. URL: https://www.langedutech.com/letjournal/index.php/let/article/view/49/36 [accessed 2024-02-01]
41. Zwilling L, Berger M. ChatGPT does not speak style!. 2024 Presented at: Computational Linguistics Poster Session at the 46th Annual Meeting of the German Linguistic Society (DGfS); February 28, 2024; Bochum, Germany.
42. AlAfnan MA, Mohdzuki SF. Do artificial intelligence chatbots have a writing style? An investigation into the stylistic features of ChatGPT-4. Journal of Artificial Intelligence and Technology 2023 May 12;3:3 [FREE Full text]

Abbreviations

AI: artificial intelligence
LLM: large language model
RoBERTa: Robustly Optimized BERT Pretraining Approach
RUB: Ruhr University Bochum

Edited by B Lesselroth; submitted 31.05.24; peer-reviewed by C Bach, L Zhu, R Corizzo; comments to author 12.08.24; revised version received 28.11.24; accepted 16.01.25; published 03.03.25.

Please cite as:
Doru B, Maier C, Busse JS, Lücke T, Schönhoff J, Enax- Krumova E, Hessler S, Berger M, Tokic M
Detecting Artificial Intelligence–Generated Versus Human-Written Medical Student Essays: Semirandomized Controlled Study
JMIR Med Educ 2025;11:e62779
URL: https://mededu.jmir.org/2025/1/e62779
doi:10.2196/62779
PMID:40053752

Original Paper

AIFM-ed Curriculum Framework for Postgraduate Family Medicine Education on Artificial Intelligence: Mixed Methods Study

Raymond Tolentino[1], BHSc, MSc; Fanny Hersson-Edery[1], MD; Mark Yaffe[1,2], MD; Samira Abbasgholizadeh-Rahimi[1,3,4,5], BEng, PhD

[1]Department of Family Medicine, Faculty of Medicine and Health Sciences, McGill University, Montreal, QC, Canada

[2]Department of Family Medicine, St. Mary's Hospital Center, Integrated University Centre for Health and Social Services of West Island of Montreal, Montreal, QC, Canada

[3]Mila-Quebec, Montreal, QC, Canada

[4]Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, QC, Canada

[5]Faculty of Dental Medicine and Oral Health Sciences, McGill University, Montreal, QC, Canada

Corresponding Author:
Samira Abbasgholizadeh-Rahimi, BEng, PhD
Department of Family Medicine
Faculty of Medicine and Health Sciences
McGill University
5858 chemin de la Côte-des-Neiges
Montreal, QC, H3S 1Z1
Canada
Phone: 1 514 399 9218
Email: samira.rahimi@mcgill.ca

Abstract

Background: As health care moves to a more digital environment, there is a growing need to train future family doctors on the clinical uses of artificial intelligence (AI). However, family medicine training in AI has often been inconsistent or lacking.

Objective: The aim of the study is to develop a curriculum framework for family medicine postgraduate education on AI called "Artificial Intelligence Training in Postgraduate Family Medicine Education" (AIFM-ed).

Methods: First, we conducted a comprehensive scoping review on existing AI education frameworks guided by the methodological framework developed by Arksey and O'Malley and Joanna Briggs Institute methodological framework for scoping reviews. We adhered to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist for reporting the results. Next, 2 national expert panels were conducted. Panelists included family medicine educators and residents knowledgeable in AI from family medicine residency programs across Canada. Participants were purposively sampled, and panels were held via Zoom, recorded, and transcribed. Data were analyzed using content analysis. We followed the Standards for Reporting Qualitative Research for panels.

Results: An integration of the scoping review results and 2 panel discussions of 14 participants led to the development of the AIFM-ed curriculum framework for AI training in postgraduate family medicine education with five key elements: (1) need and purpose of the curriculum, (2) learning objectives, (3) curriculum content, (4) organization of curriculum content, and (5) implementation aspects of the curriculum.

Conclusions: Using the results of this study, we developed the AIFM-ed curriculum framework for AI training in postgraduate family medicine education. This framework serves as a structured guide for integrating AI competencies into medical education, ensuring that future family physicians are equipped with the necessary skills to use AI effectively in their clinical practice. Future research should focus on the validation and implementation of the AIFM-ed framework within family medicine education. Institutions also are encouraged to consider adapting the AIFM-ed framework within their own programs, tailoring it to meet the specific needs of their trainees and health care environments.

(JMIR Med Educ 2025;11:e66828) doi:10.2196/66828

KEYWORDS

artificial intelligence; family medicine; curriculum; framework; postgraduate education

Introduction

The College of Family Physicians of Canada (CFPC) establishes standards for postgraduate family medicine training and its accreditation [1]. It promotes a competency-based curriculum model known as Triple C (comprehensive, continuous, and centered in family medicine) [2] based on the Canadian Medical Education Directives for Specialists (CanMEDS)—Family Medicine framework [3] and on assessment objectives for certification in family medicine [4]. To ensure that medical curricula respond to new developments in health care, education, and societal trends, they must undergo periodic review, modification, and renewal [5-9]. Accordingly, a number of new content areas have been introduced in the recent past into the family medicine curricula. They include leadership [10], social determinants of health [11], ethics [12], global health [13-15], and physician wellness and burnout [16-18]. The increasing complexity of the medical needs of an aging population, the exponential growth in medical knowledge, and an increasingly digitalized environment suggest the need for digital-mediated solutions to support medical practitioners.

Artificial intelligence (AI) and its applications have made a rapid impact on many segments of society, including medicine [19] and notably, in primary health care [20]. While there is no universal consensus on the definition of AI, the World Health Organization [21] describes it as "the performance by computer programs of tasks that are commonly associated with intelligent beings." The introduction, integration, and implementation of AI-based tools and systems into family medicine education and practice assume an adequately trained cohort of users, but to date, training of family physicians on relevant aspects of AI to ensure effective and safe implementation has been absent or inconsistent [20,22]. As such, the CFPC's Outcomes of Training project has identified digital care and health informatics as a training gap and an area for educational enhancement requiring priority attention across the 17 family medicine postgraduate programs in Canada [23,24]. There have been efforts to include AI education globally within each level of medical training. These efforts are led by national medical associations such as the UK National Health Service, the US American Medical Association, and Canada's Royal College of Physicians and Surgeons. They have released documents recommending policies for integrating AI within their respective medical educational institutions [25-27].

Initiatives of AI teaching directed at physicians already in practice include the development of a continuing professional development 3-module CFPC Learn e-course titled, "Artificial Intelligence for Family Medicine" [28]. The first module of this course reviews the basic functionality of AI with applications in family medicine, while the second module focuses on core terminology and related concepts as well as potential harms or risks associated with AI. The last module reviews the concepts of the first 2 and focuses on learning how to tell if an AI-based tool is working well [28].

Competency about a particular subject has been described as the ability to carry out a certain task or action at a basic or acceptable level [29]. Liaw et al [30] have recently proposed six competency domains for family medicine training in AI: (1) foundational knowledge (What is this tool?), (2) critical appraisal (Should I use this tool?), (3) medical decision-making (When should I use this tool?), (4) technical use (How do I use this tool?), (5) patient communication (How should I communicate with patients regarding the use of this tool?), and (6) awareness of unintended consequences (What are the "side effects" of this tool?)." These authors suggest that such competencies can be integrated within current residency training during existing sessions on health informatics or evidence-based medicine but emphasize that these competencies are a "point of departure" and must be further worked on [30].

A curriculum framework can be described as "a core policy document that describes a range of requirements, regulations and advice which should be respected by all stakeholders in the education system, and which should guide the work of schools, teachers and the developers of other curriculum documents" [31]. Curriculum frameworks allow for a visual and detailed roadmap to develop and implement a curriculum [32]. Input from an interdisciplinary team of medical educators, AI experts, end users, researchers, and curriculum designers [33] can effectively support the development of a curriculum framework for teaching AI in family medicine postgraduate training programs. Our comprehensive review of the available curriculum frameworks [34,35] highlighted that there is no framework designed specifically for family medicine residency and no paper that described a systematic approach to design one. From the 2 frameworks uncovered, one framework was incomplete, while the other framework was brief and focused on ophthalmology [34]. The ophthalmology curriculum framework lacks adaptability, as it may prove inadequate for family medicine residency due to the diverse, community-based nature of family medicine, which differs significantly from the highly technological and hospital-based focus of ophthalmology.

Considering the gaps mentioned previously and the foundational importance of curriculum frameworks in the creation of new educational structures, our objective was to design and develop a curriculum framework for AI family medicine education, that is, Artificial Intelligence Training in Postgraduate Family Medicine Education (AIFM-ed), ensuring alignment with current competencies and educational goals. To achieve this, a combination of validated methods including 2 national expert panel discussions were conducted, supplemented by a previous comprehensive systematic scoping review [34]. Developing a framework based on expert insights would help address gaps in AI education and provide an adaptable guide for family medicine educators, curriculum designers, postgraduate residency program directors, medical education researchers, and policy makers in health care education. Due to the systematic approach in designing this framework, audiences can adopt this framework to other fields and specialties, considering that our review did not find any systematically developed frameworks.

Methods

Study Design

For the construction of an AIFM-ed framework, we followed the analysis, design, development, implementation, and evaluation model for instructional design, using the first 3 activities to guide our work. We followed a two-step approach suggested by Redwood-Campbell et al [36] for framework development, wherein (1) a review of the literature was made focusing on curriculum frameworks and core competencies for AI education in medicine [34,35] and (2) a working group used qualitative or consensus methods for final development of the framework.

Our scoping review aimed to synthesize knowledge from the literature on curriculum frameworks and current educational programs that focus on the teaching and learning of AI for medical students, residents, and practicing physicians, and adhered to PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines. Details of this comprehensive study have been published elsewhere [34,35]. Our review specifically identified several AI educational curricula programs (eg, courses, workshops, webinars, and projects) and 2 curriculum frameworks for AI education, one outlining a broad framework for any level of education [37], while the other described a complete framework for ophthalmology residency education [38].

The outcome of our review was the identification of early concepts that could be applied to elements of the curriculum framework for family medicine and AI [34,35]. This initial curriculum framework was later used during the panel discussion as part of the co-development and redesigning of the framework. This discussion applied the curriculum framework structure described by Obadeji [39], which examines six common elements: (1) the need and the purpose of a curriculum or a program, (2) learning objectives and outcomes, (3) course content that will facilitate the accomplishment of the objectives or learning outcomes, (4) organization of the content, (5) implementation of curriculum, and (6) curriculum evaluation and refinement. This study examined all elements except the final element (curriculum evaluation and refinement). The initial framework was deemed successful by the expert team based on the following indicators: relevance to medical educators and curriculum designers, alignment to current family medicine competencies and educational goals, clarity of AI-specific content, and its potential for further validation. However, we acknowledge that further studies are needed.

Qualitative Methodology

The expert panel methodology follows the SRQR (Standards for Reporting Qualitative Research) checklist [40]. Expert panels help to attempt to reach consensus on controversial subjects [41,42] such as the risk of AI tools leading to reduced proficiency in independent critical thinking and clinical judgment among physicians. The use of qualitative consensus methods for curriculum development facilitates input from a wide range of stakeholders (eg, physicians and curriculum developers) in order to assess and validate expert knowledge [43]. The use of expert panel discussions to assist in creating curricula has become established in pedagogical research and development [44]. Examples within the field of medicine include discussions around social determinants of health for undergraduate medical education [45], telemedicine opportunities for postgraduate medical education [46], and geriatric oncology in continuing medical education [47].

Participant Recruitment and Sampling Strategy

Our panel size fell within the recommended average of 8 members or a median of 6 [48]. The definition of an expert in our case is flexible due to the limited knowledge and experience on this emerging topic; this is emphasized by Duncan et al [49], who state that, “[t]oo narrow a definition, however, can restrict the number of potential participants.” In our case, we chose experts according to the definitions of Fink et al [41], which state that they must be, “representative of their professional group, with either sufficient expertise not to be disputed or the power required to instigate the findings.” This was reinforced by Mead and Mosely [50], which state that, “experts can be defined in a number of ways, such as their position in a hierarchy [...] or as recommended by other participants in a study.” Therefore, from these definitions, we selected panelists based on their academic qualifications, their number and relevance of AI-related publications, professional experience within the development, implementation or research of AI, and finally, any participation in AI-specific projects or conferences.

For panel 1, we reached out to family medicine (clinical) educators from affiliated universities and professional organizations across Canada via email. Snowballing by this initial group generated the names of others known as family medicine educators. For panel 2, family medicine residents were invited from an initial group of residents who were knowledgeable and aware of AI, and that initial group helped to recruit relevant residents for this study.

Each participant voluntarily participated in the study by providing their explicit consent and agreement, which was confirmed through email correspondence. To uphold confidentiality, data were safeguarded through limited, secure data access, the disposal of audiotapes after transcription, and the anonymous analysis of transcripts.

Ethical Considerations

This study involved a panel discussion with experts, which does not require formal ethics board approval under the Economic and Social Research Council Framework for Research Ethics guidelines [51]. According to these guidelines as well as Canada’s Tri-Council Policy Statement on Ethical Conduct for Research Involving Humans, research that presents minimal risk and does not involve sensitive information may be exempt from formal ethics review [52]. This study adhered to these recognized guidelines, ensuring that all participants were treated in accordance with principles of research integrity, voluntary participation, and informed consent.

Participant Eligibility Criteria

Input from 2 different types of panelists was desired, and they were included as participants within 2 distinct expert panels.

The first included family medicine educators practicing in Canada who were somewhat knowledgeable or have expertise in AI education. No limitations were placed on years of practice experience, years of knowledge or experience in AI, language proficiency, work setting, or the types of patients for whom they provided care. The second expert panel included participants who were at the time of the study family medicine residents at McGill University, and who were somewhat knowledgeable in AI. No limitations were placed on language proficiency, years of knowledge or experience in AI, work settings, or the types of patients they provided care for.

Data Collection

We conducted a recorded session of each expert panel via Zoom (version 5.16.10; Zoom Video Communications). The use of a web-based expert panel minimizes costs associated with travel; it also mitigates potential biases linked to panelists [53]. Each web-based expert panel discussion was approximately 2 hours long, followed the same format, used congruent discussion guides, and was facilitated by 2 members of the research team (RT and SAR). The discussions began with a brief presentation given by RT on the results of the first step of the project, that is, the comprehensive scoping review in the field [34,35]. Following the presentation, each of the five elements of the curriculum framework: (1) the need and the purpose of a curriculum or a program, (2) learning objectives and outcomes, (3) course content that will facilitate the accomplishment of the objectives or learning outcomes, (4) organization of the content, and (5) implementation of the curriculum were discussed sequentially and at length. When presenting each element, participants were invited to respond and discuss their opinions and thoughts related to each element, allowing for the co-development and redesigning of the framework together.

Data Analysis

Expert panel discussion data were analyzed using content analysis strategies [54,55] as previously used in a study developed for a training model for nurses using a literature review and expert panel discussions, in which data were analyzed using a descriptive qualitative approach that includes content analysis [56]. In our work, the preparation phase included transcribing the data, immersing in the data, and obtaining a sense of whole through reading the transcript multiple times. In our study, once the recordings from the expert panel discussions were received, one of the authors (RT) listened to the entire recording and subsequently transcribed it verbatim. The next stage of data analysis was the organizing phase, in which open coding and the creating of categories were conducted along with the grouping of codes under higher-order headings. These were carried out by one of the authors (RT) and verified by the senior author (SAR).

As the analysis of data used an inductive approach, no prior coding systems were used, such that coded categories were derived directly from the data [55]. Sentences and phrases from the panelists were captured. In vivo coding was used to prioritize participants' language and perspectives, while descriptive coding aided in categorizing key themes. Two independent coders reviewed the data (RT and SAR), with discrepancies resolved through discussion between coders and the research team. Saturation was achieved when no new themes emerged during the coding of the final transcript. The final step included the presentation of the final curriculum framework, which resulted from the incidence of codes and categories and its relation to the literature. Codes and categories derived were prioritized and highlighted with how frequently they appeared during the panel discussion as well as the overlap between both groups. These highlighted findings were then compared with existing literature to either support or challenge them. If these codes and categories were supported by the literature, they were subsequently integrated into the framework.

Results

Panelists Characteristics

A total of 37 educator and resident experts were invited, 14 for the educator group and 23 for the resident group. Ultimately, 8 from the former and 6 from the latter group participated, for a total of 14 participants. Scheduling problems were the most common reasons for nonparticipation. The characteristics of those included in the expert panel discussion are displayed in Table 1.

Table 1. Characteristics of expert panel participants included.

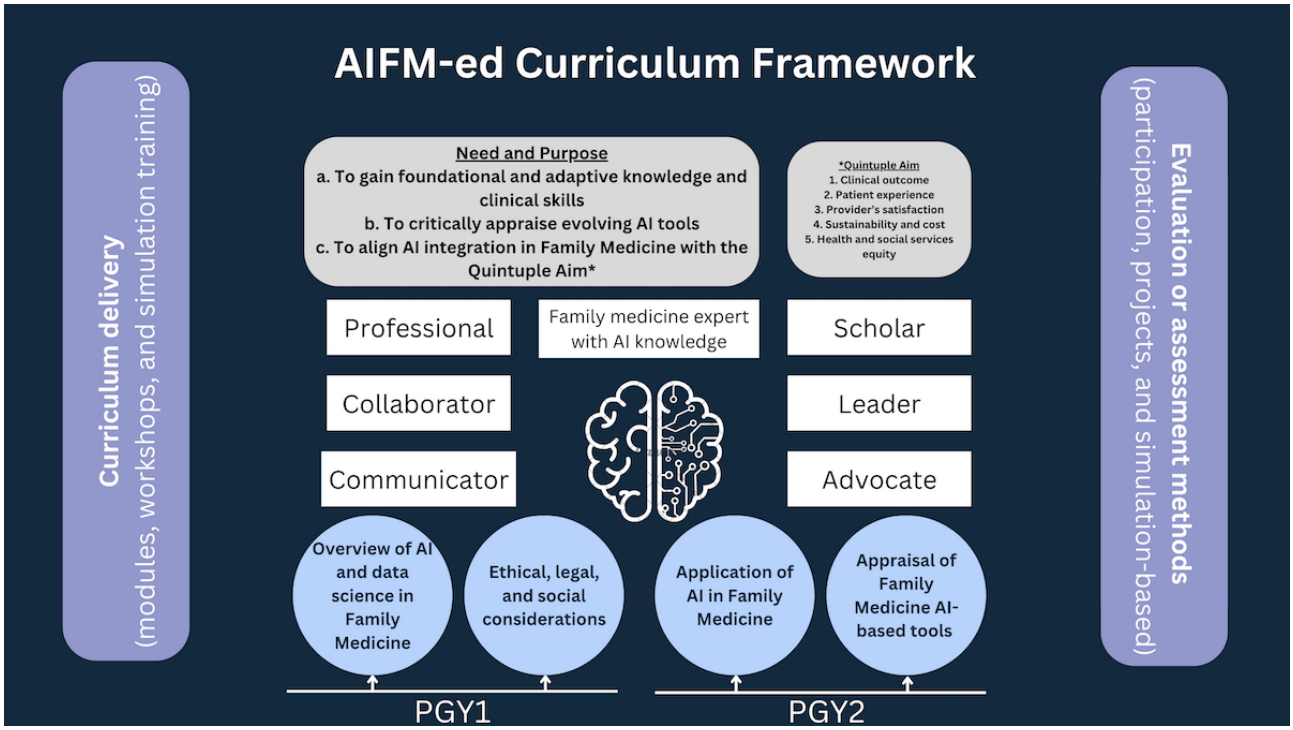
	Educator experts (n=8), n (%)	Resident experts (n=6), n (%)
Sex		
Male	3 (38)	4 (66)
Female	5 (62)	2 (33)
Educational background		
Doctoral (PhD)	7 (88)	0 (0)
Master	1 (22)	2 (33)
Bachelor or MD only	0 (0)	4 (66)
Affiliation		
McGill University	5 (62)	6 (100)
Other academic institution	3 (38)	0 (0)

Curriculum Framework for AIFM-ed

Overview

Our project has identified five elements of the curriculum framework for AI training in postgraduate family medicine education: (1) need and purpose of the curriculum, (2) learning objectives, (3) curriculum content, (4) organization of curriculum content, and (5) implementation of the curriculum. A condensed visual representation of the AIFM-ed curriculum framework is displayed in Figure 1, while each element is discussed in detail below.

Figure 1. Representation of the AIFM-ed curriculum framework. AI: artificial intelligence; AIFM-ed: Artificial Intelligence Training in Postgraduate Family Medicine Education; PGY: postgraduate year. * refers to the Quintuple Aim.



Element 1: Need and Purpose of the AIFM-ed Curriculum

When modifying a curriculum in family medicine postgraduate training, it is important to understand why it must be changed and for what purpose. Both panels discussed the current low priority of AI curricula. Residents emphasized a lack of exposure in training and practice. Both panels agreed that the integration of an AI curriculum will inevitably become imperative, recognizing its potential as an essential toolset in practice. One educator summarized this thought by saying, “AI will continue to evolve quickly, so a curriculum must be built for the future.”

To describe the need and purpose for AI education in family medicine, we co-developed the following: “The purpose of an AI curriculum for family medicine residents is for future family physicians to: (a) gain foundational and adaptive knowledge and clinical skills, (b) to critically appraise evolving AI tools, and (c) to align AI integration in Family Medicine with the quintuple aim for health care improvement (i.e., improving population health, improving the provider and patient experience, reducing costs, and advancing health equity [57,58]).”

Using various definitions of AI, the educator panel debated what constitutes AI specifically in family medicine. The term “AI-based tools” is used throughout the results of this paper as a way of describing technologies empowered or enabled with AI algorithms to support clinical practice. This term has been used in previous literature on AI in the context of family medicine training [20,30].

Element 2: AIFM-ed Learning Objectives

Learning objectives are statements that describe significant and essential learning that learners need to be familiar with and reliably demonstrate at the end of a course or educational program [59]. The following outlines the learning objectives for AI training, aligning with CanMEDS and family medicine roles. Table 2 presents each CanMEDS role on the left column [3], with their affiliated learning objectives for AI family medicine education as structured by participants during the panel on the right column. Although the learning objectives are comprehensive and their practical application for most family doctors may be limited, they are ideal for advancing the knowledge and skills of AI-empowered family physicians.

Table 2. Learning objectives discussed during panels about artificial intelligence (AI) in relation to Canadian Medical Education Directives for Specialists (CanMEDS) roles.

CanMEDS roles	The learner engaged in AI education will be able to
Family medicine expert with AI knowledge	
Family physicians are skilled generalists who should be able to understand and use technology including AI tools to provide high-quality, responsive, community-adaptive care across the lifecycle, from prevention to palliation, in multiple settings, and for diverse populations.	• Explain a basic understanding of AI and basic concepts in relation to family medicine. • Demonstrate the use of AI-based tools for family medicine by showing how to use the tool and understand the output. • Critique and decide on when to use an AI-based tool over another health care resource. • Recognize AI-based tools' perceived biases and discriminatory behavior (eg, an AI-based tool diagnosing skin conditions mainly trained on images of lighter skin tones may be less accurate in detecting conditions in individuals with darker skin tones) and the results demonstrated by AI-based tools where the learner will be able to solve and prevent further effects.
Communicator	
Family physicians foster life-long therapeutic relationships with patients and their families. This incorporates the dynamic exchanges that occur before, during, and after the medical encounter that facilitates gathering and sharing essential information for effective patient-centered health care [3].	• Explain to patients the current AI-based tool they are using and its results. • Address relevant gaps of understanding of AI tools among patients such as differing cultural perspectives and digital health literacy.
Collaborator	
Family physicians work with patients, families, communities, and other health care providers to provide safe, high-quality, patient-centered care [3].	• Practice a collaborative team-based approach, including establishing positive and continuing working relationships with relevant stakeholders in relation to developing, implementing, and improving the quality of AI-based tools.
Leader	
Family physicians must actively contribute to implementing and maintaining a high-quality health care system and take responsibility for delivering excellent patient care. This includes prioritizing and using health care resources efficiently, executing tasks collaboratively with colleagues, and contributing to ongoing quality improvement initiatives within their own practice and its management [3].	• Identify which AI-based tools are appropriate for the clinical practice of family physicians. • Allocate AI-based tools, when available, to specific tasks (eg, administrative work) in order for optimal patient care and practice management. • Analyze incidents of use of AI-based tools, appraise AI-based tools, and resolve any issues to avoid patient injury.
Advocate	
Family physicians leverage AI-driven insights to advocate for patients and communities, using their expertise to identify health needs, drive meaningful change, and mobilize resources for improved care outcomes [3].	• Extend AI-based tools and resources, when available and known, with other family physicians and family medicine communities. • Advocate for established AI-based tools, when available, to patients with the aim of improving their health outcomes.
Scholar	
Family physicians demonstrate a lifelong commitment to excellence in practice through continuous learning and teaching others; gather, combine, and evaluate evidence; and contribute to the creation and dissemination of knowledge [3].	• Participate in scholarly activities related to AI that benefit professional growth, clinical practice, and patients. • Maintain or enhance one's own knowledge and skills through professional educational activities related to AI and ongoing self-directed learning.
Professional	
Family physicians are committed to the health and well-being of their patients and society through competent medical practice; accountability to their patients, the profession, their colleagues, and society; profession-led regulation; ethical behavior; and maintenance of personal well-being [3].	• Recognize and appropriately respond to ethical, legal, and social issues encountered in practice, as it relates to AI-based tools and family medicine by communicating to the proper channels and resources (eg, AI and data experts, information technology specialists, ethics boards, and lawyers).

Element 3: AIFM-ed Curriculum Content

When developing a curriculum, a crucial task is to identify relevant subject knowledge, skills, attitudes, and behaviors that will help form the learning objectives [39]. Currently, there is no required AI education in Canadian undergraduate medical education. However, both educators and residents in our study agreed that for AI to be efficiently introduced in family medicine residency, it must be preceded by education in undergraduate

medical education. This earlier introduction of principles and concepts of AI will facilitate learning the more difficult material that is to come. The panels envisaged a basic stream of education in residency for those who had no exposure in undergraduate years. This would address fundamentals and basic knowledge of AI (eg, history, AI model development process, and core algorithms). A more advanced stream of AI education for residents would summarize the fundamentals and focus on how to use AI-based tools (applications) along with how to decide when to use and evaluate them (critical appraisal).

Residents noted that understanding how AI-based tools are used in clinical practice was the preferred content area for study, with less attention devoted to ethical, legal, and social considerations of AI. A resident put this in context, noting that they "do not need or want to learn the history of ChatGPT, but rather how to write effective prompts within this natural language processing chatbot." Table 3 summarizes the key concepts and areas of interest that family physicians should learn and content to include in the curriculum, as viewed by the participants.

Table 3. The curricular concepts and topics of relevance to family physicians.

Main curricular topic	Subtopics
Overview of AI[a] and data science in family medicine	
Providing an overview of AI definitions and concepts including machine learning as well as topics related to data science (eg, mathematics and statistics) and clinical epidemiology for family medicine.	• Review of AI (definitions and concepts) as it relates to family medicine • Introduction to AI and fundamentals of data science in family medicine • Strength and limitations of AI-based tools
Ethics, legal, and social considerations	
Understanding the ethical, legal, and social concerns of AI as it impacts family medicine clinical practice.	• Ethics, patient rights, data security, and confidentiality • Liabilities and regulatory and policy considerations • Equity, diversity, and inclusion of AI
Application of AI in family medicine	
Understanding how to choose and engage with AI-based tools in clinical settings and workflows with the ability to understand, interpret, and apply results of AI systems in clinical practice.	• Clinical practice management and operation • Preventative care and risk profiling (eg, mental health and chronic disease) • Patient self-management • Physician decision support • Physician wellness and resilience • Social determinants of health
Appraisal of family medicine AI-based tools	
Assessing and reviewing AI-based tools to ensure safe and effective integration and use in clinical practice.	• Identification of potential AI adverse effects and potential solutions • Quality improvement

[a]AI: artificial intelligence.

Element 4: Organization of AIFM-ed Curriculum Content

Family medicine postgraduate training is 24 months long in Canada. Given that the current curriculum is considered very heavy, educators and residents emphasized that the addition of another competency could be a burden to both educators and resident learners. They nonetheless agreed that AI curricula will eventually need to be added to that and an organized teaching structure would need to be established. Residents favored incorporating the teaching within the existing, already tight, 24-month core teaching, so that the benefits of longitudinal learning could be taken advantage of. The educators saw AI knowledge–based training during the first postgraduate year, followed by the development of AI-based clinical skills in the second postgraduate year. Educators proposed that if deeper AI education is needed, an additional third-year training program could be introduced for a select group of interested trainees to develop advanced AI skills in family medicine.

Element 5: Implementation of AIFM-ed Curriculum

Curriculum implementation will require the identification of appropriate resources (eg, educators and materials) along with educational strategies that will facilitate teaching activities and learner evaluation.

Curriculum Delivery

Residents highlight that AI education must be longitudinal, as it must be built upon throughout the medical education continuum. Furthermore, educators emphasized that residency is student-centered with learners coming from diverse backgrounds where they must replicate the actual tasks performed during in practice. Therefore, the learning theory of constructivism appears to be a sound and advantageous choice. This learning theory posits that learners actively construct their own learning by drawing upon their prior experiences [60].

There are several methods to implement an AI education curriculum to family medicine residents; however, there are

certain methods that are recommended by both educators and residents. In terms of learning about the knowledge and background of AI (eg, review of AI concepts or the ethical, legal, and social considerations of AI), hybrid (web-based and in-person) courses with asynchronous web-based modules, and in-person workshops, problem-solving sessions could be applied. Residents emphasized that didactic large group lectures especially in regard to a novel topic such as AI would be less engaging. The learning of such content should be considered a refresher with emphasis on the context of AI in family medicine. Both educators and residents then suggest that the in-person sessions would serve as a space for questions and answers and problem-solving activities.

To execute these educational methods, human resources (eg, AI medical educators) and material resources (eg, existing AI-based tools) are pertinent. Educators and residents highlighted that experts in the field of AI and family medicine would be ideal; however, educators emphasized that the faculty challenges such as the current number of experts are limited to provide this education. To overcome this, residents suggested that once an AI curriculum is established, further educators could be sourced from recently graduated residents who completed the AI in family medicine curriculum. With respect to material resources such as family physician–focused AI-based tools, both groups emphasized that they must be validated before use in educational settings.

Assessment and Evaluation Methods

Residents emphasize that the assessment and evaluation methods for the curriculum should be simple in context and focus on learners' participation and exposure. More specifically, learners should be able to have the capacity to demonstrate how to use AI-enabled tools and techniques in a health care setting. This can be seen through the completion of projects and problem-based and simulation-based assessments. Educators on the other hand emphasized taking into account Kirkpatrick's 4 levels of training evaluation model [61], where assessments should be directly related to the activity's learning objectives.

Discussion

The First Curriculum Framework for AI in Family Medicine (AIFM-ed)

In this study, we introduced a novel and evidence-based initial curriculum framework, that is, AIFM-ed developed for AI literacy education in family medicine postgraduate training. This systematically co-developed framework used a combination of validated methods including a comprehensive scoping review, resident and educator panel discussions, and the involvement of interdisciplinary experts in the field. During the development and cocreation of this framework, several key findings emerged. These include the crucial role of multiple resource partners and innovative practices when integrating AI educational content in family medicine education. For example, AI technology vendors specializing in health care, upcoming startups, and AI-focused organizations.

Furthermore, educators and residents stressed the importance of learning about the application of AI-based tools and simulating their use as a method of learning. Several innovative practices have already been implemented including case-based learning and flipped classroom models. Moreover, the adoption of AI-based tools can be diverse depending on its context (eg, teaching and learning and clinical practice) with several barriers and enablers. Additionally, the study identified several challenges in effectively integrating an AI curriculum framework into existing educational structures. These include the lack of AI definition standardization, the reduced urgency in practice due to the lack of time and resources, as well as the capacity to balance theoretical and practical curricular content.

Interprofessional Collaboration and Resources

During the development of the AIFM-ed curriculum framework, several resource partners were identified when discussing the implementation of AI education in family medicine. Interprofessional collaboration within multidisciplinary teams is essential in order for an AI curriculum to be effective [62]. Other researchers emphasize this sentiment when listing their recommendations of ensuring a responsible integration of AI technologies in medical education [63]. This multidisciplinary team and resource partners may include several stakeholders such as nurses, social workers, epidemiologists, AI experts, data engineers, software developers, and patients [64]. Other resource partners identified included AI technology vendors specializing in health care, upcoming startups, and AI-focused organizations. Residents brought up the concept that AI-based tools and AI in general will substantially change in the future (eg, improved tools, systems, and integrations) and thus stressed the importance of continuous partnerships with other professionals in order for relevant information and AI tools.

Educators emphasized that they were unaware of many AI-based tools for patient support and were thus apprehensive in advocating for AI-based tools. Therefore, family physicians and other primary care team members (eg, administrative staff and nurses) should share AI-based tools and resources, when available and known, with other family physician and family medicine communities. Additionally, residents have suggested that before advocating or suggesting AI-based tools, a list of recommended AI-based tools must be developed and released from a medical organization such as the CFPC. Currently, there is a scoping review and inventory that has identified and evaluated published studies that have tested or implemented AI in primary care settings [20,65]. This can be a starting point for such a list of recommended AI-based tools.

Application and Simulation of AI-Based Tools

Both educators and residents emphasize that a curriculum should focus on how to use AI-based tools (application) along with how to decide when to use and evaluate them (critical appraisal). Residents are already doing this comparatively as seen through their discussions of using ChatGPT, an AI-based chatbot launched by OpenAI that can be used as a digital consultant (eg, simple inquiries about diagnoses and treatment plans). One resident stressed that although they use ChatGPT at times for inquiries related to patient care, they are cautious of the information, as they are aware that ChatGPT can make mistakes and always consult other resources. As ChatGPT rises in

prominence, its impact on medical education has been evident through the resident panel discussion and the literature [66,67].

The incorporation of AI content in medical education has already begun with innovative practices, which include case-based learning and flipped classroom models. Case-based learning incorporates real-world AI use cases, where AI is used in clinical practice as examples for physicians [68]. Through this learning approach, students have a better understanding of the technical aspects of AI, as it allows physicians to compare their thought processes with other students and critically reflect or challenge their assumptions and biases of AI and clinical practice [68]. One study assessed the capabilities of ChatGPT within the framework of a preclerkship case-based active learning curriculum. Although the AI chatbot is not comprehensive enough to serve as a textbook, it was shown to answer questions, generate test questions, and appropriately respond to prompts in case-based learning scenarios [69]. According to a scoping review of teaching AI ethics in medical education, 5 publications reported in using case-based learning when understanding ethical challenges [70]. Resident panelists believe that simulation of these tools is beneficial, as it allows residents to enjoy the learning process and realize how these AI-based tools would operate in actual clinical settings. During these simulation sessions or case-based learning approaches, educator panelists highlighted reviewing the capabilities and basic functions of AI-based tools.

Another practical example of incorporating AI content through innovative practices is the flipped classroom model approach. Flipped classroom models can consist of web-based content supplemented by in-person classroom sessions [71], a key observation reinforced by residents of the panel discussion. One study designed and evaluated a novel AI course for medical students using a flipped classroom model, and they found that attending the course can increase self-perceived AI readiness in medical students [71]. In addition, educators have also commented on facilitating AI learning by integrating family medicine AI-based tools in quality improvement projects, which has been emphasized and recommended by other researchers [72].

Adopting AI in Education and Clinical Practice

Family physicians use AI, when implemented, primarily for diagnosis, detection, or surveillance purposes [20]. Although educators have flexibility in choosing from a wide range of AI tools, certain tools have proven to be particularly essential for effective integration. These include AI-enabled chatbots, clinical documentation support, and diagnostic decision support, which have shown to improve physicians' efficiency and accuracy in their work [73-75]. However, there have been several barriers identified in previous reviews, which have made the adoption of AI-based tools difficult [76-78]. These issues include a lack of trust among educators, students, and clinicians; insufficient training and digital literacy; and resistance to change [77].

Additional challenges include data privacy and patient safety concerns, ethical and legal issues, interoperability issues, lack of funding, and inequities in access to AI tools—particularly between rural and urban settings [79]. In contrast, several strategies and enablers have been identified in order to better facilitate the adoption of AI and its continued use. These strategies include strategies fostering interdisciplinary collaboration between educators, clinicians, and AI developers; providing targeted training programs to build AI literacy; developing high-quality datasets for diverse use cases; and creating supportive regulatory frameworks [77]. Establishing national or local community networks to share resources and best practices, while leveraging trusted relationships within these networks, can also significantly enhance confidence in and adoption of AI-based tools. To identify relevant enablers and barriers to AI adoption of a certain audience, a comprehensive, stakeholder-centered approach is essential. For example, researchers in Canada conducted in-depth interviews with primary health care and digital health stakeholders and were able to ascertain their current barriers and potential facilitators of AI [80].

It is important to note that AI systems exist in diverse contexts and content with distinct implications, risks, and ethical and legal challenges depending on their application and domain. For example, in education, AI-enabled tools using large language models may offer personalized education, but biases may be propagated, inaccurate information may be generated, or students may overrely on AI, undermining their critical thinking skills [63,81]. In addition, there is potential for the exacerbation of inequities in accessing AI tools as well as the misuse of AI-generated content. In comparison, AI-enabled tools in clinical practice, such as decision-support systems, could carry risks of incorrect or biased recommendations that may directly impact patient outcomes [82,83], thus, raising ethical concerns about patient autonomy and safety as well as legal liability in cases of harm. Therefore, the differences of AI in each domain are important to understand in order to identify appropriate safeguards. Future research should conduct comparative analyses of AI's risks, implications, and ethical and legal dimensions in educational versus clinical settings, examining factors such as accuracy, equity, accountability, and trust. These studies can inform best practices and policies to optimize AI's potential while mitigating domain-specific risks.

Curriculum Framework Challenges

During the development of this curriculum framework, there were several challenges in effectively integrating an AI curriculum framework into a family medicine residency training program. During the expert panel discussions, many experts emphasized the issue regarding the lack of standardization with the definition of AI. Although a definition of AI was chosen for the purpose of the panel, a specific and committed definition of AI within medical education has not been established [84-86]. Panelists argued that an AI definition must be properly explained to avoid confusion or misrepresentation. In relation to family medicine, a recent primer for AI in primary care was published, which provided the definition, "The field of AI is broad and rapidly expanding. The field is centred on how computers might be able to perform humanlike 'intelligent tasks,' such as summarizing large amounts of information or making inferences about a situation" [87]. The discussions regarding this framework highlight the necessity of a standardized AI definition for better development of teaching and learning content. This is especially true when specializing in different

fields of medical education, including family medicine and primary care.

There is a need to introduce AI education within family medicine; however, the low urgency and priority to integrate this type of education at the moment were noted throughout the discussions. This can be due to the lack of AI-enabled tools for family physicians currently being developed, tested, and implemented in practice [88,89]. Furthermore, some residency programs lack the appropriate AI tools or are in lower-resource settings. As a result of the minimal exposure family physicians have with AI, their motivation to learn about the topic can also be reduced. This reduced priority of AI education competes with the CFPC's 105 topics of family medicine curricula [4]. This is exacerbated by the fact that Canada is in a unique position, in which the length of residency training is only 2 years. In addition, the rapid advancement of AI introduces an extra layer of complexity. As new AI-based tools emerge and existing ones advance, educators and family physicians must frequently reassess and update their knowledge and skills. For example, the recent introduction of generative AI and generative AI tools such as ChatGPT has gained widespread popularity in medical and academic settings [90]. Thus, it is difficult to maintain a robust framework due to the inevitable rapid changes of AI in health care. Therefore, the eagerness to integrate this type of education within the curriculum should be met with caution to manage the expectations of both educators and learners.

A key observation made throughout the panel discussion was about the AI content and how much should a family physician know about AI. During the discussions, many of the participants voiced support on the application and appraisal of AI-enabled tools. This is especially challenging when residency is only 24 months, and there are no required AI educational programs presented in the Canadian undergraduate medical education system. Therefore, within the learning objectives, in regard to how much a family physician should know about AI remains undetermined. Further research must be conducted to investigate the level of AI education a family physician should be aware of. Overall, the aforementioned challenges must be addressed in order for this curriculum framework to be effectively implemented.

Future Studies

Following the analysis, design, development, implementation, and evaluation model process, researchers may move forward to the implementation and evaluation of the AIFM-ed framework. During the implementation step, an educational program such as a course or workshop can be developed with the main concepts originating from the curriculum framework. The training for family medicine is already packed; thus, the implementation of this framework will depend on several factors including the current use of AI-enabled tools in family medicine training, previous training in AI (eg, the undergraduate foundation of AI), and the capacity of experienced teachers. However, once implemented, certain success indicators will need to be evaluated to understand its impact as well as any areas for improvement. Future studies could explore indicators such as the perceived impact of the framework, degree of implementation, as well as knowledge and skill apprehensions. These indicators can be evaluated through the framework-derived educational training program according to the Kirkpatrick model [61].

Strengths and Limitations of This Study

This study had several strengths, including the formation of a national, multidisciplinary panel of family medicine educators. This diverse panel facilitated enriching discussions with varied expertise and insights, allowing for a comprehensive understanding of practical implications and current perspectives on AI education in family medicine postgraduate training. Additionally, by involving both educators and residents, the AIFM-ed curriculum framework ensures the representation from key stakeholders involved in the teaching and learning process of AI education. This co-design approach enhanced the relevance and applicability of the AIFM-ed curriculum framework. Regarding the overall development of this framework, a multi-method systematic approach was used, which includes a comprehensive systematic scoping review and multiple expert panel discussions. This approach allowed us to identify and build on existing AI curriculum topics and resources while also creating new ones. Furthermore, this structured and reproducible methodology ensures a robust foundation that can be used by other educators and researchers to develop training programs (eg, courses) following the established framework.

Despite the strengths, this study also had few limitations. First, the study was developed for programs in Canada, which limits its applicability to other countries due to the different medical education structures globally and their current relationships with AI. However, this could be a starting guide for other researchers to adapt it to their own context. Additionally, expert panel diversity was limiting, where the resident panel came from a single institution, which may further limit the generalizability of the framework. Furthermore, as the participants for the panel discussion were not randomized and were purposively recruited, the results may be subject to selection bias.

Conclusions

We co-developed an AIFM-ed framework for family medicine residency training that outlines its curricular purpose, learning objectives, AI curricular topics, delivery methods, and evaluation strategies to be used by medical institutions. The AIFM-ed curriculum framework ultimately aims to enhance the education of future family physicians, equipping them to effectively integrate AI-enabled tools into their practice and patient care. It is hoped that this framework will provide further advocacy, productivity, and gradual change within the area of curriculum development and AI medical education. Overall, medical institutions are encouraged to begin equipping future physicians with the knowledge, skills, and confidence to effectively use AI-enabled tools, as these technologies will continue to grow within the field of health care and family medicine.

Acknowledgments

The authors would like to thank the panelists for their support, time, and contributions to this research. The authors would also like to acknowledge Dr Pierre Pluye who provided key guidance to the development of this project prior to his passing in August 2023. SAR is Canada Research Chair (Tier II) in artificial intelligence and Advanced Digital Primary Health Care, received salary support from a Research Scholar Junior 1 Career Development Award from the Fonds de Recherche du Québec-Santé during part of this project, and her research program is supported by the Natural Sciences Research Council Discovery (grant 2020-05246).

Authors' Contributions

Conceptualization was led by SAR and RT, who established the study's goals, design, and research questions and obtained the funding for the project. The methodology was developed by SAR and RT. The data collection was done by RT and SAR. Data curation was managed by RT. A formal analysis was conducted by RT. The original draft was written by RT. SAR, FH-E, and MY provided critical revisions. Reviewing and editing were a collaborative effort with all authors. Supervision and overall project leadership were provided by SAR.

Conflicts of Interest

None declared.

References

1. Hennen BK. Academic family medicine in Canada. CMAJ 1993;148(9):1559-1563 [FREE Full text] [Medline: 8477381]
2. Oandasan I, Working Group on Postgraduate Curriculum Review. Advancing Canada's family medicine curriculum: Triple C. Can Fam Physician 2011;57(6):739-40, e237 [FREE Full text] [Medline: 21673223]
3. Shaw E, Oandasan I, Fowler N. CanMEDS-FM 2017: a competency framework for family physicians across the continuum. The College of Family Physicians of Canada. 2017. URL: https://www.cfpc.ca/CFPC/media/Resources/Medical-Education/CanMEDS-Family-Medicine-2017-ENG.pdf [accessed 2025-03-15]
4. Crichton T, Schultz K, Lawrence K, Donoff M, Laughlin T, Brailovsky C, et al. Assessment objectives for certification in family medicine. College of Family Physicians of Canada. 2020. URL: https://www.cfpc.ca/CFPC/media/Resources/Examinations/Assessment-Objectives-for-Certification-in-FM-full-document.pdf [accessed 2025-03-15]
5. Wojda T, Hoffman C, Jackson J, Conti T, Maier J. AI in healthcare: implications for family medicine and primary care. In: Artificial Intelligence in Medicine and Surgery—An Exploration of Current Trends, Potential Opportunities, and Evolving Threats—Volume 1. London: IntechOpen; 2023.
6. Buja LM. Medical education today: all that glitters is not gold. BMC Med Educ 2019;19(1):110 [FREE Full text] [doi: 10.1186/s12909-019-1535-9] [Medline: 30991988]
7. Jamieson S. State of the science: quality improvement of medical curricula—how should we approach it? Med Educ 2023;57(1):49-56 [FREE Full text] [doi: 10.1111/medu.14912] [Medline: 35950304]
8. Mcleod P, Steinert Y. Twelve tips for curriculum renewal. Med Teach 2015;37(3):232-238. [doi: 10.3109/0142159X.2014.932898] [Medline: 25010218]
9. Jones R, Higgs R, de Angelis C, Prideaux D. Changing face of medical curricula. Lancet 2001;357(9257):699-703. [doi: 10.1016/S0140-6736(00)04134-9] [Medline: 11247568]
10. Sultan N, Torti J, Haddara W, Inayat A, Inayat H, Lingard L. Leadership development in postgraduate medical education: a systematic review of the literature. Acad Med 2019;94(3):440-449. [doi: 10.1097/ACM.0000000000002503] [Medline: 30379659]
11. Hunter KA, Thomson B. A scoping review of social determinants of health curricula in post-graduate medical education. Can Med Ed J 2019;10(3):e61-e71. [doi: 10.36834/cmej.61709]
12. Hong DZ, Goh JL, Ong ZY, Ting JJQ, Wong MK, Wu J, et al. Postgraduate ethics training programs: a systematic scoping review. BMC Med Educ 2021;21(1):338 [FREE Full text] [doi: 10.1186/s12909-021-02644-5] [Medline: 34107935]
13. Pritchard J, Alavian S, Soogoor A, Bartels S, Hall A. Global health competencies in postgraduate medical education: a scoping review and mapping to the CanMEDS physician competency framework. Can Med Educ J 2023;14(1):70-79 [FREE Full text] [doi: 10.36834/cmej.75275] [Medline: 36998501]
14. Gupta A, Talavlikar R, Ng V, Chorny Y, Chawla A, Farrugia M, et al. Global health curriculum in family medicine: resident perspective. Can Fam Physician 2012;58(2):143-146 [FREE Full text] [Medline: 22439168]
15. Drain PK, Primack A, Hunt DD, Fawzi WW, Holmes KK, Gardner P. Global health in medical education: a call for more training and opportunities. Acad Med 2007;82(3):226-230. [doi: 10.1097/ACM.0b013e3180305cf9] [Medline: 17327707]
16. Runyan C, Savageau JA, Potts S, Weinreb L. Impact of a family medicine resident wellness curriculum: a feasibility study. Med Educ Online 2016;21:30648 [FREE Full text] [doi: 10.3402/meo.v21.30648] [Medline: 27282276]
17. Penwell-Waines L, Runyan C, Kolobova I, Grace A, Brennan J, Buck K, et al. Making sense of family medicine resident wellness curricula: a Delphi study of content experts. Fam Med 2019;51(8):670-676. [doi: 10.22454/FamMed.2019.899425] [Medline: 31269221]

18. Eckleberry-Hunt J, Van Dyke A, Lick D, Tucciarone J. Changing the conversation from burnout to wellness: physician well-being in residency training programs. J Grad Med Educ 2009;1(2):225-230 [FREE Full text] [doi: 10.4300/JGME-D-09-00026.1] [Medline: 21975983]
19. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
20. Abbasgholizadeh Rahimi S, Légaré F, Sharma G, Archambault P, Zomahoun HTV, Chandavong S, et al. Application of artificial intelligence in community-based primary health care: systematic scoping review and critical appraisal. J Med Internet Res 2021;23(9):e29839 [FREE Full text] [doi: 10.2196/29839] [Medline: 34477556]
21. Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization. 2021. URL: https://www.who.int/publications/i/item/9789240029200 [accessed 2025-03-12]
22. Upshaw TL, Craig-Neil A, Macklin J, Gray CS, Chan TCY, Gibson J, et al. Priorities for artificial intelligence applications in primary care: a Canadian deliberative dialogue with patients, providers, and health system leaders. J Am Board Fam Med 2023;36(2):210-220 [FREE Full text] [doi: 10.3122/jabfm.2022.220171R1] [Medline: 36948537]
23. Fowler N, Oandasan I, Wyman R. Preparing our future family physicians: an educational prescription for strengthening health care in changing times. College of Family Physicians of Canada. 2022. URL: https://www.cfpc.ca/CFPC/media/Resources/Education/AFM-OTP-Report.pdf [accessed 2025-03-15]
24. Fowler N, Lemire F, Oandasan I, Wyman R. The evolution of residency training in family medicine: a Canadian perspective. Fam Med 2021;53(7):595-598 [FREE Full text] [doi: 10.22454/FamMed.2021.718541] [Medline: 34000054]
25. Reznick R, Harris K, Horsley T, Hassani M. Artificial intelligence (AI) and emerging digital technologies. The Royal College of Physicians and Surgeons of Canada. 2020. URL: https://www.royalcollege.ca/content/dam/document/membership-and-advocacy/2020-task-force-report-on-ai-and-emerging-digital-technologies-e.pdf [accessed 2022-06-18]
26. Topol E. The Topol review: preparing the health care workforce to deliver the digital future. National Health Service. 2019. URL: https://topol.hee.nhs.uk/wp-content/uploads/HEE-Topol-Review-2019.pdf [accessed 2023-04-25]
27. AMA passes first policy recommendations on augmented intelligence. American Medical Association. 2018. URL: https://www.ama-assn.org/press-center/press-releases/ama-passes-first-policy-recommendations-augmented-intelligence [accessed 2023-04-25]
28. AI for family medicine. College of Family Physicians of Canada. URL: https://cfpclearn.ca/ai-for-family-medicine [accessed 2025-03-12]
29. Austin Z. Competency and its many meanings. Pharmacy (Basel) 2019 Apr 22;7(2):e12402 [FREE Full text] [doi: 10.3390/pharmacy7020037] [Medline: 31013596]
30. Liaw W, Kueper JK, Lin S, Bazemore A, Kakadiaris I. Competencies for the use of artificial intelligence in primary care. Ann Fam Med 2022;20(6):559-563 [FREE Full text] [doi: 10.1370/afm.2887] [Medline: 36443071]
31. Stabback P. What makes a quality curriculum? UNESCO. 2016. URL: https://unesdoc.unesco.org/ark:/48223/pf0000243975 [accessed 2022-07-10]
32. Training tools for curriculum development: developing and implementing curriculum frameworks. UNESCO International Bureau of Education. 2017. URL: https://unesdoc.unesco.org/ark:/48223/pf0000250052 [accessed 2022-07-10]
33. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. JMIR Med Educ 2019;5(1):e13930 [FREE Full text] [doi: 10.2196/13930] [Medline: 31199295]
34. Tolentino R, Baradaran A, Gore G, Pluye P, Abbasgholizadeh-Rahimi S. Curriculum frameworks and educational programs in AI for medical students, residents, and practicing physicians: scoping review. JMIR Med Educ 2024;10:e54793 [FREE Full text] [doi: 10.2196/54793] [Medline: 39023999]
35. Tolentino R, Baradaran A, Gore G, Pluye P, Abbasgholizadeh-Rahimi S. Curriculum frameworks and educational programs in artificial intelligence for medical students, residents, and practicing physicians: a scoping review protocol. JBI Evid Synth 2023;21(7):1477-1484. [doi: 10.11124/JBIES-22-00374] [Medline: 37434376]
36. Redwood-Campbell L, Pakes B, Rouleau K, MacDonald CJ, Arya N, Purkey E, et al. Developing a curriculum framework for global health in family medicine: emerging principles, competencies, and educational approaches. BMC Med Educ 2011;11:46 [FREE Full text] [doi: 10.1186/1472-6920-11-46] [Medline: 21781319]
37. Masters K. Artificial intelligence developments in medical education: a conceptual and practical framework. MedEdPublish (2016) 2020;9:239 [FREE Full text] [doi: 10.15694/mep.2020.000239.1] [Medline: 38058891]
38. Valikodath NG, Cole E, Ting DSW, Campbell JP, Pasquale LR, Chiang MF, et al. Impact of artificial intelligence on medical education in ophthalmology. Transl Vis Sci Technol 2021;10(7):14 [FREE Full text] [doi: 10.1167/tvst.10.7.14] [Medline: 34125146]
39. Obadeji A. Health professions education in the 21st century: a contextual curriculum framework for analysis and development. J Contemp Med Edu 2019;9(1):34. [doi: 10.5455/jcme.20181212085450]
40. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for Reporting Qualitative Research: a synthesis of recommendations. Acad Med 2014;89(9):1245-1251 [FREE Full text] [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
41. Fink A, Kosecoff J, Chassin M, Brook RH. Consensus methods: characteristics and guidelines for use. Am J Public Health 1984;74(9):979-983. [doi: 10.2105/ajph.74.9.979] [Medline: 6380323]

42. Black N, Murphy M, Lamping D, McKee M, Sanderson C, Askham J, et al. Consensus development methods: a review of best practice in creating clinical guidelines. J Health Serv Res Policy 1999;4(4):236-248. [doi: 10.1177/135581969900400410] [Medline: 10623041]
43. Vankova D, Videnova J. Delphi technique for curriculum development. 2019 Presented at: 12th annual International Conference of Education, Research and Innovation; November 11-13, 2019; Seville, Spain p. 6167-6171 URL: https://library.iated.org/view/VANKOVA2019DEL
44. Lewthwaite S, Nind M. Teaching research methods in the social sciences: expert perspectives on pedagogy and practice. Br J Educ Stud 2016;64(4):413-430. [doi: 10.1080/00071005.2016.1197882]
45. Mangold K, Bartell T, Doobay-Persaud A, Adler M, Sheehan K. Expert consensus on inclusion of the social determinants of health in undergraduate medical education curricula. Acad Med 2019;94(9):1355-1360. [doi: 10.1097/ACM.0000000000002593] [Medline: 31460933]
46. Hart A, Romney D, Sarin R, Mechanic O, Hertelendy A, Larson D, et al. Developing telemedicine curriculum competencies for graduate medical education: outcomes of a modified Delphi process. Acad Med 2022;97(4):577-585. [doi: 10.1097/ACM.0000000000004463] [Medline: 34670239]
47. Hsu T, Kessler E, Parker I, Dale W, Gajra A, Holmes H, et al. Identifying geriatric oncology competencies for medical oncology trainees: a modified Delphi consensus study. Oncologist 2020;25(7):591-597 [FREE Full text] [doi: 10.1634/theoncologist.2019-0950] [Medline: 32237179]
48. Evans C. The use of consensus methods and expert panels in pharmacoeconomic studies. Practical applications and methodological shortcomings. Pharmacoeconomics 1997;12(2 Pt 1):121-129. [doi: 10.2165/00019053-199712020-00003] [Medline: 10169665]
49. Duncan EAS, Nicol MM, Ager A. Factors that constitute a good cognitive behavioural treatment manual: a delphi study. Behav Cognit Psychother 1999;32(2):199-213. [doi: 10.1017/s135246580400116x]
50. Mead D, Moseley L. The use of the Delphi as a research approach. Nurse Res 2001;8(4):4-23. [doi: 10.7748/nr2001.07.8.4.4.c6162]
51. Framework for research ethics. UK Research and Innovation—Economic and Social Research Council. URL: https://www.ukri.org/councils/esrc/guidance-for-applicants/research-ethics-guidance/framework-for-research-ethics/ [accessed 2025-03-25]
52. Tri-council policy statement: ethical conduct for research involving humans. Government of Canada. 2018. URL: https://ethics.gc.ca/eng/policy-politique_tcps2-eptc2_2022.html [accessed 2025-04-07]
53. Khodyakov D, Hempel S, Rubenstein L, Shekelle P, Foy R, Salem-Schatz S, et al. Conducting online expert panels: a feasibility and experimental replicability study. BMC Med Res Methodol 2011;11:174 [FREE Full text] [doi: 10.1186/1471-2288-11-174] [Medline: 22196011]
54. Hsieh H, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
55. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
56. Hamid AYS, Chandra YA, Putri AF, Wakhid A, Falahaini A, Yulianingsih Y. Sustainable disaster risk reduction training model for nurses: a descriptive qualitative approach. Nurse Educ Pract 2023;69:103616. [doi: 10.1016/j.nepr.2023.103616] [Medline: 36996553]
57. Itchhaporia D. The evolution of the quintuple aim: health equity, health outcomes, and the economy. J Am Coll Cardiol 2021;78(22):2262-2264 [FREE Full text] [doi: 10.1016/j.jacc.2021.10.018] [Medline: 34823665]
58. Nundy S, Cooper LA, Mate KS. The quintuple aim for health care improvement: a new imperative to advance health equity. JAMA 2022;327(6):521-522. [doi: 10.1001/jama.2021.25181] [Medline: 35061006]
59. Chatterjee D, Corral J. How to write well-defined learning objectives. J Educ Perioper Med 2017;19(4):E610 [FREE Full text] [Medline: 29766034]
60. Badyal D, Singh T. Learning theories: the basics to learn in medical education. Int J Appl Basic Med Res 2017;7(Suppl 1):S1-S3 [FREE Full text] [doi: 10.4103/ijabmr.IJABMR_385_17] [Medline: 29344448]
61. Kirkpatrick D, Kirkpatrick J. Evaluating Training Programs: The Four Levels. Oakland, CA: Berrett-Koehler Publishers; 2006.
62. Stogiannos N, Gillan C, Precht H, Reis CSD, Kumar A, O'Regan T, et al. A multidisciplinary team and multiagency approach for AI implementation: a commentary for medical imaging and radiotherapy key stakeholders. J Med Imaging Radiat Sci 2024;55(4):101717. [doi: 10.1016/j.jmir.2024.101717] [Medline: 39067309]
63. Knopp MI, Warm EJ, Weber D, Kelleher M, Kinnear B, Schumacher DJ, et al. AI-enabled medical education: threads of change, promising futures, and risky realities across four potential future worlds. JMIR Med Educ 2023;9:e50373 [FREE Full text] [doi: 10.2196/50373] [Medline: 38145471]
64. Kueper JK, Emu M, Banbury M, Bjerre LM, Choudhury S, Green M, et al. Artificial intelligence for family medicine research in Canada: current state and future directions: report of the CFPC AI Working Group. Can Fam Physician 2024;70(3):161-168 [FREE Full text] [doi: 10.46747/cfp.7003161] [Medline: 38499374]
65. Inventory of Artificial Intelligence Resources in Family Medicine Education. Rahimi's Lab. URL: https://rahimislab.ca/inventory

66. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT—reshaping medical education and clinical management. Pak J Med Sci 2023;39(2):605-607 [FREE Full text] [doi: 10.12669/pjms.39.2.7653] [Medline: 36950398]
67. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023;11(6):887 [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
68. Ossa LA, Rost M, Lorenzini G, Shaw DM, Elger BS. A smarter perspective: learning with and from AI-cases. Artif Intell Med 2023;135:102458 [FREE Full text] [doi: 10.1016/j.artmed.2022.102458] [Medline: 36628794]
69. Sauder M, Tritsch T, Rajput V, Schwartz G, Shoja MM. Exploring generative artificial intelligence-assisted medical education: assessing case-based learning for medical students. Cureus 2024;16(1):e51961 [FREE Full text] [doi: 10.7759/cureus.51961] [Medline: 38333501]
70. Weidener L, Fischer M. Teaching AI ethics in medical education: a scoping review of current literature and practices. Perspect Med Educ 2023;12(1):399-410 [FREE Full text] [doi: 10.5334/pme.954] [Medline: 37868075]
71. Laupichler MC, Hadizadeh DR, Wintergerst MWM, von der Emde L, Paech D, Dick EA, et al. Effect of a flipped classroom course to foster medical students' AI literacy with a focus on medical imaging: a single group pre-and post-test study. BMC Med Educ 2022;22(1):803 [FREE Full text] [doi: 10.1186/s12909-022-03866-x] [Medline: 36397110]
72. Krive J, Isola M, Chang L, Patel T, Anderson M, Sreedhar R. Grounded in reality: artificial intelligence in medical education. JAMIA Open 2023;6(2):ooad037 [FREE Full text] [doi: 10.1093/jamiaopen/ooad037] [Medline: 37273962]
73. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183(6):589-596 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
74. Jain A, Way D, Gupta V, Gao Y, de Oliveira Marinho G, Hartford J, et al. Development and assessment of an artificial intelligence-based tool for skin condition diagnosis by primary care physicians and nurse practitioners in teledermatology practices. JAMA Netw Open 2021;4(4):e217249 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.7249] [Medline: 33909055]
75. Lee C, Britto S, Diwan K. Evaluating the impact of artificial intelligence (AI) on clinical documentation efficiency and accuracy across clinical settings: a scoping review. Cureus 2024;16(11):e73994. [doi: 10.7759/cureus.73994] [Medline: 39703286]
76. Ahmed MI, Spooner B, Isherwood J, Lane M, Orrock E, Dennison A. A systematic review of the barriers to the implementation of artificial intelligence in healthcare. Cureus 2023;15(10):e46454 [FREE Full text] [doi: 10.7759/cureus.46454] [Medline: 37927664]
77. Hassan M, Kushniruk A, Borycki E. Barriers to and facilitators of artificial intelligence adoption in health care: scoping review. JMIR Hum Factors 2024;11:e48633 [FREE Full text] [doi: 10.2196/48633] [Medline: 39207831]
78. Nair M, Svedberg P, Larsson I, Nygren JM. A comprehensive overview of barriers and strategies for AI implementation in healthcare: mixed-method design. PLoS One 2024;19(8):e0305949 [FREE Full text] [doi: 10.1371/journal.pone.0305949] [Medline: 39121051]
79. Singh RP, Hom GL, Abramoff MD, Campbell JP, Chiang MF, AAO Task Force on Artificial Intelligence. Current challenges and barriers to real-world artificial intelligence adoption for the healthcare system, provider, and the patient. Transl Vis Sci Technol 2020;9(2):45 [FREE Full text] [doi: 10.1167/tvst.9.2.45] [Medline: 32879755]
80. Terry AL, Kueper JK, Beleno R, Brown JB, Cejic S, Dang J, et al. Is primary health care ready for artificial intelligence? What do primary health care stakeholders say? BMC Med Inform Decis Mak 2022;22(1):237 [FREE Full text] [doi: 10.1186/s12911-022-01984-6] [Medline: 36085203]
81. Perkins M, Pregowska A. The role of artificial intelligence in higher medical education and the ethical challenges of its implementation. AIH 2025;2(1):1-13 [FREE Full text] [doi: 10.36922/aih.3276]
82. Jiang L, Wu Z, Xu X, Zhan Y, Jin X, Wang L, et al. Opportunities and challenges of artificial intelligence in the medical field: current application, emerging problems, and problem-solving strategies. J Int Med Res 2021;49(3):3000605211000157 [FREE Full text] [doi: 10.1177/03000605211000157] [Medline: 33771068]
83. Yin J, Ngiam KY, Teo HH. Role of artificial intelligence applications in real-life clinical practice: systematic review. J Med Internet Res 2021;23(4):e25759 [FREE Full text] [doi: 10.2196/25759] [Medline: 33885365]
84. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019;5(2):e16048 [FREE Full text] [doi: 10.2196/16048] [Medline: 31793895]
85. Grunhut J, Wyatt AT, Marques O. Educating future physicians in artificial intelligence (AI): an integrative review and proposed changes. J Med Educ Curric Dev 2021;8:23821205211036836 [FREE Full text] [doi: 10.1177/23821205211036836] [Medline: 34778562]
86. Lee J, Wu AS, Li D, Kulasegaram KM. Artificial intelligence in undergraduate medical education: a scoping review. Acad Med 2021;96(11S):S62-S70. [doi: 10.1097/ACM.0000000000004291] [Medline: 34348374]
87. Waljee AK, Higgins PDR. Machine learning in medicine: a primer for physicians. Am J Gastroenterol 2010;105(6):1224-1226. [doi: 10.1038/ajg.2010.173] [Medline: 20523307]

88. Blease C, Kaptchuk TJ, Bernstein MH, Mandl KD, Halamka JD, DesRoches CM. Artificial intelligence and the future of primary care: exploratory qualitative study of UK general practitioners' views. J Med Internet Res 2019;21(3):e12802 [FREE Full text] [doi: 10.2196/12802] [Medline: 30892270]
89. Irfan F. Artificial intelligence: help or hindrance for family physicians? Pak J Med Sci 2021;37(1):288-291 [FREE Full text] [doi: 10.12669/pjms.37.1.3351] [Medline: 33437293]
90. Mousavi M, Shafiee S, Harley JM, Cheung JCK, Abbasgholizadeh Rahimi S. Performance of generative pre-trained transformers (GPTs) in certification examination of the College of Family Physicians of Canada. Fam Med Com Health 2024 May 28;12(Suppl 1):e002626. [doi: 10.1136/fmch-2023-002626] [Medline: 38806403]

Abbreviations

AI: artificial intelligence
AIFM-ed: Artificial Intelligence Training in Postgraduate Family Medicine Education
CanMEDS: Canadian Medical Education Directives for Specialists
CFPC: College of Family Physicians of Canada
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
SRQR: Standards for Reporting Qualitative Research

Edited by B Lesselroth; submitted 24.09.24; peer-reviewed by K Thompson, G Evangelinos; comments to author 12.12.24; revised version received 04.02.25; accepted 25.02.25; published 25.04.25.

Please cite as:
Tolentino R, Hersson-Edery F, Yaffe M, Abbasgholizadeh-Rahimi S
AIFM-ed Curriculum Framework for Postgraduate Family Medicine Education on Artificial Intelligence: Mixed Methods Study
JMIR Med Educ 2025;11:e66828
URL: https://mededu.jmir.org/2025/1/e66828
doi:10.2196/66828
PMID:

Original Paper

Automated Evaluation of Reflection and Feedback Quality in Workplace-Based Assessments by Using Natural Language Processing: Cross-Sectional Competency-Based Medical Education Study

Jeng-Wen Chen[1,2,3,4*], MSc, MD; Hai-Lun Tu[5*], PhD; Chun-Hsiang Chang[1,2*], MSc, MD; Wei-Chung Hsu[2], MD, PhD; Pa-Chun Wang[6,7,8], MD, PhD; Chun-Hou Liao[9], MD, PhD; Mingchih Chen[3,10], PhD

[1]Department of Otolaryngology–Head and Neck Surgery, Cardinal Tien Hospital, Fu Jen Catholic University, New Taipei City, Taiwan
[2]Department of Otolaryngology–Head and Neck Surgery, National Taiwan University Hospital and Children's Hospital, Taipei, Taiwan
[3]Department of Hospital Management, Graduate Institute of Business Administration, Fu Jen Catholic University, New Taipei City, Taiwan
[4]Department of Education and Research, Cardinal Tien Junior College of Healthcare and Management, New Taipei City, Taiwan
[5]Department of Library and Information Science, Fu-Jen Catholic University, New Taipei City, Taiwan
[6]Cathay General Hospital, Department of Otolaryngology, Taipei, Taiwan
[7]School of Medicine, Fu-Jen Catholic University, New Taipei City, Taiwan
[8]Department of Medical Research, China Medical University Hospital, China Medical University, Taichung, Taiwan
[9]Department of Surgery, Division of Urology, Cardinal Tien Hospital and School of Medicine, Fu Jen Catholic University, New Taipei City, Taiwan
[10]Artificial Intelligence Development Center, Fu Jen Catholic University, New Taipei City, Taiwan
[*]these authors contributed equally

Corresponding Author:
Jeng-Wen Chen, MSc, MD
Department of Otolaryngology–Head and Neck Surgery
Cardinal Tien Hospital
Fu Jen Catholic University
362, ZhongZheng Rd
Xindian Dist
New Taipei City, 23148
Taiwan
Phone: 886 2 22193391 ext 67451
Fax: 886 2 22195821
Email: 086365@mail.fju.edu.tw

Abstract

Background: Competency-based medical education relies heavily on high-quality narrative reflections and feedback within workplace-based assessments. However, evaluating these narratives at scale remains a significant challenge.

Objective: This study aims to develop and apply natural language processing (NLP) models to evaluate the quality of resident reflections and faculty feedback documented in Entrustable Professional Activities (EPAs) on Taiwan's nationwide Emyway platform for otolaryngology residency training.

Methods: This 4-year cross-sectional study analyzes 300 randomly sampled EPA assessments from 2021 to 2025, covering a pilot year and 3 full implementation years. Two medical education experts independently rated the narratives based on relevance, specificity, and the presence of reflective or improvement-focused language. Narratives were categorized into 4 quality levels—effective, moderate, ineffective, or irrelevant—and then dichotomized into high quality and low quality. We compared the performance of logistic regression, support vector machine, and bidirectional encoder representations from transformers (BERT) models in classifying narrative quality. The best performing model was then applied to track quality trends over time.

Results: The BERT model, a multilingual pretrained language model, outperformed other approaches, achieving 85% and 92% accuracy in binary classification for resident reflections and faculty feedback, respectively. The accuracy for the 4-level classification

was 67% for both. Longitudinal analysis revealed significant increases in high-quality reflections (from 70.3% to 99.5%) and feedback (from 50.6% to 88.9%) over the study period.

Conclusions: BERT-based NLP demonstrated moderate-to-high accuracy in evaluating the narrative quality in EPA assessments, especially in the binary classification. While not a replacement for expert review, NLP models offer a valuable tool for monitoring narrative trends and enhancing formative feedback in competency-based medical education.

(JMIR Med Educ 2025;11:e81718) doi:10.2196/81718

KEYWORDS

competency-based medical education; entrustable professional activities; otolaryngology; residency; workplace-based assessment; reflection; feedback; Emyway platform

Introduction

Medical education has undergone a fundamental transformation, with competency-based medical education (CBME) emerging as a central paradigm [1]. In contrast to traditional time-based models that focus on the completion of predetermined curricula over fixed durations, CBME emphasizes the direct assessment of learner's abilities to perform core professional activities safely and effectively in authentic clinical environments [2,3]. This outcomes-oriented approach aims to ensure that physicians are not only knowledgeable but also clinically competent, adaptable, and equipped to address the evolving complexities of patient care [4-6].

The field of otorhinolaryngology–head and neck surgery underscores the urgency of this educational shift, given its demand for proficiency in complex surgical procedures and nuanced clinical decision-making [7,8]. In response, the Taiwan Society of Otorhinolaryngology–Head and Neck Surgery (TSO-HNS) launched a structured competency framework in 2020, introducing 11 Entrustable Professional Activities (EPAs) as benchmarks for assessing resident performance (TSO-HNS Entrustable Professional Activities Assessment Framework for Resident Physician Training, second edition; see Multimedia Appendix 1). To support the systematic implementation of these EPAs, the Emyway digital platform was adopted in 2021, enabling more structured, transparent, and objective competency evaluations [9]. Central to Emyway is the integration of workplace-based assessment (WBA), which promotes continuous learning through direct observation, self-reflection, formative feedback, and performance appraisal in real-world clinical settings [10,11]. Unlike traditional assessments, WBAs offer dynamic, individualized insights that inform both clinical decision-making and technical skill development [9].

A key challenge in CBME is bridging the gap between assessment and learning. Reflection and feedback play complementary roles in this process. When aligned, feedback shapes the focus of reflection, and reflection deepens engagement with feedback, turning assessments into learning opportunities. However, prior studies show that reflections often remain descriptive, and feedback lacks specificity, limiting their combined educational value [12,13]. Evaluating the quality of both processes is therefore essential to understanding how WBAs contribute to learning. A growing body of evidence underscores the role of high-quality reflections and feedback in reinforcing core competencies and enhancing learning outcomes [14,15]. However, the quality of these narrative components within WBAs—particularly in otolaryngology residency programs and in multilingual training environments—remains insufficiently studied.

A major challenge in the implementation of CBME is managing the substantial volume of narrative data generated through WBAs [11]. On digital platforms such as Emyway, thousands of EPA evaluations are recorded, rendering manual review impractical. Traditional assessment methods that rely on human interpretation are time-consuming, resource-intensive, and susceptible to variability, limiting their ability to yield consistent and meaningful insights from large datasets [16]. Overcoming this challenge requires innovative strategies to ensure that narrative reflections and feedback remain relevant, specific, and actionable—supporting continuous learning and improvement in residency training [17,18].

This study aims to address the challenge of evaluating narrative data in CBME by applying natural language processing (NLP) to systematically assess the quality of resident reflections and faculty feedback recorded within the Emyway platform. To capture these distinct but interrelated processes at scale, we applied NLP models to evaluate reflection and feedback separately, allowing for a clearer analysis of their respective contributions to CBME. We hypothesize that NLP can provide an objective, consistent, and scalable method for evaluating the effectiveness of narrative assessments, offering valuable insights into how feedback contributes to residents' competency development [16,19]. By leveraging NLP, this study seeks to improve the relevance, specificity, and actionability of reflections and feedback, thereby enhancing the guidance residents receive for their professional growth [19-22]. Resident reflections and faculty feedback are distinct constructs: reflections involve personal self-assessment, while feedback represents external evaluation from faculty. Although different, they occur simultaneously within the same WBA encounter. This study therefore examines both while ensuring that the NLP models and evaluation rubrics for reflections and feedback were developed and analyzed independently. Ultimately, this approach aims to bridge the gap between assessment and learning, strengthen CBME implementation, and support the development of a more robust otolaryngology residency training system.

Methods

Ethical Considerations

This study adheres to established ethical standards for medical education research. Informed consent was obtained actively. Participants were required to read the "Training-Related Data Collection and Privacy Information" and click an "I agree" button before accessing the Emyway platform. The participants did not receive any compensation for their participation. The system includes built-in data protection mechanisms to prevent confidential information from being displayed. All data were deidentified prior to analysis, with personal identifiers removed, and access was restricted to the research team through secure, password-protected servers. The study protocol was reviewed and approved by the institutional review board of Cardinal Tien Hospital (CTH-112-2-1-002).

Study Design and Setting

This cross-sectional study examines the quality of resident reflections and faculty feedback recorded in the Emyway platform of TSO-HNS between 2021 and 2025. Emyway is a nationwide digital platform designed to support CBME by systematically collecting workplace-based EPA assessments from otolaryngology residency programs across Taiwan [9]. Basic clinical information, encounter descriptions, resident reflections, and subsequent faculty feedback and ad hoc entrustment decisions were collected within a single standardized electronic form on the Emyway platform [9]. The primary objective of this study was to evaluate the narrative quality of resident reflections and faculty feedback by using NLP algorithms, with the goal of improving assessment reliability and enhancing the educational value of feedback in clinical training.

Data Collection and Sample Selection

We selected 300 EPA assessment entries from the Emyway national database, covering the period from 2021 to 2025. Each entry included structured fields such as the EPA title, clinical diagnosis, and narrative components authored by both residents and faculty [9]. To ensure diversity and representativeness, we employed stratified random sampling across training years, resident levels, and EPA categories. To reduce potential bias related to temporal improvements in narrative quality, we used cross-validation and ensured a balanced distribution of entries across earlier and later phases of implementation. Only complete assessments containing both resident reflections and faculty feedback were included in the final analysis.

Narrative Quality Assessment

Two medical education experts—one a physician-educator specializing in otolaryngology residency training and the other a senior faculty developer with expertise in educational measurement and feedback assessment—independently evaluated the quality of resident reflections and faculty feedback by using a structured rubric based on the core principles of CBME. Narratives were evaluated using established rubrics developed by Solano et al [17] and Ötleş et al [18], which have been previously validated in surgical residency programs and were adopted in our study without modification to ensure consistency with the existing literature. The rubric assesses 3 key dimensions: relevance, specificity, and either reflection content (for resident narratives) or actionability (for faculty feedback). Relevance evaluates the alignment of the narrative with the EPA and the clinical context. Specificity measures the clarity and detail with which strengths, weaknesses, or areas for improvement were identified. Reflection content assesses the presence of self-directed learning goals in resident narratives, while actionability examines whether faculty feedback provided clear, constructive guidance to support resident development. The analysis of interrater reliability showed a fair to moderate agreement in the 4-level classification and a substantial to almost perfect agreement in the 2-level classification (Table S1 in Multimedia Appendix 2). In cases where the 2 expert raters had discrepancies in their ratings, a third reviewer (the corresponding author) adjudicated and made the final decision to ensure consistency and accuracy in the gold standard dataset.

Based on the evaluation criteria, narratives were categorized into 4 quality levels (Table 1): effective, moderate, ineffective, and irrelevant. Effective narratives were both relevant and specific; resident reflections demonstrated meaningful insight, and faculty feedback included actionable guidance. Moderate narratives maintained relevance but demonstrated only one additional element—either specificity or reflection content for residents or actionability for faculty. Ineffective narratives were superficially related to the EPA but lacked depth, with vague language and an absence of both specificity and meaningful reflection or guidance. Irrelevant narratives were off-topic, superficial, or disconnected from the clinical context. In this study, "high quality" refers to the combined category in the 2-level classification (encompassing both effective and moderate narratives) and "low quality" refers to ineffective and irrelevant narratives, whereas "effective" denotes the highest category within the 4-level classification.

Table 1. Classification of the quality levels in residents' reflections and faculty feedback.

Characteristics according to the 4-level classification[a]	Quality of narrative content				
	Effective[b]	Moderate[b]	Moderate[b]	Ineffective[c]	Irrelevant[c]
Relevance	Yes	Yes	Yes	Yes	No
Specificity	Yes	Yes	No	No	N/A[d]
Reflection content in residents' reflections	Yes	No	Yes	No	N/A
Action plan in faculty feedback	Yes	No	Yes	No	N/A

[a]In the 4-level classification, the categories are effective (highest quality), moderate, ineffective, and irrelevant.

[b]The combined group of effective and moderate narratives was classified as high quality per the 2-level classification.

[c]Ineffective and irrelevant narratives were classified as low quality per the 2-level classification.

[d]N/A: not applicable.

NLP Framework

To enhance the scalability and objectivity of narrative assessment, NLP techniques were applied to analyze resident reflections and faculty feedback. Two independent NLP models were developed and trained separately for reflections and feedback, ensuring that the classification processes remained independent while allowing both dimensions to be examined within the same WBA encounter. Three supervised machine learning models were implemented for classification: logistic regression (LR) [23], support vector machine (SVM) [24], and bidirectional encoder representations from transformers (BERT) [25], which is a state-of-the-art deep learning model for natural language understanding.

Data Preprocessing and Feature Extraction

For traditional machine learning models such as LR and SVM, text preprocessing included tokenization using CKIPtagger for Chinese language segmentation, followed by transformation into term frequency–inverse document frequency feature vectors. In contrast, the BERT model processed raw text inputs directly, structured as a combination of context, EPA title, diagnosis, and either reflection or feedback. This approach leveraged BERT's ability to generate contextualized embeddings without requiring additional preprocessing.

Model Training and Evaluation

To evaluate model performance, the dataset was randomly divided into a training set (80%) and a validation set (20%). Both fine-grained (4-level) and binary (2-level) classification models were developed to assess the impact of classification granularity. LR and SVM models were implemented using the *scikit-learn* library, while the BERT model was fine-tuned using the *simpletransformers* library with the pretrained BERT-base-multilingual-uncased model. BERT was trained for 10 epochs with a learning rate of 2e-5. The code used for training all the models is provided in Multimedia Appendix 3.

Performance Metrics and Narrative Quality Trend Analysis

We evaluated model performance by using standard metrics, including accuracy, precision, recall, and F_1-score. We generated confusion matrices to visualize classification outcomes and identify patterns of misclassification. The analysis aimed to assess the accuracy of distinguishing high-quality and low-quality reflections and feedback, compare the performance across different machine learning models, and explore longitudinal trends in the narrative quality by using the best performing model throughout the study period from 2021 to 2025.

Results

Overall Model Performance

Across the study period, the majority of EPA assessments were complete, containing both resident reflections and faculty feedback. Specifically, 90.1% (1422/1580) were complete in the pilot year (2021-2022), 95.1% (9939/10,447) in 2022-2023, 96.7% (10,601/10,966) in 2023-2024, and 97.1% (12,139/12,497) in 2024-2025. In total, 34,101 out of 35,490 assessments (96.1%) were complete and included in the final analysis. Table 2 presents the expert-assessed quality distribution of 300 randomly selected EPA entries, comprising resident reflections and faculty feedback, used for developing and validating the NLP models.

Table 3 summarizes the prediction outcomes from the 3 models evaluated in the study. The NLP-based classification models demonstrated substantial accuracy in assessing the quality of both resident reflections and faculty feedback, with the BERT model consistently outperforming the LR and SVM models. Specifically, for resident reflections, the BERT model achieved an accuracy of 85% for the 2-level classification and 67% for the more granular 4-level classification. Performance was even stronger for faculty feedback evaluation, where the BERT model attained an accuracy of 92% in the 2-level classification and maintained a 67% accuracy for the 4-level classification. Additionally, precision, recall, and F_1-scores showed consistent patterns across these evaluations, supporting the robustness and reliability of the BERT model.

Table 2. Distribution of expert-assessed quality of 300 randomly selected Entrustable Professional Activity entries (resident reflections and faculty feedback) for natural language processing model development and validation.

Classification/quality rating	Resident reflections (n=300), n (%)	Faculty feedback (n=300), n (%)
4-level classification		
Effective	134 (44.7)	168 (56)
Moderate	86 (28.7)	28 (9.3)
Ineffective	49 (16.3)	24 (8)
Irrelevant	31 (10.3)	80 (26.7)
2-level classification		
High-quality	220 (73.3)	196 (65.3)
Low-quality	80 (26.7)	104 (34.7)

Table 3. Prediction results of the residents' reflections and faculty feedback by the 3 models in the study.

Narrative content, model	4-level classification				2-level classification			
	Accuracy (%)	Precision (%)	Recall (%)	F_1-score	Accuracy (%)	Precision (%)	Recall (%)	F_1-score
Resident reflections								
LR[a]	63	66	63	64	80	83	80	81
SVM[b]	60	63	60	60	85	85	85	85
BERT[c]	67	67	67	65	85	85	85	85
Faculty feedback								
LR	63	55	63	59	78	78	78	78
SVM	63	54	63	54	78	81	78	76
BERT	67	65	67	64	92	92	92	92

[a]LR: logistic regression.

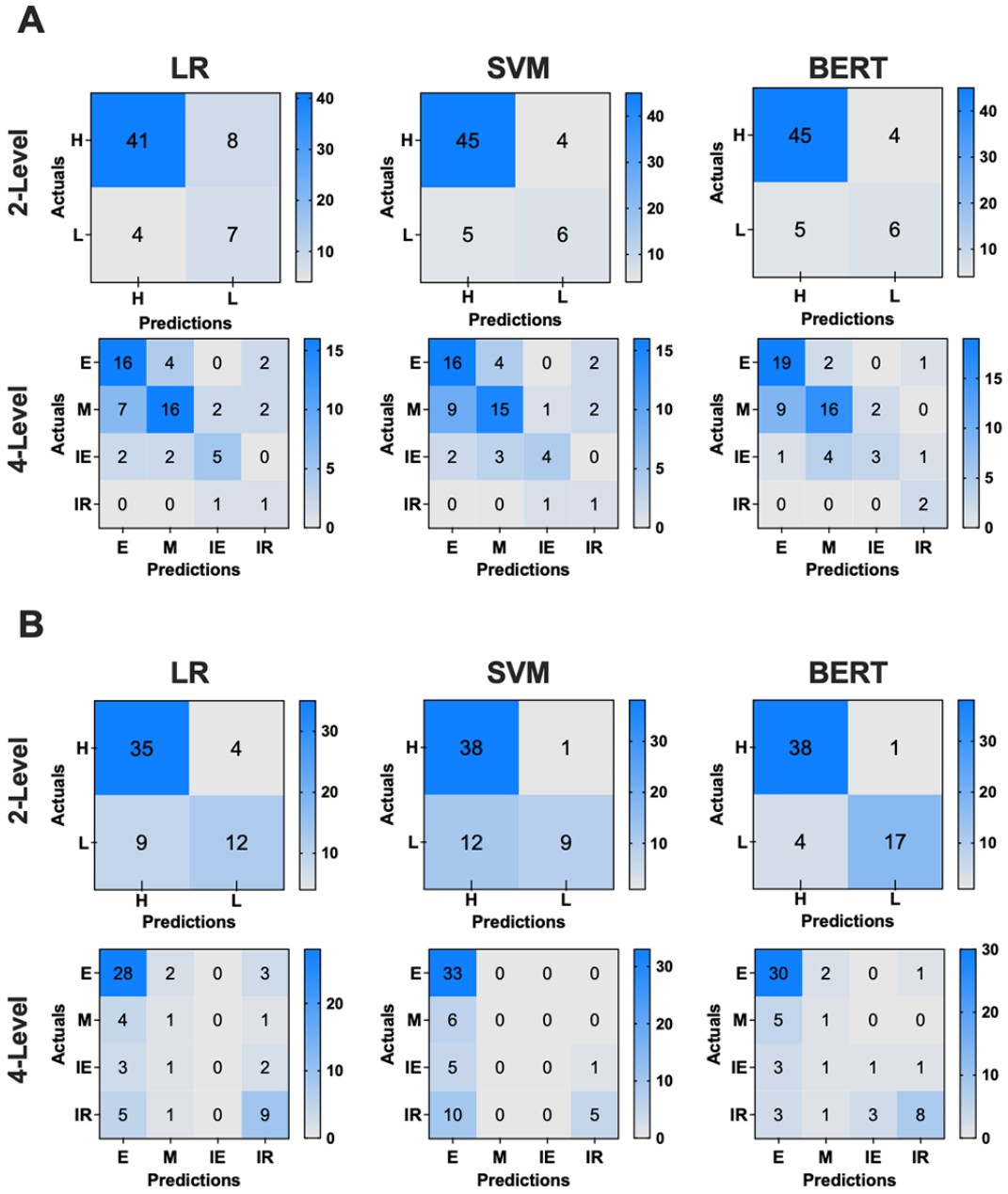
[b]SVM: support vector machine.

[c]BERT: bidirectional encoder representations from transformers.

Confusion Matrix Analysis

To further assess model performance, confusion matrices were generated (Figure 1). The BERT model exhibited fewer misclassifications than LR and SVM, particularly in distinguishing between effective and moderate narratives. In contrast, LR and SVM frequently misclassified effective narratives as moderate or irrelevant, reflecting their limitations in detecting subtle contextual cues. Notably, BERT's superior classification capability was most evident in faculty feedback, where its accuracy surpassed 90%, demonstrating its potential to improve automated assessment reliability in competency-based education frameworks. To illustrate the model's interpretability and limitations, Table S2 in Multimedia Appendix 4 presents anonymized examples of correctly classified and misclassified narratives.

Figure 1. Confusion matrices illustrating the classification performance of 3 natural language processing models—LR, SVM, and BERT—in evaluating the quality of resident reflections (A) and faculty feedback (B). The x-axis represents predicted categories, and the y-axis represents actual expert ratings. For the 2-level classification, narratives were categorized as high quality (H) or low quality (L). For the 4-level classification, the categories are effective (E), moderate (M), ineffective (IE), and irrelevant (IR). Numbers within each cell indicate the count of narratives, while shading intensity reflects frequency (darker=higher count). Compared with LR and SVM, BERT demonstrated fewer misclassifications and stronger performance in distinguishing between adjacent categories, particularly for faculty feedback. BERT: bidirectional encoder representations from transformers; LR: logistic regression; SVM: support vector machine.

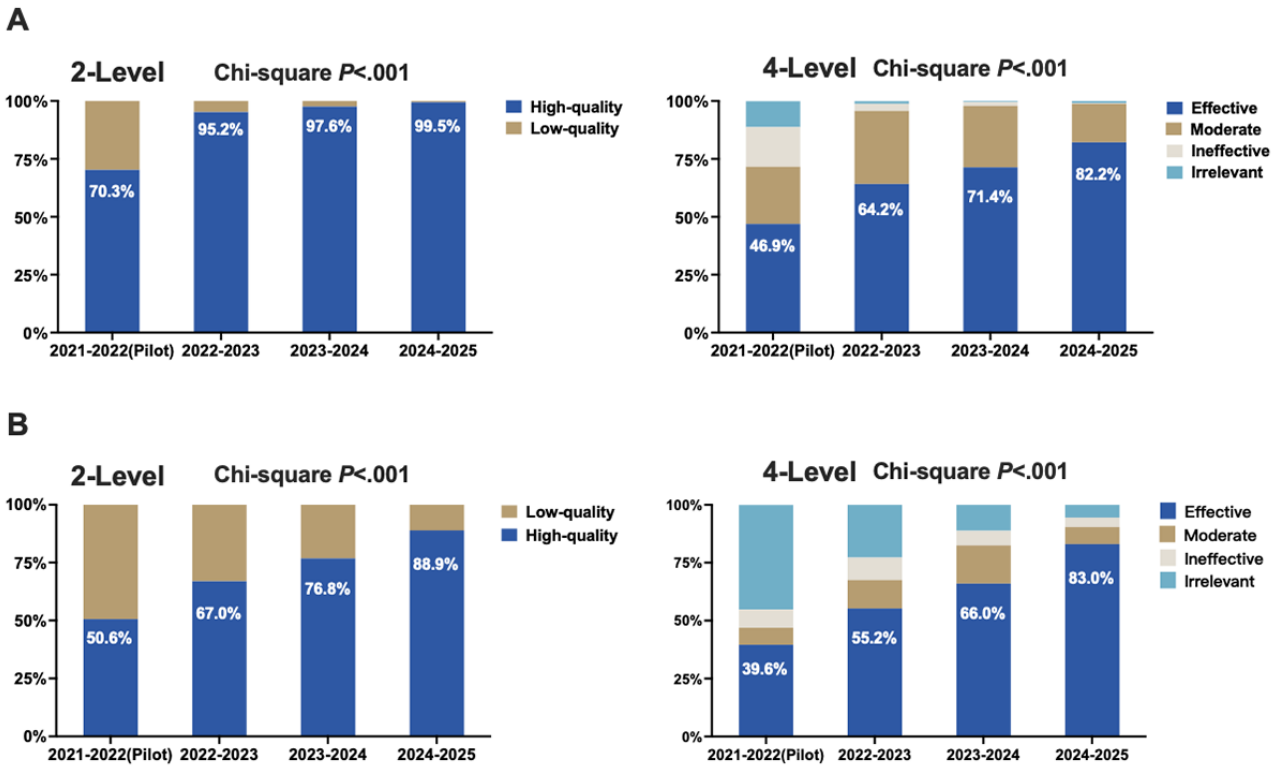


Two-Level and Four-Level Quality Classification Outcomes in the Emyway Platform

Figure 2 illustrates the longitudinal trends in the narrative quality of resident reflections and faculty feedback, as classified by the BERT model using both 2-level and 4-level rating algorithms, across 4 academic years: the pilot year (2021-2022) through 2024-2025. Detailed distributions of frequencies and percentages are presented in Table S3 of Multimedia Appendix 5.

In the 2-level classification, the proportion of high-quality resident reflections increased from 70.3% to 99.5%, while high-quality faculty feedback increased from 50.6% to 88.9% over the study period. Chi-square analyses confirmed that these improvements were statistically significant (P<.001 for both groups), reflecting meaningful enhancement in the quality of narrative documentation. Similarly, in the 4-level classification, the proportion of "effective" resident reflections increased from 46.9% to 82.2%, and "effective" faculty feedback increased from 39.6% to 83%. These gains were also statistically significant (P<.001), suggesting a sustained and substantive improvement in narrative quality over time, likely associated with the ongoing implementation of structured EPA frameworks and digital feedback systems.

Figure 2. Longitudinal trends in the quality of narrative assessments from 2021 to 2025, as classified by the bidirectional encoder representations from transformers model. Panel A displays resident reflections; panel B displays faculty feedback. In each panel, the left graph shows the 2-level classification (high quality vs low quality), and the right graph shows the 4-level classification (effective, moderate, ineffective, irrelevant). The x-axis represents academic years, with 2021-2022 as the pilot year, followed by 3 full implementation years. The y-axis indicates the percentage distribution of the narratives. Over time, both resident reflections and faculty feedback showed a significant increase in the proportion of high-quality and effective narratives.



Discussion

Principal Findings

This study demonstrates the utility of NLP, specifically the BERT algorithm, in evaluating the narrative quality within WBAs in otolaryngology residency training. The BERT model achieved high accuracy in the binary classification—85% for resident reflections and 92% for faculty feedback—supporting its potential as a scalable, objective adjunct to manual evaluation. Notably, narrative quality improved significantly over the study period, with high-quality reflections increasing from 70.3% to 99.5% and high-quality faculty feedback from 50.6% to 88.9%. These findings highlight the potential of NLP to enhance quality assurance and longitudinal monitoring in CBME.

Compared to traditional manual qualitative analysis, NLP offers unique advantages [26]. Although human raters can capture contextual nuance and interpret implicit meaning, their assessments are time-intensive and subject to interrater variability. In contrast, NLP enables consistent, rapid, and scalable evaluation across large datasets [27,28]. Prior research by Akbasli et al [29] has demonstrated the feasibility of applying fine-tuned language models to non-English and multilingual medical texts. Our findings further support this approach, showing that integrating structured contextual inputs such as EPA titles, clinical diagnoses, and narrative components substantially enhance model accuracy. With adequate structured contextual inputs, BERT approximates human interpretive depth while retaining the efficiency and objectivity of automation.

This approach should also be interpreted through the lens of the educational assessment theory. Beyond its statistical performance, the application of NLP algorithms in this study aligns closely with established educational assessment theories and feedback quality frameworks. The structured rubric used to generate the gold standard—encompassing relevance, specificity, and either having reflection content or actionability—reflects the core principles found in frameworks such as the Feedback Quality Instrument [30-32] and the R2C2 model (relationship building, exploring reactions, exploring content, coaching for change) [14,33,34]. These frameworks emphasize that effective feedback and reflection must be contextually relevant, sufficiently specific, and actionable to promote self-regulated learning and professional growth. By incorporating these dimensions into the training data, BERT's decision-making process operationalizes these theoretical constructs, mapping narrative text to empirically validated quality indicators. In this way, the model does not merely classify text based on linguistic patterns but also embeds the pedagogical priorities of CBME and EPA assessment. This alignment ensures that automated scoring supports the same developmental goals as expert human raters, enabling the model to serve as a theoretically grounded, scalable complement to manual evaluation.

However, it is important to clarify that the R2C2 model is a coaching framework designed to structure feedback

conversations rather than an evaluation rubric for written comments. In this study, R2C2 was referenced as a conceptual lens to underscore the coaching potential embedded in high-quality narrative feedback and not as a scoring tool. Recent literature has emphasized its role in facilitating meaningful faculty–learner interactions in WBAs [35,36]. Our findings on the quality of written reflections and feedback should therefore be viewed as complementary to, rather than substitutive of, coaching frameworks such as R2C2, providing a stronger foundation for effective feedback dialogue.

In addition to methodological contributions, our findings suggest practical applications for residency programs. NLP outputs could be integrated into dashboards that track reflection and feedback quality over time, enabling program directors to identify gaps and design targeted faculty development workshops. At the same time, residents could receive timely, formative, reflective prompts into the quality of their reflections. By embedding these tools into CBME frameworks, narrative data can serve not only as an assessment record but also as a resource to strengthen feedback culture and support continuous coaching.

Comparison With Previous Studies

The superior performance of BERT relative to traditional machine learning models such as LR and SVM is a key contribution of this study. For instance, previous work by Ötleş et al [18] reported a mean accuracy of 0.64 by using SVM for the 4-level classification of surgical feedback, which improved to 0.83 when simplified to binary classification. Similarly, Solano et al [17] achieved an overall accuracy of 0.83 by using NLP but noted limitations in sensitivity (0.37), suggesting challenges in detecting lower quality feedback. In contrast, our BERT-based model achieved 85% accuracy for resident reflections and 92% for faculty feedback in binary classification, with balanced precision and recall scores. These results highlight BERT's superior ability to contextualize text and detect nuanced linguistic patterns. Unlike traditional models, BERT effectively interprets the complex, often implicit nature of reflective narratives, validating its use in educational quality assessment within clinical training contexts [37]. This capacity is particularly valuable, as reflective writing in medical education is typically layered, context-sensitive, and difficult to assess using rule-based or shallow models [38,39].

Although the 4-level classification achieved only moderate accuracy, its outputs can still inform educational practice. Even without perfect distinction between adjacent categories, the model can highlight patterns of lower quality narratives that may warrant attention. For instance, faculty development dashboards could flag programs or individuals generating a higher proportion of ineffective or moderate entries, prompting targeted coaching or workshops. These applications position the model as a supportive tool for monitoring and guiding feedback culture, complementing human judgment rather than replacing it.

Unlike prior studies that emphasized cross-sectional performance [17,18], this research provides longitudinal evidence of NLP's ability to track and support improvements in feedback quality over time. Consistent with earlier findings, the model maintained high specificity, particularly in identifying low-quality narratives—a valuable feature for faculty development and system-level monitoring. Although the 4-level classification performance remained moderate (67% accuracy), this aligns with known challenges in distinguishing subtle qualitative gradations and highlights areas for future enhancement.

The sustained improvement in the reflection quality across the study period underscores the value of structured WBA systems such as those implemented through the Emyway platform. These systems provide clear expectations and guidance, promoting deeper engagement, self-awareness, and professional development [40]. This observation aligns with literature indicating that structured reflection fosters clinical reasoning, self-regulated learning, and long-term growth [41-44].

Faculty feedback quality also improved substantially, increasing in specificity, relevance, and actionability. While still trailing resident reflections in overall quality, the upward trajectory from 50.6% to 88.9% suggests growing familiarity with EPA-based frameworks and greater faculty engagement. These findings reinforce the importance of structured systems in supporting effective feedback practices. NLP tools, in this context, can function as educational dashboards—tracking feedback quality across programs and timeframes, flagging low-quality entries, and informing faculty development and institutional policy.

It is important to note that reflection quality and feedback quality were not conflated in this study; rather, they were modeled separately using independent rubrics and NLP training processes. Presenting them together highlights how these complementary elements of the same assessment encounter can be studied in parallel to inform faculty development and resident learning.

We selected BERT over commercial large language models such as ChatGPT for both practical and performance-based reasons. As an open-source model, BERT is accessible to academic institutions without licensing constraints, facilitating integration into resource-limited settings. Moreover, internal comparisons indicated that ChatGPT, while powerful, lacked discriminative precision in this context and frequently defaulted to mid-range classifications (Multimedia Appendix 6). In contrast, BERT demonstrated greater reliability and accuracy, particularly when provided with structured contextual information.

Generalizability

Although our findings highlight the utility of BERT-based NLP within Taiwan's structured otolaryngology training system, their generalizability to other specialties, languages, and international contexts remains uncertain. Narrative style, cultural norms, and feedback practices vary widely across training environments, potentially affecting model performance. To ensure validity in non-Chinese language settings, rubric recalibration would be needed to align evaluation criteria with local educational practices and expectations. Furthermore, although multilingual pretrained models such as BERT provide a strong foundation, language-specific fine-tuning with locally

generated narrative data would be required to capture semantic nuances and ensure accurate classification. These adaptations highlight the importance of international replication and validation, which will be essential to confirm generalizability and extend the impact of NLP-assisted evaluation across medical specialties and cultural contexts.

The use of open-source NLP tools such as BERT also carries important ethical and practical implications. Although these models provide scalability, accessibility, and adaptability for educational use, they raise concerns about confidentiality, data security, and potential bias. To ensure responsible application, future implementation should include secure data management, careful local fine-tuning, and ongoing evaluation of fairness so that such tools enhance rather than compromise educational integrity.

Limitations

Despite encouraging results in binary classification, several limitations should be noted. First, the model's 67% accuracy in the 4-level classification reflects the inherent difficulty of distinguishing subtle qualitative differences in narrative assessments. Overlap in language used across adjacent categories—such as moderate and ineffective—poses challenges for both human raters and machine learning models. This limitation is common in educational NLP research and underscores the need for larger, more diverse training datasets, domain-specific model fine-tuning, and potentially incorporating contextual metadata (eg, resident level or case type). Although model performance stabilized during cross-validation, suggesting that the sample was adequate for the study objectives, larger datasets could further strengthen robustness. Moreover, the limited sample size may have contributed to weaker performance in the 4-level classification. Future strategies to address this limitation include expanding the dataset as the Emyway platform accumulates more entries, exploring data augmentation and domain-adaptive pretraining, and pursuing cross-institutional collaborations to increase sample diversity. These steps would strengthen model robustness and improve its ability to support nuanced educational decision-making. Although 4-level predictions should be interpreted with caution, they can still offer valuable insights for faculty development and formative assessment when combined with human judgment.

Second, as with all text-based evaluations, important nonverbal cues and dynamic interpersonal interactions are not captured. Future work could extend beyond text-based analysis by integrating audio and video data with NLP. Multimodal inputs would capture tone, pacing, and nonverbal cues, complementing narrative content and offering a more holistic view of feedback interactions. This approach could strengthen competency-based medical education by providing richer insights to guide faculty development and resident learning.

Third, although improvements were observed in the narrative quality, this study did not directly measure faculty engagement or sustained educational change. Future research should examine how NLP-generated insights might be incorporated into faculty development initiatives and longitudinal assessment strategies to determine whether they enhance faculty participation and support lasting improvements in feedback and reflection quality.

Finally, the possibility of a Hawthorne effect should be considered. The awareness of being evaluated may have influenced improvements in reflection and feedback quality [45,46]. Complementary qualitative research such as interviews or focus groups with residents and faculty could elucidate underlying motivations and perceptions, providing a richer perspective on behavioral change.

Conclusions

This study demonstrates that BERT-based NLP, when applied with structured contextual inputs, can effectively evaluate the quality of multilingual resident reflections and faculty feedback in WBAs. The model achieved moderate to high accuracy, particularly in binary classification, suggesting its utility as a scalable adjunct to human evaluation. While not a substitute for expert judgment, NLP can facilitate large-scale monitoring of narrative quality and enhance the analysis of formative feedback in CBME. The progressive improvement in the narrative quality over 4 years highlights the value of structured EPA frameworks and digital platforms such as Emyway in promoting reflective practice and faculty development. Future research should explore the generalizability of this approach across medical specialties and investigate the integration of multimodal data to further enhance assessment validity and educational outcomes.

Acknowledgments

The authors are grateful to Taiwan Society of Otorhinolaryngology-Head and Neck Surgery and all its faculties and resident physicians for utilizing the Joint Commission of Taiwan's Emyway platform. The authors also thank the information technology team of Dalin Tzu Chi Hospital for their support with the platform. Additionally, the authors are grateful for the administrative assistance provided by Chiu-Ping Wang, Shu-Hwei Fan, Uan-Shr Jan, and Wan-Ning Luo in this project. They received no additional compensation for their contributions. This study was supported by the National Science and Technology Council of the Republic of China (Taiwan) under grants NSTC 109-2511-H-567-001-MY2, NSTC 110-2511-H-567-001-MY2, NSTC 112-2410-H-567-001-MY3, and in part, funded by Cardinal Tien Hospital under grants CTH110AK-2220 and CTH111AK-2221. The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Data Availability

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

Authors' Contributions

Conceptualization: J-WC, C-HC, W-CH, P-WC
Data curation: J-WC, H-LT, C-HC
Methodology/formal analysis/validation: J-WC, H-LT, W-CH, P-CW
Project administration: W-CH, C-HL, MC, P-CW
Funding acquisition: J-WC, C-HC
Visualization: C-HL, MC, J-WC
Writing – original draft: J-WC, H-LT, C-HC
Writing – review & editing: J-WC, H-LT, C-HC, W-CH, P-CW, C-HL, and MC

Conflicts of Interest

None declared.

Multimedia Appendix 1
Taiwan Society of Otorhinolaryngology–Head and Neck Surgery Entrustable Professional Activities Assessment Framework for Resident Physician Training, second edition.
[PDF File (Adobe PDF File), 795 KB - mededu_v11i1e81718_app1.pdf]

Multimedia Appendix 2
Quantified agreement results (interrater reliability) for expert scoring.
[PDF File (Adobe PDF File), 54 KB - mededu_v11i1e81718_app2.pdf]

Multimedia Appendix 3
Logistic regression, support vector machine, and bidirectional encoder representations from transformers codes in the Google Colaboratory.
[PDF File (Adobe PDF File), 1265 KB - mededu_v11i1e81718_app3.pdf]

Multimedia Appendix 4
Sample outputs from the bidirectional encoder representations from transformers model for classifying narrative quality in resident reflections and faculty feedback.
[PDF File (Adobe PDF File), 177 KB - mededu_v11i1e81718_app4.pdf]

Multimedia Appendix 5
Distribution of numbers (percentages) of 4-level and 2-level quality ratings for resident reflections and faculty feedback across pilot year (2021-2022), 2022-2023, 2023-2024, and 2024-2025.
[PDF File (Adobe PDF File), 61 KB - mededu_v11i1e81718_app5.pdf]

Multimedia Appendix 6
Detailed process and results for evaluating resident reflections and faculty feedback quality by using ChatGPT-4o.
[PDF File (Adobe PDF File), 273 KB - mededu_v11i1e81718_app6.pdf]

References

1. Chen JX, Yu SE, Miller LE, Gray ST. A needs assessment for the future of otolaryngology education. Otolaryngol Head Neck Surg 2023 Jul;169(1):192-193. [doi: 10.1177/01945998221128292] [Medline: 36125895]
2. Kovatch KJ, Prince MEP, Sandhu G. Weighing entrustment decisions with patient care during residency training. Otolaryngol Head Neck Surg 2018 Jun;158(6):1024-1027 [FREE Full text] [doi: 10.1177/0194599818764652] [Medline: 29558240]
3. Lucey CR, Thibault GE, ten Cate O. Competency-based, time-variable education in the health professions. Academic Medicine 2018;93(3S):S1-S5. [doi: 10.1097/acm.0000000000002080]
4. Wagner N, Fahim C, Dunn K, Reid D, Sonnadara RR. Otolaryngology residency education: a scoping review on the shift towards competency-based medical education. Clin Otolaryngol 2017 Jun;42(3):564-572. [doi: 10.1111/coa.12772] [Medline: 27754613]

5. Chiang Y, Yu H, Chung H, Chen J. Implementing an entrustable professional activities programmatic assessments for nurse practitioner training in emergency care: a pilot study. Nurse Educ Today 2022 Aug;115:105409. [doi: 10.1016/j.nedt.2022.105409] [Medline: 35636245]
6. Fu C, Huang C, Yang Y, Liao W, Huang S, Chang W, et al. Developing an entrustable professional activity for providing health education and consultation in occupational therapy and examining its validity. BMC Med Educ 2024 Jun 28;24(1):705 [FREE Full text] [doi: 10.1186/s12909-024-05670-1] [Medline: 38943116]
7. Huynh PP, Malkin BD, Wang KH. Otolaryngology resident education: beyond procedural case logs-a 10-year single institutional review. Otolaryngol Head Neck Surg 2025 Mar;172(3):1077-1084. [doi: 10.1002/ohn.1082] [Medline: 39756016]
8. Singer MC. The future of otolaryngology training threatened: the negative impact of residency training reforms. Otolaryngol Head Neck Surg 2010 Mar;142(3):303-305. [doi: 10.1016/j.otohns.2009.12.010] [Medline: 20172370]
9. Guo F, Chen Y, Hsu W, Wang P, Chen M, Chen J. EMYWAY workplace-based entrustable professional activities assessments in otolaryngology residency training: a nationwide experience. Otolaryngol Head Neck Surg 2025 Apr;172(4):1242-1253. [doi: 10.1002/ohn.1104] [Medline: 39739526]
10. Norcini J, Burch V. Workplace-based assessment as an educational tool: AMEE Guide No. 31. Med Teach 2007 Nov;29(9):855-871. [doi: 10.1080/01421590701775453] [Medline: 18158655]
11. Ahle SL, Eskender M, Schuller M, Carnes E, Chen X, Koehler J, et al. The quality of operative performance narrative feedback: a retrospective data comparison between end of rotation evaluations and workplace-based assessments. Ann Surg 2022 Mar 01;275(3):617-620. [doi: 10.1097/SLA.0000000000003907] [Medline: 32511125]
12. Archer JC. State of the science in health professional education: effective feedback. Med Educ 2010 Jan;44(1):101-108. [doi: 10.1111/j.1365-2923.2009.03546.x] [Medline: 20078761]
13. Watling CJ, Ginsburg S. Assessment, feedback and the alchemy of learning. Med Educ 2019 Jan;53(1):76-85. [doi: 10.1111/medu.13645] [Medline: 30073692]
14. Faucett EA, McCrary HC, Barry JY, Saleh AA, Erman AB, Ishman SL. High-quality feedback regarding professionalism and communication skills in otolaryngology resident education. Otolaryngol Head Neck Surg 2018 Jan;158(1):36-42. [doi: 10.1177/0194599817737758] [Medline: 29065274]
15. Fernandes RD, de Vries I, McEwen L, Mann S, Phillips T, Zevin B. Evaluating the quality of narrative feedback for entrustable professional activities in a surgery residency program. Ann Surg 2024 Dec 01;280(6):916-924. [doi: 10.1097/SLA.0000000000006308] [Medline: 38660808]
16. Spadafore M, Yilmaz Y, Rally V, Chan TM, Russell M, Thoma B, et al. Using natural language processing to evaluate the quality of supervisor narrative comments in competency-based medical education. Acad Med 2024 May 01;99(5):534-540. [doi: 10.1097/ACM.0000000000005634] [Medline: 38232079]
17. Solano QP, Hayward L, Chopra Z, Quanstrom K, Kendrick D, Abbott KL, et al. Natural language processing and assessment of resident feedback quality. J Surg Educ 2021;78(6):e72-e77. [doi: 10.1016/j.jsurg.2021.05.012] [Medline: 34167908]
18. Ötleş E, Kendrick DE, Solano QP, Schuller M, Ahle SL, Eskender MH, et al. Using natural language processing to automatically assess feedback quality: findings from 3 surgical residencies. Acad Med 2021 Oct 01;96(10):1457-1460. [doi: 10.1097/ACM.0000000000004153] [Medline: 33951682]
19. Burke HB, Hoang A, Lopreiato JO, King H, Hemmer P, Montgomery M, et al. Assessing the ability of a large language model to score free-text medical student clinical notes: quantitative study. JMIR Med Educ 2024 Jul 25;10:e56342 [FREE Full text] [doi: 10.2196/56342] [Medline: 39118469]
20. Van Ostaeyen S, De Langhe L, De Clercq O, Embo M, Schellens T, Valcke M. Automating the identification of feedback quality criteria and the CanMEDS roles in written feedback comments using natural language processing. Perspect Med Educ 2023;12(1):540-549 [FREE Full text] [doi: 10.5334/pme.1056] [Medline: 38144670]
21. Dine CJ, Shea JA, Clancy CB, Heath JK, Pluta W, Kogan JR. Finding the needle in the haystack: can natural language processing of students' evaluations of teachers identify teaching concerns? J Gen Intern Med 2025 Jan;40(1):119-123. [doi: 10.1007/s11606-024-08990-6] [Medline: 39167336]
22. Le KDR, Tay SBP, Choy KT, Verjans J, Sasanelli N, Kong JCH. Applications of natural language processing tools in the surgical journey. Front Surg 2024;11:1403540 [FREE Full text] [doi: 10.3389/fsurg.2024.1403540] [Medline: 38826809]
23. Hosmer JD, Lemeshow S, Sturdivant R. Applied Logistic Regression. Hoboken, New Jersey: John Wiley & Sons; 2013.
24. Hearst M, Dumais S, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intell Syst Their Appl 1998 Jul 10;13(4):18-28. [doi: 10.1109/5254.708428]
25. Devin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 2-7; Minneapolis, Minnesota p. 4171-4186.
26. Deiner MS, Honcharov V, Li J, Mackey TK, Porco TC, Sarkar U. Large language models can enable inductive thematic analysis of a social media corpus in a single prompt: human validation study. JMIR Infodemiology 2024 Aug 29;4:e59641 [FREE Full text] [doi: 10.2196/59641] [Medline: 39207842]
27. Jacennik B, Zawadzka-Gosk E, Moreira JP, Glinkowski WM. Evaluating patients' experiences with healthcare services: extracting domain and language-specific information from free-text narratives. Int J Environ Res Public Health 2022 Aug 17;19(16):10182 [FREE Full text] [doi: 10.3390/ijerph191610182] [Medline: 36011816]

28. Khanbhai M, Anyadi P, Symons J, Flott K, Darzi A, Mayer E. Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. BMJ Health Care Inform 2021 Mar;28(1):e100262 [FREE Full text] [doi: 10.1136/bmjhci-2020-100262] [Medline: 33653690]
29. Akbasli IT, Birbilen AZ, Teksam O. Leveraging large language models to mimic domain expert labeling in unstructured text-based electronic healthcare records in non-english languages. BMC Med Inform Decis Mak 2025 Mar 31;25(1):154 [FREE Full text] [doi: 10.1186/s12911-025-02871-6] [Medline: 40165165]
30. Amirzadeh S, Rasouli D, Dargahi H. Assessment of validity and reliability of the feedback quality instrument. BMC Res Notes 2024 Aug 16;17(1):227 [FREE Full text] [doi: 10.1186/s13104-024-06881-x] [Medline: 39152449]
31. Johnson CE, Keating JL, Leech M, Congdon P, Kent F, Farlie MK, et al. Development of the Feedback Quality Instrument: a guide for health professional educators in fostering learner-centred discussions. BMC Med Educ 2021 Jul 12;21(1):382 [FREE Full text] [doi: 10.1186/s12909-021-02722-8] [Medline: 34253221]
32. Bok HGJ, Teunissen PW, Favier RP, Rietbroek NJ, Theyse LFH, Brommer H, et al. Programmatic assessment of competency-based workplace learning: when theory meets practice. BMC Med Educ 2013 Sep 11;13:123 [FREE Full text] [doi: 10.1186/1472-6920-13-123] [Medline: 24020944]
33. Sargeant J, Lockyer JM, Mann K, Armson H, Warren A, Zetkulic M, et al. The R2C2 model in residency education. Academic Medicine 2018;93(7):1055-1063. [doi: 10.1097/acm.0000000000002131]
34. Sargeant J, Lockyer J, Mann K, Holmboe E, Silver I, Armson H, et al. Facilitated reflective performance feedback. Academic Medicine 2015;90(12):1698-1706. [doi: 10.1097/acm.0000000000000809]
35. Patocka C, Cooke L, Ma IWY, Ellaway RH. Untangling feedback: mapping the patterns behind the practice. Med Educ. Online ahead of print 2025 Apr 07. [doi: 10.1111/medu.15706] [Medline: 40194907]
36. Ramani S, Armson H, Hanmore T, Lee-Krueger R, Könings KD, Roze des Ordons A, et al. Could the R2C2 feedback and coaching model enhance feedback literacy behaviors: a qualitative study exploring learner-preceptor feedback conversations. Perspect Med Educ 2025;14(1):9-19 [FREE Full text] [doi: 10.5334/pme.1368] [Medline: 39831131]
37. Babu A, Boddu SB. BERT-based medical chatbot: enhancing healthcare communication through natural language understanding. Explor Res Clin Soc Pharm 2024 Mar;13:100419 [FREE Full text] [doi: 10.1016/j.rcsop.2024.100419] [Medline: 38495953]
38. Preiksaitis C, Ashenburg N, Bunney G, Chu A, Kabeer R, Riley F, et al. The role of large language models in transforming emergency medicine: scoping review. JMIR Med Inform 2024 May 10;12:e53787 [FREE Full text] [doi: 10.2196/53787] [Medline: 38728687]
39. Zhang K, Meng X, Yan X, Ji J, Liu J, Xu H, et al. Revolutionizing health care: the transformative impact of large language models in medicine. J Med Internet Res 2025 Jan 07;27:e59069 [FREE Full text] [doi: 10.2196/59069] [Medline: 39773666]
40. Ginsburg S, Stroud L, Brydges R, Melvin L, Hatala R. Dual purposes by design: exploring alignment between residents' and academic advisors' documents in a longitudinal program. Adv Health Sci Educ Theory Pract 2024 Nov;29(5):1631-1647. [doi: 10.1007/s10459-024-10318-2] [Medline: 38438699]
41. Cheung WJ, Bhanji F, Gofton W, Hall AK, Karpinski J, Richardson D, et al. Design and implementation of a national program of assessment model - integrating entrustable professional activity assessments in Canadian specialist postgraduate medical education. Perspect Med Educ 2024;13(1):44-55 [FREE Full text] [doi: 10.5334/pme.956] [Medline: 38343554]
42. Khan SB, Maart R. Clinical assessment strategies for competency-based education in prosthetic dentistry. J Dent Educ 2025 Mar;89(3):375-382. [doi: 10.1002/jdd.13746] [Medline: 39436275]
43. Chan TM, Dowling S, Tastad K, Chin A, Thoma B. Integrating training, practice, and reflection within a new model for Canadian medical licensure: a concept paper prepared for the Medical Council of Canada. Can Med Educ J 2022 Aug;13(4):68-81 [FREE Full text] [doi: 10.36834/cmej.73717] [Medline: 36091730]
44. Rogers SL, Priddis LE, Michels N, Tieman M, Van Winkle LJ. Applications of the reflective practice questionnaire in medical education. BMC Med Educ 2019 Feb 07;19(1):47 [FREE Full text] [doi: 10.1186/s12909-019-1481-6] [Medline: 30732611]
45. Sedgwick P, Greenwood N. Understanding the Hawthorne effect. BMJ 2015 Sep 04;351:h4672 [FREE Full text] [doi: 10.1136/bmj.h4672] [Medline: 26341898]
46. Demetriou C, Hu L, Smith TO, Hing CB. Hawthorne effect on surgical studies. ANZ J Surg 2019 Dec;89(12):1567-1576. [doi: 10.1111/ans.15475] [Medline: 31621178]

Abbreviations

BERT: bidirectional encoder representations from transformers
CBME: competency-based medical education
EPA: Entrustable Professional Activity
LR: logistic regression
NLP: natural language processing
SVM: support vector machine
TSO-HNS: Taiwan Society of Otorhinolaryngology–Head and Neck Surgery

WBA: workplace-based assessment

Edited by J Eriksen; submitted 01.08.25; peer-reviewed by S Valanci, CT Hsiao, LA Lee, T Hanmore; comments to author 25.08.25; revised version received 13.09.25; accepted 01.10.25; published 22.10.25.

Please cite as:
Chen JW, Tu HL, Chang CH, Hsu WC, Wang PC, Liao CH, Chen M
Automated Evaluation of Reflection and Feedback Quality in Workplace-Based Assessments by Using Natural Language Processing: Cross-Sectional Competency-Based Medical Education Study
JMIR Med Educ 2025;11:e81718
URL: https://mededu.jmir.org/2025/1/e81718
doi:10.2196/81718
PMID:

Original Paper

How AI Is Transforming Medical Education: Bibliometric Analysis

Youyang Wang[1*], MD; Chuheng Chang[2*], MD; Wen Shi[3], MD; Huiting Liu[4], MD; Xiaoming Huang[1], MD; Yang Jiao[1], MD, MPH

[1]Department of General Practice (General Internal Medicine), Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China

[2]Department of Radiation Oncology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[3]Department of Gastroenterology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China

[4]Department of Infectious Diseases, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Yang Jiao, MD, MPH
Department of General Practice (General Internal Medicine)
Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College
No.1 Shuaifuyuan, Dongcheng District
Beijing, 100730
China
Phone: 1 10 69155645
Email: peterpumch@163.com

Abstract

Background: Artificial intelligence (AI) is increasingly being integrated into medical education. As AI technologies continue to evolve, they are expected to enable more sophisticated student tutoring, performance evaluation, and reforms of curricula. However, medical education entities have been ill-prepared to embrace this technological revolution, and there is anxiety concerning its potential harm to the community.

Objective: To explore research trends in the field and identify future directions for AI-enabled medical education, we conducted a systematic bibliometric analysis focusing on temporal trajectories in the field.

Methods: Documents were collected from the Web of Science and Scopus databases covering the period from 2000 to 2024. A multistep search strategy combining information retrieval, a definitive journal list, and cocitation analysis was used to identify relevant publications. Journal and author impact were assessed using both publication and citation metrics. Research trends and hot spots were examined through citation burst detection, frequency analysis, and co-occurrence networks, with a color gradient used to indicate the average occurrence year of keywords. The citation lineage structure of the field was evaluated using a k-means clustering-based analysis of cocitation networks to trace influential references.

Results: Our analysis revealed a significant increase in publications since 2021, with foundational works emerging as early as 2019. Influential journals in this domain included *JMIR Medical Education*, *Anatomical Sciences Education*, and *Medical Education*. The evolving research trajectory exhibited a shift from conventional computer-assisted learning tools toward generative AI platforms. Earlier applications of AI in medical education were predominantly concentrated at the undergraduate level, indicating substantial potential for expansion into graduate and continuing medical education. Furthermore, limited cocitation connections were observed between recent generative AI research and conventional medical AI studies, and investigations into medical students' attitudes toward generative AI remain scarce.

Conclusions: There are critical needs for (1) interdisciplinary studies that intentionally integrate generative AI with foundational medical AI work and (2) involving medical educators and students in AI development. Future research should focus on building theoretical frameworks and collaborative projects that connect these currently separate domains to foster a more cohesive knowledge base.

(JMIR Med Educ 2025;11:e75911) doi:10.2196/75911

KEYWORDS

artificial intelligence; AI; medical education; bibliometric analysis; research trends; generative artificial intelligence; generative AI

Introduction

The integration of artificial intelligence (AI) into medical education is not just a passing trend but a transformative leap that promises to revolutionize the quality, efficiency, and accuracy of medical learning and practice. This integration not only addresses the challenges posed by the sheer volume and complexity of medical knowledge but can also offer personalized learning experiences tailored to individual students' needs and abilities [1]. Nevertheless, there is still concern about AI's role in the future of medical education. Ethical issues, including student privacy; bias inherent in the data or algorithms; and model explainability, transparency, and accountability, are legitimate concerns [2]. The possibility of being replaced by AI or losing control over machines also generates distrust among stakeholders in medical education [3]. While advanced, AI tools often focus on the technical aspects of education delivery, neglecting the nuanced intricacies of human cognition and learning [4].

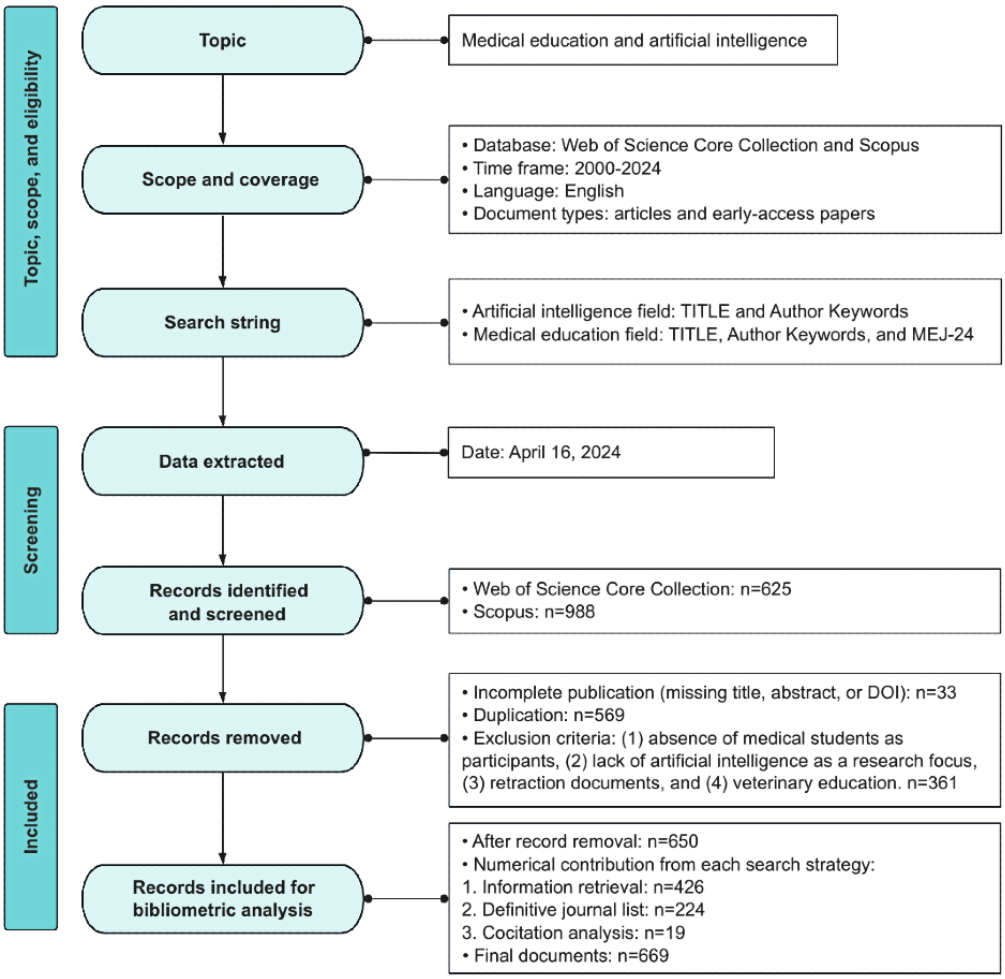
While the scholarly conversation on this topic is growing rapidly, the existing literature lacks a comprehensive, data-driven map of its intellectual structure and evolution. Narrative reviews and bibliometric studies in adjacent fields such as general AI in health care or digital education tools exist [5,6], but a dedicated analysis that tracks the evolution of research hubs within this specific domain is missing. Key questions remain unanswered: What are the foundational papers that have shaped this field? How have research themes evolved over time? What are the current dominant clusters of knowledge, and where is the frontier heading?

This study conducted a systematic bibliometric analysis to delineate the structural and temporal dynamics of medical education technology research. Adopting the methodological framework by Maggio et al [7], we systematically analyzed journal and author impact through publication and citation metrics; identified research trends and hot spots using citation burst detection, keyword frequency analysis, and co-occurrence networks supplemented by average occurrence year; and traced influential reference lineages via k-means clustering of cocitation networks. The objective was to provide an evidence-based map of the field's evolution, highlighting key areas of impact and emerging thematic directions to inform strategic research and educational planning.

Methods

This was a systematic bibliometric analysis, and the study selection process is illustrated in Figure 1 [7], which shows the identification, screening, and inclusion and exclusion criteria of documents.

Figure 1. Study flowchart. *The Medical Education Journal List 24 (MEJ-24) was proposed by Maggio et al [7] and is provided in Multimedia Appendix 1.



Ethical Considerations

This study did not involve human or animal participants and did not require patient consent. The ethics committee of Peking Union Medical College Hospital waived the ethical application because only published data were used. Patients or the public were not involved in the design, conduct, reporting, or dissemination plans of our research.

Data Extraction

Documents were extracted from the Web of Science and Scopus databases. To refine the strategy, we used 3 complementary approaches to delineate the field of AI in medical education following the framework established by Maggio et al [7].

The first approach was information retrieval. Initially, search terms were identified by reviewing existing systematic reviews and key publications in the fields of AI and medical education [7,8]. This process established a baseline set of keywords, including "artificial intelligence," "machine learning," and "medical education." The preliminary search string was then evaluated by a medical librarian and domain experts to ensure its appropriateness and was refined based on their feedback. This refinement involved incorporating more technical terms (eg, "deep learning," "natural language processing," and "transformer models") to better capture emerging trends. Furthermore, Boolean operators and truncation (eg, "educat* OR curricul* OR teach*") were applied to enhance search breadth. The revised search string was deployed using field searching techniques, specifically within the title and author keyword fields. The detailed search strategy is provided in Multimedia Appendix 2.

The second approach was definitive journals. Core medical education journals were identified using the Medical Education Journal List 24 (MEJ-24), an empirically derived set of journals validated through collaboration between medical education researchers and bibliometric experts [7]. For the search strategy, we conducted field searches across the 24 journals of the MEJ-24 using the journal source field. For AI-related terms, we searched the author keyword and title fields following the same information retrieval approach detailed previously. The MEJ-24 journals can be found in Multimedia Appendix 1.

The third approach was journal cocitation analysis. The journal cocitation analysis was conducted using the document pool retrieved through the 2 search strategies described previously (ie, information retrieval and the definitive journal list). Following data selection (as detailed in the Data Selection section), 650 documents were identified. A reference co-occurrence matrix was then constructed (see the Bibliometric Analysis section for details), and the top 30 most cited references among these articles were quantified. Notably, 63% (19/30) of these highly cited references had not been captured by the

original search string. These 19 articles were subsequently added to the cohort.

Documents were collected from January 1, 2000, to April 16, 2024. Only articles in English, reviews, and early-access papers were included.

Data Selection

The extraction approach yielded 1613 documents (n=625, 38.75% from Web of Science and n=988, 61.25% from Scopus). Documents lacking titles, abstracts, or digital object identifiers (DOIs) were removed, as were duplicates identified across the document pools using DOI matching. The inclusion criteria required that documents focus on the application of AI or computer-assisted techniques within medical education and present educational interventions or evaluations of AI tools in medical training contexts. Exclusions comprised retracted publications and documents addressing specific off-topic technical or noneducational themes. These included machine learning methods with nomenclature overlaps but no substantive educational relevance (eg, "teacher-student curriculum learning algorithm"), medically focused topics unrelated to AI in education (eg, "computer vision syndrome"), general online or remote education literature without specific AI components, and publications related to veterinary education. Our literature screening and selection process was conducted by 2 independent reviewers. Interrater reliability was quantified using the Cohen κ, which yielded a value of 0.85 based on 38.56% (622/1613) agreed inclusions, indicating strong agreement between reviewers. Discrepancies in inclusion decisions were resolved through consultation with 2 domain experts. Of the 107 articles over which there was initial disagreement, 28 (26.2%) were selected for inclusion after expert review, resulting in a total of 650 articles. A subsequent cocitation analysis of these 650 articles identified the 30 most cited references, 19 (63%) of which were not captured in the original search. These 19 articles were added to the cohort, bringing the final total to 669 articles (Multimedia Appendix 3). Information retrieval from databases contributed 63.7% (426/669) of the articles, the definitive journal list method contributed 33.5% (224/669) of the articles, and cocitation analysis contributed 2.8% (19/669) of the articles.

Data Processing

We imported raw export files in CSV format from both Scopus and Web of Science into the R environment (version 4.4.2; R Foundation for Statistical Computing) and standardized column names across datasets to ensure compatibility. Formatting inconsistencies in titles, abstracts, keywords, and author lists were resolved to create consistent records across both databases. Following the data selection process (as described in the Data Selection section), the preprocessed dataset was manually screened by 2 independent reviewers using Microsoft Excel. Discrepancies regarding inclusion were resolved through consultation with domain experts.

During keyword normalization, terms were standardized by converting them to lower case, reconciling plural and singular forms, expanding abbreviations, and removing special characters. For example, variants such as "3d computer-models," "3d models," and "3d reconstruction" were merged into the unified term "3d model." Keywords such as "artificial intelligence," "medical students," "computer," and "technology" were removed due to their limited analytical value. A complete mapping of these keyword transformations is provided in Multimedia Appendix 4.

Finally, we constructed a comprehensive analysis-ready dataset containing standardized fields for DOI, title, abstract, keywords, citation counts, authors, journals, publication year, and references. This final matrix formed the basis for all subsequent bibliometric analyses.

Bibliometric Analysis

Quantitative assessments of authors, journals, countries, and keywords were conducted using all available data under an inclusion threshold of 1. Collaboration networks between countries and keyword co-occurrence networks were constructed using VOSviewer (version 1.6.19; Centre for Science and Technology Studies, Leiden University) [9]. To enhance interpretability and network clarity, stricter thresholds were applied for network visualizations: keywords were retained only if they occurred at least 4 times, and country coauthorships were included only with a minimum of 5 occurrences. A keyword tree map was generated using the R package *Bibliometrix* (version 4.1.2) [10], with box sizes proportional to the keyword occurrence frequency. Keyword burst detection was conducted using CiteSpace (version 6.3.R1) [11] with the following parameters: g-index scale factor=2000, gamma=1, and minimum duration=1. The gamma parameter regulates the weight distribution in the burst strength calculation. A value of 1 ensures a balanced weighting scheme that neither overemphasizes recent citation spikes nor discounts significant earlier activity. The minimum duration parameter set to 1 requires that any identified citation burst persist for at least 1 year.

To evaluate journal impact, we considered both publication volume and citation metrics, and visualized the results using bubble plots. The x-axis represented $\log_{10}$(total publications+1), whereas the y-axis corresponded to $\log_{10}$(total citations+1). Bubble size reflected the average number of citations per publication, and color indicated the mean publication year, together providing a multidimensional perspective on journal influence.

To assess the influence of authors, we applied fractional counting to both publication and citation counts. For each author *A* who published *k* articles, we defined the following metrics:

(1)

(2)

where *mi* denotes the number of authors for article *i* and *ci* represents the citation count for article *i*. This fractional counting approach prevents the inflation of metrics for authors who frequently publish in large collaborations, thereby ensuring a

more equitable representation of individual scholarly contributions compared to whole-counting methods.

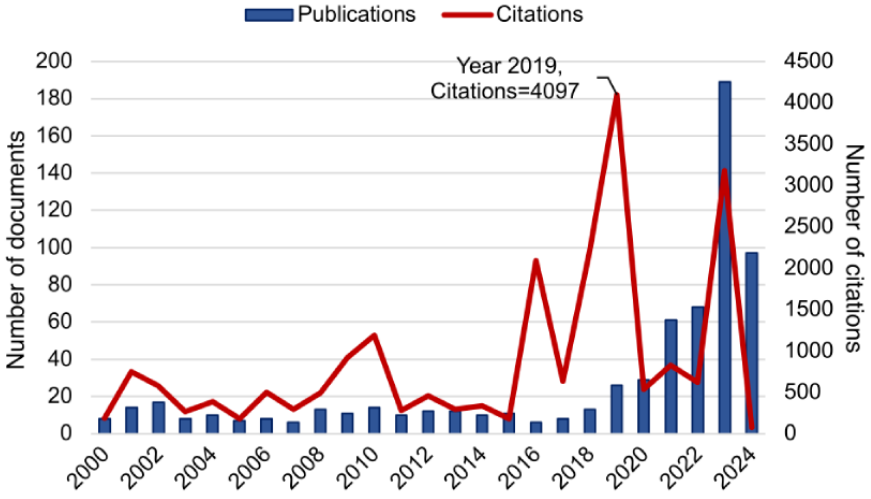
The cocitation clustering analysis was conducted using the R package *ComplexHeatmap* (version 2.12.1) [12]. A reference matrix was constructed, with rows representing cited references and columns representing citing documents. To focus on the most influential references, two inclusion criteria were applied: (1) cited references must be among the top 30 most frequently cited based on citation counts, and (2) citing documents must reference at least 2 of these top 30 cited references. This approach ensured the identification of meaningful cocitation patterns rather than isolated references. K-means clustering (k=3) was applied to the citing documents, with the algorithm repeated 500 times to ensure stability. The choice of 3 clusters was empirically determined to balance interpretability and granularity. Clusters were interpreted by examining cocitation patterns (which references frequently appeared together) and temporal trends (how citation networks evolved over time).

Results

Influential Documents, Journals, and Authors in the Field of AI in Medical Education

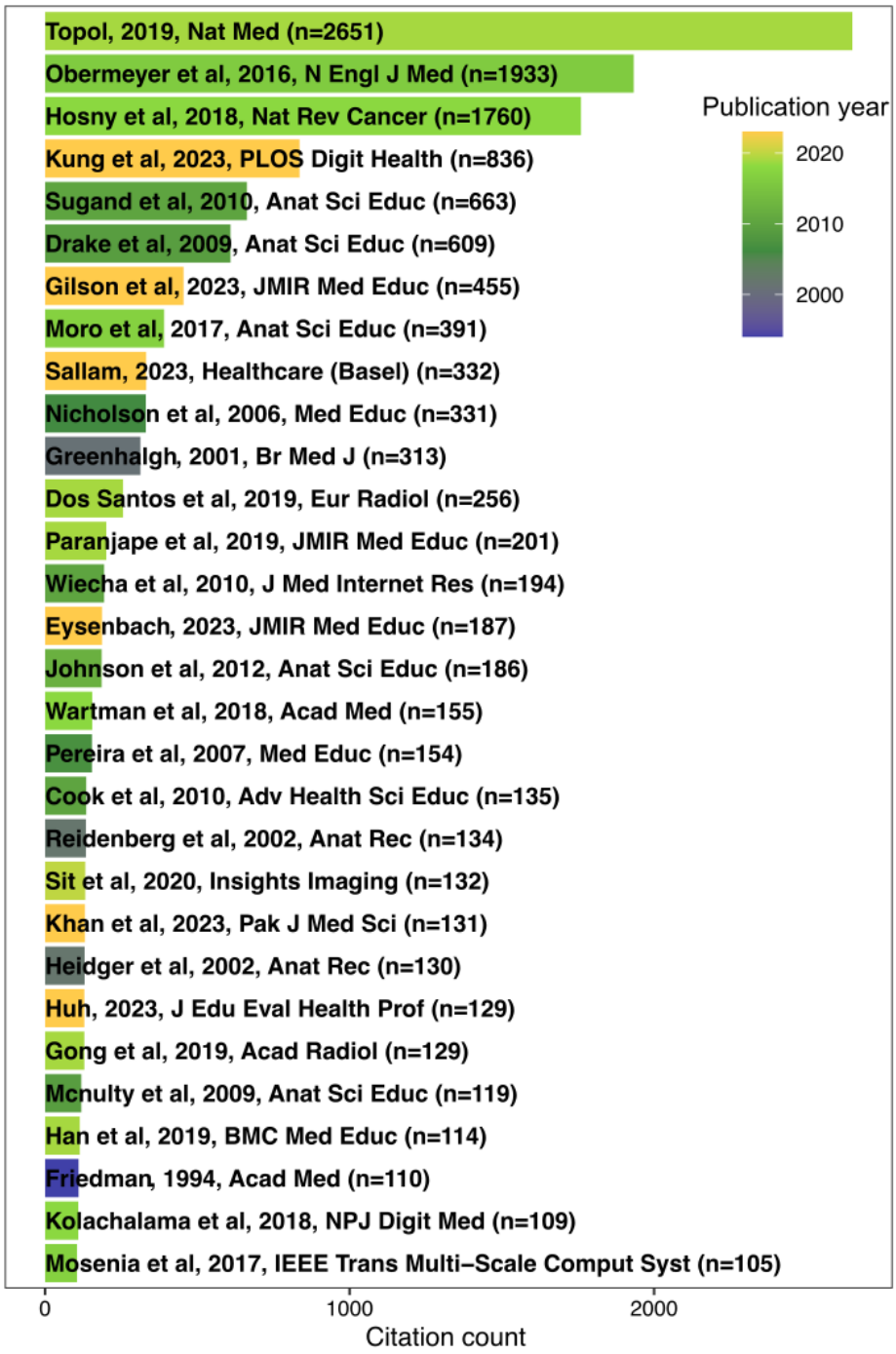
Our bibliometric analysis encompassed 669 documents from 269 journals authored by 3296 individuals across 295 countries, highlighting the global scope of research in this field. Figure 2 shows a notable increase in publications on AI in medical education since 2021. It is worth noting that, although publications from 2019 were relatively limited in number, they received substantial citations, suggesting that key conceptual foundations were proposed during this period. Figure 3 shows the top 30 most cited documents. Several highly influential works were published in 2019, most prominently the landmark article "High-Performance Medicine: The Convergence of Human and Artificial Intelligence" by Topol [13], which garnered 2651 citations. Other notable contributions included "Medical Students' Attitude Towards Artificial Intelligence: A Multicentre Survey" by Pinto Dos Santos et al [14] (cited 256 times) and "Introducing Artificial Intelligence Training in Medical Education" by Paranjape et al [15] (cited 201 times), further underscoring that 2019 marked the emergence of critical discussions on AI integration. More recent highly cited works reflected growing interest in generative AI models, such as "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models" by Kung et al [16] (cited 836 times) and "How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment" by Gilson et al [17] (cited 455 times).

Figure 2. Trends in publication and citation numbers.



Next, we evaluated the contribution of journals, authors, and countries within the field of AI in medical education (Figure 4). Relying solely on citation metrics may overlook journals that make consistent and foundational contributions to the field. To address this limitation, we applied a log-transformation to both citation and publication counts, allowing for a more balanced and comparable assessment across journals. Our analysis identified the following as the most influential journals in AI in medical education: *JMIR Medical Education* (publications: 67/669, 10%; citations: n=1575), *Anatomical Sciences Education* (publications: 48/669, 7.2%; citations: n=3317), *Medical Education* (publications: 35/669, 5.2%; citations: n=1467), *BMC Medical Education* (publications: 36/669, 5.4%; citations: n=456), *Academic Medicine* (publications: 23/669, 3.4%; citations: n=669), and *Medical Teacher* (publications: 28/669, 4.2%; citations: n=326).

Figure 3. Top 30 most cited documents.

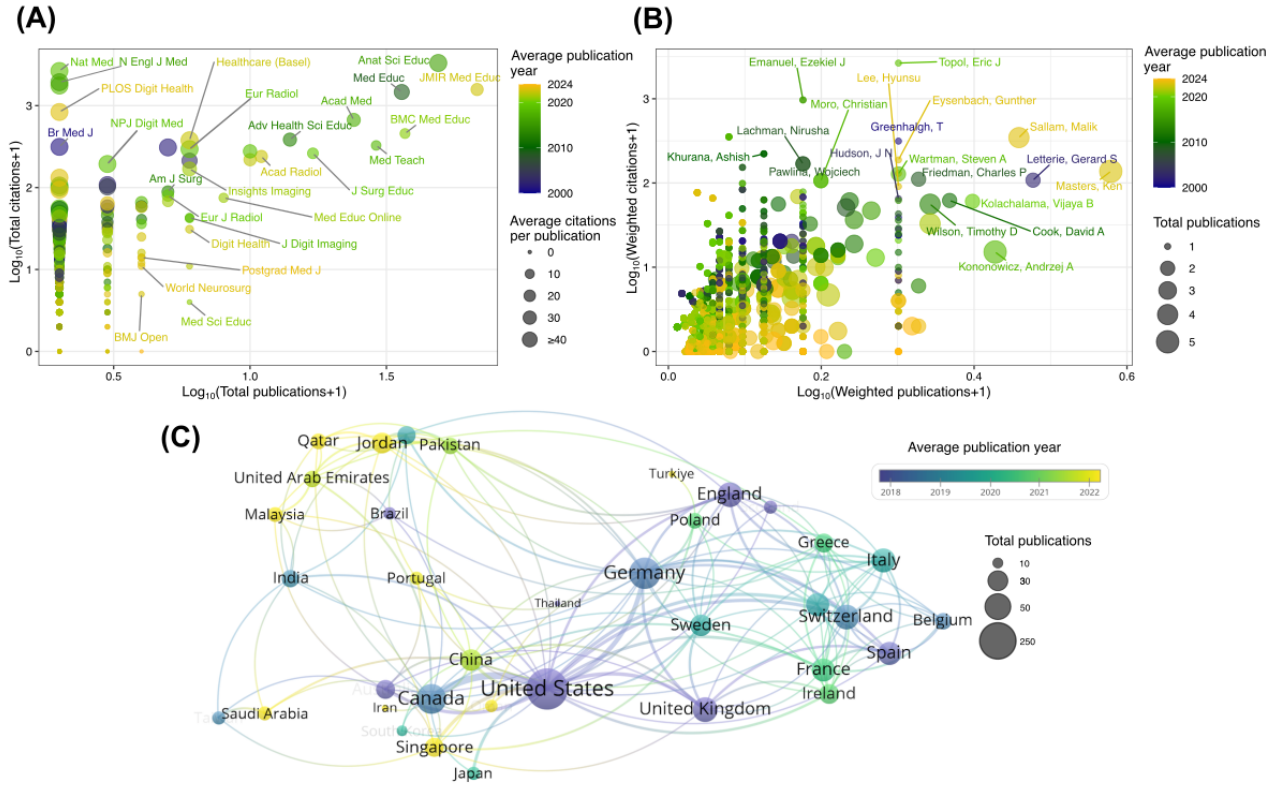


To assess the scholarly impact of authors while preventing the inflation of metrics for authors who frequently publish in large collaborations, we applied weighted counting to both publication and citation counts. Analogous to the normalization applied in journal-level analysis, we also used a $\log_{10}$ transformation to weighted citation and weighted publication counts at the author level. The results indicated that the following authors achieved high weighted publication and citation scores, reflecting their substantial contributions to the AI in medical education research field: Ken Masters (publications: 5/669, 0.7%; citations: n=140), Malik Sallam (publications: 4/669, 0.6%; citations: n=389), Gerard Letterie (publications: 2/669, 0.3%; citations: n=107), Steven Wartman (publications: 2/669, 0.3%; citations: n=254), Charles Friedman (publications: 2/669, 0.3%; citations: n=110), Kolachalama Vijaya (publications: 2/669, 0.3%; citations: n=114), Timothy Wilson (publications: 4/669, 0.6%; citations: n=176), David Cook (publications: 2/669, 0.3%; citations: n=151), and Andrzej Konowicz (publications: 5/669, 0.7%; citations: n=50).

To examine global collaboration patterns in the field, we constructed a country-level collaboration network, which revealed a central role for researchers based in the United States. The United States not only produced the highest volume of publications but also formed the most extensive collaborative ties with other countries. Similarly, several European nations—including Germany, the United Kingdom, and

France—emerged as major contributors. By incorporating temporal information through the average publication year for each country, we observed that non-Western countries such as China, Qatar, and Malaysia are increasingly active in the field, reflecting its expanding global engagement.

Figure 4. Journal, author, and country contributions visualized through bubble plots and a collaboration network. (A) For journals, bubble size reflects average citations per publication. (B) For authors, bubble size shows total publications. (C) In the country collaboration network, node size shows total publications, and line thickness represents partnership strength. Bubble and node colors reflect average publication year.



Identification of Research Hot Spots Using Keyword Analysis

To identify research hot spots over time, a keyword citation burst analysis was conducted. This method reveals when specific keywords become popular, how long they stay important, and the point at which their influence declines [18]. Figure 5 illustrates the top 10 keywords with the strongest citation bursts. Early bursts around 2010 featured terms such as "computer-assisted instruction" and "computer simulation," largely within the context of gross anatomy courses, particularly in undergraduate education. Subsequent trends highlighted the expansion of AI-enabled techniques into computer-assisted surgery. In 2023, research attention shifted toward "machine learning" and "large language models," with the generative pretrained transformer model standing out as one of the most promising developments likely to shape the future of the field.

A keyword tree map was constructed to visualize the frequency of keywords across the collected documents (Figure 6). The most common keywords included "ChatGPT," "computer-assisted instruction," and "machine learning." Frequent applications for AI in medical education were associated with anatomy, curriculum design, simulation, and image analysis. In addition, topics such as ethics, health care, and digital health also emerged as focal areas within AI-based educational research.

To gain deeper insights into conceptual relationships and their evolution, we developed keyword co-occurrence networks. Temporal evolution of the co-occurrence network aligned with the burst citation analysis (Figure 7). Computer-assisted learning was linked to various teaching techniques, including virtual microscopy, patient simulation, and 3D models. More recently, with the development of machine learning approaches, research interest shifted toward wider applications, such as curriculum reform, big data analysis, health care practice, and surgical training. The network further highlights natural language processing and large language models as current research hot spots, particularly in applications such as answering multiple-choice tests in medical licensing examinations, providing feedback systems, and serving as conversational agents. We then compared co-occurrence patterns across undergraduate education (occurrence: n=29; average occurrence year 2013, SD 7.8), graduate education (occurrence: n=9; average occurrence year 2016, SD 8.5), and continuing medical education (occurrence: n=10; average occurrence year 2016, SD 7.2). The results revealed that anatomy education, examinations, and problem-based learning were prominent in undergraduate medical education; graduate education emphasized communication, innovation, and decision support; and continuing medical education focused on clinical practice and digital health. Notably, large language models such as ChatGPT permeated all 3 educational stages.

Figure 5. Keyword burst citation detection analysis indicating the year of emergence, burst strength, and the start and end years of each keyword's citation surge.

Figure 6. Tree map of author keywords sized by their frequency of occurrence in the literature.

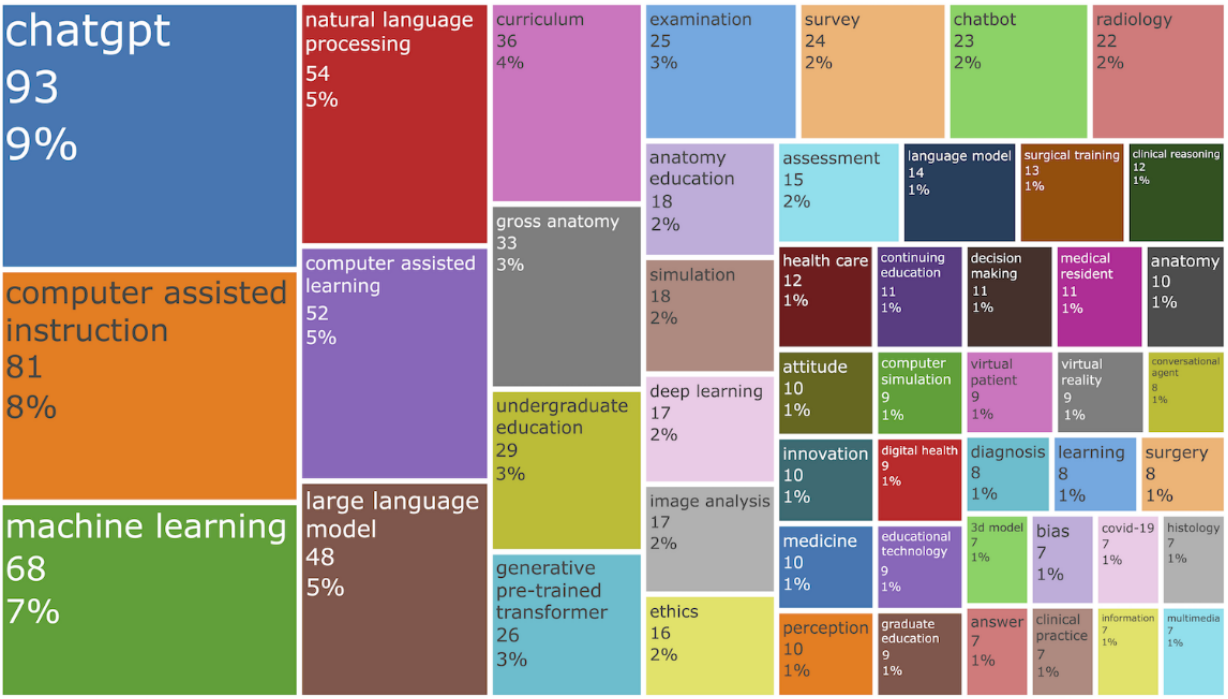
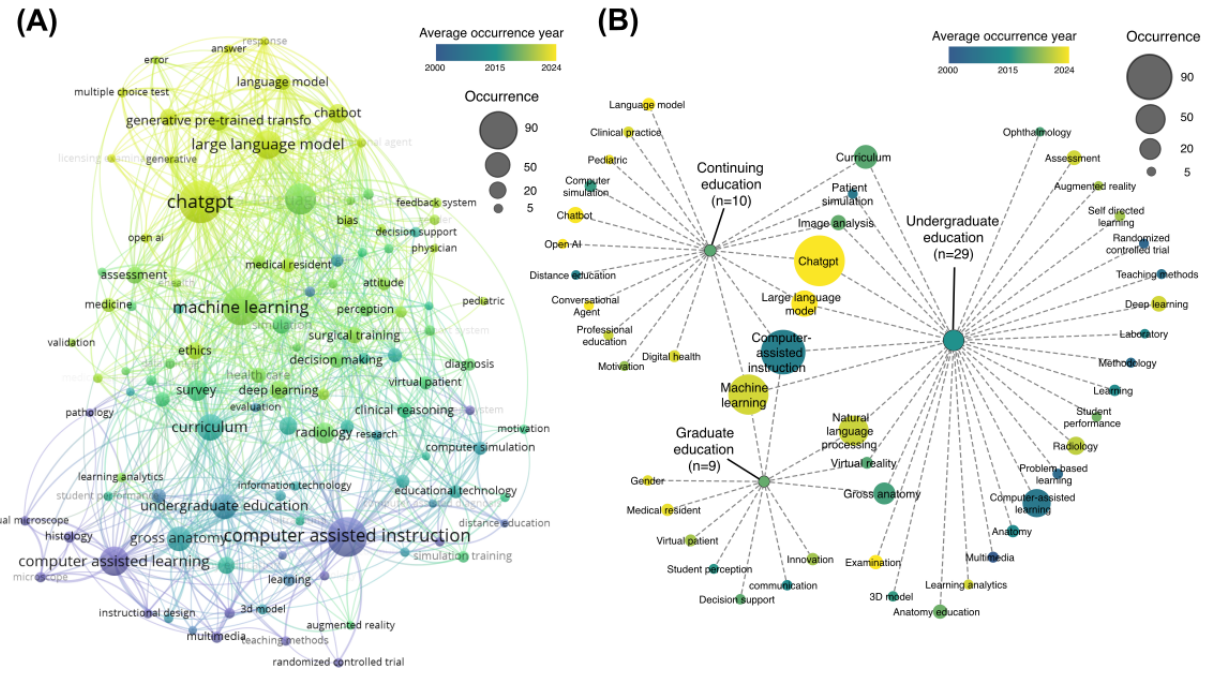


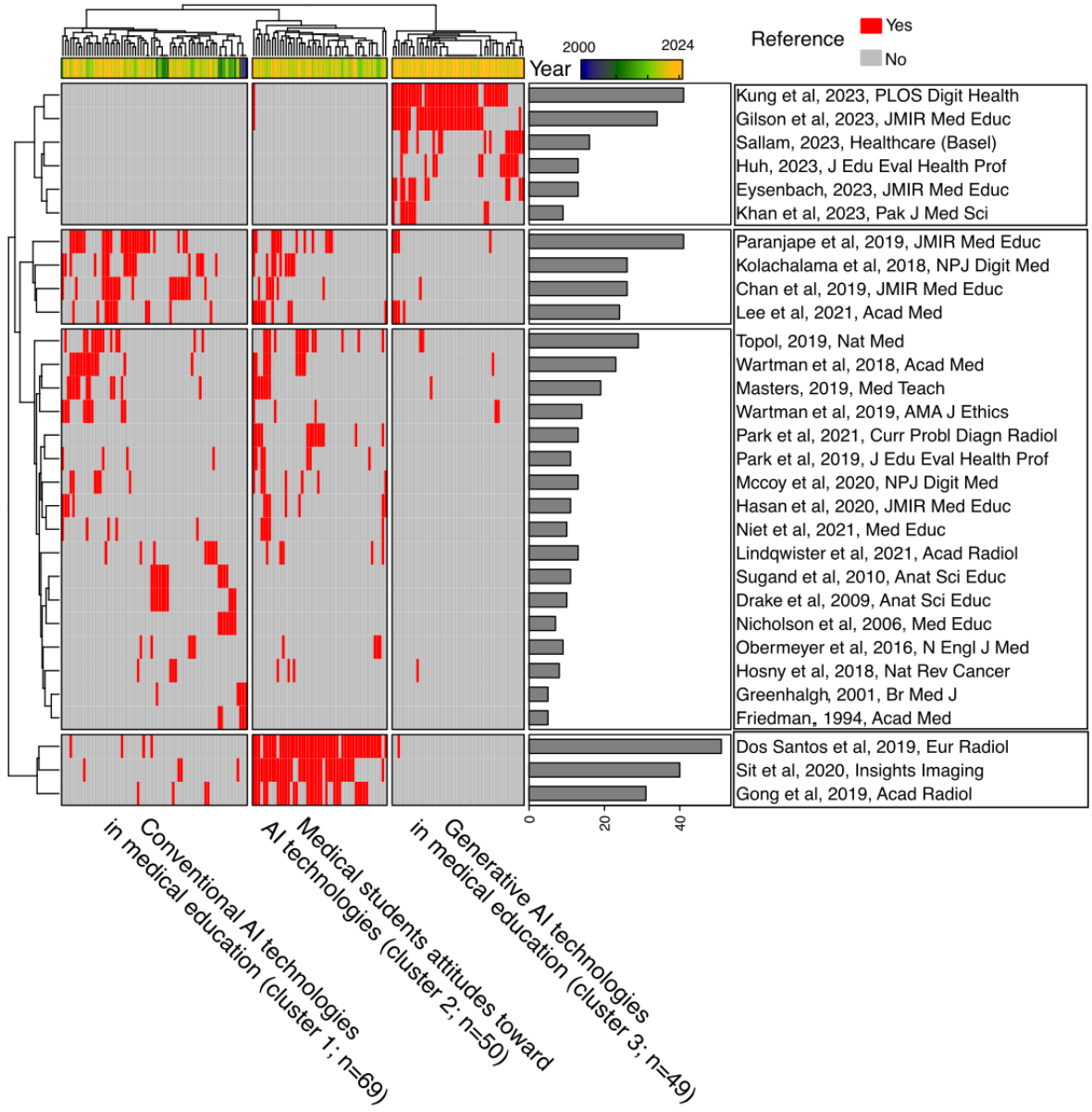
Figure 7. Keyword co-occurrence network. (A) Node size corresponds to keyword frequency, and node color represents the average occurrence year (blue represents older terms; yellow indicates more recent terms). (B) Magnified view of the network highlighting keywords associated with undergraduate, graduate, and continuing medical education.



Cocitation Analysis of References

Next, we examined the cocitation structure among the documents. Studies with shared research interest tended to cite similar foundational references, revealing their intellectual lineages [11]. A reference co-occurrence matrix was constructed, where rows represented cited references and columns represented citing articles. To capture meaningful cocitation patterns rather than isolated references, we included only the top 30 most frequently cited references and required that citing documents reference at least 2 of these highly cited works. After applying these criteria, 25.1% (168/669) of the articles remained for subsequent analysis. This approach identified 3 principal thematic clusters within the literature (Figure 8). The first was conventional AI technologies in medical education (cluster 1; citing documents: 69/168, 41.1%). References in this cluster advocated for structural innovations in medical training, with an emphasis on integrating computational thinking and AI literacy (Paranjape et al [15], Chan and Zary [2], Kolachalama and Garg [19], Topol [13], Wartman and Combs [20,21], and Masters [22]). The second thematic cluster was medical students' attitudes toward AI technologies (cluster 2; citing documents: 50/168, 29.8%). Studies in this cluster used quantitative surveys and qualitative methods to evaluate learner engagement with emerging tools (Pinto Dos Santos et al [14], Gong et al [23], and Sit et al [3]). The third thematic cluster was generative AI technologies in medical education (cluster 3; citing documents: 49/168, 29.2%). A rapidly growing cluster, it addressed implementation challenges and validation frameworks for generative AI models (Sallam [1], Huh [24], Kung et al [16], and Gilson et al [17]). Notably, while recent publications showed substantial emphasis on generative AI applications (cluster 3), few studies comparatively evaluated generative AI technologies against earlier conventional AI education reforms (cluster 1) or prioritized investigating student attitudes (cluster 2) toward these new tools.

Figure 8. The cocitation structure of the literature. Rows represent cited references, whereas columns denote citing articles. Red cells indicate the presence of a cocitation relationship, and the temporal gradient (blue represents older terms; yellow indicates more recent terms) highlights the chronological evolution of citation patterns. The accompanying bar plot reflects the cumulative citation frequency of each reference (Multimedia Appendix 3). AI: artificial intelligence.



Discussion

Emerging Trends and Global Contributions in AI-Driven Medical Education Research

The analysis of influential works in AI-based medical education revealed that 2019 was a critical turning point for the field. Seminal publications by Topol [13], Pinto Dos Santos et al [14], and Paranjape et al [15] laid the conceptual foundation for AI integration into medical training, sparking sustained scholarly dialogue. Their high citation rates reflect broad recognition of issues such as digital competency development, curricular reform, and learner attitudes toward AI—themes that remain relevant today. The more recent surge in citations for works on large language models (eg, Kung et al [16] and Gilson et al [17]) highlights a shift toward generative AI applications, particularly in knowledge assessment and adaptive learning platforms. This pattern aligns with global trends in AI adoption, where emerging technologies rapidly influence educational paradigms.

Our journal-level analysis further demonstrates the importance of using normalized metrics to evaluate influence. While journals such as *Nature Medicine* and the *New England Journal of Medicine* achieved high citation counts per article—often due to overarching prestige or coverage of breakthrough topics—specialized journals such as *JMIR Medical Education*, *Anatomical Sciences Education*, and *Medical Education* provided consistent intellectual contributions that shaped the field structurally. This result is consistent with previous literature findings that *JMIR Medical Education* and *Medical Teacher*, among others, are ranked as the top most productive journals

[25]. The log-transformation approach helped mitigate biases toward low-output, high-impact journals, offering a more equitable assessment of sustained influence.

At the author level, weighted counting methods revealed key contributors without overrepresenting researchers involved predominantly in large collaborations. Scholars such as Ken Masters, Malik Sallam, and Steven Wartman produced focused yet high-impact work, suggesting that quality and thematic consistency can achieve significant scholarly recognition even with modest publication numbers.

Geographically, the United States and Western European nations currently lead in output and collaboration reach. However, the rising activity from countries such as China, Qatar, and Malaysia signals a broadening of research capacity and interest beyond traditional hubs. This decentralization may foster diverse perspectives and context-specific AI applications in global medical education.

Evolution of Thematic Focus and Implications of AI in Medical Education

The integration of AI into medical education has undergone distinct thematic shifts over the past 2 decades, reflecting both technological advancements and evolving pedagogical priorities. In the early 2000s, efforts centered on computer-assisted instruction, leveraging multimedia, 3D models, and virtual reality to enhance anatomy instruction and patient simulation [26,27]. These tools initially focused on cognitive skill development but often overlooked social and psychomotor competencies essential to clinical practice. Our keyword analysis validated this trajectory, showing early citation bursts around 2010 for terms such as “computer-assisted instruction” and “computer simulation,” particularly within undergraduate anatomy education. Subsequent developments saw AI applications expand into computer-assisted surgery and more integrated educational supports, including systems for personalized learning and clinical reasoning simulation [15,19]. Our co-occurrence network analysis further delineated these thematic alignments across different training stages. The undergraduate education phase, with an average keyword occurrence year of 2013, emphasized foundational topics such as anatomy, examinations, and problem-based learning. In contrast, graduate education, with an average keyword occurrence year of around 2016, reflected a shift toward communication skills and clinical decision support. Meanwhile, continuing medical education, also centered around 2016, focused strongly on clinical practice and digital health. Together, these patterns illustrate AI’s expanding role in supporting the entire continuum of medical training.

A significant shift occurred around 2023, as observed in both our burst detection and co-occurrence networks, with research attention toward “machine learning” and “large language models.” Keywords such as “ChatGPT” and “generative AI” now dominate the literature, reflecting a new wave of innovation aimed at conversational agents, feedback systems, and assessment tools, particularly in answering multiple-choice questions and simulating tutor-learner interactions. Notably, large language models have permeated all levels of medical education, suggesting their transformative potential across undergraduate, graduate, and continuing training contexts [17,28].

Challenges in AI-Driven Medical Education and Future Directions

AI can automate routine tasks, alleviating the cognitive load on learners’ working memory. This process of “cognitive off-loading” enables students to dedicate more mental resources to mastering higher-order, complex skills [29]. Consequently, modern medical curricula should transition from rote memorization to emphasizing higher-order skills such as computational thinking and AI literacy [21]. However, keyword analysis indicates that current AI in medical education remains predominantly focused at the undergraduate level, with keyword frequencies of 29 for undergraduate education compared to only 9 for graduate education and 10 for continuing medical education. This disparity suggests significant potential for expanding AI integration into both graduate and continuing medical education [30]. Proposed reforms include providing foundational training in statistics, offering hands-on experience with open-source tools, and integrating AI components into existing courses [2,31]. A structured, stage-specific educational approach is recommended across all training levels: premedical education should introduce AI concepts in entrance examinations, medical training should require coursework in data science and AI ethics, clinical rotations should incorporate AI decision support tools (eg, radiology AI), and postgraduate training should include continuing education modules on emerging AI applications [13,15]. Nevertheless, overreliance on AI carries the risk of eroding fundamental medical skills among students [29]. A balanced approach is essential to ensure that AI serves as a complement to rather than a replacement for core clinical competencies.

Cocitation analysis revealed that cluster 1 corresponded to curriculum reforms and recommendations proposed by medical education experts for structural innovations in medical training [2,13,15,19-22], whereas cluster 2 comprised studies exploring students’ attitudes toward AI applications in medical education [3,14,23]. However, there is a noticeable gap between recent research on generative AI (cluster 3) and the earlier AI in medical education studies (clusters 1 and 2), indicating that integrative frameworks incorporating learning theories and user feedback remain underdeveloped. Future research should prioritize the development of unified theoretical models that incorporate diverse learning perspectives into generative AI systems, as well as establish assessment frameworks to evaluate the effectiveness of generative AI in achieving intended educational outcomes. Moreover, the perceptions and attitudes of both teachers and students are critical for technological advancement, underscoring the importance of a co-design approach in developing educational tools [20]. Therefore, it is essential to evaluate the efficacy of AI integration into existing curricula using multidimensional metrics, including surveys and learning analytics [14,32].

We also observed that the keyword “ethics” (occurrence: n=16) represented a significant theme in AI in medical education, frequently co-occurring with terms such as “big data,” “bias,” and “privacy.” Data privacy and model reliability remained

major concerns, especially when processing a large volume of sensitive information [2]. Furthermore, skepticism persists due to issues such as potential algorithmic bias and AI's limited adaptability to diverse cultural and contextual settings within medical education [33]. To address these challenges, AI systems should emphasize explainability and incorporate human oversight, enabling educators to review and override unjustified decisions [34]. Robust data governance and ethical guidelines must form the foundation of AI deployment, whereas interdisciplinary collaboration helps ensure that AI solutions are aligned with educational goals, whereas interdisciplinary collaboration helps ensure that AI solutions are aligned with educational goals [19]. Involving both instructors and students in the development and validation of AI tools can lead to more adaptive and context-aware systems that effectively complement human expertise and deliver targeted, AI-enhanced educational support.

Limitations

Potential methodological limitations should be considered, including language bias, the exclusion of gray literature, and the artificial temporal boundary set for the analysis (2000-2024). Due to the inherent constraints of bibliometric methods, our findings should be interpreted as exploratory in nature. It is also important to note that the identified trends and research hot spots are derived from publication keywords. To address these limitations, future studies should use a more comprehensive content analysis framework. This approach will facilitate a deeper exploration of broader trends in AI applications, such as their objectives, methodologies, and major outcomes, thereby yielding richer insights than those of traditional bibliometric techniques.

Comparison With Prior Work

Previous bibliometric studies have explored AI applications in medical education [6,25]. Many of their findings can be reproduced, such as author impact, identification of influential journals, international collaboration patterns, and keyword frequency analyses. However, our study introduced several novel methodological approaches. First, by using the search strategy proposed by Maggio et al [7], we identified a broader corpus of literature in the domain of AI in medical education. Second, we investigated the temporal shift in the average publication year across journals, authors, and countries, as well as the evolution in research themes. This approach allowed for dynamic tracking of knowledge contributions and thematic trends over time. Third, while Wang et al [6] summarized the top 10 most cited references in their dataset, a list that shows considerable overlap with our reference matrix, we extended this analysis by visualizing the cocitation network structure using a clustered heat map. This revealed distinct clustering among citing articles, offering deeper structural insights into the intellectual foundations of the field. Our cocitation methodology is designed to be generalizable and applicable to other bibliometric research domains.

Conclusions

As AI technologies continue to evolve, they are likely to enable more sophisticated simulations, personalized learning experiences, and data-driven insights that may transform medical education. Our bibliometric analysis provides an exploratory overview of current research trends and collaborations in this emerging domain, offering preliminary insights that can guide future, more in-depth investigations. Future research in AI-based medical education should prioritize expanding into graduate and continuing medical training, bridging the disciplinary gap between generative AI and established medical AI research through integrated frameworks, and deepening inquiry into learner and educator perceptions to ensure the responsible and effective implementation of these technologies.

Acknowledgments

The New Medical Education Reform Project 2024 at Peking Union Medical College (2024bkjg051), Youth Education Scholar Program (2023zlgc0706), and China Medical Board (CMB-JYCXGG04) provided financial support for language editing and publication. The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files (Multimedia Appendices 1 to 4).

Authors' Contributions

Conceptualization: YW (lead), CC (equal)
Data curation: WS
Formal analysis: CC (lead), YW (supporting)
Funding acquisition: YJ (lead), WS (equal)
Investigation: YW
Methodology: CC
Project administration: YW (lead), CC (equal)
Resources: CC

Supervision: HL
Validation: XH
Visualization: CC (lead), YW (supporting)
Writing—original draft: CC (lead), YW (supporting)
Writing—review and editing: YJ (lead), XH (supporting), HL (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Medical Education Journal List 24.
[XLSX File (Microsoft Excel File), 11 KB - mededu_v11i1e75911_app1.xlsx]

Multimedia Appendix 2
Detailed search strings for data extraction.
[XLSX File (Microsoft Excel File), 12 KB - mededu_v11i1e75911_app2.xlsx]

Multimedia Appendix 3
Dataset of the summarized 669 documents included in the analysis.
[XLSX File (Microsoft Excel File), 542 KB - mededu_v11i1e75911_app3.xlsx]

Multimedia Appendix 4
Mapping of keyword transformations.
[XLSX File (Microsoft Excel File), 17 KB - mededu_v11i1e75911_app4.xlsx]

References

1. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887 [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
2. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. JMIR Med Educ 2019 Jun 15;5(1):e13930 [FREE Full text] [doi: 10.2196/13930] [Medline: 31199295]
3. Sit C, Srinivasan R, Amlani A, Muthuswamy K, Azam A, Monzon L, et al. Attitudes and perceptions of UK medical students towards artificial intelligence and radiology: a multicentre survey. Insights Imaging 2020 Mar 05;11(1):14 [FREE Full text] [doi: 10.1186/s13244-019-0830-7] [Medline: 32025951]
4. Liefooghe B, van Maanen L. Three levels at which the user's cognition can be represented in artificial intelligence. Front Artif Intell 2022 Jan 13;5:1092053 [FREE Full text] [doi: 10.3389/frai.2022.1092053] [Medline: 36714204]
5. Tabassum A, Ghaznavi I, Abd-Alrazaq A, Qadir J. Exploring the application of AI and extended reality technologies in metaverse-driven mental health solutions: scoping review. J Med Internet Res 2025 Aug 19;27:e72400 [FREE Full text] [doi: 10.2196/72400] [Medline: 40829151]
6. Wang S, Yang L, Li M, Zhang X, Tai X. Medical education and artificial intelligence: web of science-based bibliometric analysis (2013-2022). JMIR Med Educ 2024 Oct 10;10:e51411 [FREE Full text] [doi: 10.2196/51411] [Medline: 39388721]
7. Maggio LA, Ninkov A, Frank JR, Costello JA, Artino Jr AR. Delineating the field of medical education: bibliometric research approach(es). Med Educ 2022 Apr 02;56(4):387-394 [FREE Full text] [doi: 10.1111/medu.14677] [Medline: 34652832]
8. Chang C, Shi W, Wang Y, Zhang Z, Huang X, Jiao Y. The path from task-specific to general purpose artificial intelligence for medical diagnostics: a bibliometric analysis. Comput Biol Med 2024 Apr;172:108258 [FREE Full text] [doi: 10.1016/j.compbiomed.2024.108258] [Medline: 38467093]
9. van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 2010 Aug 31;84(2):523-538 [FREE Full text] [doi: 10.1007/s11192-009-0146-3] [Medline: 20585380]
10. Aria M, Cuccurullo C. bibliometrix : an R-tool for comprehensive science mapping analysis. J Informetr 2017 Nov;11(4):959-975. [doi: 10.1016/j.joi.2017.08.007]
11. Chen C. CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature. J Am Soc Inf Sci 2005 Dec 14;57(3):359-377 [FREE Full text] [doi: 10.1002/asi.20317]
12. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 2016 Sep 15;32(18):2847-2849. [doi: 10.1093/bioinformatics/btw313] [Medline: 27207943]
13. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan 7;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]

14. Pinto Dos Santos D, Giese D, Brodehl S, Chon SH, Staab W, Kleinert R, et al. Medical students' attitude towards artificial intelligence: a multicentre survey. Eur Radiol 2019 Apr 6;29(4):1640-1646. [doi: 10.1007/s00330-018-5601-1] [Medline: 29980928]
15. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019 Dec 03;5(2):e16048 [FREE Full text] [doi: 10.2196/16048] [Medline: 31793895]
16. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Mar 9;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
17. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the united states medical licensing examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Mar 08;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
18. Kleinberg J. Bursty and hierarchical structure in streams. Data Min Knowl Discov 2003 Oct;7(4):373-397. [doi: 10.1023/a:1024940629314]
19. Kolachalama VB, Garg PS. Machine learning and medical education. NPJ Digit Med 2018 Sep 27;1(1):54 [FREE Full text] [doi: 10.1038/s41746-018-0061-1] [Medline: 31304333]
20. Wartman SA, Combs CD. Medical education must move from the information age to the age of artificial intelligence. Acad Med 2018 Aug;93(8):1107-1109. [doi: 10.1097/ACM.0000000000002044] [Medline: 29095704]
21. Wartman SA, Combs CD. Reimagining medical education in the age of AI. AMA J Ethics 2019 Mar 01;21(2):E146-E152 [FREE Full text] [doi: 10.1001/amajethics.2019.146] [Medline: 30794124]
22. Masters K. Artificial intelligence in medical education. Med Teach 2019 Apr 21;41(9):976-980. [doi: 10.1080/0142159x.2019.1595557] [Medline: 31007106]
23. Gong B, Nugent JP, Guest W, Parker W, Chang PJ, Khosa F, et al. Influence of artificial intelligence on Canadian medical students' preference for radiology specialty: a national survey study. Acad Radiol 2019 Apr;26(4):566-577. [doi: 10.1016/j.acra.2018.10.007] [Medline: 30424998]
24. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. J Educ Eval Health Prof 2023 Jan 11;20:1 [FREE Full text] [doi: 10.3352/jeehp.2023.20.1] [Medline: 36627845]
25. Li R, Wu T. Evolution of artificial intelligence in medical education from 2000 to 2024: bibliometric analysis. Interact J Med Res 2025 Jan 30;14:e63775 [FREE Full text] [doi: 10.2196/63775] [Medline: 39883926]
26. Sugand K, Abrahams P, Khurana A. The anatomy of anatomy: a review for its modernization. Anat Sci Educ 2010;3(2):83-93. [doi: 10.1002/ase.139] [Medline: 20205265]
27. Fidler BD. Use of a virtual patient simulation program to enhance the physical assessment and medical history taking skills of doctor of pharmacy students. Curr Pharm Teach Learn 2020 Jul;12(7):810-816. [doi: 10.1016/j.cptl.2020.02.008] [Medline: 32540042]
28. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, et al. Toward expert-level medical question answering with large language models. Nat Med 2025 Mar 08;31(3):943-950. [doi: 10.1038/s41591-024-03423-7] [Medline: 39779926]
29. Abdulnour RE, Gin B, Boscardin CK. Educational strategies for clinical supervision of artificial intelligence use. N Engl J Med 2025 Aug 21;393(8):786-797. [doi: 10.1056/nejmra2503232] [Medline: 40834302]
30. Gordon M, Daniel M, Ajiboye A, Uraiby H, Xu NY, Bartlett R, et al. A scoping review of artificial intelligence in medical education: BEME guide no. 84. Med Teach 2024 Feb 29;46(4):446-470. [doi: 10.1080/0142159x.2024.2314198] [Medline: 38423127]
31. Wang C, Li S, Lin N, Zhang X, Han Y, Wang X, et al. Application of large language models in medical training evaluation-using ChatGPT as a standardized patient: multimetric assessment. J Med Internet Res 2025 Jan 01;27:e59435 [FREE Full text] [doi: 10.2196/59435] [Medline: 39742453]
32. Sakelaris PG, Novotny KV, Borvick MS, Lagasca GG, Simanton EG. Evaluating the use of artificial intelligence as a study tool for preclinical medical school exams. J Med Educ Curric Dev 2025 Feb 24;12:23821205251320150 [FREE Full text] [doi: 10.1177/23821205251320150] [Medline: 40008118]
33. Masters K. Ethical use of artificial intelligence in health professions education: AMEE guide no. 158. Med Teach 2023 Mar 13;45(6):574-584. [doi: 10.1080/0142159x.2023.2186203] [Medline: 36912253]
34. Cohen IG, Babic B, Gerke S, Xia Q, Evgeniou T, Wertenbroch K. How AI can learn from the law: putting humans in the loop only on appeal. NPJ Digit Med 2023 Aug 25;6(1):160 [FREE Full text] [doi: 10.1038/s41746-023-00906-8] [Medline: 37626155]

Abbreviations

AI: artificial intelligence
DOI: digital object identifier
MEJ-24: Medical Education Journal List 24

Edited by J Eriksen; submitted 13.04.25; peer-reviewed by A Syahid, M Iniesta; comments to author 28.07.25; revised version received 31.08.25; accepted 27.10.25; published 18.11.25.

Please cite as:
Wang Y, Chang C, Shi W, Liu H, Huang X, Jiao Y
How AI Is Transforming Medical Education: Bibliometric Analysis
JMIR Med Educ 2025;11:e75911
URL: https://mededu.jmir.org/2025/1/e75911
doi:10.2196/75911
PMID:41252190

Original Paper

Evaluating ChatGPT-4o as an Educational Support Tool for the Emergency Management of Dental Trauma: Randomized Controlled Study Among Students

Franziska Haupt[1], PD, Dr med dent; Tina Rödig[1], Prof Dr; Paula Liersch[1], Dr med dent

Department of Preventive Dentistry, Periodontology and Cariology, University Medical Center Göttingen, Göttingen, Lower Saxony, Germany

Corresponding Author:
Franziska Haupt, PD, Dr med dent
Department of Preventive Dentistry, Periodontology and Cariology
University Medical Center Göttingen
Robert-Koch-Str. 40
Göttingen, Lower Saxony, 37075
Germany
Phone: 49 5513960888
Email: franziska.haupt@med.uni-goettingen.de

Abstract

Background: Digital tools are increasingly used to support clinical decision-making in dental education. However, the accuracy and efficiency of different support tools, including generative artificial intelligence, in the context of dental trauma management remain underexplored.

Objective: This study aimed to evaluate the accuracy of various information sources (chatbot, textbook, mobile app, and no support tool) in conveying clinically relevant educational content related to decision-making in the primary care of traumatically injured teeth. Additionally, the effect of the input strategy on the chatbot's output response was evaluated.

Methods: Fifty-nine dental students with limited prior experience in dental trauma were randomly assigned to one of 4 groups: chatbot (based on generative pretrained transformer [GPT]-4o, n=15), digital textbook (n=15), mobile app (AcciDent app 3.5, n=15), and control group (no support tool, n=14). Participants answered 25 dichotomous questions in a digital examination format using the information source allocated to their group. The primary outcome measures were the percentage of correct responses and the time required to complete the examination. Additionally, for the group using ChatGPT-4o, the quality of prompts and the clarity of chatbot responses were independently evaluated by 2 calibrated examiners using a 5-point Likert scale. Statistical analyses included nonparametric analyses using Kruskal-Wallis tests and mixed-effects regression analyses with an α level of .05.

Results: All support tools led to a significantly higher accuracy compared with the control group (P<.05), with mean accuracies of 87.47% (SD 5.63%), 86.40% (SD 5.19%), and 86.40% (SD 6.38%) for the textbook, the AcciDent app, and ChatGPT-4o, respectively. The groups using the chatbot and the mobile app required significantly less time than the textbook group (P<.05). Within the ChatGPT-4o group, higher prompt quality was associated with greater clarity of the chatbot's responses (odds ratio 1.44, 95% CI 1.13-1.83, P<.05), which in turn increased the likelihood of students selecting the correct answers (odds ratio 1.89, 95% CI 1.26-2.80, P<.05).

Conclusions: ChatGPT-4o and the AcciDent app can serve dental students as an accurate and time-efficient support tool in dental trauma care. However, the performance of ChatGPT-4o varies with the precision of the input prompt, underscoring the necessity for users to critically evaluate artificial intelligence–generated responses.

Trial Registration: OSF Registries 10.17605/OSF.IO/XW62J; https://osf.io/xw62j/overview

(JMIR Med Educ 2025;11:e80576) doi:10.2196/80576

KEYWORDS

decision–making support tool; dental trauma education; large language model; mobile app; prompting strategy

Introduction

Dental trauma is defined as an acute mechanical injury to teeth and surrounding structures, commonly resulting from accidents or falls. With a global prevalence of 25%-30%, dental trauma represents a significant public health issue, affecting approximately 1 billion individuals and ranking as the fifth most common type of injury worldwide [1,2]. The incidence of dental trauma has shown an upward trend in recent years, leading to considerable health care expenditures [3]. In Germany alone, the estimated annual costs related to dental trauma range between €200 and €550 million (€1=US $1.03) [4].

Immediate and appropriate treatment is crucial to prevent long-term complications and to reduce treatment-related costs [4]. Inadequate or inappropriate initial care is a major contributing factor to poor outcomes following dental trauma. Therefore, timely and accurate diagnosis combined with evidence-based primary treatment is essential to optimizing functional and aesthetic prognoses [5,6]. To standardize the management of dental trauma, several clinical practice guidelines have been published [7]. Among these, the guidelines developed by the International Association of Dental Traumatology (IADT) provide clearly structured recommendations, which, when followed, have been associated with improved clinical outcomes and a reduced incidence of complications [8-10]. Unfortunately, it has been shown that there is substantial evidence of educational shortcomings in the field of dental trauma [11,12], which is further underscored by studies reporting varying, partly insufficient knowledge among dental professionals [13,14].

According to a recent systematic review, mobile health approaches have gained increasing popularity among both patients and health care professionals [15]. Mobile health refers to the emerging field of health care that uses mobile-based technologies, such as smartphone apps, social media platforms, and artificial intelligence (AI), for disease prevention, health education, and clinical decision support [15,16]. To date, several mobile apps have been developed, including 3 based on the IADT guidelines (ToothSoS, AcciDent, and Injured tooth) [15]. Given the growing availability and integration of health care technologies, it appears likely that students tend to compensate for knowledge deficits in dental trauma management by turning to digital solutions [17].

Moreover, AI-driven systems have gained increasing relevance in medical and dental practice in recent years [18-20]. Advances in AI have accelerated the development of large language models (LLMs), such as generative pretrained transformer (GPT) [21], which are increasingly recognized for their potential to assist in medical practice [22]. However, when using chatbots based on LLMs, obtaining high-quality answers largely depends on how users interact with the chatbot [23,24]. For this purpose, the emerging field of prompt engineering aims to systematically develop prompts to enhance the performance of LLMs [23,25]. Nevertheless, awareness among practitioners regarding the importance of effective prompting and the potential impact of various prompting techniques on the quality of generated output remains limited [23].

Previous research in dental traumatology has evaluated the accuracy of chatbot-generated responses in simulated dental trauma scenarios. Ozden et al [26] reported that the performance of such models varied significantly depending on the specific algorithm. In their evaluation, ChatGPT-3.5 demonstrated an accuracy of 51%, whereas Google Bard achieved a slightly higher accuracy of 64% [26]. More recent findings, however, suggest substantial advancements in the performance of newer LLMs. A recently published study assessing the emergency management of tooth avulsion found that ChatGPT-4 attained an accuracy of up to 96.3% in responding to dichotomous (yes/no) clinical questions [27].

Although AI apps in health care are expanding rapidly, further research is necessary to evaluate their reliability and clinical applicability in dentistry [28]. Given their ability to provide rapid, accessible information, LLMs may represent a valuable resource for managing dental trauma in emergency settings. Moreover, they are likely to be used by students as a supplementary tool during exam preparation. Nevertheless, as previously noted, the effectiveness of chatbot responses remains highly dependent on the specificity and clarity of user input. Therefore, the primary objectives of this study were to evaluate the accuracy of 3 different support tools compared with a control group by assessing the percentage of correct responses and to record the total time required. The null hypotheses stated that there would be no significant differences among the 4 groups in terms of response accuracy or time efficiency. As a secondary objective, the study examined whether the precision of the input prompt submitted to the chatbot influenced the clarity of its responses (output), and whether this output clarity affected the students' likelihood of selecting the correct answer.

Methods

Ethical Considerations

The study was approved by the local ethics committee of the University Medical Center Göttingen, Germany (number 28/10/24) and conducted in accordance with the Declaration of Helsinki (Registration on OSF Registries osf.io/xw62j). Dental students in their final academic year were asked to participate in the study. All of them had signed informed consent prior to participation. The researchers had no access to personal information such as sex, age, or prior academic performance. All data were pseudonymized before being subjected to statistical analysis. Participants did not receive any financial compensation. As an incentive, the 5 best-performing students in each group were awarded a book prize.

Study Design

A consecutive randomization procedure was applied to dental students in their final academic year, comprising participants from the 9th (n=26) and 10th semesters (n=33), resulting in a total sample of 59 students. Selection was conducted in a blinded manner, ordered by ascending matriculation numbers. Students were pseudorandomly allocated into 4 groups, ensuring that each group included an equal number of students from the 9th and 10th semesters. This approach ensured allocation concealment and minimized potential selection bias. None of the students had received formal education in dental

traumatology, ensuring a homogenous baseline knowledge across groups. They were not informed about the study content or its design until shortly before the start of the study, in order to prevent individual preparation. At this point, informed consent was required for participation. Based on group allocation, each student was granted access to one specific support tool: control (no tool, n=14), a digital textbook (n=15) [29], a mobile app (AcciDent app 3.5, n=15), or a GPT-4o-based chatbot (OpenAI, n=15). Students using the digital textbook [29] were instructed to use the keyword-based search function. Using a computer-based assessment tool (CAMPUS version 1.4 Rev 7090; IMS-3, Institut für Kommunikations-und Prüfungsforschung gGmbH), all students completed 25 questions presented as declarative statements, which they were instructed to classify as either true or false, using only their assigned support tool (Table S1 in Multimedia Appendix 1).

To ensure standardization and minimize potential bias, all participants in the GPT-4o chatbot group were provided with newly created accounts that had no prior usage history. Each assessment session was initiated with a fresh login and no prior interactions, effectively eliminating session memory effects. Students accessed the chatbot via a browser-based interface under supervised conditions. No other prompts or unrelated queries were entered during the assessment. This setup ensured that all responses were generated based solely on the test items and not influenced by prior context, history, or personalized adaptation of the model.

The 25 dichotomous assessment questions were developed by one specialized endodontist and one general dentist with clinical expertise in endodontics. The questionnaire was based on current IADT guidelines [5,6,9], which provide evidence-based recommendations for dental trauma management.

To eliminate time-related variability within the chatbot group, the assessment was conducted simultaneously for all study participants on December 11, 2024. During the experiment, the total time required to complete the questionnaire was recorded for each student. Participants were informed that their working time would be recorded, but no time limits were imposed in order to prevent influencing their performance. All answers were pseudonymized and transferred to an Excel spreadsheet (Microsoft Corp) for statistical analysis.

Item Analysis and Scoring Criteria

To ensure content validity, the items were reviewed by 5 dental education experts prior to the experiment. All questions were reviewed to confirm they could be answered using one of the assessed supporting tools. For the item quality analysis, all items were evaluated using the difficulty index (P), the discrimination index (D), and internal consistency reliability (Kuder Richardson formula 20; KR-20). In addition, group-specific difficulty and discrimination rates for the control group were calculated in order to better reflect the discriminatory power of the items in relation to baseline student knowledge. The difficulty index ranges between 0% and 100%, with the latter representing the lowest difficulty. Based on established guidelines in educational measurement [30,31], the ideal difficulty of items should be at a point on the difficulty scale midway between 100% and the chance level difficulty (50% for true-false items). This means the ideal difficulty index (P) value should be about 0.75 for dichotomous questions [31]. The discrimination index is the ability of an item to differentiate between students of higher and lower abilities and ranges between 0 and 1 [32]. The KR-20 reliability coefficient is a measure of reliability for a test with binary variables, ranging from 0 to 1, where 0 is no reliability and 1 is perfect reliability. In general, a score of more than 0.5 is usually considered reasonable.

A prompting strategy analysis was conducted within the chatbot group. For this purpose, all chat transcripts were exported and subsequently stored in pseudonymized form as PDF files. Based on previous publications, 2 independent raters evaluated each interaction using a 5-point Likert scale (1=strongly disagree, 2=disagree, 3=neutral, 4=agree, and 5=strongly agree) [33,34]. In cases of disagreement between the 2 raters, a consensus was achieved by reevaluation. This process ensured the achievement of a unanimous result while maintaining the integrity and objectivity of a dual-rater system. The scale was specifically developed in alignment with established prompting techniques, including role prompting, chain-of-thought prompting, knowledge-guided prompting, and output format specification. The following aspects were assessed: (1) the extent to which the student's prompt was focused, specific, and aligned with effective prompting strategies (input evaluation), and (2) the clarity and persuasiveness of the chatbot's response (output evaluation; Table 1). Furthermore, the Flesch Reading Ease was used to assess the readability and complexity of the answers provided by ChatGPT-4o, with results presented as a numerical value ranging from 1 to 100 [35].

Table 1. Predefined criteria for evaluating the input prompt and the output response given by the chatbot. For each applicable criterion, one point was assigned, and the corresponding value on the Likert scale was applied.

Evaluated element	Predefined criteria for chat evaluation
Input prompt	Focused question, inclusion of context or domain-specific details, precise requirements for the output (such as format), reasoning, reference to evidence
Output	Content-related, clear and persuasive, with justification, use of specific medical terms, and reference to guidelines

Statistical Analysis

As the data were not normally distributed, Kruskal-Wallis tests were applied to analyze potential differences between groups in terms of the percentage of correct responses and the total time required. Multiple post hoc pairwise comparisons were performed (using Dunn tests), and the resulting *P* values were adjusted using the Bonferroni-Holm method. To assess the internal agreement within each group, Fleiss κ coefficients were computed [36].

Regarding the prompting strategy, interrater agreement was calculated using weighted Cohen κ. Based on the evaluations of input and output quality, 2 regression analyses were carried out. (1) An ordinal regression model with a random intercept was applied to analyze the effect of the input prompt on the ordinal outcome variable output response. The student was included as a random effect to account for interindividual variability. Confidence intervals were computed using profile likelihood estimation. (2) A logistic regression model with a random intercept for the student was applied to examine the effect of the output response on the students' likelihood of selecting the correct dichotomous answer. Confidence intervals were estimated using profile likelihood methods.

All statistical analyses were performed using the software R (version 4.4.2; R Foundation for Statistical Computing). The predetermined α level was .05.

Results

The participant flow diagram is shown in Figure 1. Regarding the item analysis, the KR-20 reliability coefficient was 0.59. The results of the item difficulty index and the discrimination index are summarized in Table 2 and presented in greater detail in Table S2 in Multimedia Appendix 2.

Figure 1. Research design and participant allocation into the groups.

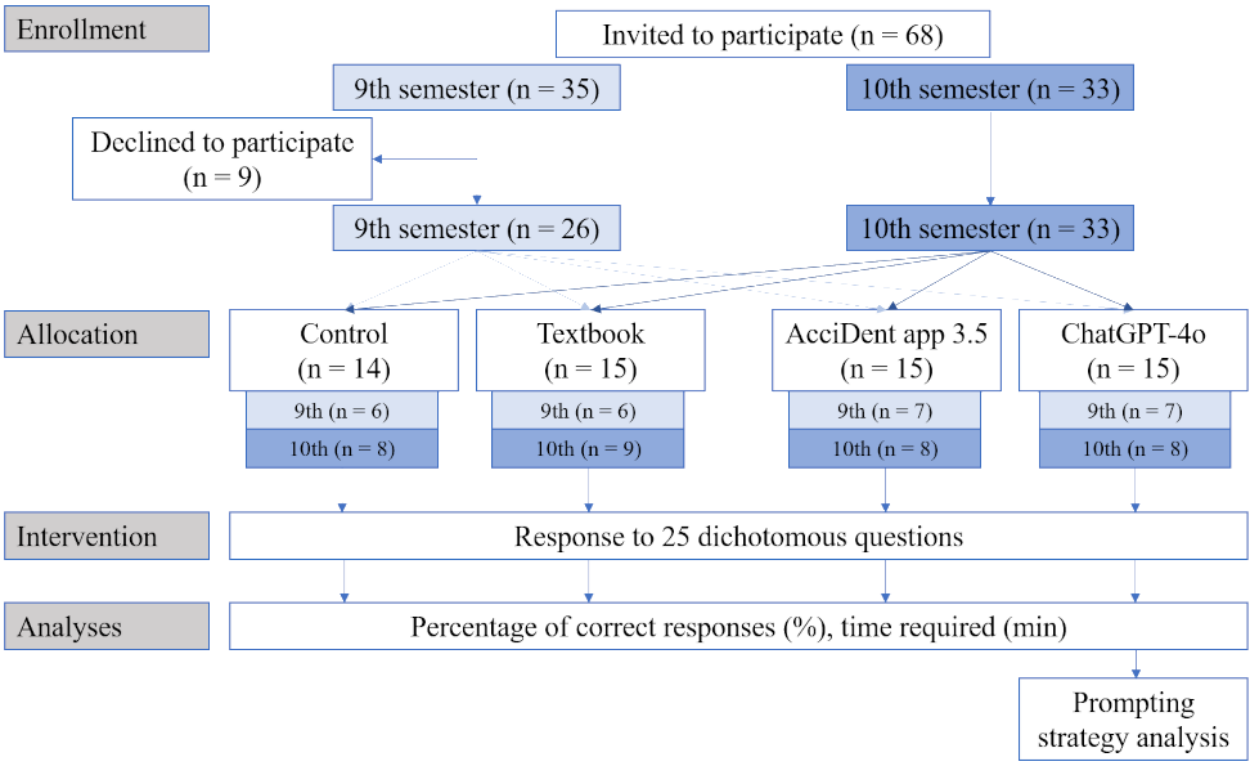


Table 2. Mean and SD and median and range of item difficulty and discrimination indices among all groups and for the control group across all 25 items.

	Total		Control group	
	Mean (SD)	Median (range)	Mean (SD)	Median (range)
Difficulty index	0.84 (0.16)	0.92 (0.47 to 1.0)	0.77 (0.25)	0.86 (0.29 to 1.0)
Discrimination index	0.18 (0.18)	0.12 (0.01 to 0.56)	0.20 (0.29)	0.07 (–0.20 to 1.0)

The control group achieved the lowest percentage of correct responses, with statistically significant differences compared with all 3 groups having used an information source ($P_{Textbook}$=.006, P_{App}=.02, $P_{Chatbot}$=.008; Figure 2, Table 3). In terms of time required, the control group demonstrated the fastest performance, when compared with all groups using a support tool ($P_{Textbook}$<.001, P_{App}<.001, $P_{Chatbot}$=.005). Of these, the textbook group required the most time on average, with significant differences to the AcciDent app group (P=.02) and the chatbot group (P=.001). The latter showed no significant difference between each other (P=.37). Internal agreement was highest within the chatbot group, followed by the AcciDent app group, both showing moderate agreement. (Figure 3 and Table 3).

Figure 2. Boxplots illustrating the percentage of correct responses among the 4 experimental groups. Small letters indicate statistically significant differences among groups.

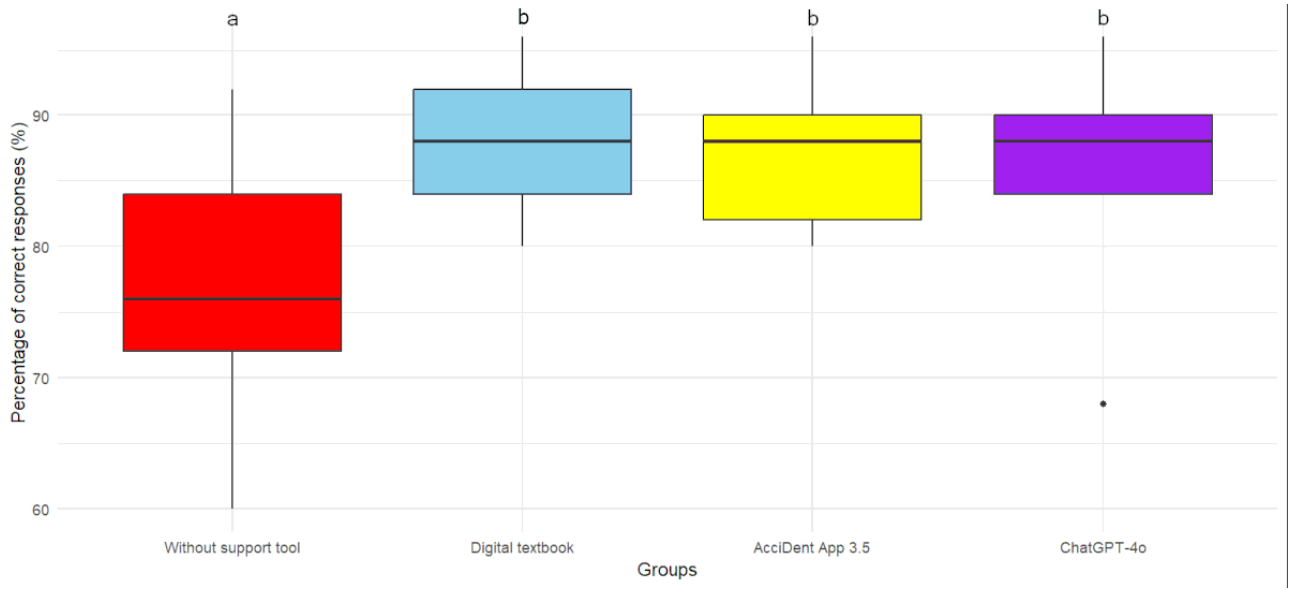
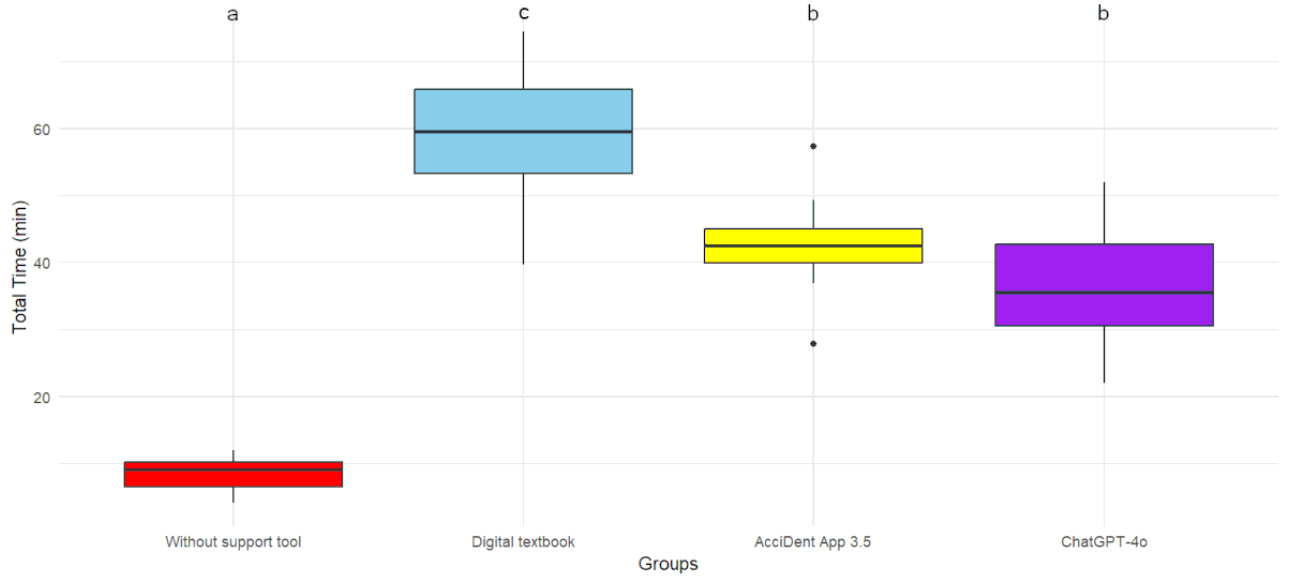


Table 3. Statistics regarding the percentage of correct answers and the total time required, as well as the Fleiss κ coefficient for the experimental groups. Small letters indicate statistically significant differences among groups.

	Percentage of correct responses		Time required (min)		Fleiss κ
	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)	
Control	76.57 (8.71)	76 (72-84)[a]	8.54 (2.45)	9.03 (6.55-10.20)[a]	0.28
Textbook	87.47 (5.63)	88 (84-92)[b]	59.60 (10.39)	59.50 (53.62-65.90)[c]	0.24
AcciDent app 3.5	86.40 (5.19)	88 (82-90)[b]	42.65 (6.44)	42.52 (39.89-45.03)[b]	0.41
ChatGPT-4o	86.40 (6.38)	88 (84-90)[b]	36.56 (10.03)	35.53 (30.58-42.78)[b]	0.47

[a,b,c]Statistically significant differences among groups.

Figure 3. Boxplots illustrating the time required to complete the questionnaire among the 4 experimental groups. Small letters indicate statistically significant differences among groups.



In the quality analysis, the input prompts posed by the students received a median rating of 2 (IQR 2-3, range 1-5), while the chatbot's output responses were rated with a median of 4 (IQR 3-4, range 1-5). Cohen κ indicated a high level of interrater agreement, with values of 0.98 and 0.82 for the input prompts and the output responses, respectively.

The ordinal regression analysis revealed a significant effect of the input prompt on the output response with an effect estimate of 0.3647 (P=.003). The estimated (odds ratio 1.44, 95% CI 1.13-1.83) suggests that for each unit increase in prompt precision, the odds of being in a higher category for the output response increased by 44%.

The logistic regression model demonstrated a significant effect of the chatbot's output response on the selection of the correct answer (effect estimate=0.6345, P=.002). The odds ratio of 1.89 (95% CI 1.26-2.80) indicates an 89% increase in the odds of correctness for each unit increase in clarity of the output response.

According to the Flesch Reading Ease, the answers given by ChatGPT-4o had a mean value of 36.20 (SD 9.13), classifying them as rather difficult to read.

Discussion

Choice of Support Tools and Principal Findings

Despite the availability of regularly updated clinical practice guidelines [5,6,9], studies report substantial variation in dentists' knowledge regarding appropriate treatment approaches [13,14], which may, at least in part, be attributed to insufficient education on this topic during undergraduate training. It is likely that such knowledge gaps are increasingly being addressed through the use of digital tools, as they offer rapid access to relevant information and align with modern learning habits. Accordingly, the primary aim of this study was to compare different support tools in terms of their accuracy in clinical decision-making situations. Considering the upward trend of digital health technologies while still accounting for students' preferences, we included 2 technologically advanced support tools (mobile app and chatbot) and a digital textbook representing a more traditional information source. To the best of our knowledge, this is the first study to compare different support tools specifically addressing emergency treatment questions related to dental trauma. Overall, the results indicated that the use of a support tool significantly increased accuracy compared with the control group, with no significant differences observed among the 3 sources. In recent years, the mobile health approach has gained increasing popularity, not only in daily clinical practice among health care professionals, but also as a means of delivering educational content [15,17]. In the field of dental traumatology, several mobile apps have been introduced. However, information regarding the dental trauma management according to the type of injury varied among the mobile apps [37]. Three of the apps were based on the widely accepted IADT guidelines, with the AcciDent app being the most popular and commonly used in Germany. Therefore, it was selected for inclusion in this study. In line with our results, previous studies have demonstrated that the use of a dental trauma app as a decision support tool significantly improves dental students` accuracy in answering multiple-choice questions related to the management of injured primary teeth, compared with a control group without access to informational resources [38,39].

Although this study indicated that the use of any of the evaluated support tools increases performance, the time required to get an answer differed significantly. Students using the mobile app and the chatbot completed the questionnaire within a mean time of 42.65 and 36.56 minutes, respectively, whereas those using the digital textbook required significantly more time (59.60 min). Moreover, it is important to note that the digital textbook included a keyword-based search function, which facilitates information retrieval. Consequently, it can be assumed that the use of a paper-based textbook would likely have resulted in even longer completion times.

Item Quality

For the purpose of the study, students of the final academic year were asked to answer 25 dichotomous questions. The item quality analysis revealed an internal consistency (KR-20) of 0.59. While this value is below the commonly accepted threshold of 0.70 for high reliability, it may still be considered acceptable in the context of this exploratory study. The value likely reflects the intentional heterogeneity of the item content, which was designed to cover a broad range of clinically relevant topics regarding the emergency treatment of traumatic dental injuries. Since the primary aim was not to measure a single underlying construct, but rather to assess the accuracy of different support tools across a variety of question types, a lower internal consistency is expected and does not compromise the interpretability of the results. In terms of discrimination appropriateness, a difficulty analysis was conducted. While many items in the full dataset fell into the "easy" category (P_{mean}=0.84, D_{mean}=0.18), several still demonstrated substantial variation between groups, underscoring their discriminatory utility at the group level. To better reflect the discriminatory power of the items in relation to baseline student knowledge, difficulty, and discrimination indices were also calculated exclusively for the control group without support tool. Within this subgroup, the values were more balanced (P_{mean}=0.77, D_{mean}=0.20), suggesting that item performance was more aligned with typical student knowledge and less confounded by tool-specific effects. Moreover, since the purpose of the assessment was not to achieve maximum discrimination among individuals but rather to evaluate the accuracy of information derived from different sources, the inclusion of items reflecting clinically relevant knowledge, even when classified as "easy," is justifiable.

Performance of LLM-Based Support Tools

Recent advances in AI have accelerated the development of LLMs [21]. As these models continue to develop and become more accessible, their integration into health care settings has expanded considerably. Numerous studies have investigated LLMs as an assistive tool in medicine and dentistry regarding various domains, including treatment recommendations [34], emergency treatments [40], radiological and clinical diagnostics [41-43], education [24,44-46], and decision-making [47,48]. In the context of dental traumatology, several chatbots have been examined with regard to their accuracy in answering topic-specific questions [26,49-52]. These studies either simulated patient-initiated queries or potential questions posed by dental professionals. Moreover, methodological differences among these studies are evident in the type of question analyzed: patient-initiated questions were predominantly open-ended or

case-related [49,50], while questions directed to professionals were assessed using multiple-choice or dichotomous formats [26,51,52]. With respect to the latter, reported accuracies for dichotomous questions varied widely (ranging from 10% to 80.81%) depending on the chatbot used (Google Bard, Google Gemini, Copilot F, Copilot P, ChatGPT-3.5, ChatGPT-4). In this study, ChatGPT-4o showed a mean accuracy of 86.40%. These discrepancies may also be explained by differences in the types of questions posed, as well as by continuous learning and training of the underlying models, which complicate direct comparisons between studies conducted at different time points.

Role of Prompt Design

It has been demonstrated that he accuracy of recommendations by LLMs is largely contingent upon the input's precision, correctness, and reasoning [24,34]. Various prompting strategies have been introduced and assessed to enhance the output response of LLMs [23,25,53,54]. These include prompting strategies like zero-shot, few-shot, or the thought-generation [23]. Further subtypes of prompting techniques have been described to enhance the precision and relevance of model responses. Role prompting, for example, instructs the model to adopt a specific perspective (such as responding to a clinical expert) in order to generate a domain-specific output. Chain-of-thought prompting encourages step-by-step reasoning to enhance the clarity of the response. Moreover, knowledge-guided prompting directs the model to rely on established guidelines or scientific sources, improving factual accuracy. Finally, the determination of the output format constrains the structure of the response, for example, by instructing the model to answer only with correct or false, or to respond in bullet points or tables [23,53]. Although prompt engineering is an emerging area of research, many users remain unaware of how the precision of a prompt can influence the model's output [34]. A recent review on the innovation and application of LLMs in dentistry noted that even in LLM-related research, only a few studies have examined the impact of the input prompt design on the output quality [38,55,56]. In this study, the quality of the input prompt and the output response was assessed using Likert scales based on predefined criteria derived from the prompting strategies described above. Our results indicated that most students were not aware of the relevance of prompt formulation, resulting in a significant influence of the prompt quality on the output response. Even though the integration of AI-assisted tools in educational settings and, therefore, students' familiarity with LLMs, may differ between countries, the importance of the input prompt can be considered universal. However, prompts should be locally adapted, since treatment guidelines or disease prevalence may vary across regions, which may otherwise lead to inappropriate treatment suggestions or implausible differential diagnoses.

Role of Digital Tools in Learning

Previous research has demonstrated that a substantial proportion of medical and dental students rely on digital learning tools, frequently favoring them over conventional textbooks [57]. This tendency is especially pronounced in the context of short-term exam preparation, where efficiency and accessibility are key considerations [58,59]. Consequently, it is highly likely that students prefer to use apps or other digital tools to address knowledge gaps in clinical situations rather than relying on traditional textbooks, particularly when immediate support is required.

Although the groups using ChatGPT-4o and the AcciDent app performed equally in terms of accuracy, it is worth noting that the content provided by a mobile app can be controlled by its developers, for example, by aligning it with international guidelines, which increases its reliability and safety. In contrast to mobile apps, which often require regular updates and may involve purchase costs, one advantage of the use of LLM-based chatbots is the continuous access to a broad and dynamically evolving source of information. Nevertheless, educators and students should be encouraged to critically evaluate the AI-generated feedback in order to develop their clinical reasoning and problem-solving skills, rather than relying unreflectively on the model's output. Excessive dependence on AI may otherwise impair the development of independent clinical judgment and practical skills [60]. Therefore, conventional textbooks remain essential, as they provide reliable fundamental knowledge to understand and internalize clinical concepts over the long term, and help to mitigate the risks and limitations associated with LLMs.

Moreover, adequate application of LLMs includes a precise prompt design, as this directly influences the clarity and accuracy of AI-generated responses. In medical and dental education, this ability can be regarded as a core component of LLM literacy, which is teachable and measurable [61,62]. Aligning this competency with established frameworks, such as digital health literacy or AI-related competencies in health professions education, reinforces its curricular and theoretical relevance [61-63]. Ultimately, an acquired, ideally curriculum-integrated, competence in using LLMs enables students and professionals to engage with these tools in a safe, informed, and reflective manner.

Strengths and Limitations

In general, the format of the question has a significant impact on chatbot performance [51]. In this study, the dichotomous format was chosen for its simplicity, ease of automated scoring, and efficiency in covering a broad content range within a limited timeframe. It also minimized the influence of distractor quality, which can significantly affect the validity of multiple-choice items. Poorly constructed distractors—such as implausible or irrelevant answer choices—can reduce item discrimination and compromise test reliability, particularly in AI-assisted educational settings [64,65]. However, the use of dichotomous questions is not without limitations. Prior research has shown that the performance of LLMs can vary significantly depending on the question type, with multiple-choice formats potentially offering a more nuanced reflection of clinical reasoning processes [66,67]. Moreover, true/false formats have a higher chance of guessing (50%), with limited possibility to reflect complex clinical scenarios, and a reduced ability to discriminate between different levels of understanding [68]. Future studies may benefit from incorporating multiple-choice or open-ended formats as well as clinical vignettes to better evaluate critical

thinking and complex decision-making skills, especially in the context of AI-assisted learning [69,70].

Regarding more complex question formats, different approaches have been explored. For instance, ChatGPT-3.5 was shown to respond adequately to patient-related advice-seeking questions across medical scenarios, although without providing appropriate or personalized advice [71]. In internal medicine, another study assessed the diagnostic accuracy and the ability of stating differential diagnoses of ChatGPT-3.5 and ChatGPT-4 using case vignettes. The authors highlighted the potential utility of LLMs as a supplementary tool [72]. This finding was supported by a similar study on orofacial pain [73]. However, although GPT-4 can augment diagnostic workflows, particularly in primary care or educational settings, it does not yet outperform clinical experts [73]. With regard to the impact of LLM use on physicians' diagnostic reasoning when working with clinical vignettes in internal, family, or emergency medicine, it was demonstrated that providing physicians with an LLM as a diagnostic aid did not significantly improve clinical reasoning compared with conventional resources [74]. Interestingly, however, the LLM alone performed significantly better when given standardized zero-shot prompts than when used as a supportive tool by physicians, highlighting the importance of input prompt quality [74]. As mentioned before, the results of this study are limited to a question format with a relatively high likelihood of guessing. Even within this "easier-to-answer" format, however, we were able to demonstrate the impact of prompt formulation. It is reasonable to assume that in more complex clinical case scenarios, the quality of the input prompt will become even more critical.

This study involved a relatively small sample of 59 final-year dental students, a limitation primarily dictated by the cohort size of the final academic year. While the sample size constrains generalizability to some extent, the decision to restrict participation to senior students was intentional. This group was presumed to represent students with adequate clinical knowledge to comprehend the clinical implications of traumatic dental injuries, despite not having formal education on the topic. Including only students at the end of their undergraduate education ensured a more uniform baseline of dental knowledge and minimized potential bias due to major curricular differences across earlier academic years. Although our results are limited to the included cohort, it can be assumed that all evaluated support tools are helpful in trauma-related questions or clinical situations. However, users with less clinical experience and prior knowledge may face a greater risk of relying on incorrect chatbot responses. For these groups in particular, precise formulation of input prompts and critical appraisal of AI-generated output will be especially important. Our results demonstrated that any of the support tools increased the students' performance; however, there were differences regarding the total time required. It can be assumed that users who were more familiar with a specific tool required less time, as its handling was easier and more intuitive for them. Students' familiarity with the tools was not assessed prior to the study, which should be considered a limitation. However, participants were randomly allocated to the groups to minimize this potential influence.

One of the strengths of the study is that the students' input entries to the chatbot were considered and evaluated. Using clearly defined criteria, we achieved a high interrater agreement (Cohen κ) for both the input and the output quality, demonstrating the straightforward and reproducible application of the predefined Likert scale. Unlike previous studies that assessed chatbot accuracy and consistency using identical prompts, this study accounted for both the interindividual variability in question formulation and decision-making based on chatbot recommendations, a factor often disregarded in earlier evaluations [26,51,52].

Conclusions

The present findings indicate that all evaluated support tools improved the students' performance when answering dental trauma questions. Both ChatGPT-4o and the AcciDent app demonstrated high accuracy while requiring less time than the digital textbook group. However, the performance of ChatGPT-4o depends on the precision and specificity of the input prompt, an aspect students should be made aware of when using it as an information source. In general, AI-based apps have the potential to transform dental education by enhancing learning experiences and supporting instructional practices. At the same time, educators and students should be encouraged to critically evaluate the AI-generated feedback in order to address the current risks and limitations associated with the use of LLMs.

Acknowledgments

A total of 20 textbooks were kindly provided by Quintessence Publishing to support the implementation of the study. The authors would like to thank Dr Fabian Kück, Department of Medical Statistics, University Medical Center Göttingen, Göttingen, for the valuable support with the statistical analysis. We acknowledge support by the Open Access Publication Funds of the Göttingen University.

Data Availability

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

Authors' Contributions

FH contributed to the conceptualization and formal analysis of the study and, together with TR and PL, participated in the investigation, methodology, and resource provision. FH and PL were responsible for project administration. FH led the writing

of the original draft, with TR and PL providing supporting contributions. All 3 authors—FH, TR, and PL—contributed equally to the review and editing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Authors' translation of the used questionnaire, which was originally distributed in German.
[DOCX File , 25 KB - mededu_v11i1e80576_app1.docx]

Multimedia Appendix 2
Detailed item analysis.
[DOCX File , 14 KB - mededu_v11i1e80576_app2.docx]

Multimedia Appendix 3
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 50005 KB - mededu_v11i1e80576_app3.pdf]

References

1. Glendor U. Aetiology and risk factors related to traumatic dental injuries--a review of the literature. Dent Traumatol 2009 Feb;25(1):19-31. [doi: 10.1111/j.1600-9657.2008.00694.x] [Medline: 19208007]
2. Petti S, Glendor U, Andersson L. World traumatic dental injury prevalence and incidence, a meta-analysis-one billion living people have had traumatic dental injuries. Dent Traumatol 2018 Apr;34(2):71-86. [doi: 10.1111/edt.12389] [Medline: 29455471]
3. Lam R. Epidemiology and outcomes of traumatic dental injuries: a review of the literature. Aust Dent J 2016 Mar;61 Suppl 1:4-20 [FREE Full text] [doi: 10.1111/adj.12395] [Medline: 26923445]
4. S2k-Leitlinie (Langfassung) - therapie des dentalen traumas bleibender zähne [S2k guideline: therapy of dental trauma to permanent teeth]. German Society of Oral and Maxillofacial Surgery (DGMKG). 2022. URL: https://register.awmf.org/de/leitlinien/detail/083-004 [accessed 2025-04-14]
5. Bourguignon C, Cohenca N, Lauridsen E, Flores MT, O'Connell AC, Day PF, et al. International association of dental traumatology guidelines for the management of traumatic dental injuries: 1. fractures and luxations. Dent Traumatol 2020 Aug;36(4):314-330 [FREE Full text] [doi: 10.1111/edt.12578] [Medline: 32475015]
6. Fouad AF, Abbott PV, Tsilingaridis G, Cohenca N, Lauridsen E, Bourguignon C, et al. International association of dental traumatology guidelines for the management of traumatic dental injuries: 2. avulsion of permanent teeth. Dent Traumatol 2020 Aug;36(4):331-342. [doi: 10.1111/edt.12573] [Medline: 32460393]
7. Saikia A, Patil SS, Ms M, Cv D, Sabarish R, Pandian S, et al. Systematic review of clinical practice guidelines for traumatic dental injuries. Dent Traumatol 2023;39(4):371-380. [doi: 10.1111/edt.12838] [Medline: 36920339]
8. Day PF, Gregg TA, Ashley P, Welbury RR, Cole BO, High AS, et al. Periodontal healing following avulsion and replantation of teeth: a multi-centre randomized controlled trial to compare two root canal medicaments. Dent Traumatol 2012;28(1):55-64. [doi: 10.1111/j.1600-9657.2011.01053.x] [Medline: 21988960]
9. Levin L, Day PF, Hicks L, O'Connell A, Fouad AF, Bourguignon C, et al. International association of dental traumatology guidelines for the management of traumatic dental injuries: general introduction. Dent Traumatol 2020;36(4):309-313 [FREE Full text] [doi: 10.1111/edt.12574] [Medline: 32472740]
10. Bücher K, Neumann C, Thiering E, Hickel R, Kühnisch J, International Association of Dental Traumatology. Complications and survival rates of teeth after dental trauma over a 5-year period. Clin Oral Investig 2013;17(5):1311-1318. [doi: 10.1007/s00784-012-0817-y] [Medline: 22886460]
11. Berlin-Broner Y, Kiani Z, Levin L. Dental trauma education among north american dental schools: results from multi-center interviews with dental educators. Dent Traumatol 2025. [doi: 10.1111/edt.13070] [Medline: 40364557]
12. O'Connell AC, Olegário IC. International teaching practices in dental trauma education. Dent Traumatol 2024;40(2):152-160. [doi: 10.1111/edt.12906] [Medline: 37915297]
13. Re D, Augusti D, Paglia G, Augusti G, Cotti E. Treatment of traumatic dental injuries: evaluation of knowledge among Italian dentists. Eur J Paediatr Dent 2014;15(1):23-28. [Medline: 24745588]
14. Zhao Y, Gong Y. Knowledge of emergency management of avulsed teeth: a survey of dentists in Beijing, China. Dent Traumatol 2010;26(3):281-284. [doi: 10.1111/j.1600-9657.2010.00877.x] [Medline: 20572844]
15. Walia T, Muthu MS, Saikia A, Anthonappa R, Satyanarayana MS. A systematic search, heuristic evaluation and analysis of dental trauma mobile applications. Dent Traumatol 2024;40(5):511-521. [doi: 10.1111/edt.12964] [Medline: 38651781]
16. Mobile technologies for oral health: an implementation guide. World Health Organization & International Telecommunication Union. 2021. URL: https://iris.who.int/handle/10665/345255 [accessed 2025-04-17]

17. Walia T, Muthu MS, Saikia A, Shetty RM, Anthonappa RP. Are mobile health applications for traumatic dental injuries effective? a systematic review of their impact on diagnosis, prevention, management, and education. Eur Arch Paediatr Dent 2025. [doi: 10.1007/s40368-025-01071-0] [Medline: 40560354]
18. Chau RCW, Thu KM, Yu OY, Hsung RT, Lo ECM, Lam WYH. Performance of generative artificial intelligence in dental licensing examinations. Int Dent J 2024;74(3):616-621 [FREE Full text] [doi: 10.1016/j.identj.2023.12.007] [Medline: 38242810]
19. Feher B, Tussie C, Giannobile WV. Applied artificial intelligence in dentistry: emerging data modalities and modeling approaches. Front Artif Intell 2024;7:1427517 [FREE Full text] [doi: 10.3389/frai.2024.1427517] [Medline: 39109324]
20. Turosz N, Chęcińska K, Chęciński M, Sielski M, Sikora M. Evaluation of dental panoramic radiographs by artificial intelligence compared to human reference: a diagnostic accuracy study. J Clin Med 2024;13(22):6859 [FREE Full text] [doi: 10.3390/jcm13226859] [Medline: 39598002]
21. Sufi F. Generative pre-trained transformer (GPT) in research: a systematic review on data augmentation. Information 2024;15(2):99. [doi: 10.3390/info15020099]
22. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. N Engl J Med 2023;388(13):1201-1208. [doi: 10.1056/NEJMra2302038] [Medline: 36988595]
23. Schulhoff S, Ilie M, Balepur N. The prompt report: a systematic survey of prompt engineering techniques. arXiv.2406.06608 2025. [doi: 10.5860/choice.189890]
24. Yan C, Li Z, Liang Y, Shao S, Ma F, Zhang N, et al. Assessing large language models as assistive tools in medical consultations for Kawasaki disease. Front Artif Intell 2025;8:1571503 [FREE Full text] [doi: 10.3389/frai.2025.1571503] [Medline: 40231209]
25. Wang L, Chen X, Deng X, Wen H, You M, Liu W, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. NPJ Digit Med 2024;7(1):41 [FREE Full text] [doi: 10.1038/s41746-024-01029-4] [Medline: 38378899]
26. Ozden I, Gokyar M, Ozden ME, Sazak Ovecoglu H. Assessment of artificial intelligence applications in responding to dental trauma. Dent Traumatol 2024;40(6):722-729. [doi: 10.1111/edt.12965] [Medline: 38742754]
27. Mustuloğlu Ş, Deniz BP. Evaluation of chatbots in the emergency management of avulsion injuries. Dent Traumatol 2025;41(4):437-444. [doi: 10.1111/edt.13041] [Medline: 39865377]
28. Khanagar SB, Al-Ehaideb A, Vishwanathaiah S, Maganur PC, Patil S, Naik S, et al. Scope and performance of artificial intelligence technology in orthodontic diagnosis, treatment planning, and clinical decision-making - a systematic review. J Dent Sci 2021;16(1):482-492 [FREE Full text] [doi: 10.1016/j.jds.2020.05.022] [Medline: 33384838]
29. Krastl G, Weiger R, Filippi A. Zahntrauma - Therapieoptionen für die Praxis [dental trauma – treatment options for clinical practice]. Berlin: Quintessenz Publishing; 2020.
30. Downing S, Haladyna T. Handbook of Test Development. New Jersey: Lawrence Erlbaum Associates; 2006.
31. Ebel R, Frisbie D. Essentials of Educational Measurement. 5th Edition. Englewood Cliffs: Prentice-Hall; 1991.
32. Date AP, Borkar AS, Badwaik RT, Siddiqui RA, Shende TR, Dashputra AV. Item analysis as tool to validate multiple choice question bank in pharmacology. Int J Basic Clin Pharmacol 2019;8(9):1999. [doi: 10.18203/2319-2003.ijbcp20194106]
33. Hadjiathanasiou A, Goelz L, Muhn F, Heinz R, Kreißl L, Sparenberg P, et al. Artificial intelligence in neurovascular decision-making: a comparative analysis of ChatGPT-4 and multidisciplinary expert recommendations for unruptured intracranial aneurysms. Neurosurg Rev 2025;48(1):261. [doi: 10.1007/s10143-025-03341-3] [Medline: 39982556]
34. Truhn D, Weber CD, Braun BJ, Bressem K, Kather JN, Kuhl C, et al. A pilot study on the efficacy of GPT-4 in providing orthopedic treatment recommendations from MRI reports. Sci Rep 2023;13(1):20159 [FREE Full text] [doi: 10.1038/s41598-023-47500-2] [Medline: 37978240]
35. Flesch R. A new readability yardstick. J Appl Psychol 1948;32(3):221-233. [doi: 10.1037/h0057532] [Medline: 18867058]
36. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282 [FREE Full text] [Medline: 23092060]
37. Loureiro JM, Jural LA, Soares TRC, Risso PA, Fonseca-Gonçalves A, Magno MB, et al. Critical appraisal of the information available on traumatic dental injuries found in applications. Dent Traumatol 2022;38(1):77-87. [doi: 10.1111/edt.12715] [Medline: 34698435]
38. Huh AJH, Chen J, Bakland L, Goodacre C. Comparison of different clinical decision support tools in aiding dental and medical professionals in managing primary dentition traumatic injuries. Pediatr Emerg Care 2022;38(2):e534-e539. [doi: 10.1097/PEC.0000000000002409] [Medline: 34009888]
39. Machado JP, Lam XT, Chen J. Use of a clinical decision support tool for the management of traumatic dental injuries in the primary dentition by novice and expert clinicians. Dent Traumatol 2018;34(2):120-128. [doi: 10.1111/edt.12390] [Medline: 29476702]
40. Günay S, Öztürk A, Özerol H, Yiğit Y, Erenler AK. Comparison of emergency medicine specialist, cardiologist, and chat-GPT in electrocardiography assessment. Am J Emerg Med 2024;80:51-60. [doi: 10.1016/j.ajem.2024.03.017] [Medline: 38507847]

41. Dehdab R, Brendlin A, Werner S, Almansour H, Gassenmaier S, Brendel JM, et al. Evaluating ChatGPT-4V in chest CT diagnostics: a critical image interpretation assessment. Jpn J Radiol 2024;42(10):1168-1177. [doi: 10.1007/s11604-024-01606-3] [Medline: 38867035]
42. Mendonça de Moura JD, Fontana CE, Reis da Silva Lima VH, de Souza Alves I, André de Melo Santos P, de Almeida Rodrigues P. Comparative accuracy of artificial intelligence chatbots in pulpal and periradicular diagnosis: a cross-sectional study. Comput Biol Med 2024;183:109332. [doi: 10.1016/j.compbiomed.2024.109332] [Medline: 39471663]
43. de Araujo BMDM, de Jesus Freitas PF, Deliga Schroder AG, Küchler EC, Baratto-Filho F, Ditzel Westphalen VP, et al. PAINe: An artificial intelligence-based virtual assistant to aid in the differentiation of pain of odontogenic versus temporomandibular origin. J Endod 2024;50(12):1761-1765.e2. [doi: 10.1016/j.joen.2024.09.008] [Medline: 39342988]
44. Elkarmi R, Abu-Ghazaleh S, Sonbol H, Haha O, Al-Haddad A, Hassona Y. ChatGPT for parents' education about early childhood caries: a friend or foe? Int J Paediatr Dent 2025;35(4):717-724. [doi: 10.1111/ipd.13283] [Medline: 39533165]
45. Danesh A, Pazouki H, Danesh F, Danesh A, Vardar-Sengul S. Artificial intelligence in dental education: ChatGPT's performance on the periodontic in-service examination. J Periodontol 2024;95(7):682-687. [doi: 10.1002/JPER.23-0514] [Medline: 38197146]
46. Uribe SE, Maldupa I, Kavadella A, El Tantawi M, Chaurasia A, Fontana M, et al. Artificial intelligence chatbots and large language models in dental education: worldwide survey of educators. Eur J Dent Educ 2024;28(4):865-876. [doi: 10.1111/eje.13009] [Medline: 38586899]
47. Özbay Y, Erdoğan D, Dinçer GA. Evaluation of the performance of large language models in clinical decision-making in endodontics. BMC Oral Health 2025;25(1):648 [FREE Full text] [doi: 10.1186/s12903-025-06050-x] [Medline: 40296000]
48. Lee JK, Choi S, Park S, Hwang S, Cho D. Evaluation of six large language models for clinical decision support: application in transfusion decision-making for rhd blood-type patients. Ann Lab Med 2025;45(5):520-529 [FREE Full text] [doi: 10.3343/alm.2024.0588] [Medline: 40289855]
49. Guven Y, Ozdemir OT, Kavan MY. Performance of artificial intelligence chatbots in responding to patient queries related to traumatic dental injuries: a comparative study. Dent Traumatol 2025;41(3):338-347. [doi: 10.1111/edt.13020] [Medline: 39578674]
50. Johnson AJ, Singh TK, Gupta A, Sankar H, Gill I, Shalini M, et al. Evaluation of validity and reliability of AI Chatbots as public sources of information on dental trauma. Dent Traumatol 2025;41(2):187-193. [doi: 10.1111/edt.13000] [Medline: 39417352]
51. Kuru HE, Aşık A, Demir DM. Can artificial intelligence language models effectively address dental trauma questions? Dent Traumatol 2025;41(5):567-580. [doi: 10.1111/edt.13063] [Medline: 40170270]
52. Portilla ND, Garcia-Font M, Nagendrababu V, Abbott PV, Sanchez JAG, Abella F. Accuracy and consistency of gemini responses regarding the management of traumatized permanent teeth. Dent Traumatol 2025;41(2):171-177. [doi: 10.1111/edt.13004] [Medline: 39460511]
53. Li J, Deng Y, Sun Q, Zhu J, Tian Y, Li J, et al. Benchmarking large language models in evidence-based medicine. IEEE J Biomed Health Inform 2025;29(9):6143-6156. [doi: 10.1109/JBHI.2024.3483816] [Medline: 39437276]
54. Sivarajkumar S, Kelley M, Samolyk-Mazzanti A, Visweswaran S, Wang Y. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: algorithm development and validation study. JMIR Med Inform 2024;12:e55318 [FREE Full text] [doi: 10.2196/55318] [Medline: 38587879]
55. Umer F, Batool I, Naved N. Innovation and application of large language models (LLMs) in dentistry - a scoping review. BDJ Open 2024;10(1):90 [FREE Full text] [doi: 10.1038/s41405-024-00277-6] [Medline: 39617779]
56. Russe M, Rau A, Ermer M, Rothweiler R, Wenger S, Klöble K, et al. A content-aware chatbot based on GPT 4 provides trustworthy recommendations for Cone-Beam CT guidelines in dental imaging. Dentomaxillofac Radiol 2024;53(2):109-114 [FREE Full text] [doi: 10.1093/dmfr/twad015] [Medline: 38180877]
57. Bjurström MF, Lundkvist E, Sturesson LW, Borgquist O, Lundén R, Fagerlund MJ, et al. Digital learning resource use among Swedish medical students: insights from a nationwide survey. BMC Med Educ 2025;25(1):849 [FREE Full text] [doi: 10.1186/s12909-025-07446-7] [Medline: 40500719]
58. Al Shmanee M, Issa M, Alkholy H, Alnaqbi A, Awadallah A, Hassan H, et al. Medical students' preferences of study resources: physical vs digital resources. Cureus 2024;16(3):e56196 [FREE Full text] [doi: 10.7759/cureus.56196] [Medline: 38618352]
59. Scott K, Morris A, Marais B. Medical student use of digital learning resources. Clin Teach 2018;15(1):29-33. [doi: 10.1111/tct.12630] [Medline: 28300343]
60. Ali K, Barhom N, Tamimi F, Duggal M. ChatGPT-A double-edged sword for healthcare education? implications for assessments of dental students. Eur J Dent Educ 2024;28(1):206-211. [doi: 10.1111/eje.12937] [Medline: 37550893]
61. Claman D, Sezgin E. Artificial intelligence in dental education: opportunities and challenges of large language models and multimodal foundation models. JMIR Med Educ 2024;10:e52346 [FREE Full text] [doi: 10.2196/52346] [Medline: 39331527]
62. Vrdoljak J, Boban Z, Vilović M, Kumrić M, Božić J. A review of large language models in medical education, clinical decision support, and healthcare administration. Healthcare (Basel) 2025;13(6):603 [FREE Full text] [doi: 10.3390/healthcare13060603] [Medline: 40150453]

63. Wamala Andersson S, Gonzalez MP. Digital health literacy-a key factor in realizing the value of digital transformation in healthcare. Front Digit Health 2025;7:1461342 [FREE Full text] [doi: 10.3389/fdgth.2025.1461342] [Medline: 40538571]
64. Haladyna TM, Downing SM. How many options is enough for a multiple-choice test item? Educational and Psychological Measurement 1993;53(4):999-1010. [doi: 10.1177/0013164493053004013]
65. Tarrant M, Ware J. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. Med Educ 2008;42(2):198-206. [doi: 10.1111/j.1365-2923.2007.02957.x] [Medline: 18230093]
66. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How Does ChatGPT perform on the united states medical licensing examination (USMLE)? the implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
67. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
68. Brady A. Assessment of learning with multiple-choice questions. Nurse Educ Pract 2005;5(4):238-242. [doi: 10.1016/j.nepr.2004.12.005] [Medline: 19038205]
69. Burton RF. Multiple‐choice and true/false tests: myths and misapprehensions. Assess Eval High Educ 2005;30(1):65-72. [doi: 10.1080/0260293042003243904]
70. Patel VL, Groen GJ, Arocha JF. Medical expertise as a function of task difficulty. Mem Cognit 1990;18(4):394-406. [doi: 10.3758/bf03197128] [Medline: 2381318]
71. Nastasi AJ, Courtright KR, Halpern SD, Weissman GE. A vignette-based evaluation of ChatGPT's ability to provide appropriate and equitable medical advice across care contexts. Sci Rep 2023;13(1):17885 [FREE Full text] [doi: 10.1038/s41598-023-45223-y] [Medline: 37857839]
72. Hirosawa T, Kawamura R, Harada Y, Mizuta K, Tokumasu K, Kaji Y, et al. ChatGPT-generated differential diagnosis lists for complex case-derived clinical vignettes: diagnostic accuracy evaluation. JMIR Med Inform 2023;11:e48808 [FREE Full text] [doi: 10.2196/48808] [Medline: 37812468]
73. Vueghs C, Shakeri H, Renton T, Van der Cruyssen F. Development and evaluation of a gpt4-based orofacial pain clinical decision support system. Diagnostics (Basel) 2024;14(24):2835 [FREE Full text] [doi: 10.3390/diagnostics14242835] [Medline: 39767196]
74. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open 2024;7(10):e2440969 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.40969] [Medline: 39466245]

Abbreviations

AI: artificial intelligence
GPT: generative pretrained transformer
IADT: International Association of Dental Traumatology
KR-20: Kuder Richardson formula 20
LLM: large language model

Edited by A Stone; submitted 18.07.25; peer-reviewed by K Bitter, AL Hillebrecht, J Schmidt; comments to author 12.09.25; revised version received 30.09.25; accepted 07.10.25; published 20.11.25.

Please cite as:
Haupt F, Rödig T, Liersch P
Evaluating ChatGPT-4o as an Educational Support Tool for the Emergency Management of Dental Trauma: Randomized Controlled Study Among Students
JMIR Med Educ 2025;11:e80576
URL: https://mededu.jmir.org/2025/1/e80576
doi:10.2196/80576
PMID:

Original Paper

Assessing Pharmacists' Use and Perception of AI Chatbots in Pharmacy Practice: Cross-Sectional Survey Study

Anly Li[1], PharmD; Amy Heck Sheehan[2], PharmD; Christopher Giuliano[3], PharmD, MPH; Paul Dobry[4], PharmD; Paul Walker[5], PharmD; Jennifer Philips[6], PharmD; Joseph Jordan[7], PharmD

[1]Regulatory Pharmaceutical Fellow in Drug Information, Purdue University College of Pharmacy, West Lafayette, IN, United States

[2]Professor of Pharmacy Practice, Purdue University College of Pharmacy, West Lafayette, IN, United States

[3]Clinical Professor of Pharmacy Practice, Wayne State University Eugene Applebaum College of Pharmacy and Health Sciences, Detroit, MI, United States

[4]Clinical Pharmacy Specialist, Drug Information, Detroit Receiving Hospital, Detroit, MI, United States

[5]Clinical Professor Emeritus, University of Michigan College of Pharmacy, Ann Arbor, MI, United States

[6]Clinical Professor of Pharmacy Practice, Director of Drug Information Group, University of Illinois Chicago College of Pharmacy, Chicago, IL, United States

[7]Professor of Pharmacy Practice, Butler University College of Pharmacy and Health Sciences, Indianapolis, IN, United States

Corresponding Author:
Anly Li, PharmD
Regulatory Pharmaceutical Fellow in Drug Information
Purdue University College of Pharmacy
575 Stadium Mall Dr
West Lafayette, IN, 47907
United States
Phone: 1 7347943680
Email: anlyl@umich.edu

Abstract

Background: The use of artificial intelligence (AI)–based large language model chatbots such as ChatGPT has become increasingly popular in many disciplines. However, concerns exist regarding ethics, legal considerations, accuracy, and reproducibility with its use in health care practice, education, and research.

Objective: This study aimed to assess current perceptions and use of AI chatbots in pharmacy practice from the perspective of pharmacist preceptors and determine factors that may influence the use of AI chatbots in practice.

Methods: A cross-sectional survey of pharmacy practice preceptors from Indiana, Illinois, and Michigan was conducted using the validated Technology Acceptance Model Edited to Assess ChatGPT Adoption (TAME-ChatGPT) survey tool to collect information regarding current use of AI chatbots and factors associated with use, including ease of use, perceived risk, technology or social influences, anxiety, and perceived usefulness.

Results: A total of 194 responses (194/1877, 10.34% response rate) were received. Approximately one-third (n=59, 30.4%) of respondents reported having used an AI chatbot, with 51.5% (n=100) indicating that they planned to start or would continue using chatbots in the future. In practice, common uses for AI chatbots included summarizing information (n=90, 46.4%), letter of recommendation writing (n=64, 32.9%), and obtaining disease state information (n=63, 32.5%). The 2 main constructs associated with the use of chatbots identified from the TAME-ChatGPT tool included perceived risk of using AI and attitude toward AI. Factors that predicted pharmacists' current use of AI chatbots included positive attitude toward technology (odds ratio [OR] 3.64, 95% CI 2.08-6.36), coworker use of AI (OR 7.41, 95% CI 2.64-20.8), and working in academia (OR 5.62, 95% CI 1.30-24.23).

Conclusions: Most pharmacist respondents had not used an AI chatbot and were unlikely to make patient care decisions based on information from a chatbot. The TAME-ChatGPT survey is validated for assessing chatbot use and attitudes among pharmacists, and future studies using this survey tool can guide the implementation of chatbots into pharmacy practice.

(JMIR Med Educ 2025;11:e71767) doi:10.2196/71767

KEYWORDS

chatbot; artificial intelligence; pharmacy; health care practice; survey; technology

Introduction

Artificial Intelligence Chatbots

In November 2022, the artificial intelligence (AI)–based large language model chatbot ChatGPT was launched [1]. It quickly gained popularity, amassing over 100 million monthly users in just 2 months [2]. Since the release of ChatGPT, several other AI chatbots have debuted, including Gemini, Microsoft Bing AI, and Copilot. This expansion in AI chatbots has led to the diversification of their use in fields outside of the technology industry. ChatGPT has been explored for use in health care practice, education, and research, and specialized chatbots such as OpenEvidence and Dougall GPT have emerged, catering specifically to clinicians and health care workers [3-6].

Use of Artificial Intelligence in Healthcare

A survey evaluating the use and perceptions of ChatGPT in health care professionals was conducted at Northwestern University [7]. In addition to gathering information regarding applications of ChatGPT in health care, the survey focused on perceptions related to use in health care research, education, and practice. Participants reported uncertainty about the use of ChatGPT due to its recent release, although many respondents indicated an interest in future use. Similarly, in a survey of health care workers in Saudi Arabia, participants expressed concerns about ChatGPT use in health care, including lack of credibility and concerns about inaccurate medical information [8]. Despite these concerns, most respondents still anticipated a positive impact of ChatGPT on the future of health care practice, including medical decision-making, patient and family support, and medical research appraisal. Studies conducted in Jordan, Saudi Arabia, and the United Arab Emirates have reported similar findings among pharmacists [9-11].

These findings were reinforced in a 2023 systematic review that examined 60 studies focusing on the applications of and concerns about ChatGPT in health care practice, education, and research [6]. Potential uses for ChatGPT in practice included clinical documentation, creation of personalized care plans, improved health literacy, and provision of patient education. In education, ChatGPT was used in the creation of personalized learning tools and writing clinical cases. In research, ChatGPT was found to be beneficial in improving writing efficiency. However, in nearly all the studies, there were reported concerns with ChatGPT, including ethical, legal, and copyright issues; inaccuracy; and limited reproducibility. Specific to pharmacy, the use of ChatGPT has been explored with respect to drug information, medication therapy management, patient education, and adverse drug reaction assessment, with the general consensus that it has potential as a supplementary tool but is not yet able to handle complex problems [12-15].

On the basis of the findings from previously conducted surveys and research studies, it is evident that there is current and future potential for the use of AI technology in health care. However, information about the opinions and perceptions of pharmacists regarding the utility of AI chatbots in practice is limited. Pharmacists have been poorly represented in most surveys of health professionals conducted to date, and most studies regarding pharmacist perceptions have been conducted outside of the United States. Understanding the current perceptions, use, and barriers to use of chatbots will help inform and identify future roles for chatbots in pharmacy practice. Pharmacy preceptors are at the forefront of practice and education in the United States; therefore, this study aimed to assess current perceptions and use of AI chatbots in pharmacy practice from the perspective of pharmacist preceptors and determine factors that may influence the use of AI chatbots according to the Technology Acceptance Model Edited to Assess ChatGPT Adoption (TAME-ChatGPT) tool [16].

Methods

Overview

A cross-sectional survey was conducted using a convenience sample of pharmacy practice preceptors from Purdue University College of Pharmacy, University of Michigan College of Pharmacy, Wayne State University Eugene Applebaum College of Pharmacy and Health Sciences, University of Illinois Chicago Retzky College of Pharmacy, and Butler University College of Pharmacy and Health Sciences. Current pharmacist preceptors were recruited through their respective offices of experiential education. Pharmacists were excluded from the study if they had not precepted students in the previous year given the recent introduction of AI chatbots or if they did not practice in one of the affiliated states of Illinois, Indiana, or Michigan. Participants were instructed to only take the survey once.

Questionnaire Design and Administration

The survey instrument was built in Qualtrics XM (Qualtrics International Inc) and was primarily based on the TAME-ChatGPT survey tool, a validated tool adapted from the technology acceptance model (TAM) for assessing health care students' attitudes toward ChatGPT [16]. Before the first survey item, respondents were asked a screening question to ensure that they met the study inclusion criteria. The first section gathered information regarding pharmacists' practice setting and current use of AI chatbots. The second section contained questions from the TAME-ChatGPT that were adapted to detect an association with different factors of the TAM (perceived ease of use, perceived risk, technology or social influence, anxiety, perceived usefulness, and behavior) and the use of AI chatbots. The survey concluded with questions regarding demographic information and tasks in which preceptors would recommend the use of AI. The instrument was pilot-tested by 3 faculty members to ensure adequate formatting, comprehension, clarity, and completeness of the survey. The final survey instrument is available in Multimedia Appendix 1.

An email invitation to complete the survey instrument was sent to 1877 pharmacy preceptors on February 12, 2024. The survey link was accessible until April 22, 2024, and 3 reminder emails were sent out before the survey closed. Only fully completed surveys were included in the final analysis.

Data Analysis

Data were described using means and SDs for continuous variables and medians and frequencies for nominal variables. Univariable analysis was conducted using the student 2-tailed *t* test for continuous variables and chi-square test for nominal

variables. Normality was assessed using visual inspection of *Q*-*Q* plots.

An exploratory factor analysis was conducted to assess construct validity of the TAME-ChatGPT instrument. First, correlation matrices, the Bartlett test, and the Kaiser-Meyer-Olkin measure of sampling adequacy were assessed to ensure that proceeding with factor analysis was appropriate. Multicollinearity was assessed by examining the determinant of the correlation in the correlation matrix with a <0.00001 cutoff. If multicollinearity was present, items with *r* values greater than 0.8 were considered for removal. The number of factors included was based on performing parallel analysis in which an eigenvalue cutoff was determined from randomly generated correlation matrices and then compared with eigenvalues from the data. Factors were retained if the dataset eigenvalue was greater than the corresponding random eigenvalue. Promax rotation was used to allow for correlation of factors. The Cronbach α was calculated to assess internal consistency.

Multivariable analysis was conducted using logistic regression to assess the association between preceptor characteristics and the components of the TAME-ChatGPT with the outcome of use of AI or future use of AI. Variables included the TAME-ChatGPT constructs and factors that significantly predicted current or future AI use. The number of variables included in the model was limited based on the number of participants who responded that they used AI. Goodness of fit was evaluated using the Hosmer-Lemeshow test (*P*>.05 indicating model fit) along with −2 log likelihood ratio, with smaller values indicating improved fit. As a secondary analysis, the model's ability to discriminate between those who reported AI use and those who did not was evaluated using a receiver operating characteristic curve. The area under the curve was used as a summary measure of the model's discrimination. SPSS (version 29; IBM Corp) was used to conduct the data analysis. A *P* value of <.05 was considered statistically significant.

Ethical Considerations

This study was determined to be exempt research by the institutional review boards at all participating institutions—Butler University, Purdue University, University of Illinois Chicago, University of Michigan, and Wayne State University. The survey responses were anonymous and confidential, and all responses were stored without any identifiers. After survey completion, respondents were directed to an optional link to receive a US $10 Amazon gift card as compensation for participation.

Results

Overview

A total of 235 responses were received. Responses were excluded from the analysis if they were incomplete (n=10, 4.3%); if the respondent practiced outside of Indiana, Illinois, or Michigan (n=4, 1.7%); or if the respondent had not been a preceptor in the previous year (n=27, 11.5%). The final response number was 194, which represented an overall response rate of 10.3% (194/1877). Respondent demographics can be found in Table 1, and they were generally representative of pharmacy preceptors with the exception of years and area of practice based on data from the American Association of Colleges of Pharmacy 2024 preceptor survey [17]. Of the 194 responses that met the inclusion criteria, 59 (30.4%) indicated that the respondents had used an AI chatbot before. A total of 51.5% (100/194) of the respondents indicated that they would continue or plan to start using chatbots in the future.

Among those who had used AI chatbots before and those who had not, demographics were similar with respect to age, gender, and years of practice (Table 2). Most respondents (146/194, 75.3%) indicated that they were unlikely to make a patient care decision based on information from a chatbot. Of the respondents who had used AI chatbots (59/194, 30.4%), there was a substantially higher percentage who practiced in academia and a lower percentage who practiced in the community setting than among those who had not used AI chatbots. Additionally, respondents who had coworkers who used AI chatbots or an institutional AI policy were more likely to have used an AI chatbot themselves.

ChatGPT was the most frequently reported chatbot used (55/59, 93%), followed by Bing Chat (10/59, 17%) and Google Bard, now Gemini (7/59, 12%). The most common uses for AI chatbots in practice included summarizing information (31/59, 53%), letter of recommendation writing (20/59, 34%), and obtaining disease state information (14/59, 24%). Among respondents who had not used AI chatbots in practice before (135/194, 69.6%), 63% (85/135) selected not knowing how to use them effectively as the reason for disuse. Other common reasons for disuse included preference for other resources (80/135, 59.3%), lack of credibility or trust (63/135, 46.7%), and concerns of plagiarism (43/135, 31.9%).

All 194 respondents received the survey including questions about the respondents' recommendations for use of AI chatbots by pharmacists and pharmacy students in practice. The top recommendations for use for pharmacists included administrative purposes (92/194, 47.4%), summarizing information (90/194, 46.4%), creating meeting agendas (74/194, 38.1%), letter of recommendation writing (64/194, 33%), and obtaining disease state information (63/194, 32.5%); 15.5% (30/194) of respondents indicated that they would not recommend that pharmacists use AI chatbots in practice. In comparison, the top uses recommended for students were summarizing information (61/194, 31.4%), creating meeting agendas (45/194, 23.2%), obtaining disease state information (42/194, 21.6%), administrative purposes (42/194, 21.6%), and conducting literature searches (38/194, 19.6%). In total, 40.2% (78/194) of the respondents indicated that they would not recommend that students use AI chatbots in practice.

Table 1. Demographics of pharmacy preceptor respondents.

	Survey respondents (n=194), n (%)	Respondents to the American Association of Colleges of Pharmacy survey [17] (n=4739), n (%)
Gender		
Woman	142 (73.2)	2609 (56.6)
Man	48 (24.7)	1665 (36.1)
Nonbinary or third gender	1 (0.5)	NR[a]
Preferred not to self-describe	3 (1.5)	327 (7.1)
Age (years), mean (SD)	38.7 (9.4)	NR
Years of practice		
0-5	55 (28.4)	588 (12.6)
6-10	51 (26.3)	951 (20.4)
11-15	32 (16.5)	922 (19.8)
>15	56 (28.9)	2097 (45.1)
Area of practice		
Hospital	112 (57.7)	1521 (32.9)
Community	29 (14.9)	1251 (27.1)
Academia	16 (8.2)	172 (3.7)
Ambulatory	16 (8.2)	717 (15.5)
Drug information	8 (4.1)	NR
Managed care	7 (3.6)	85 (1.8)
Long-term care	3 (1.5)	NR
Industry	1 (0.5)	88 (1.9)
Other	14 (7.2)	604 (13.1)
Specialty	6 (3.1)	NR
Consulting	1 (0.5)	NR
Infusion	1 (0.5)	NR
Research	1 (0.5)	NR
Medication safety	1 (0.5)	NR

[a]NR: not reported.

Table 2. Comparison of demographics between respondents who had and had not used artificial intelligence (AI) chatbots.

	Had used AI chatbots (n=59), n (%)	Had not used AI chatbots (n=135), n (%)	*P* value
Age (years), mean (SD)	38.7 (9.1)	38.7 (9.5)	>.99
Gender			.40
Woman	39 (66.1)	103 (76.3)	
Man	19 (32.2)	29 (21.5)	
Nonbinary or third gender	0 (0.0)	1 (0.7)	
Preferred not to self-describe	1 (1.7)	2 (1.5)	
Years of practice			.41
0-5	15 (25.4)	40 (29.6)	
6-10	16 (27.1)	62 (45.9)	
11-15	11 (18.6)	21 (15.6)	
>15	17 (28.8)	39 (28.9)	
Area of practice			.002
Hospital	38 (64.4)	74 (54.8)	
Community	4 (6.8)	25 (18.5)	
Ambulatory	12 (20.3)	24 (17.8)	
Academia	11 (18.6)	5 (3.7)	
Specialty	0 (0.0)	6 (4.4)	
Coworker use of AI chatbots			<.001
Yes	25 (42.4)	9 (6.7)	
No	9 (15.3)	69 (51.1)	
Unsure	25 (42.4)	57 (42.2)	
AI policy at practice site			.02
Yes	11 (18.6)	8 (5.9)	
No	30 (50.8)	82 (60.7)	
Unsure	18 (30.5)	45 (33.3)	
"How likely would you be to make a health care (ie, patient care or treatment) related recommendation based on the information an AI Chatbot (eg, ChatGPT) provides you?"			.33
Extremely unlikely	32 (54.2)	57 (42.2)	
Somewhat unlikely	16 (27.1)	41 (30.4)	
Neither likely nor unlikely	9 (15.3)	25 (18.5)	
Somewhat likely	2 (3.4)	12 (8.9)	
Extremely likely	0 (0.0)	0 (0.0)	
"How likely would you be to make a policy related decision based on the information an AI Chatbot (eg, ChatGPT) provides you?"			.25
Extremely unlikely	17 (28.8)	49 (36.3)	
Somewhat unlikely	15 (25.4)	37 (27.4)	
Neither likely nor unlikely	16 (27.1)	34 (25.2)	
Somewhat likely	11 (18.6)	12 (8.9)	
Extremely likely	0 (0.0)	3 (2.2)	

Factors Associated With AI Use

The first 13 items of the TAME-ChatGPT were included in the primary exploratory factor analysis. These items included all survey respondents, with an overall rate of AI use of 30.4% (59/194). The overall data were appropriate for conducting factor analysis (Bartlett test χ^2_{78}=1288.7; *P*<.001), and the Kaiser-Meyer-Olkin value (0.82) indicated that sampling was

adequate. There was no concern for multicollinearity, and parallel analysis identified 2 factors as the optimal number, with an eigenvalue cutoff of 1.31. The eigenvalues for the 2 factors were 4.49 and 1.97, which explained 49.7% of the cumulative variance. These 2 factors were classified as attitude toward the technology and perceived risk. Descriptive statistics for these constructs are provided in Table 3, and the pattern matrix is shown in Table 4. Including a third factor (eigenvalue=1.155), similar to the original TAME-ChatGPT validation study, resulted in an explanation of 57.1% of the cumulative variance. However, factors 2 and 3 were largely correlated (r=0.57), and many of the items were correlated with both factors 2 and 3. Cronbach α values for the 2 constructs were good at 0.86 for both comfort with technology and perceived risk.

In the secondary exploratory analysis, all items were included for the 30.4% (59/194) of respondents who had previously used AI. The overall data were appropriate for conducting factor analysis (Bartlett test χ^2_{528}=1301.5; P<.001), and the Kaiser-Meyer-Olkin value (0.70) indicated that sampling was adequate. There was concern for multicollinearity, and therefore, items 4, 14, and 31 were removed. Parallel analysis identified 4 factors as the optimal number, with an eigenvalue cutoff of 2.08. The eigenvalues for the 4 factors were 9.33, 4.36, 2.96, and 2.26, which explained 57.3% of the cumulative variance. The 4 factors identified included attitude toward the technology, perceived usefulness, perceived risk, and ease of use. Descriptive statistics for these constructs are provided in Table 3, and the pattern matrix is shown in Table 5. Cronbach α values for the 4 constructs were very reliable for attitude toward the technology (0.92), perceived usefulness (0.92), perceived risk (0.84), and ease of use (0.83).

Table 3. Technology Acceptance Model Edited to Assess ChatGPT Adoption constructs.

	Score, mean (SD)	*P* value
Perceived risk[a]		<.001
Had used AI[b] chatbots before	27.92 (6.32)	
Had not used AI chatbots before	30.59 (6.01)	
Attitude toward technology[c]		<.001
Had used AI chatbots before	19.19 (3.63)	
Had not used AI chatbots before	15.07 (4.06)	
Attitude toward technology among users[d]	44.53 (10.02)	—[e]
Perceived usefulness among users[f]	14.08 (5.80)	—
Perceived ease of use among users[g]	11.29 (2.54)	—
Perceived risk of use among users[h]	33.75 (7.98)	—

[a]Possible range from 8 to 40, with higher scores indicating lower perceived risk and a score of 24 indicating a neutral attitude.

[b]AI: artificial intelligence.

[c]Possible range from 5 to 25, with higher scores indicating positive attitude and a score of 15 indicating a neutral attitude.

[d]Possible range from 13 to 65, with higher scores indicating positive attitude and a score of 39 indicating a neutral attitude.

[e]Not applicable.

[f]Possible range from 6 to 30, with higher scores indicating higher perceived usefulness and a score of 18 indicating a neutral attitude.

[g]Possible range from 3 to 15, with higher scores indicating perceived ease of use and a score of 9 indicating a neutral attitude.

[h]Possible range from 5 to 50, with higher scores indicating lower perceived risk and a score of 27.5 indicating a neutral attitude.

Table 4. Pattern matrix of the 2 inferred factors for all respondents irrespective of previous artificial intelligence (AI) chatbot use.

Item	Perceived risk	Attitude toward technology
"I am concerned about the reliability of the information provided by AI chatbots."	0.435	<0.400
"I am concerned that using AI chatbots is considered plagiarism."	0.617	<0.400
"I fear relying too much on AI chatbots may decrease my critical thinking skills."	0.585	<0.400
"I am concerned about the potential security risks of using AI chatbots."	0.798	<0.400
"I am afraid of becoming too dependent on technology like AI chatbots."	0.618	<0.400
"I am afraid that using AI chatbots would result in a lack of originality in my work."	0.688	<0.400
"I am afraid that the use of the AI chatbots would be a violation of workplace policies."	0.655	<0.400
"I am concerned about the potential privacy risks that might be associated with using AI chatbots."	0.786	<0.400
"I am enthusiastic about using technology, such as AI chatbots for learning, practice, and research."	<0.400	0.887
"I believe technology, such as AI chatbots is an important tool for workplace success."	<0.400	0.844
"I think that technology like AI chatbots is attractive and fun to use."	<0.400	0.871
"I am always open to learning about new technologies like AI chatbots."	<0.400	0.672
"I trust the opinions of my friends or colleagues about using AI chatbots."	<0.400	0.444

Logistic regression was conducted to predict current and future use of AI (Table 6). Factors that remained significant in the model that predicted current AI use included positive attitude toward technology, coworker use of AI, and the respondent practicing in academia. Factors that predicted future use of AI included perceived risk, positive attitude toward technology, and coworker use.

Table 5. Pattern of the 4 inferred factors for only those who had used artificial intelligence (AI) chatbots before.

Item	Attitude toward technology	Perceived usefulness	Ease of use	Perceived risk
"I am concerned about the reliability of the information provided by AI chatbots."	<0.400	<0.400	<0.400	0.530
"I am concerned that using AI chatbots is considered plagiarism."	<0.400	<0.400	<0.400	0.652
"I fear relying too much on AI chatbots may decrease my critical thinking skills."	<0.400	<0.400	<0.400	0.457
"I am afraid of becoming too dependent on technology like AI chatbots."	<0.400	<0.400	<0.400	0.746
"I am afraid that using AI chatbots would result in a lack of originality in my work."	<0.400	<0.400	<0.400	0.762
"I am afraid that the use of the AI chatbots would be a violation of workplace policies."	<0.400	<0.400	<0.400	0.582
"I am concerned about the potential privacy risks that might be associated with using AI chatbots."	<0.400	<0.400	<0.400	0.750
"I am enthusiastic about using technology, such as AI chatbots for learning, practice, and research."	0.514	<0.400	<0.400	<0.400
"I believe technology, such as AI chatbots is an important tool for workplace success."	0.611	<0.400	<0.400	<0.400
"I think that technology like AI chatbots is attractive and fun to use."	0.609	<0.400	<0.400	<0.400
"I am always open to learning about new technologies like AI chatbots."	0.498	<0.400	<0.400	<0.400
"I trust the opinions of my friends or colleagues about using AI chatbots."	<0.400	<0.400	<0.400	−0.571
"For me, AI chatbots are a convenient method for accessing medical information."	<0.400	0.913	<0.400	<0.400
"For me, AI chatbots are a reliable source of accurate medical information."	<0.400	0.855	<0.400	<0.400
"AI chatbots help me in better understanding of difficult medical topics and concepts."	<0.400	0.878	<0.400	<0.400
"AI chatbots make it easier for me to complete tasks in my workplace."	0.784	<0.400	<0.400	<0.400
"I recommend AI chatbots to my colleagues to facilitate their work."	0.807	<0.400	<0.400	<0.400
"AI chatbots are more useful than other sources of medical information that I have used previously."	<0.400	0.843	<0.400	<0.400
"I think that using AI chatbots has helped to improve my overall workplace performance."	0.737	<0.400	<0.400	<0.400
"I have used tools similar to AI chatbots in the past in my workplace."	0.679	<0.400	<0.400	<0.400
"I spontaneously find myself using AI chatbots when I need medical information for my work."	<0.400	0.731	<0.400	<0.400
"I often use AI chatbots as a source of medical information in my workplace."	<0.400	0.766	<0.400	<0.400
"I appreciate the convenience and efficiency that AI chatbots provide for my work."	0.690	<0.400	<0.400	<0.400
"I think that relying on technology like AI chatbots can disrupt my critical thinking skills."	<0.400	<0.400	<0.400	0.674
"I appreciate the accuracy and reliability of the medical information provided by AI chatbots."	<0.400	0.821	<0.400	<0.400
"I believe that using AI chatbots can save time and effort in my workplace."	0.828	<0.400	<0.400	<0.400

Item	Attitude toward technology	Perceived usefulness	Ease of use	Perceived risk
"It does not take a long time to learn how to use AI chatbots."	<0.400	<0.400	0.782	<0.400
"Using AI chatbots does not require extensive technical knowledge."	<0.400	<0.400	0.828	<0.400
"I do not face many difficulties when using AI chatbots."	<0.400	<0.400	0.787	<0.400
"The positive experiences of others have encouraged me to use AI chatbots."	0.486	<0.400	<0.400	<0.400
"I believe that people I know have improved their workplace performance as a result of using AI chatbots."	0.796	<0.400	<0.400	<0.400
"I think using AI chatbots is important for me to keep up with my peers professionally."	0.542	<0.400	<0.400	<0.400

Table 6. Predictors of current and future artificial intelligence (AI) chatbot use.

	OR[a] (95% CI)	*P* value
Current use of AI[b]		
Perceived risk	0.98 (0.61-1.56)	.94
Attitude toward technology	3.64 (2.08-6.36)	<.001
Coworker use of AI	7.41 (2.64-20.80)	<.001
AI policy present	2.72 (0.80-9.20)	.11
Academia	5.62 (1.30-24.23)	.02
Community	0.52 (0.14-1.90)	.32
Future use of AI[c]		
Perceived risk	0.63 (0.41-0.96)	.03
Attitude toward technology	4.11 (2.42-6.97)	<.001
Coworker use of AI	33.00 (5.02-216.76)	<.001
AI policy present	2.24 (0.60-8.43)	.23
Academia	2.06 (0.42-10.21)	.38
Community	1.09 (0.41-2.89)	.86

[a]OR: odds ratio.

[b]Hosmer and Lemeshow *P*=.15; area under the curve 0.85.

[c]Hosmer and Lemeshow *P*=.71; area under the curve 0.87.

Discussion

Approximately one-third of pharmacy preceptors (59/194, 30.4%) reported use of an AI chatbot, with approximately half (100/194, 51.5%) indicating that they planned to start or would continue using chatbots in the future. Consistent with findings from other studies, we found that most respondents were unlikely to make patient care decisions based on information provided by an AI chatbot. However, they did report use for administrative tasks such as summarizing information and writing letters of recommendation. To our knowledge, this is the first study using the TAME-ChatGPT assessment tool among pharmacists. The findings show that this tool is valid and reliable for assessing pharmacists' attitudes toward chatbots and their use in pharmacy practice. Pharmacists' attitudes toward chatbots were largely influenced by their attitude toward the technology and their perceived risk related to use of the technology. Furthermore, among pharmacists who had used chatbots, attitudes toward use of chatbots were affected by the same 2 factors plus their perceived usefulness and ease of use. Positive attitudes toward technology, having coworkers who use AI, and working in academia predicted current use of AI chatbots, whereas factors predicting future use of the technology included perceived risk, positive attitudes toward technology, and coworker use of AI chatbots. Our results affirm that, when adopting ChatGPT and other AI chatbots, it is important to consider perceptions of risk, usefulness, and ease of use, as well as the users' attitudes toward technology.

Significantly more pharmacists practicing in academia have used chatbots than pharmacists in other practice settings, perhaps because these pharmacists are often at the forefront of exploring new technologies and their applications in pharmacy practice and may be more comfortable with adopting new technologies. While academic pharmacists may be at the forefront of chatbot adoption, the use of AI chatbots is expected to increase across all pharmacy practice settings as the technology becomes more

refined and its benefits become more apparent. Since the time of our data collection, more advanced AI chatbots have become available, such as OpenEvidence, which may provide more targeted information for pharmacists to use in practice.

However, most pharmacists surveyed (135/194, 69.6%) had not used an AI chatbot, with nearly two-thirds of them (85/135, 63%) stating that their reason for not using chatbots was that they did not know how to use them effectively; 46.7% (63/135) cited lack of credibility or trust in chatbots as the reason for lack of use. AI chatbots are relatively new technologies in health care, and many pharmacists may not have had sufficient exposure or training to use them effectively. Furthermore, there is currently no standardized approach to training pharmacists or other health care providers on the use of AI chatbots; thus, knowledge and skills across the profession may be inconsistent. Knowledge on how to use AI chatbots will likely improve as they become more prevalent in health care settings. Pharmacists will gain hands-on experience with these tools, become more comfortable using them, and learn to integrate them into their daily workflows. Preceptors are in an ideal position to help guide students who are just learning about AI capabilities and limitations on the optimal use of this new technology. Health care organizations, educational institutions, and postgraduate training programs should consider including training on the appropriate use of AI and risks of inappropriate use.

Our findings differ somewhat from those of the work by Sallam et al [16], who validated the TAME-ChatGPT in a sample of Jordanian health care students and identified factors affecting their attitudes toward ChatGPT and use of ChatGPT. Attitudes of Jordanian students toward ChatGPT were influenced by an additional third factor, anxiety related to fear of ChatGPT; attitudes toward use of ChatGPT were affected by behavior as a fourth factor instead of attitudes toward technology [16]. The differences are likely due to differences in the populations studied; US pharmacists and Jordanian students represent very different populations based on culture, age, practice experience, and other characteristics. Different approaches to factor analysis may also contribute to differences.

Risk perception is known to be a key factor affecting decision-making, and perception of risks associated with chatbots significantly influenced pharmacists' attitudes toward chatbots and their use [18]. The credibility of AI chatbots in health care is a complex and evolving issue. Reliability and accuracy can vary significantly between chatbots, and not all AI chatbots are reliable sources of information. Inaccuracies, hallucinations, potential for biased responses, and the inability of chatbots to provide nuanced or context-specific information have been documented [12,19-21]. These potential barriers highlight the need for health care professional scrutiny and oversight of chatbot responses. As the technologies continue to develop, ongoing research and validation will be crucial to establish and maintain the credibility of AI chatbots in health care applications and minimize their potential for harm.

Some of the limitations of this study include the sample size as only pharmacist preceptors in the Midwest who were affiliated with the participating colleges were surveyed, so the results may not be representative of pharmacy preceptors across the United States. However, based on comparison with the 2024 American Association of Colleges of Pharmacy preceptor survey, outside of years and area of practice, the surveyed sample seems to be fairly representative of preceptors in the United States [17]. Our response rate was low, which may be due to technology-related factors (ie, the email going to the spam folder) and may have resulted in selection bias in the sense that preceptors with experience using AI may have been more likely to respond. As the survey was anonymous, we could not explore differences between respondents and nonrespondents. It is possible that the use of AI at the time of our survey overestimated use in the target population. Additionally, the AI and chatbot space is constantly evolving, and although the survey was conducted recently, perceptions can change rapidly as more people are exposed to chatbots and start using them in their daily work. Future studies could expand the surveyed population to include pharmacist preceptors nationwide or include all pharmacists, along with following up on respondents over time to evaluate changes in responses. Additionally, the validated survey can be readministered following training and education on AI chatbots to determine how perceptions of AI technology have changed.

At the time of our study, it had been almost 2 years since AI chatbots were introduced, and pharmacist preceptors were still hesitant to use the new technology, with only approximately one-third of respondents (59/194, 30.4%) indicating that they had used a chatbot in practice. Pharmacist preceptors were hesitant to use the technology for clinical decisions and were uncertain about their place in practice. This study also demonstrated that the TAME-ChatGPT survey is a reliable and validated tool that can be used to assess pharmacists' attitude toward and use of chatbots. Constructs from the TAM, including attitude toward technology, perceived usefulness, ease of use, and perceived risk of use, as well as practice setting and coworker use, can determine and predict pharmacist use of AI chatbots. Future studies with this validated tool can be used to guide the implementation of chatbots into pharmacy practice and help inform policymakers and organization leaders on the education and training needed to promote the safe and effective use of AI chatbots in pharmacy practice.

Acknowledgments

The authors would like to acknowledge Dr Margie E Snyder, Dr Darren Covington, Dr Sarah E Vordenberg, and Dr Faria Munir for their assistance in this project.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey instrument.
[PDF File (Adobe PDF File), 140 KB - mededu_v11i1e71767_app1.pdf]

References

1. Introducing ChatGPT. OpenAI. URL: https://openai.com/blog/chatgpt [accessed 2024-07-17]
2. Vogels E. A majority of Americans have heard of ChatGPT, but few have tried it themselves. Pew Research Center. 2023 May 24. URL: https://www.pewresearch.org/short-reads/2023/05/24/a-majority-of-americans-have-heard-of-chatgpt-but-few-have-tried-it-themselves/ [accessed 2024-07-17]
3. Welcome to Dougall GPT. Dougall GPT. URL: https://dougallgpt.com/user/login [accessed 2024-07-17]
4. Home page. OpenEvidence. URL: https://www.openevidence.com/ [accessed 2024-07-17]
5. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023 Jul 12;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
6. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare 2023 Mar 19;11(6):887 [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
7. Hosseini M, Gao CA, Liebovitz DM, Carvalho AM, Ahmad FS, Luo Y, et al. An exploratory survey about using ChatGPT in education, healthcare, and research. PLoS ONE 2023 Oct 5;18(10):e0292216 [FREE Full text] [doi: 10.1371/journal.pone.0292216]
8. Temsah MH, Aljamaan F, Malki KH, Alhasan K, Altamimi I, Aljarbou R, et al. ChatGPT and the future of digital health: a study on healthcare workers' perceptions and expectations. Healthcare (Basel) 2023 Jun 21;11(13):1812. [doi: 10.3390/healthcare11131812] [Medline: 37444647]
9. Abu Hammour K, Alhamad H, Al-Ashwal FY, Halboup A, Abu Farha R, Abu Hammour A. ChatGPT in pharmacy practice: a cross-sectional exploration of Jordanian pharmacists' perception, practice, and concerns. J Pharm Policy Pract 2023 Oct 03;16(1):115 [FREE Full text] [doi: 10.1186/s40545-023-00624-2] [Medline: 37789443]
10. Alghitran A, AlOsaimi HM, Albuluwi A, Almalki E, Aldowayan A, Alharthi R, et al. Integrating ChatGPT as a tool in pharmacy practice: a cross-sectional exploration among pharmacists in Saudi Arabia. Integr Pharm Res Pract 2025 Mar 17;14:31-43 [FREE Full text] [doi: 10.2147/iprp.s500689]
11. Jairoun AA, Al-Hemyari SS, Shahwan M, Alnuaimi GR, Ibrahim N, Jaber AA. Capturing pharmacists' perspectives on the value, risks, and applications of ChatGPT in pharmacy practice: a qualitative study. Explor Res Clin Soc Pharm 2024 Dec;16:100518 [FREE Full text] [doi: 10.1016/j.rcsop.2024.100518] [Medline: 40046775]
12. Lima TD, Bonafé M, Baby AR, Visacri MB. ChatGPT in pharmacy practice: disruptive or destructive innovation? A scoping review. Sci Pharm 2024 Oct 21;92(4):58 [FREE Full text] [doi: 10.3390/scipharm92040058]
13. Khatri S, Sengul A, Moon J, Jackevicius CA. Accuracy and reproducibility of ChatGPT responses to real‐world drug information questions. J Am Coll Clin Pharm 2025 Apr 22;8(6):432-438. [doi: 10.1002/jac5.70038]
14. Roosan D, Padua P, Khan R, Khan H, Verzosa C, Wu Y. Effectiveness of ChatGPT in clinical pharmacy and the role of artificial intelligence in medication therapy management. J Am Pharm Assoc (2003) 2024;64(2):422-8.e8 [FREE Full text] [doi: 10.1016/j.japh.2023.11.023] [Medline: 38049066]
15. Huang X, Estau D, Liu X, Yu Y, Qin J, Li Z. Evaluating the performance of ChatGPT in clinical pharmacy: a comparative study of ChatGPT and clinical pharmacists. Br J Clin Pharmacol 2024 Jan;90(1):232-238 [FREE Full text] [doi: 10.1111/bcp.15896] [Medline: 37626010]
16. Sallam M, Salim NA, Barakat M, Al-Mahzoum K, Al-Tammemi AB, Malaeb D, et al. Assessing health students' attitudes and usage of ChatGPT in Jordan: validation study. JMIR Med Educ 2023 Sep 05;9:e48254 [FREE Full text] [doi: 10.2196/48254] [Medline: 37578934]
17. American Association of Colleges of Pharmacy 2024 preceptor survey. American Association of Colleges of Pharmacy. URL: https://www.aacp.org/sites/default/files/2024-10/2024-preceptor-survey-national-summary-report.pdf [accessed 2024-07-17]
18. Dergaa I, Ben Saad H, Glenn JM, Amamou B, Ben Aissa M, Guelmami N, et al. From tools to threats: a reflection on the impact of artificial-intelligence chatbots on cognitive health. Front Psychol 2024;15:1259845 [FREE Full text] [doi: 10.3389/fpsyg.2024.1259845] [Medline: 38629037]
19. Williams DJ, Noyes JM. How does our perception of risk influence decision-making? Implications for the design of risk information. Theor Issues Ergonomics Sci 2007 Jan 23;8(1):1-35 [FREE Full text] [doi: 10.1080/14639220500484419]
20. Deng J, Lin Y. The benefits and challenges of ChatGPT: an overview. Front Comput Intell Syst 2023 Jan 05;2(2):81-83 [FREE Full text] [doi: 10.54097/fcis.v2i2.4465]
21. Xue J, Wang Y, Wei C, Liu X, Woo J, Kuo CJ. Bias and fairness in chatbots: an overview. APSIPA Trans Signal Inf Process 2023 Sep 16;13(2):e102 [FREE Full text] [doi: 10.1561/116.00000064]

Abbreviations

AI: artificial intelligence
OR: odds ratio
TAM: technology acceptance model
TAME-ChatGPT: Technology Acceptance Model Edited to Assess ChatGPT Adoption

Edited by R Pellegrino; submitted 26.01.25; peer-reviewed by C Wang, I Akpan; comments to author 13.06.25; revised version received 04.08.25; accepted 21.08.25; published 21.11.25.

Please cite as:
Li A, Sheehan AH, Giuliano C, Dobry P, Walker P, Philips J, Jordan J
Assessing Pharmacists' Use and Perception of AI Chatbots in Pharmacy Practice: Cross-Sectional Survey Study
JMIR Med Educ 2025;11:e71767
URL: https://mededu.jmir.org/2025/1/e71767
doi:10.2196/71767
PMID:

Original Paper

Enhancing Large Language Models for Improved Accuracy and Safety in Medical Question Answering: Comparative Study

Dingqiao Wang[1,2], PhD; Jinguo Ye[1], BSc; Jingni Li[1], MD; Jiangbo Liang[1], MD; Qikai Zhang[1], BSc; Qiuling Hu[1], BSc; Caineng Pan[1], BSc; Dongliang Wang[1], PhD; Zhong Liu[1], PhD; Wen Shi[1], PhD; Mengxiang Guo[3*], PhD; Fei Li[1*], PhD; Wei Du[2*], PhD; Ying-Feng Zheng[1*], PhD

[1]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science, Guangdong Provincial Clinical Research Center for Ocular Diseases, Guangzhou, China

[2]Department of Ophthalmology, Eighth Affiliated Hospital of Sun Yat-sen University, Shenzhen, Guangdong, China

[3]Guangzhou Women and Children's Medical Center, Guangzhou, China

[*]these authors contributed equally

Corresponding Author:
Ying-Feng Zheng, PhD
State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science, Guangdong Provincial Clinical Research Center for Ocular Diseases
07 Jinsui Road
Guangzhou, 510060
China
Phone: 86 13922286455
Email: zhyfeng@mail.sysu.edu.cn

Abstract

Background: Large language models (LLMs) offer the potential to improve virtual patient-physician communication and reduce health care professionals' workload. However, limitations in accuracy, outdated knowledge, and safety issues restrict their effective use in real clinical settings. Addressing these challenges is crucial for making LLMs a reliable health care tool.

Objective: This study aimed to evaluate the efficacy of Med-RISE, an information retrieval and augmentation tool, in comparison with baseline LLMs, focusing on enhancing accuracy and safety in medical question answering across diverse clinical domains.

Methods: This comparative study introduces Med-RISE, an enhanced version of a retrieval-augmented generation framework specifically designed to improve question-answering performance across wide-ranging medical domains and diverse disciplines. Med-RISE consists of 4 key steps: query rewriting, information retrieval (providing local and real-time retrieval), summarization, and execution (a fact and safety filter before output). This study integrated Med-RISE with 4 LLMs (GPT-3.5, GPT-4, Vicuna-13B, and ChatGLM-6B) and assessed their performance on 4 multiple-choice medical question datasets: MedQA (US Medical Licensing Examination), PubMedQA (original and revised versions), MedMCQA, and EYE500. Primary outcome measures included answer accuracy and hallucination rates, with hallucinations categorized into factuality (inaccurate information) or faithfulness (inconsistency with instructions) types. This study was conducted between March 2024 and August 2024.

Results: The integration of Med-RISE with each LLM led to a substantial increase in accuracy, with improvements ranging from 9.8% to 16.3% (mean 13%, SD 2.3%) across the 4 datasets. The enhanced accuracy rates were 16.3%, 12.9%, 13%, and 9.8% for GPT-3.5, GPT-4, Vicuna-13B, and ChatGLM-6B, respectively. In addition, Med-RISE effectively reduced hallucinations, with reductions ranging from 11.8% to 18% (mean 15.1%, SD 2.8%), factuality hallucinations decreasing by 13.5%, and faithfulness hallucinations decreasing by 5.8%. The hallucination rate reductions were 17.7%, 12.8%, 18%, and 11.8% for GPT-3.5, GPT-4, Vicuna-13B, and ChatGLM-6B, respectively.

Conclusions: The Med-RISE framework significantly improves the accuracy and reduces the hallucinations of LLMs in medical question answering across benchmark datasets. By providing local and real-time information retrieval and fact and safety filtering, Med-RISE enhances the reliability and interpretability of LLMs in the medical domain, offering a promising tool for clinical practice and decision support.

(JMIR Med Educ 2025;11:e70190) doi:10.2196/70190

KEYWORDS

large language models; ChatGPT; medical question answering; health care communication; retrieval-augmented generation

Introduction

Large language models (LLMs), such as ChatGPT, have emerged as a powerful paradigm in natural language processing [1]. With its humanlike conversational capabilities, ChatGPT has become a potent tool for medical question answering (QA), improving virtual patient-physician communication and reducing health care professionals' workload [2-7].

Previous studies have focused on assessing the performance of LLMs in medical QA on standard datasets, such as MedMCQA, MedQA, and PubMedQA, and in specific specialties such as surgery, oncology, ophthalmology, and radiology [8,9]. The accuracy of models such as GPT-3.5 (50%-60%) and GPT-4 (70%-80%) is insufficient for clinical application, underscoring the need for further enhancements of LLMs regarding domain-specific medical knowledge [10-18]. Furthermore, the "hallucination" phenomenon in LLMs, which leads to factually inaccurate or irrelevant content, presents serious risks to patient care [19-23]. In medical practice, these hallucinations can lead to erroneous information, unsupported diagnoses, or inappropriate treatments. Consequently, many studies that have assessed LLM performance in medical applications highlight that major challenges in applying LLMs to clinical settings are insufficient accuracy; outdated knowledge; and potential safety issues, including hallucinations and bias [18,24,25]. Enhancing accuracy, timeliness, and safety is essential to make LLMs reliable for health care, ultimately improving patient outcomes.

Retrieval-augmented generation (RAG) is a promising strategy to enhance the accuracy of medical QA tasks and reduce hallucinations [26-29]. RAG improves LLMs' responses by retrieving relevant documents from external knowledge, grounding the responses in factual information and increasing their reliability [30]. Previous studies have investigated the use of external knowledge to augment LLMs in medical domains, such as Almanac [31] and RECTIFIER [32]. However, most of these studies that have used retrieval techniques are based on small, predownloaded datasets, leading RAG models to be applied mostly in specific disciplines rather than broadly across medical fields [33-36]. In our previous study, we developed an RAG framework called RISE to improve LLMs' performance in diabetes-related QA, achieving significant enhancements compared to the base LLMs [37]. Building on this framework, further exploration and refinement of knowledge augmentation techniques are warranted to expand its application across broader medical domains.

In this study, we introduced Med-RISE, an enhanced version of the RAG framework specifically designed to improve performance in medical QA tasks across broader medical domains. We assessed the impact of integrating the Med-RISE framework with the GPT-3.5, GPT-4, Vicuna-13B, and ChatGLM-6B LLMs and quantitatively evaluated its improvements in accuracy and reductions in hallucinations in the reasoning process using standard medical datasets, including MedQA (US Medical Licensing Examination [USMLE]), PubMedQA, and MedMCQA.

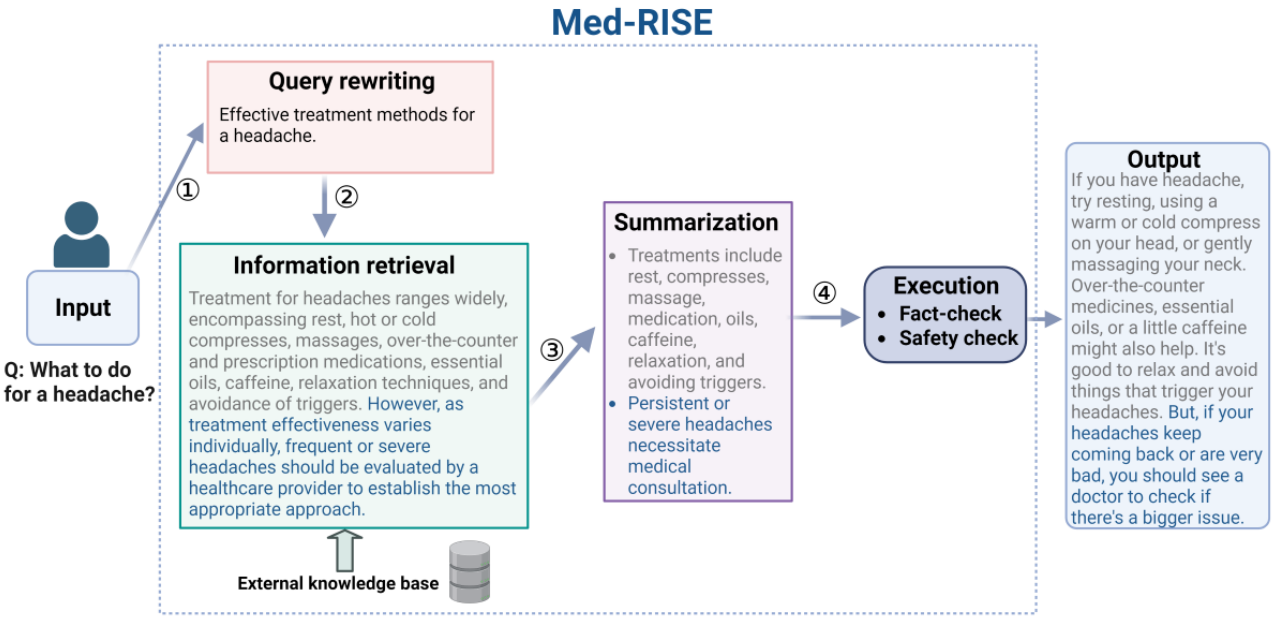
Methods

The Framework of Med-RISE

Overview

In this study, we developed an advanced framework called Med-RISE to enhance the performance of LLMs in medical QA tasks. Building on our previous work on the RISE framework [37], which focused on diabetes-related QA in a single specialty domain, Med-RISE expands the retrieval database and incorporates more authoritative medical sources, making it suitable for diverse medical domains. The Med-RISE implementation requires 8 V100 graphics processing units (GPUs; 32 GB each) and 4 TB of storage for the medical database and achieves response times of 5 to 20 seconds using parallel retrieval strategies. The Med-RISE framework involves 4 key steps: query rewriting, information retrieval, summarization, and execution (Figure 1). The framework is implemented using the Python programming language (Python Software Foundation) on the Ubuntu operating system (Canonical Ltd). We provide open-source code (Multimedia Appendix 1) that supports both web deployment and mobile app integration.

Figure 1. The Med-RISE framework: (1) the query rewriting step refines user queries using advanced large language models for better retrieval; (2) the information retrieval step searches for relevant context from an expanded medical local database and external real-time academic sources; (3) the summarization step summarizes the retrieved information into concise key points; and (4) the execution step generates the final answer, with fact-checking and safety validation to filter out incorrect, biased, or unsafe information before output, ensuring accuracy and safety.



Query Rewriting

In this step, Med-RISE first identifies the intent of the user or the patient's query through the following prompt: "Please identify the intent of the patient's query and select the most appropriate category from a list of predefined intents." The 50 intent categories were developed by our team through analysis of clinical practice patterns and validated by clinical physicians (YZ). The LLM selects the single most appropriate intent category using natural language processing. The complete list of the 50 intent categories is provided in Multimedia Appendix 1.

Med-RISE uses LLMs such as GPT-3.5, GPT-4, Vicuna-13B, and ChatGLM-6B to refine the initial query. The refinement process involves addressing typographical errors, introducing related terms, expanding the range of possible matches, and enhancing overall retrieval precision. This step ensures that the rewritten query accurately captures the user's intent and is well suited for subsequent information retrieval.

Information Retrieval

Med-RISE uses a dual retrieval approach combining local database and real-time academic sources to ensure comprehensive and current medical information. The system first converts the rewritten query into vector representation using the OpenAI Text-Embedding-ADA-002 model and then applies the Faiss similarity search algorithm to identify the 5 most relevant documents from the local database. Simultaneously, the framework performs automatic online retrieval through the Google search engine, filtering results using a predefined white list of 200 authoritative medical websites and academic institutions. This white list encompasses major medical journals (*The New England Journal of Medicine*, *Lancet*, and *The Journal of the American Medical Association*), medical databases (PubMed and Cochrane Library), professional medical organizations (the World Health Organization, Centers for Disease Control and Prevention, and National Institutes of Health), specialized medical societies, and official websites of medical schools and teaching hospitals. When local and real-time retrieval results are consistent, the information is integrated and synthesized. When discrepancies arise, priority is given to real-time retrieved information to ensure up-to-date medical knowledge. The filtered documents are then analyzed and integrated by the LLMs before proceeding to the summarization and execution steps.

To optimize retrieval latency, Med-RISE implements several strategies. For local database retrieval, the system uses 8 V100 GPUs (32 GB each) with parallel processing capabilities and optimized indexing methods such as IndexIVFFlat. For real-time retrieval, asynchronous parallel processing across multiple GPUs and network stability enhancements are used to minimize response delays, achieving response times of 5 to 20 seconds.

Summarization

In this step, Med-RISE summarizes the retrieved content into a clear and concise format, focusing on key points and eliminating redundancy. The model is prompted with the following: "You are an assistant skilled in organizing text. Summarize the following content clearly and briefly, keeping the important points and removing repeated information." This step ensures that the content is well suited for generating an effective response.

Execution

The final step involves generating the final answer from the summarized information using the LLMs. Before sharing with the user, the answer undergoes fact-checking and safety validation to filter out any biased, unsafe, or incorrect information, ensuring that the response is accurate, neutral, and safe.

For fact-checking, the retrieved information is broken down into individual claims, which are then verified against external knowledge sources to confirm their accuracy using the following prompt: "As an AI medical assistant, your current task is to break down your last response in the conversation into independent claims. Do not include claims about opinions, or subjective personal experiences." These claims are then verified against external knowledge sources to confirm their accuracy using the following prompt: "As an AI medical assistant, your current task is to fact-check the claims based on the external knowledge provided. You should label each claim as true or false. After that, output all the true claims without saying anything else. Now it is your turn to fact-check the claims based on the external knowledge."

For safety validation, 24 predefined rules were developed based on medical ethics principles, patient safety requirements, and responsible artificial intelligence guidelines. The rules were tested in pilot experiments and then reviewed and refined by 2 clinical experts (MG and FL) before finalization. The model is given the following prompt: "The final response should comply with the following 24 guidelines." These rules prevent specific medical diagnoses, treatment advice, and harmful generalizations and ensure that the responses recommend professional consultation when appropriate.

This 2-step process ensures that the final summarized content is filtered for safety and correctness before being presented to the user.

Local Database

The local database for Med-RISE was constructed using 5 main sources: PubMed, StatPearls, 15 widely used medical textbooks, clinical practice guidelines, and Wikipedia. Specifically, PubMed provides 60,000 carefully selected high-quality biomedical articles (2015-2024; impact factor of ≥5; *Journal Citation Reports* quartiles 1-2; ≥50 citations) covering a wide range of biomedical research [37,38]. StatPearls serves as a clinical decision support tool, and we included more than 3000 publicly available StatPearls articles. The 15 medical textbooks are widely used in medical education and are the main references for USMLE preparation. Clinical practice guidelines from major medical societies, including the American College of Physicians and Chinese Medical Association, were included as main references for clinical decision support. Wikipedia, an open-source encyclopedia, provides general medical knowledge to support broader medical topics. These sources cover various medical disciplines, including anatomy, pharmacology, pathology, pediatrics, psychiatry, internal medicine, and gynecology (Multimedia Appendix 2).

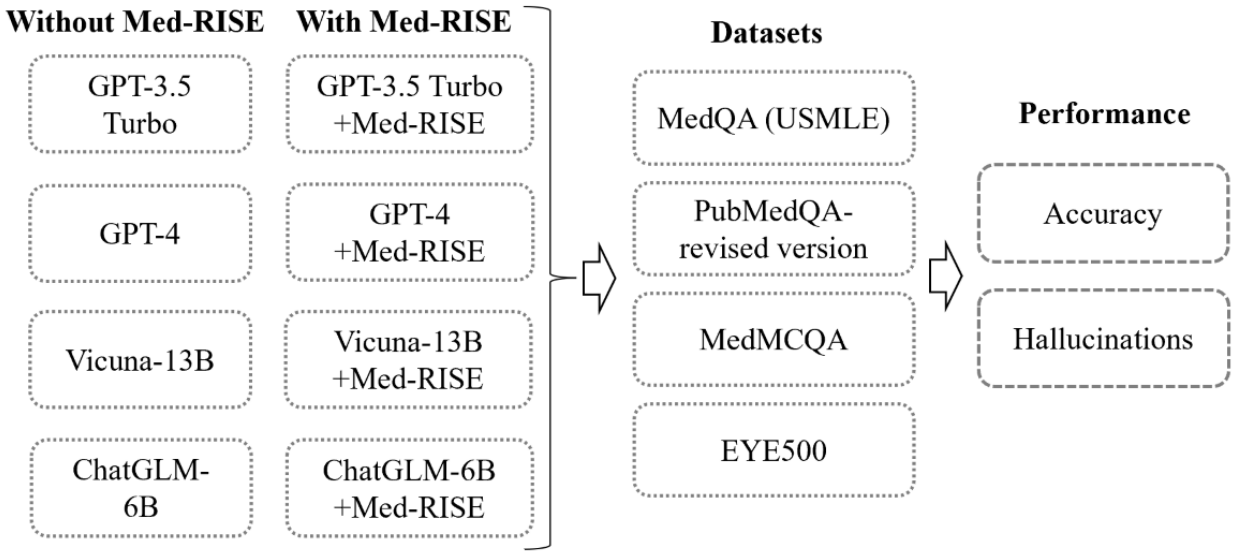
The collected documents (from the 5 main sources) underwent preprocessing to eliminate unstructured or extraneous information before being segmented into smaller units. Embeddings for these segments were generated using the OpenAI Text-Embedding-ADA-002 model and then indexed using the Faiss similarity search algorithm for efficient retrieval. The process for generating embeddings and retrieving data follows a methodology similar to that of our previous RISE framework [37]. Due to the stable nature of foundational medical knowledge, the local database is updated annually through manual curation by our team according to our local database inclusion criteria.

After receiving a user's query, the system transforms the rewritten query into an embedding vector. This vector is then matched against the database using cosine similarity to identify the 5 most relevant segments, which form the knowledge context for the query.

Study Design

As shown in Figure 2, our study focused on assessing how the Med-RISE framework influences the effectiveness of LLMs in answering medical questions. We selected 4 representative LLMs: proprietary models GPT-3.5 and GPT-4 and open-source LLMs Vicuna-13B and ChatGLM-6B. These models were accessed and used between March 2024 and August 2024. These models have distinct characteristics in terms of performance, organization, source of training data, and parameter size (Table S1 in Multimedia Appendix 3). By testing Med-RISE on these diverse LLMs, we aimed to demonstrate the effectiveness and versatility of the Med-RISE framework.

Figure 2. Flowchart of the overall study design. This study examined the accuracy and hallucination rates of large language models across 4 medical datasets with and without the Med-RISE framework. USMLE: US Medical Licensing Examination.



Ethical Considerations

The Zhongshan Ophthalmic Centre ethics committee (Guangzhou, China) approved this study (2024KYPJ124). As this research did not involve the collection of patient information or data, the ethics approval document includes an exemption from informed consent.

Validation Assessment

In the accuracy evaluation, we compared the performance of various LLMs with and without the application of the Med-RISE framework. This comparison was conducted on multiple-choice questions derived from several datasets: MedQA (USMLE), PubMedQA, MedMCQA, and EYE500. Accuracy was defined as the proportion of questions answered correctly out of the total number of questions compared to the standard answers provided in each dataset.

We assessed the presence of hallucinations in the reasoning process of the LLMs using the chain-of-thought (CoT) approach. Hallucinations are defined as content that is either factually incorrect or unrelated to the given information and are categorized into 2 main types: factuality hallucinations and faithfulness hallucinations [39]. Factuality hallucinations occur when the generated content conflicts with real-world information, which includes factual inconsistency (contradicting real-world information) and factual fabrication (generating content that cannot be verified using real-world information). Faithfulness hallucinations involve the inconsistency between the generated content and user instructions, input context, or internal logical coherence, which includes instruction inconsistency (output deviating from user instructions), context inconsistency (output contradicting contextual information), and logical inconsistency (inconsistency between reasoning steps and the final answer). The hallucination rate refers to the percentage of generated responses that included any form of hallucination. Two clinical physicians (DW and WS) independently evaluated the presence of each specific hallucination type, with disagreements resolved through consensus by a senior physician (YZ). Furthermore, we evaluated the proportion of each specific hallucination type.

Medical Datasets Used

In this study, we evaluated the performance of the Med-RISE framework using 4 medical datasets: MedQA (USMLE) [40], MedMCQA [41], PubMedQA [42], and EYE500 (Table S2 in Multimedia Appendix 3). MedQA (USMLE), MedMCQA, and EYE500 are multiple-choice medical board exam question datasets, whereas PubMedQA is a medical reading comprehension dataset [21,43]. For each dataset, we randomly selected 200 questions to reduce computational costs while maintaining evaluation comprehensiveness as preliminary testing showed comparable accuracy rates between the full datasets and subsampled questions.

The MedQA (USMLE) dataset, sourced from the USMLE, consists of challenging multiple-choice questions designed to assess medical knowledge. The MedMCQA dataset consists of diverse multiple-choice questions from the All India Institute of Medical Science and National Eligibility cum Entrance Test (Postgraduate) exam, encompassing various health care topics and medical subjects. These datasets cover diverse medical disciplines, including clinical medicine (internal medicine, surgery, pediatrics, and psychiatry), basic medical sciences (anatomy, physiology, and pathology), and specialized fields. We also introduce the EYE500 dataset, explicitly developed for this study in collaboration with ophthalmologists from Zhongshan Ophthalmic Center, Sun Yat-sen University (Multimedia Appendix 4).

The PubMedQA dataset, a medical reading comprehension dataset, contains questions based on biomedical article abstracts from the PubMed database, requiring models to answer questions with “yes,” “no,” or “maybe” based on the provided abstract information. However, the accuracy of the answers may be affected by the quality and interpretation of the original

studies. To ensure the reliability of the 200 questions and answers used in this study, we had 2 medical experts (DW and WS) independently reverify them. In cases of disagreement, a senior physician (YZ) made the final decision, creating the PubMedQA—revised version (Multimedia Appendix 5).

Prompts

We used 2 prompting strategies: zero-shot prompting for accuracy assessment (eg, "Please provide the answer to this question") and CoT prompting for hallucination evaluation to obtain step-by-step reasoning processes (eg, "Please use step-by-step reasoning to analyze this medical question systematically") [43].

Statistical Analysis

The data analysis was conducted using the SPSS software (version 25.0; IBM Corp). Chi-square tests were used to compare accuracy and hallucination rates. Two clinicians assessed hallucinations (>5 years of experience each), with disagreements resolved by a senior physician. Statistical analysis was conducted in consultation with Jin Ling, a professional biostatistician from Zhongshan Ophthalmic Center. Statistical significance was set at a *P* value of <.05.

Results

Enhanced Accuracy of LLMs Across All Datasets With Med-RISE Integration

Overview

This section presents a comparative analysis of the performance of the 4 LLMs—GPT-3.5, GPT-4, Vicuna-13B, and ChatGLM-6B—on the MedQA (USMLE), PubMedQA, PubMedQA—revised version, MedMCQA, and EYE500 medical datasets. The integration of Med-RISE with each of these LLMs significantly improved accuracy (Table 1 and Figure 3).

The incorporation of Med-RISE into each LLM led to a substantial increase in accuracy across the 4 datasets. Med-RISE enhanced overall accuracy rates, with improvements ranging from 9.8% to 16.3% (mean 13%, SD 2.3%) across the 4 models. Specifically, GPT-3.5 accuracy improved from 47.6% to 63.9% (improvement of 16.3%), GPT-4 accuracy improved from 67.9% to 80.8% (improvement of 12.9%), Vicuna-13B accuracy improved from 28% to 41% (improvement of 13%), and ChatGLM-6B accuracy improved from 37.8% to 47.6% (improvement of 9.8%).

Table 1. Accuracy of the large language models on biomedical question answering datasets.[a]

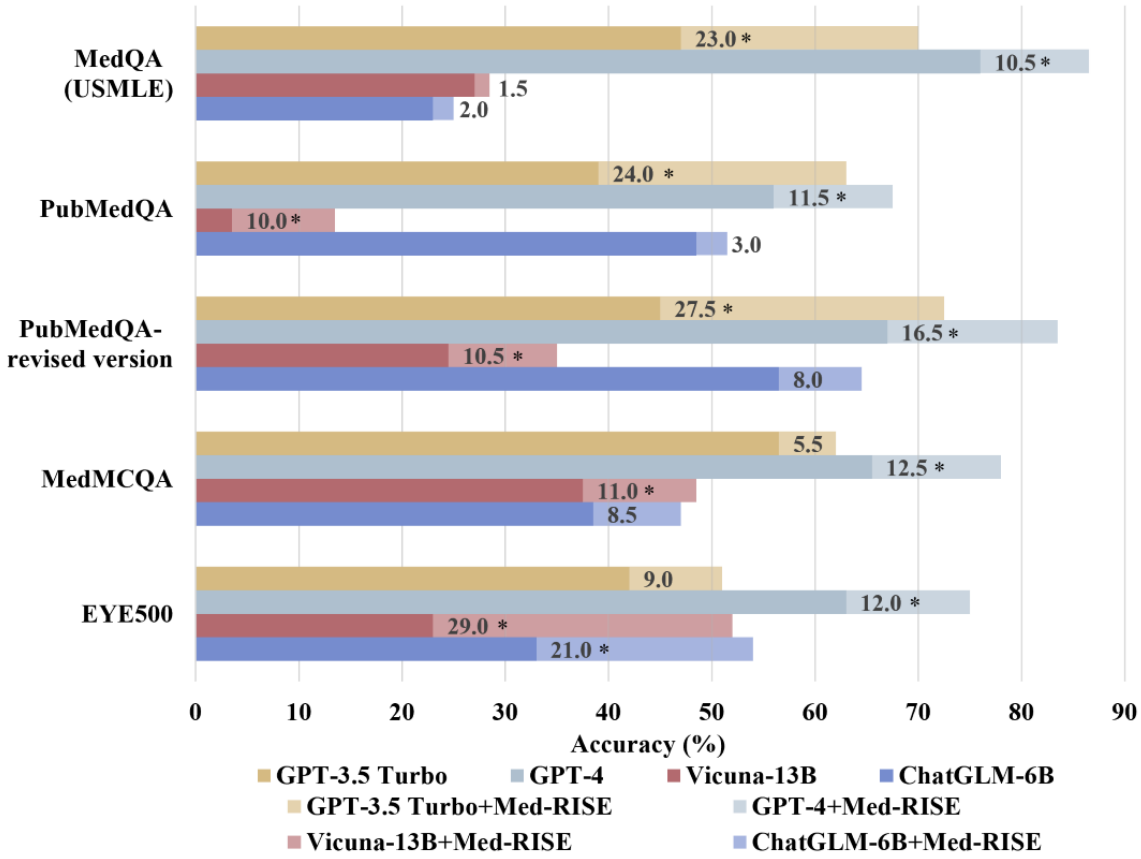
Dataset	GPT-3.5 Turbo, n (%)			GPT-4, n (%)			Vicuna-13B, n (%)			ChatGLM-6B, n (%)		
	Without Med-RISE	With Med-RISE	*P* value	Without Med-RISE	With Med-RISE	*P* value	Without Med-RISE	With Med-RISE	*P* value	Without Med-RISE	With Med-RISE	*P* value
MedQA (USMLE[b])	94 (47.0)	140 (70.0)	<.001	152 (76.0)	173 (86.5)	.02	54 (27.0)	57 (28.5)	.74	46 (23.0)	50 (25.0)	.64
PubMedQA	78 (39.0)	126 (63.0)	<.001	112 (56.0)	135 (67.5)	.02	7 (3.5)	27 (13.5)	.001	97 (48.5)	103 (51.5)	.55
PubMedQA—revised version	90 (45.0)	145 (72.5)	<.001	134 (67.0)	167 (83.5)	<.001	49 (24.5)	70 (35.0)	.02	113 (56.5)	129 (64.5)	.10
MedMCQA	113 (56.5)	124 (62.0)	.26	131 (65.5)	156 (78.0)	.01	75 (37.5)	97 (48.5)	.03	77 (38.5)	94 (47.0)	.09
EYE500	84 (42.0)	102 (51.0)	.07	126 (63.0)	150 (75.0)	.009	46 (23.0)	104 (52.0)	<.001	66 (33.0)	108 (54.0)	<.001
Total[c]	381 (47.6)	511 (63.9)	.001	543 (67.9)	646 (80.8)	.01	224 (28.0)	328 (41.0)	.02	302 (37.8)	381 (47.6)	.04

[a]The *P* values correspond to the chi-square tests comparing each language model's accuracy rate with and without Med-RISE.

[b]USMLE: US Medical Licensing Examination.

[c]Represents the combined results from 4 datasets (MedQA, PubMedQA—revised version, MedMCQA, and EYE500) with 200 questions each (n=800).

Figure 3. Bar plot of enhanced accuracy in the large language models with Med-RISE. Each bar shows the accuracy of each model on a specific dataset, with the numbers representing the difference in accuracy before and after Med-RISE integration. *P<.05. USMLE: US Medical Licensing Examination.



GPT-3.5

The integration of Med-RISE led to an accuracy increase from 47% to 70% (a 23% improvement; P<.001) in MedQA (USMLE), from 39% to 63% (a 24% improvement; P<.001) in PubMedQA, from 45% to 72.5% (a 27.5% improvement; P<.001) in PubMedQA—revised version, from 56.5% to 62% (a 5.5% improvement; P=.26) in MedMCQA, and from 42% to 51% (a 9% improvement; P=.07) in EYE500.

GPT-4

After incorporating Med-RISE, GPT-4 significantly improved in accuracy across all datasets. The accuracy on the MedQA (USMLE) dataset increased from 76% to 86.5% (a 10.5% improvement; P=.02). For PubMedQA, accuracy increased from 56% to 67.5% (an 11.5% improvement; P=.02). For PubMedQA—revised version, accuracy increased from 67% to 83.5% (a 16.5% improvement; P<.001). For MedMCQA, accuracy increased from 65.5% to 78% (a 12.5% improvement; P=.01). For EYE500, there was a substantial increase in accuracy from 63% to 75% (a 12% improvement; P=.009).

Vicuna-13B

With the application of Med-RISE, Vicuna-13B improved in accuracy from 27% to 28.5% (a 1.5% improvement; P=.74) in MedQA (USMLE), from 3.5% to 13.5% (a 10% improvement; P=.001) in PubMedQA, from 24.5% to 35% (a 10.5% improvement; P=.02) in PubMedQA—revised version, from 37.5% to 48.5% (an 11% improvement; P=.03) in MedMCQA, and from 23% to 52% (a 29% improvement; P<.001) in EYE500.

ChatGLM-6B

The implementation of Med-RISE with ChatGLM-6B resulted in an increase in accuracy from 23% to 25% (a 2% improvement; P=.64) in MedQA (USMLE), from 48.5% to 51.5% (a 3% improvement; P=.55) in PubMedQA, from 56.5% to 64.5% (an 8% improvement; P=.10) in PubMedQA—revised version, from 38.5% to 47% (an 8.5% improvement; P=.09) in MedMCQA, and from 33% to 54% (a 21% improvement; P<.001) in EYE500.

Assessing Hallucinations in the Reasoning Process of LLMs via CoT

The reasoning process of the LLMs was evaluated using the CoT method, specifically focusing on the tendency of these models to generate hallucinations. It was observed that all LLMs were prone to this issue, albeit to varying extents. The average hallucination rates were 78.6% for Vicuna-13B, 67.3% for ChatGLM-6B, and 56.3% for GPT-3.5, with the lowest being that for GPT-4 at 35.3% (Table 2).

Integrating Med-RISE into these LLMs substantially reduced hallucination occurrences, with reductions ranging from 11.8% to 18% (mean 15.1%, SD 2.8%) across 4 datasets: MedQA (USMLE), PubMedQA—revised version, MedMCQA, and

EYE500 (Figure 4). Specifically, GPT-3.5 hallucination rates decreased from 56.3% to 38.6% (reduction of 17.7%), GPT-4 hallucination rates decreased from 35.3% to 22.5% (reduction of 12.8%), Vicuna-13B hallucination rates decreased from 78.6% to 60.6% (reduction of 18%), and ChatGLM-6B hallucination rates decreased from 67.3% to 55.5% (reduction of 11.8%). Furthermore, the application of Med-RISE to the LLMs even led to a reduction in hallucinations exceeding 20% in certain cases. In the MedQA (USMLE) dataset, GPT-3.5 exhibited a significant reduction in hallucinations of 23.5% (P<.001) after Med-RISE integration. In the EYE500 dataset, Vicuna-13B and ChatGLM-6B exhibited significant reductions in hallucinations of 32.5% (P<.001) and 20.5% (P<.001), respectively, after Med-RISE integration. Similarly, the PubMedQA—revised version revealed a substantial decrease in hallucinations of 31% (P<.001) for GPT-3.5 and 22% (P<.001) for GPT-4 when applying Med-RISE.

Table 2. Incidence of hallucinations in chain of thought among the large language models.[a]

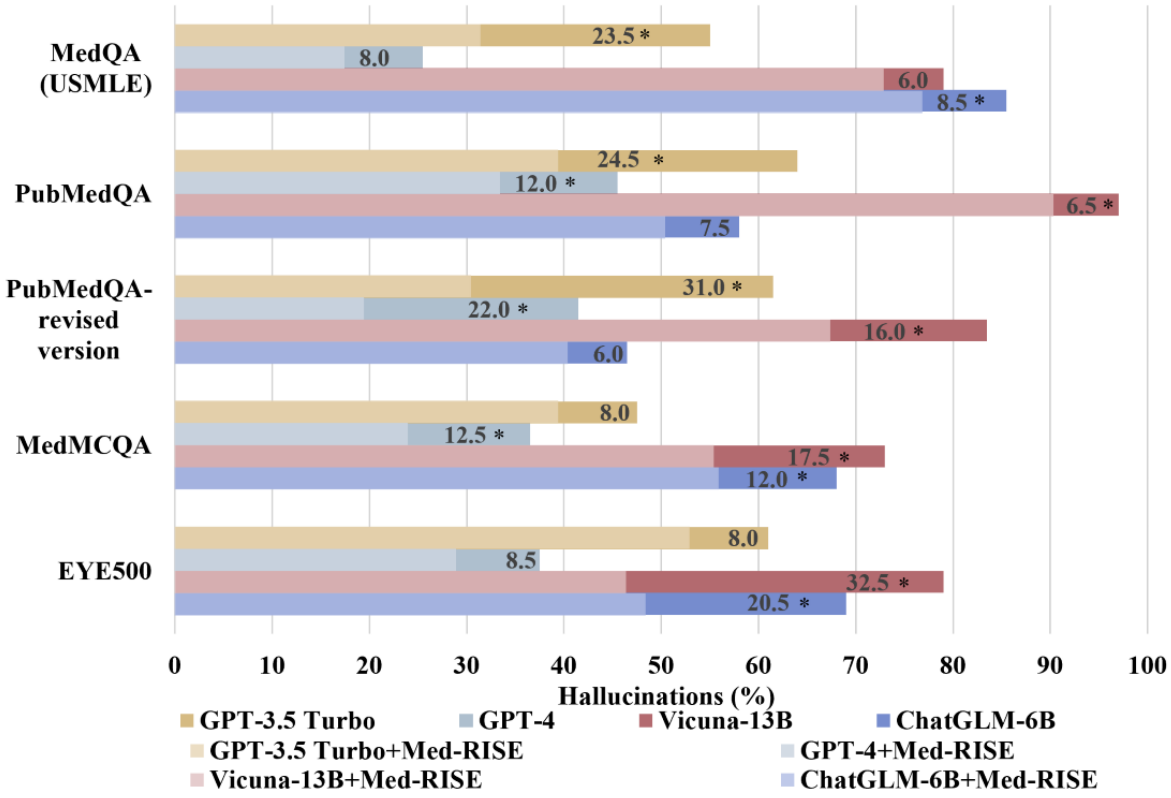
Dataset	GPT-3.5 Turbo, n (%)			GPT-4, n (%)			Vicuna-13B, n (%)			ChatGLM-6B, n (%)		
	Without Med-RISE	With Med-RISE	*P* value	Without Med-RISE	With Med-RISE	*P* value	Without Med-RISE	With Med-RISE	*P* value	Without Med-RISE	With Med-RISE	*P* value
MedQA (USMLE[b])	110 (55.0)	63 (31.5)	<.001	51 (25.5)	35 (17.5)	.05	158 (79.0)	146 (73.0)	.16	171 (85.5)	154 (77.0)	.03
PubMedQA	128 (64.0)	79 (39.5)	<.001	91 (45.5)	67 (33.5)	.01	194 (97.0)	181 (90.5)	.007	116 (58.0)	101 (50.5)	.13
PubMedQA—revised version	123 (61.5)	61 (30.5)	<.001	83 (41.5)	39 (19.5)	<.001	167 (83.5)	135 (67.5)	<.001	93 (46.5)	81 (40.5)	.27
MedMCQA	95 (47.5)	79 (39.5)	.11	73 (36.5)	48 (24.0)	.007	146 (73.0)	111 (55.5)	<.001	136 (68.0)	112 (56.0)	.01
EYE500	122 (61.0)	106 (53.0)	.11	75 (37.5)	58 (29.0)	.07	158 (79.0)	93 (46.5)	<.001	138 (69.0)	97 (48.5)	<.001
Total[c]	450 (56.3)	309 (38.6)	.009	282 (35.3)	180 (22.5)	.003	629 (78.6)	485 (60.6)	<.001	538 (67.3)	444 (55.5)	.02

[a]The *P* values correspond to the chi-square tests comparing each language model's hallucination rate with and without Med-RISE integration.

[b]USMLE: US Medical Licensing Examination.

[c]Represents the combined results from 4 datasets (MedQA, PubMedQA—revised version, MedMCQA, and EYE500) with 200 questions each (n=800).

Figure 4. Bar plot of hallucination reduction in the large language models with Med-RISE. Each bar shows the hallucination rate of each model, with the numbers representing the difference in hallucination rate before and after Med-RISE integration. *P<.05. USMLE: US Medical Licensing Examination.



Reduction in Different Hallucination Types Through Med-RISE

Table 3 presents the proportion of different types of hallucinations observed in CoT from the 4 different LLMs with and without Med-RISE integration. For this analysis, 50 questions were randomly selected from the 4 datasets, and these questions were answered by the 4 models with and without Med-RISE, resulting in 200 question-answer pairs for each set. The hallucinations were categorized into 2 main types: factuality hallucinations and faithfulness hallucinations, with further subtypes for each.

The results showed that the application of Med-RISE led to reductions in both main categories of hallucinations, with factuality hallucinations decreasing by 13.5% and faithfulness hallucinations decreasing by 5.8%. When examining specific subtypes, factuality hallucinations showed reductions in both factual inconsistency (42/200, 21%) and factual fabrication (12/200, 6%). Similarly, faithfulness hallucinations demonstrated reductions across instruction inconsistency (8/200, 4%), context inconsistency (15/200, 7.5%), and logical inconsistency (12/200, 6%). Representative examples of these different hallucination types are shown in Textboxes S1 and S2 in Multimedia Appendix 3.

Table 3. Proportion of different types of hallucinations of the large language models in chain of thought (N=200)[a].

Type of hallucination and pattern	Without Med-RISE, n (%)	With Med-RISE, n (%)	Change, n (%)
Factuality hallucination			
Factual inconsistency	85 (42.5)	43 (21.5)	42 (21.0)
Factual fabrication	38 (19.0)	26 (13.0)	12 (6.0)
Faithfulness hallucination			
Instruction inconsistency	23 (11.5)	15 (7.5)	8 (4.0)
Context inconsistency	29 (14.5)	14 (7.0)	15 (7.5)
Logical inconsistency	47 (23.5)	35 (17.5)	12 (6.0)

[a]A total of 50 questions were randomly chosen from the 4 datasets. Each group had 200 question-answer pairs.

Discussion

Principal Findings

This study demonstrates the effectiveness of Med-RISE across multiple medical disciplines in enhancing the medical QA capabilities of LLMs. Our evaluation shows that Med-RISE achieved accuracy improvements ranging from 9.8% to 16.3% (mean 13%, SD 2.3%) across all tested LLMs and datasets while significantly reducing hallucinations, with reductions ranging from 11.8% to 18% (mean 15.1%, SD 2.8%), factuality hallucinations decreasing by 13.5%, and faithfulness hallucinations decreasing by 5.8%. These comprehensive improvements in both accuracy and hallucination mitigation demonstrate Med-RISE's capability in addressing 2 fundamental challenges of medical LLMs: knowledge and response reliability.

Comparison With Prior Work

Compared to previous RAG frameworks [44,45], which often depend on static and more limited knowledge bases [46,47], Med-RISE represents a significant advancement through its dynamic, real-time knowledge retrieval across multiple medical disciplines. Building on RISE [37], which we previously implemented for diabetes-specific QA, Med-RISE extends these capabilities across broader medical domains with enhanced safety verification. Previous approaches—such as the focus on diffuse large B-cell lymphoma by Soong et al [35] using 500 PubMed articles, the liver disease–specific LiVersa by Ge et al [34], the Almanac framework by Zakka et al [31] for treatment guidelines, and the medical textbook augmentation by Wang et al [26]—have demonstrated limited application as they were constrained by their fixed knowledge bases. More importantly, even with RAG implementation, LLMs still face inherent challenges regarding errors and hallucinations. Med-RISE uniquely addresses this through its additional layer of accuracy and safety verification before output generation, significantly reducing error rates in medical responses. This dual-stage approach—combining dynamic knowledge retrieval with accuracy and safety verification—makes Med-RISE a more reliable and generalizable tool for medical QA than conventional RAG frameworks.

Research Implications

The integration of Med-RISE with LLMs achieved 2 critical improvements in medical QA performance: average accuracy increases of 13% across 4 general medical datasets and a 15% reduction in hallucinations, with the most substantial improvement in factual inconsistency (42/200, 21% reduction). These significant improvements stem from Med-RISE's novel mechanism that combines real-time medical knowledge retrieval with accuracy and safety verification filters, effectively reducing both outdated information and factual errors. Med-RISE functions as an external augmentation framework that enhances domain-specific performance without modifying underlying model parameters, offering cost-effective adaptation compared to expensive model retraining while enabling broader institutional adoption. In medical applications, even small improvements in accuracy can have profound implications given the field's sensitivity to precision. Each percentage point of enhanced accuracy directly impacts the quality of diagnoses, treatment strategies, and patient education. Medical misinformation can lead to severe consequences, from compromised clinical decision-making to patient safety risks [48]. Therefore, Med-RISE's substantial improvements in both accuracy and reliability represent crucial advancements for deploying LLMs in health care settings.

Future Directions

Med-RISE demonstrates significant potential for advancing clinical practice through its robust framework for medical applications. Its effectiveness stems from 2 key components: real-time knowledge retrieval across medical disciplines enabling transparent responses and verification of accuracy and safety. This proven performance allows for Med-RISE's integration across various health care processes—from preconsultation screening and surgical consent to patient decision support, postoperative follow-up, disease consultation, and public health education. Our study also revealed that, while Med-RISE achieved improvements across all tested models, smaller-parameter models exhibited greater performance variability across different datasets, whereas larger models demonstrated more consistent gains. Future work may consider using larger models as the foundational framework to achieve more stable performance. Currently, Med-RISE achieves response times of 5 to 20 seconds through parallel retrieval processing (using 8 V100 GPUs of 32 GB each). For faster responses, performance can be enhanced through higher-performance GPUs, increased parallel processing, and improved network speeds.

Limitations

This study has several limitations. It adopted random subsets from each dataset, which may lead to performance variations and might not fully reflect the outcomes of a comprehensive analysis of the entire dataset. Furthermore, Med-RISE's impact on response time, computational requirements, and user experience remains to be explored. Future work should focus on expanding knowledge diversity and incorporating reinforcement learning methods such as proximal policy optimization [49] and direct preference optimization [50] to enhance Med-RISE's alignment with human values and preferences, further improving response quality. Comprehensive evaluations that include feedback from clinical users and health care professionals in real clinical settings will also be essential to validate the framework's effectiveness and refine its practical implementation.

Conclusions

In conclusion, Med-RISE enhances the clinical application of LLMs through its unique integration of real-time knowledge retrieval and safety verification mechanisms. By significantly improving accuracy, reducing hallucinations, and enabling transparent responses through knowledge-grounded generation, the framework ensures safer and more reliable medical information delivery across diverse medical scenarios. Med-RISE demonstrates strong potential for advancing health care applications from patient consultation to public health education while maintaining essential safety standards.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (grants 82171034 and 81721003).

Data Availability

The code for the retrieval component of the Med-RISE framework is publicly available (Multimedia Appendix 1). A sample of the Med-RISE local database (n=500) is included in Multimedia Appendix 2. A subset of the EYE500 dataset (n=50) is provided in Multimedia Appendix 4. The revised PubMedQA dataset (PubMedQA—revised version; n=200) used in this study is available with open access in Multimedia Appendix 5. The full code and local database of Med-RISE in this study are available from the corresponding author on reasonable request.

Authors' Contributions

Dingqiao W, JY, and J Li contributed equally as the first authors. MG, FL, WD, and Y-FZ contributed equally as co-corresponding authors. MG, FL, WD, and Y-FZ contributed to conceptualization. Dingqiao W, JY, and J Li performed the methodology. The investigation was done by J Liang, QZ, QH, CP, Dongliang W, ZL, and WS. Dingqiao W, JY, and J Li wrote the original draft. MG, FL, WD, and Y-FZ handled the writing—review and editing. MG, FL, WD, and Y-FZ performed supervision. Final approval of the version to be published was done by all authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Code for the Med-RISE framework.
[ZIP File (Zip Archive), 18 KB - mededu_v11i1e70190_app1.zip]

Multimedia Appendix 2
Sample of Med-RISE local dataset (n=500).
[XLSX File (Microsoft Excel File), 225 KB - mededu_v11i1e70190_app2.xlsx]

Multimedia Appendix 3
Supplementary tables and figures.
[DOCX File , 37 KB - mededu_v11i1e70190_app3.docx]

Multimedia Appendix 4
Sample of EYE500 dataset (n=50).
[XLSX File (Microsoft Excel File), 17 KB - mededu_v11i1e70190_app4.xlsx]

Multimedia Appendix 5
PubMedQA_revised version (n=200).
[XLSX File (Microsoft Excel File), 128 KB - mededu_v11i1e70190_app5.xlsx]

References

1. Introducing ChatGPT. OpenAI. URL: https://openai.com/index/chatgpt/ [accessed 2024-08-15]
2. Else H. Abstracts written by ChatGPT fool scientists. Nature 2023 Jan;613(7944):423. [doi: 10.1038/d41586-023-00056-7] [Medline: 36635510]
3. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. JAMA 2023 Mar 14;329(10):842-844 [FREE Full text] [doi: 10.1001/jama.2023.1044] [Medline: 36735264]
4. Mihalache A, Huang RS, Popovic MM, Muni RH. Performance of an upgraded artificial intelligence chatbot for ophthalmic knowledge assessment. JAMA Ophthalmol 2023 Aug 01;141(8):798-800. [doi: 10.1001/jamaophthalmol.2023.2754] [Medline: 37440220]
5. Sharma P, Parasa S. ChatGPT and large language models in gastroenterology. Nat Rev Gastroenterol Hepatol 2023 Aug;20(8):481-482. [doi: 10.1038/s41575-023-00799-8] [Medline: 37253794]

6. Huang H, Zheng O, Wang D, Yin J, Wang Z, Ding S, et al. ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model. Int J Oral Sci 2023 Jul 28;15(1):29 [FREE Full text] [doi: 10.1038/s41368-023-00239-y] [Medline: 37507396]
7. Tariq R, Malik S, Khanna S. Evolving landscape of large language models: an evaluation of ChatGPT and Bard in answering patient queries on colonoscopy. Gastroenterology 2024 Jan;166(1):220-221. [doi: 10.1053/j.gastro.2023.08.033] [Medline: 37634736]
8. Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA 2025 Jan 28;333(4):319-328. [doi: 10.1001/jama.2024.21700] [Medline: 39405325]
9. Armbruster J, Bussmann F, Rothhaas C, Titze N, Grützner PA, Freischmidt H. "Doctor ChatGPT, can you help me?" The patient's perspective: cross-sectional study. J Med Internet Res 2024 Oct 01;26:e58831 [FREE Full text] [doi: 10.2196/58831] [Medline: 39352738]
10. Jahani Yekta MM. The general intelligence of GPT–4, its knowledge diffusive and societal influences, and its governance. Meta Radiol 2024 Jun;2(2):100078. [doi: 10.1016/j.metrad.2024.100078]
11. Ali R, Tang OY, Connolly ID, Zadnik Sullivan PL, Shin JH, Fridley JS, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. Neurosurgery 2023 Dec 01;93(6):1353-1365. [doi: 10.1227/neu.0000000000002632] [Medline: 37581444]
12. Antaki F, Milad D, Chia MA, Giguère C, Touma S, El-Khoury J, et al. Capabilities of GPT-4 in ophthalmology: an analysis of model entropy and progress towards human-level medical question answering. Br J Ophthalmol 2024 Oct 20;108(10):1371-1378. [doi: 10.1136/bjo-2023-324438] [Medline: 37923374]
13. Miao J, Thongprayoon C, Garcia Valencia OA, Krisanapan P, Sheikh MS, Davis PW, et al. Performance of ChatGPT on nephrology test questions. Clin J Am Soc Nephrol 2024 Jan 01;19(1):35-43. [doi: 10.2215/CJN.0000000000000330] [Medline: 37851468]
14. Pushpanathan K, Lim ZW, Er Yew SM, Chen DZ, Hui'En Lin HA, Lin Goh JH, et al. Popular large language model chatbots' accuracy, comprehensiveness, and self-awareness in answering ocular symptom queries. iScience 2023 Nov 17;26(11):108163 [FREE Full text] [doi: 10.1016/j.isci.2023.108163] [Medline: 37915603]
15. Gupta R, Park JB, Herzog I, Yosufi N, Mangan A, Firouzbakht PK, et al. Applying GPT-4 to the plastic surgery inservice training examination. J Plast Reconstr Aesthet Surg 2023 Dec;87:78-82. [doi: 10.1016/j.bjps.2023.09.027] [Medline: 37812847]
16. Merlino DJ, Brufau SR, Saieed G, Van Abel KM, Price DL, Archibald DJ, et al. Comparative assessment of otolaryngology knowledge among large language models. Laryngoscope 2025 Feb;135(2):629-634. [doi: 10.1002/lary.31781] [Medline: 39305216]
17. Tian S, Jin Q, Yeganova L, Lai P, Zhu Q, Chen X, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. ArXiv Preprint posted online on October 17, 2023 [FREE Full text] [Medline: 37904734]
18. Longwell JB, Hirsch I, Binder F, Gonzalez Conchas GA, Mau D, Jang R, et al. Performance of large language models on medical oncology examination questions. JAMA Netw Open 2024 Jun 03;7(6):e2417641 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.17641] [Medline: 38888919]
19. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023 Mar 30;388(13):1233-1239. [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
20. Moor M, Banerjee O, Abad ZS, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature 2023 Apr;616(7956):259-265. [doi: 10.1038/s41586-023-05881-4] [Medline: 37045921]
21. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023 Aug;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
22. Webster P. Medical AI chatbots: are they safe to talk to patients? Nat Med 2023 Nov 08;29(11):2677-2679. [doi: 10.1038/s41591-023-02535-w] [Medline: 37684542]
23. Tan TF, Thirunavukarasu AJ, Campbell JP, Keane PA, Pasquale LR, Abramoff MD, et al. Generative artificial intelligence through ChatGPT and other large language models in ophthalmology: clinical applications and challenges. Ophthalmol Sci 2023 Dec;3(4):100394 [FREE Full text] [doi: 10.1016/j.xops.2023.100394] [Medline: 37885755]
24. Schubert MC, Wick W, Venkataramani V. Performance of large language models on a neurology board-style examination. JAMA Netw Open 2023 Dec 01;6(12):e2346721 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.46721] [Medline: 38060223]
25. Small WR, Wiesenfeld B, Brandfield-Harvey B, Jonassen Z, Mandal S, Stevens ER, et al. Large language model-based responses to patients' in-basket messages. JAMA Netw Open 2024 Jul 01;7(7):e2422399 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.22399] [Medline: 39012633]
26. Wang C, Ong J, Wang C, Ong H, Cheng R, Ong D. Potential for GPT technology to optimize future clinical decision-making using retrieval-augmented generation. Ann Biomed Eng 2024 May 02;52(5):1115-1118. [doi: 10.1007/s10439-023-03327-6] [Medline: 37530906]
27. Wang Y, Ma X, Chen W. Augmenting black-box LLMs with medical textbooks for biomedical question answering. ArXiv Preprint posted online on September 5, 2023 [FREE Full text] [doi: 10.18653/v1/2024.findings-emnlp.95]

28. Lozano A, Fleming SL, Chiang CC, Shah N. Clinfo.ai: an open-source retrieval-augmented large language model system for answering medical questions using scientific literature. Pac Symp Biocomput 2024;29:8-23 [FREE Full text] [Medline: 38160266]
29. Munikoti S, Acharya A, Wagle S, Horawalavithana S. Evaluating the effectiveness of retrieval-augmented large language models in scientific document reasoning. ArXiv Preprint posted online on November 7, 2023 [FREE Full text] [doi: 10.18653/v1/2024.sdp-1.8]
30. Zhao S, Yang Y, Wang Z, He Z, Qiu LK, Qiu L. Retrieval augmented generation (RAG) and beyond: a comprehensive survey on how to make your LLMs use external data more wisely. ArXiv Preprint posted online on September 23, 2024 [FREE Full text]
31. Zakka C, Shad R, Chaurasia A, Dalal AR, Kim JL, Moor M, et al. Almanac-retrieval-augmented language models for clinical medicine. NEJM AI 2024 Feb 25;1(2):10.1056/aioa2300068 [FREE Full text] [doi: 10.1056/aioa2300068] [Medline: 38343631]
32. Unlu O, Shin J, Mailly CJ, Oates MF, Tucci MR, Varugheese M, et al. Retrieval augmented generation enabled generative pre-trained transformer 4 (GPT-4) performance for clinical trial screening. medRxiv Preprint posted online on February 8, 2024 [FREE Full text] [doi: 10.1101/2024.02.08.24302376] [Medline: 38370719]
33. Luo MJ, Pang J, Bi S, Lai Y, Zhao J, Shang Y, et al. Development and evaluation of a retrieval-augmented large language model framework for ophthalmology. JAMA Ophthalmol 2024 Sep 01;142(9):798-805. [doi: 10.1001/jamaophthalmol.2024.2513] [Medline: 39023885]
34. Ge J, Sun S, Owens J, Galvez V, Gologorskaya O, Lai JC, et al. Development of a liver disease-specific large language model chat interface using retrieval-augmented generation. Hepatology 2024 Nov 01;80(5):1158-1168. [doi: 10.1097/HEP.0000000000000834] [Medline: 38451962]
35. Soong D, Sridhar S, Si H, Wagner JS, Sá AC, Yu CY, et al. Improving accuracy of GPT-3/4 results on biomedical data using a retrieval-augmented language model. PLOS Digit Health 2024 Aug;3(8):e0000568. [doi: 10.1371/journal.pdig.0000568] [Medline: 39167594]
36. Miao J, Thongprayoon C, Suppadungsuk S, Garcia Valencia OA, Cheungpasitporn W. Integrating retrieval-augmented generation with large language models in nephrology: advancing practical applications. Medicina (Kaunas) 2024 Mar 08;60(3):445 [FREE Full text] [doi: 10.3390/medicina60030445] [Medline: 38541171]
37. Wang D, Liang J, Ye J, Li J, Li J, Zhang Q, et al. Enhancement of the performance of large language models in diabetes education through retrieval-augmented generation: comparative study. J Med Internet Res 2024 Nov 08;26:e58041 [FREE Full text] [doi: 10.2196/58041] [Medline: 39046096]
38. Jin Q, Leaman R, Lu Z. PubMed and beyond: biomedical literature search in the age of artificial intelligence. EBioMedicine 2024 Feb;100:104988 [FREE Full text] [doi: 10.1016/j.ebiom.2024.104988] [Medline: 38306900]
39. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans Inf Syst 2025 Jan 24;43(2):1-55. [doi: 10.1145/3703155]
40. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Appl Sci 2021;11(14):6421 [FREE Full text] [doi: 10.3390/app11146421]
41. Pal A, Umapathi LK, Sankarasubbu M. MedMCQA: a large-scale multi-subject multi-choice dataset for medical domain question answering. In: Proceedings of the Conference on Health, Inference, and Learning. 2022 Presented at: PMLR 2022; April 7-8, 2022; Virtual Event URL: https://proceedings.mlr.press/v174/pal22a.html
42. Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. PubMedQA: a dataset for biomedical research question answering. ArXiv Preprint posted online on September 13, 2019 [FREE Full text] [doi: 10.18653/v1/d19-1259]
43. Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? Patterns (N Y) 2024 Mar 08;5(3):100943 [FREE Full text] [doi: 10.1016/j.patter.2024.100943] [Medline: 38487804]
44. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020 Presented at: NIPS'20; December 6-12, 2020; Vancouver, BC.
45. Mao Y, He P, Liu X, Shen Y, Gao J, Han J, et al. Generation-augmented retrieval for open-domain question answering. ArXiv Preprint posted online on September 17, 2020 [FREE Full text] [doi: 10.18653/v1/2021.acl-long.316]
46. Yu W. Retrieval-augmented generation across heterogeneous knowledge. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop. 2022 Presented at: NAACL 2022; July 10-15, 2022; Seattle, WA URL: https://aclanthology.org/2022.naacl-srw.7/ [doi: 10.18653/v1/2022.naacl-srw.7]
47. Feng Z, Feng X, Zhao D, Yang M, Qin B. Retrieval-generation synergy augmented large language models. ArXiv Preprint posted online on October 8, 2023 [FREE Full text] [doi: 10.1109/icassp48485.2024.10448015]
48. Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. Bull World Health Organ 2020;98:251-256 [FREE Full text] [doi: 10.2471/blt.19.237487]
49. Zheng R, Dou S, Gao S, Hua Y, Shen W, Wang B, et al. Secrets of RLHF in large language models part I: PPO. ArXiv Preprint posted online on July 11, 2023 [FREE Full text] [doi: 10.48550/arXiv.2307.04964]

50. Rafailov R, Sharma A, Mitchell E, Manning CD, Ermon S, Finn C. Direct preference optimization: your language model is secretly a reward model. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. 2023 Presented at: NIPS '23; December 10-16, 2023; New Orleans, LA URL: https://dl.acm.org/doi/10.5555/3666122.3668460

Abbreviations

CoT: chain-of-thought
GPU: graphics processing unit
LLM: large language model
QA: question answering
RAG: retrieval-augmented generation
USMLE: United States Medical Licensing Examination

Edited by B Lesselroth; submitted 18.12.24; peer-reviewed by S Mohanadas, T Abdullahi, W Kim; comments to author 20.06.25; revised version received 09.09.25; accepted 30.09.25; published 02.12.25.

Please cite as:
Wang D, Ye J, Li J, Liang J, Zhang Q, Hu Q, Pan C, Wang D, Liu Z, Shi W, Guo M, Li F, Du W, Zheng YF
Enhancing Large Language Models for Improved Accuracy and Safety in Medical Question Answering: Comparative Study
JMIR Med Educ 2025;11:e70190
URL: https://mededu.jmir.org/2025/1/e70190
doi:10.2196/70190
PMID:

Original Paper

Large Language Model–Based Patient Simulation to Foster Communication Skills in Health Care Professionals: User-Centered Development and Usability Study

Ahmed Elhilali[1]; Andy Suy-Huor Ngo[1]; Daniel Reichenpfader[1], MSc; Kerstin Denecke[1], Dr rer nat

Institute Patient-centered Digital Health, Bern University of Applied Sciences, Biel, Switzerland

Corresponding Author:
Kerstin Denecke, Dr rer nat
Institute Patient-centered Digital Health
Bern University of Applied Sciences
Quellgasse 21
Biel, 2501
Switzerland
Phone: 41 32321 67 94
Email: kerstin.denecke@bfh.ch

Abstract

Background: Case-based learning using standardized patients is a key method for teaching communication skills in medicine. Besides logistical and financial hurdles, standardized patients portrayed by actors cannot cover the complete diversity of sociodemographic factors of patients. Large language models (LLMs) show promise for creating scalable patient simulations and could probably cover a broader diversity of factors. They could also be integrated into the continuous training of future health care professionals' communication and interaction skills.

Objective: This study aimed to introduce the system architecture of a digital tool that leverages LLMs to simulate patient conversations for medical education, focusing specifically on medical history taking. Through an explorative analysis, we aimed to assess the tool's usability and examine differences between LLMs in simulating patient encounters.

Methods: We followed a user-centered design process, gathering initial requirements from 2 medical students. We then developed a fully functional web prototype using a Python Flask backend and a PostgreSQL database, integrating 5 LLMs from OpenAI, Anthropic, and xAI. The system includes an artificial intelligence–assisted case vignette generator and a dynamic patient simulator. For the explorative analysis of the prototype, we conducted a task-based usability test with 5 medical students, measuring their experience using the System Usability Scale (SUS) questionnaire and qualitative questions. We then conducted an explorative analysis in which 4 practicing physicians evaluated the simulation quality of 3 models (Grok 3, GPT-4, and Claude 3 Opus) across 7 criteria on a 5-point Likert scale.

Results: Usability testing yielded a mean SUS score of 91.5 (SD 8.40), indicating high perceived usability in a small formative sample. Students praised the system's simplicity and intuitive design but noted the absence of a formal conclusion and performance feedback, expressing a desire for a "didactic loop" to maximize learning. The models showed limitations in simulating uncertainties and memory lapses, responding to follow-up questions, and producing natural conversational flow. They perform well in simulating a coherent symptom profile, in using patient-like language, and in describing a realistic timeline and symptom progression. The differences among the models were not statistically significant. Ratings showed limited discriminative reliability (Kendall W=0-0.19, ie, very low) and a ceiling effect, with most scores clustered at 4-5, constraining interpretation; all group differences should therefore be viewed as exploratory.

Conclusions: We successfully developed a highly usable patient simulation tool that serves as a foundation for further development. Our results show that while the tool could be effective for communication training, its full potential will only be realized by integrating an automated feedback mechanism to create a complete didactic loop, as requested by the test users. Future work should assess in more depth the differences among the models in simulating psychosocial patient characteristics.

(JMIR Med Educ 2025;11:e81271) doi:10.2196/81271

KEYWORDS

chatbot; large language model; medical education; patient simulation; vignette

Introduction

Background

Effective communication between physicians and patients is an established core competency in medical education and practice. The demand for more patient-centered medical decision-making requires that the various needs of patients must be identified in conversations [1]. Nevertheless, current guidelines for doctor-patient communication show deficits in communication skills, especially in dealing with psychosocial factors and cultural diversity, which are increasingly shaping everyday clinical practice [2].

Case-based learning is a well-established pedagogical strategy in medical education, designed to foster students' clinical reasoning by engaging them with authentic patient scenarios [3]. Unlike didactic teaching methods, case-based learning promotes active, contextualized learning, encouraging students to analyze complex clinical situations and develop vital decision-making skills [4]. A cornerstone of case-based learning—particularly for practicing communication competencies—is the use of standardized patients. Traditionally, these standardized patients are represented by trained actors who take the role of patients with specific medical histories and personalities [5]. The simulated interactions offer students a safe, low-risk environment in which they can practice conducting patient interviews, make mistakes, and receive direct feedback—all of which have been shown to significantly enhance clinical communication skills [6].

A fundamental element in constructing realistic patient simulations is the clinical vignette—a concise, precise description of a hypothetical scenario [7]. In medical education, vignettes are commonly used to represent complex clinical situations, assess diagnostic reasoning, and evaluate students' decision-making [8]. For a vignette to serve as a solid foundation for patient simulation, it must encompass not only primary medical information, such as current symptoms and medical history, but also secondary factors that influence communication, including demographic details, health literacy, socioeconomic status, and personality traits [9].

Standardized patients have proven effective in teaching basic conversation techniques, but they have significant limitations. The availability and scalability of standardized patient programs are limited by high logistical and financial costs, standardization between different actors is difficult to ensure, and the authentic simulation of complex psychosocial contexts is only possible to a limited extent for logistical reasons (eg, foreign languages, involvement of children, or emotionally extremely stressful topics such as domestic violence). Technical simulation dolls are primarily suitable for training procedural skills but offer only limited possibilities for training complex communicative interactions that require emotional sensitivity and adaptive responses [10]. Digital technologies now offer promising alternatives. Recent advances in large language models (LLMs) have enabled the dynamic generation of text based on patterns learned from vast corpora, making it possible to simulate dialog-based patient interactions. Interaction with LLMs is facilitated through user and system prompts—carefully engineered instructions that ensure the model has all relevant information to perform the desired task accurately.

Related Work and Contribution

The importance of psychosocial factors influencing doctor-patient communication has increased significantly in recent years. Factors such as migration background, language barriers, socioeconomic status, and cultural differences can hinder communication, promote misunderstandings, and impair therapy adherence and equal health opportunities [11,12]. Research on institutional mistrust, which can arise from historical discrimination and negative experiences and has a lasting impact on the quality of treatment, is particularly relevant [13]. Experienced discrimination in a medical context is directly linked to increased skepticism toward health care institutions and reduced use of care [14,15]. Dealing with these psychosocial factors during patient-doctor interactions becomes more important in medical education. This work aims to create an interactive tool that can be used to train communication and interaction skills in this context.

The potential of LLMs for the simulation of standardized patients has recently been demonstrated by multiple studies. In medical training, Cook et al [16] demonstrated the viability of using GPT-4 (OpenAI) for simulating standardized patients, evaluated based on anthropomorphism, clinical accuracy, and adaptability. The researchers also showed that the LLM can score the quality of the medical assessment. Similarly, Cook et al [16] compared 2 models (GPT-4.0-Turbo and GPT-3.5-Turbo [OpenAI]) for patient simulation and demonstrated their capability of simulating dialogues, representing patient preferences, and providing personalized feedback. In a second, preliminary study, Cook [17] also performed patient simulation and limited testing using Claude (Anthropic), performing "exceptionally well." Öncü et al [18] applied GPT-4o (OpenAI) to create an environment for intern physicians to practice case-management skills. Borg et al [19] combined the LLM-based simulation (GPT-3.5-Turbo [OpenAI]) with social robotics, comparing this setup with a conventional, computer-based simulation for fostering clinical reasoning skills in medical students. The mixed methods study showed that 15 students perceived the LLM-enhanced variant as more authentic and providing a beneficial overall learning effect. Targeting nursing education, Benfatah et al [20] showed that a small sample of 12 nursing students embraced the LLM-based interaction and recognized its value in training.

LLMs have not only demonstrated their potential to simulate virtual patients but also seem capable of generating the underlying clinical vignettes efficiently. For example, Coskun et al [21] conducted a randomized, controlled experiment for vignette generation using ChatGPT-3.5 (OpenAI) and showed that the quality of vignettes is comparable to those created by human authors. Another 2 studies demonstrate the vignette generation capability of LLMs for Japanese specifically [22,23].

We can recognize that LLM-based vignette generation and virtual patient simulation have recently received increased attention in medical informatics research. However, previous research seems to focus on each task separately (ie, vignette generation, patient simulation). Furthermore, none of the studies

mentioned above applied a user-centered development approach but rather performed experiments measuring the capability of the LLM-based simulation. Therefore, none of the studies focused on the integration of a simulation tool into existing learning processes of medical students. For example, Cook [17] used a simple, text-only Python interface.

We furthermore note that most of the studies were conducted using commercial models provided by OpenAI and do not include a comparison to other providers or models except Cook [17], who conducted limited testing with Claude (Anthropic).

Goals of the Study

Existing LLM applications in the medical education context focus primarily on medical content or general communication scenarios and neglect the systematic integration of psychosocial contexts. With this paper, we want to introduce a digital tool that leverages LLMs to simulate patient conversations based on clinical case vignettes that can be used for medical education purposes. In contrast to other patient simulators, we want to integrate the psychosocial context and offer the opportunity to train communication with patients with varying psychosocial contexts. To support the generation of case vignettes, we propose a human-in-the-loop approach consisting of a comprehensive case template and a vignette generator. In this paper, we describe the user-centered design process, the technical implementation of the tool, and the first results from usability testing. Furthermore, our goal is to study differences of LLMs in simulating patients based on case vignettes in an exploratory study. The simulation tool focuses specifically on the process of collecting the medical history.

Methods

Requirement Collection

The system development followed principles of user-centered design. Medical students, who are the targeted user group, were involved in the requirement collection process and in the usability testing. Explorative model evaluation involved medical doctors.

The initial system requirements were identified through a literature analysis and qualitative interviews with medical students as prospective end users. A total of 2 third-year medical students participated in the interviews (2/9 contacted), recruited from the social network of the authors. Both were in the 6th semester of their studies of medicine. To accelerate data collection and focus on the discussion, early functional prototypes were presented to the participants as a basis for discussion. The requirements were qualitatively analyzed and prioritized based on their perceived relevance as indicated by the students, with specific focus on the usefulness for learning clinical communication skills. This process yielded the following key functional requirements:

1. Vignette management: the system must provide a wide range of patient cases, ideally sortable by medical specialty and year of study.
2. Organization: case vignettes should be organized using a folder system inspired by familiar learning platforms.
3. Artificial intelligence (AI)–based simulation: patient simulations should be powered by generative AI to enable unpredictable and realistic dialogues.
4. Interaction design: the simulated patient should not reveal information directly but instead respond only to targeted inquiries, encouraging natural conversation flow.

Template for Case Vignettes

Building on the work of Reichenpfader and Denecke [9], a highly structured vignette template was developed to ensure realistic and diverse patient representations. Each vignette is technically stored as a JSON object within a JSONB field in the PostgreSQL (version 17; PostgreSQL Global Development Group) database. The structure is defined in the Python (version 3.11.9; Python Software Foundation) backend, which dynamically generates the user interface for vignette creation. The template comprises four main categories: (1) demographics, includes name, age, gender, occupation, and educational background; (2) medical history covers current symptoms, preexisting conditions, medications, allergies, and family history; (3) personality and communication describes health literacy, communication style, personality traits, and emotional state; and (4) social factors capture social support systems, socioeconomic status, cultural background, and language proficiency.

To facilitate thematic organization of the generated vignettes, the classification system of the Canadian Emergency Department Information System (CEDIS) Working Group [24] was implemented. Its suitability for use in medical education was confirmed by the participating students during the requirement analysis phase. Example vignettes can be found in Multimedia Appendix 1.

System Architecture

Overview

Our fully functional, web-based prototype for LLM-supported patient simulation includes the following core components: (1) AI-assisted case vignette generator—users can create vignettes manually by specifying values in a structured vignette form or have a complete vignette generated by interacting with an LLM using a “meta-prompt.” All vignettes can be organized in a folder structure (Figure 1). (2) Dynamic patient simulator—users can select a vignette and conduct a medical history gathering with a simulated patient (Figure 2).

Figure 1. Library of case vignettes. Each vignette is represented by a rectangle. Filtering along Canadian Emergency Department Information System categories is possible.

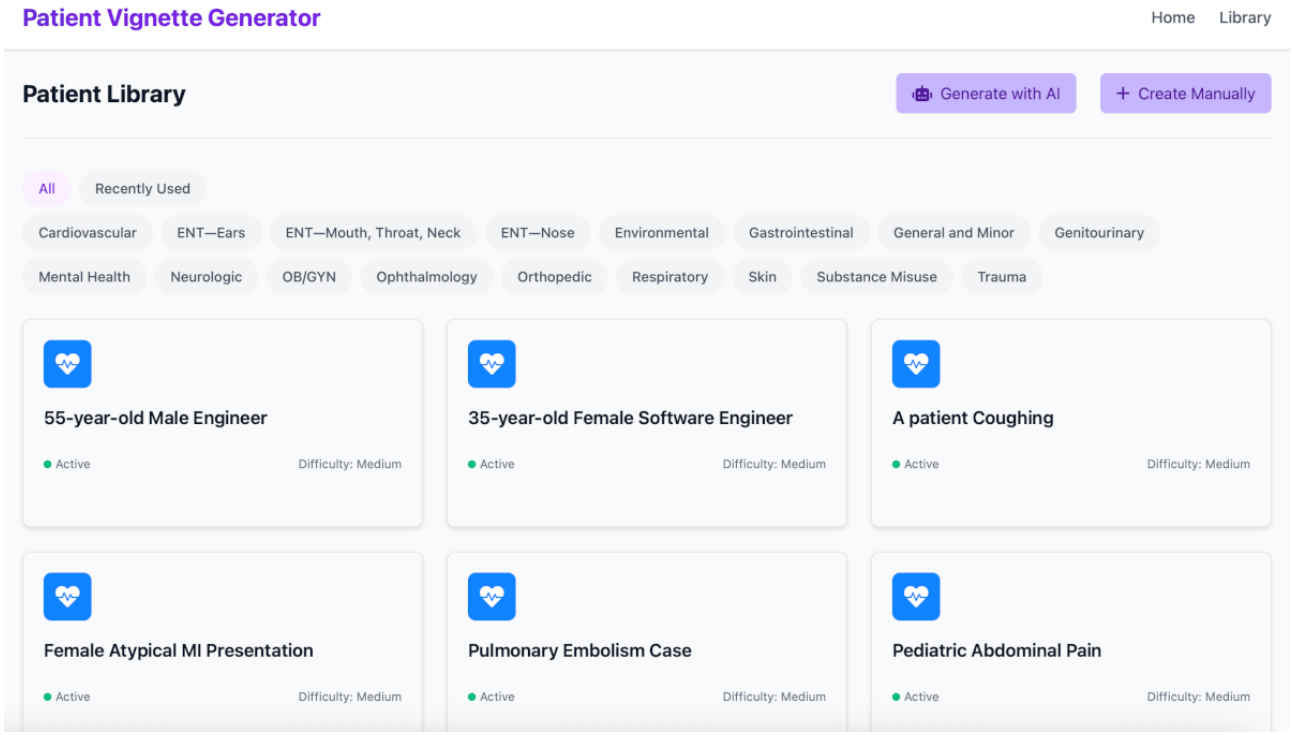
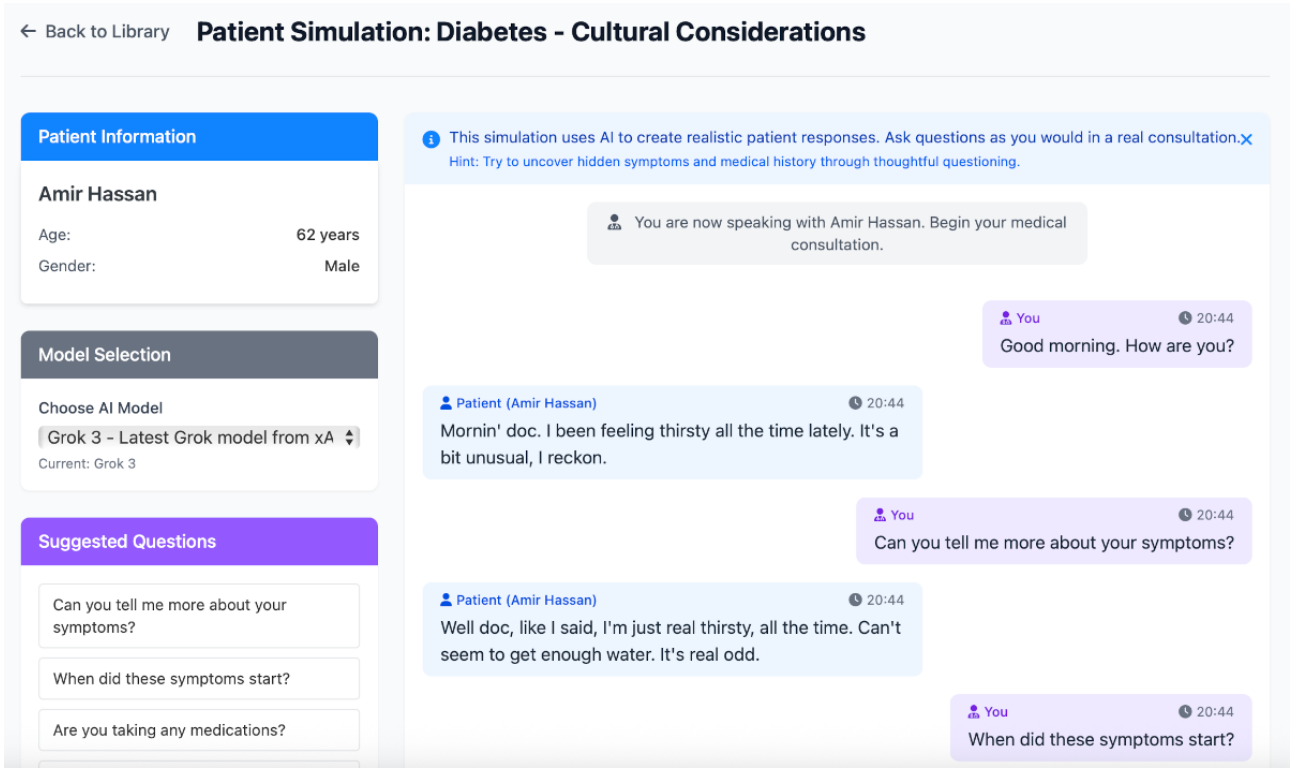


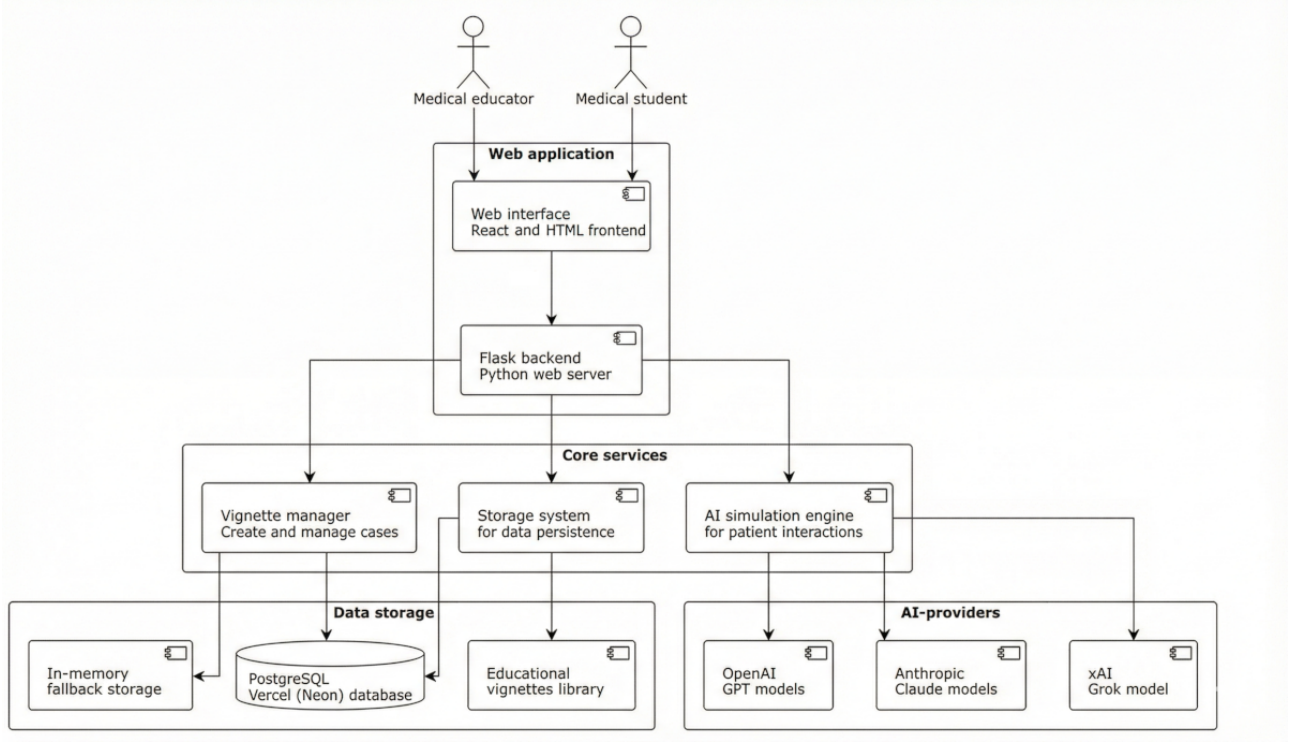
Figure 2. User interface for the interactive simulation tool. On the left, patient information is shown, together with the selected model and suggested questions as guidance for the interaction. On the right, the simulation chat is shown.



Access to the current version of the prototype will be provided upon request.

The system is implemented using a multilevel client-server architecture (Figure 3). Users interact with a web application consisting of a reactive frontend and a backend built with Flask in Python. The backend orchestrates several core services—a vignette management system for handling case definitions, a storage layer for data persistence, and an AI simulation engine that encapsulates communication with external AI services (OpenAI, Anthropic, xAI). The following technologies were used for implementing the prototype:

Figure 3. System architecture showing the users, the web application, the core components, data storage, and external artificial intelligence (AI) services.



1. Backend: Python (version 3.11.9) with the lightweight web framework Flask (version 3.0.2). Database access is abstracted via SQLAlchemy (version 2.0.41).
2. Database: PostgreSQL (version 17) was chosen for reliable and scalable data management.
3. Frontend: the user interface is implemented using standard web technologies (HTML5, CSS3).
4. Deployment: the application is deployed via the Vercel platform [25].

To ensure technological breadth and flexibility, the prototype developed in this work integrates 5 LLMs from 3 leading AI providers. These include GPT-4 and GPT-3.5-Turbo by OpenAI, Claude 3 Opus and Claude 3 Sonnet by Anthropic, and Grok 3 by xAI.

In the following, we provide more details on the two components.

AI-Assisted Case Vignette Generator

This component enables the generation of case vignettes using an LLM. It constructs a prompt that instructs an LLM to assume the role of a medical education expert and to output a case vignette in a predefined JSON format. The implementation incorporates multiple technical elements:

1. Strict output format: the prompt enforces a rigid JSON schema, including a naming convention of the form "[age]-year-old [gender] [occupation]," explicitly excluding any medical diagnoses.
2. Medical categorization: the LLM is provided with 17 predefined categories from the CEDIS. Internally, these are enriched with priority levels (CEDIS_PRIORITY) and icon mappings (CEDIS_ICONS) used for the visual representation of vignettes.
3. Clinical accuracy: the prompt includes explicit instructions to ensure realistic clinical content, such as appropriate medications, relevant comorbidities, and age-appropriate conditions.

The response from the LLM is processed by a parsing function, which uses boundary detection to extract the JSON object and validate the required fields. The generated tags are cross-checked against existing categories in the database. In case of generation failure, a fallback mechanism uses regex-based extraction to create a minimal vignette. An example LLM prompt with a generated case vignette is provided in Multimedia Appendix 2.

Dynamic Patient Simulator

The dynamic patient simulator produces highly personalized prompts for simulating realistic patient dialogues. At the core of the prompts is a strict role instruction, reinforced multiple times, ensuring that the LLM remains consistently in character as the patient: "You must stay in the character of the patient at all times. Never break character."

A key mechanism for realistic medical history taking is gradual symptom disclosure. The prompt explicitly instructs the LLM to reveal information incrementally. This is technically implemented through the hierarchical structuring of symptoms into primary and secondary groups, where secondary information may only be disclosed upon several conversation turns.

Patient behavior is dynamically modeled based on case vignette data through the following mechanisms:

1. Language modeling: the language complexity is adapted to the patient's health literacy and educational background. Patient simulations with lower literacy levels are instructed to use everyday language, analogies, and occasional

grammatical errors, while those with higher competence may use appropriate medical terminology.

2. Personality modeling: 12 predefined personality traits (eg, anxious, stoic) are mapped to specific behavioral instructions. For example, an anxious patient is guided to frequently seek reassurance, while a stoic one is prompted to minimize or downplay symptoms. These traits may interact to produce nuanced behaviors.
3. Anxiety-level integration: a numeric anxiety score (1-5) is translated into detailed behavioral descriptions ranging from "very calm and composed" to "extremely anxious and distressed."

To implement this behavioral control, advanced prompt engineering techniques are used, including negative instructions (eg, "NEVER reveal all symptoms at once") and few-shot learning. The latter is realized through the inclusion of concrete dialogue examples within the prompt to guide the model toward the desired conversational style [26-28]. An example prompt generated from a case vignette is provided in Multimedia Appendix 1.

System Evaluation

Formative Usability Test

To get feedback on the prototype and to identify aspects for improvement, we conducted a formative task-based usability test. It was conducted with 5 medical students; 1 student was in the 12th semester, 2 were in the 10th semester, and 2 were in the 6th semester. Participants were asked to complete 4 standardized tasks that cover the core functionalities of the application: manual vignette creation, AI-assisted vignette generation, organizing vignettes using folders, and conducting a patient simulation.

To quantitatively assess usability, the standardized System Usability Scale (SUS) was used [29]. Following the usability test, the participants were asked to answer a questionnaire with open questions about their user experience, specific strengths, weaknesses, and suggestions for improvement. These qualitative data were analyzed using a thematic analysis conducted by 1 coauthor.

Exploratory Language Model Comparison

Following the usability evaluation, a comparative analysis was conducted to assess the performance and differences of different LLMs in simulating patients based on case vignettes. The aim of this evaluation was to determine which model produces the most realistic and coherent patient simulations for medical history taking. A quantitative approach was applied. The participants consisted of 4 practicing physicians from the Department of Infectious Diseases at the University Hospital Basel, who served as domain experts. The group consists of 2 senior physicians and 2 assistant doctors, all aged 30 years or older, with an equal gender distribution (2 men and 2 women). Each expert conducted 3 simulated history-taking interviews using the same set of case vignettes, with each simulation powered by a different LLM. The evaluation was carried out using the developed web application. The evaluation cases were selected considering their clinical relevance to the participating physicians:

- Case 1: community-acquired pneumonia
- Case 2: complicated urinary tract infection
- Case 3: cellulitis in a diabetic patient

For the comparative evaluation of simulation quality, a targeted selection of models was made (GPT-4, Claude 3 Opus, and Grok 3). This selection represents a cross-section of state-of-the-art generative language models (in May 2025) and was intended to capture a range of capabilities in natural language understanding and dialogue generation relevant to realistic patient simulations.

A custom questionnaire was developed to rate the simulation quality across 7 criteria, using a 5-point Likert scale (ranging from "Strongly disagree" to "Strongly agree"). The evaluated criteria were (1) coherence of the symptom profile, (2) natural conversational flow, (3) patient-like language, (4) responsiveness to follow-up questions, (5) realistic emotional expression, (6) realistic uncertainty and conversational lapses, and (7) plausible time and symptom progression.

Participants received detailed written instructions. For each of the 3 cases, they were asked to launch the corresponding simulation, select the assigned LLM, and conduct a 2-5-minute history-taking interview in their preferred language. Immediately after each simulation, they completed the corresponding evaluation form. It is important to notice that each LLM was tested with a single case vignette, chosen to minimize expert workload. Neither the selected model nor the case was blinded to the participants.

Data were analyzed using Python 3 with SciPy and scikit-posthocs libraries. Normality was assessed using Shapiro-Wilk tests; given violations of normality assumptions (all $P<.001$), nonparametric procedures were used. The primary analysis used the Friedman test to compare ratings across the 3 models, accounting for the repeated-measures design. Post hoc pairwise comparisons used Wilcoxon signed-rank tests with Holm-Bonferroni sequential correction ($k=3$ comparisons, family-wise $\alpha=.05$) and Cliff δ for pairwise comparisons. Interrater reliability was assessed using the Kendall coefficient of concordance (W).

Ethical Considerations

All participants provided informed consent to take part in the study, agreeing to the anonymized analysis and publication of their responses. The study focused on quality assurance and usability evaluation of a patient simulator prototype and did not involve patient data or interventions. Accordingly, formal ethics committee approval was not required in accordance with institutional and national guidelines [30]. None of the participants received any compensation for their participation in the study.

Results

Results From Usability Testing

The analysis of the SUS assessment was performed using an online calculator. The individual SUS scores from the 5 participants were 82.5, 82.5, 95.0, 97.5, and 100. Based on these values, the mean SUS score was calculated as 91.5 (SD 8.4).

According to established benchmarks, a SUS score above 90 is generally considered "excellent" usability, in our case, high perceived usability in a small formative sample [31]. A post hoc analysis of the sample size was conducted based on the observed variability in SUS scores. With 5 participants (mean SUS score 91.5, SD 8.40), the 95% CI for the mean SUS score was 81.1-101.9, corresponding to a margin of error of ±10.4 points. While the sample size offers limited precision, this degree of uncertainty is acceptable for a formative usability evaluation, the objective of which is to identify significant usability issues rather than to achieve high statistical accuracy.

The thematic analysis of the open questions revealed consistent subject areas. The simplicity and intuitive usability of the system were unanimously cited as key strengths. The participants described the navigation as clear and well-organized and positively emphasized the appealing, simple visual design. There was a consensus that interaction with the system could be learned very quickly and without significant training.

At the same time, clear potential for improvement was identified. The most frequently mentioned point of criticism was the lack of a formal conclusion after the simulation. The students wanted feedback on their interaction, which is currently not provided, and a summary of the case in order to maximize the learning effect. One tester, who is about to graduate, noted that "only the case history is not enough" for them to use the tool regularly and therefore rated the application as "rather not" helpful for their studies. Furthermore, ideas for future extensions were mentioned, such as a note field or the integration of laboratory values.

Statistical Analysis of Model Comparison

Experimental Design

A within-participants experimental design was used to evaluate the performance of 3 LLMs (Grok 3, GPT-4, and Claude 3 Opus). Four independent raters (R1-R4, health professionals) assessed each model across 3 distinct use cases, answering 7 evaluation questions (Q1-Q7) for each case. Each rater evaluated all 3 models, with one model assigned to each case. This resulted in a balanced design with 84 total observations (N=84). Responses were recorded on a 5-point Likert scale (1=lowest; 5=highest), though the observed range was restricted to 2-5, with no ratings of 1 recorded.

Descriptive Statistics

The quantitative evaluation by 4 health professionals provided the differences in the LLM results as shown in Table 1 and Figure 4. Ratings showed limited discriminative reliability with an intraclass correlation coefficient (2, 1) of 0.165, and a pronounced ceiling effect, with most scores clustered between 4 and 5, constraining interpretation. Accordingly, all group differences should therefore be viewed as exploratory. For the 3 criteria, bigger differences in ratings can be recognized. We can see that Grok 3 and Claude 3 Opus performed better regarding realistic uncertainties and memory lapses than GPT-4. Regarding responsiveness to follow-up questions and natural conversation flow, Grok 3 was less good than the other two LLMs.

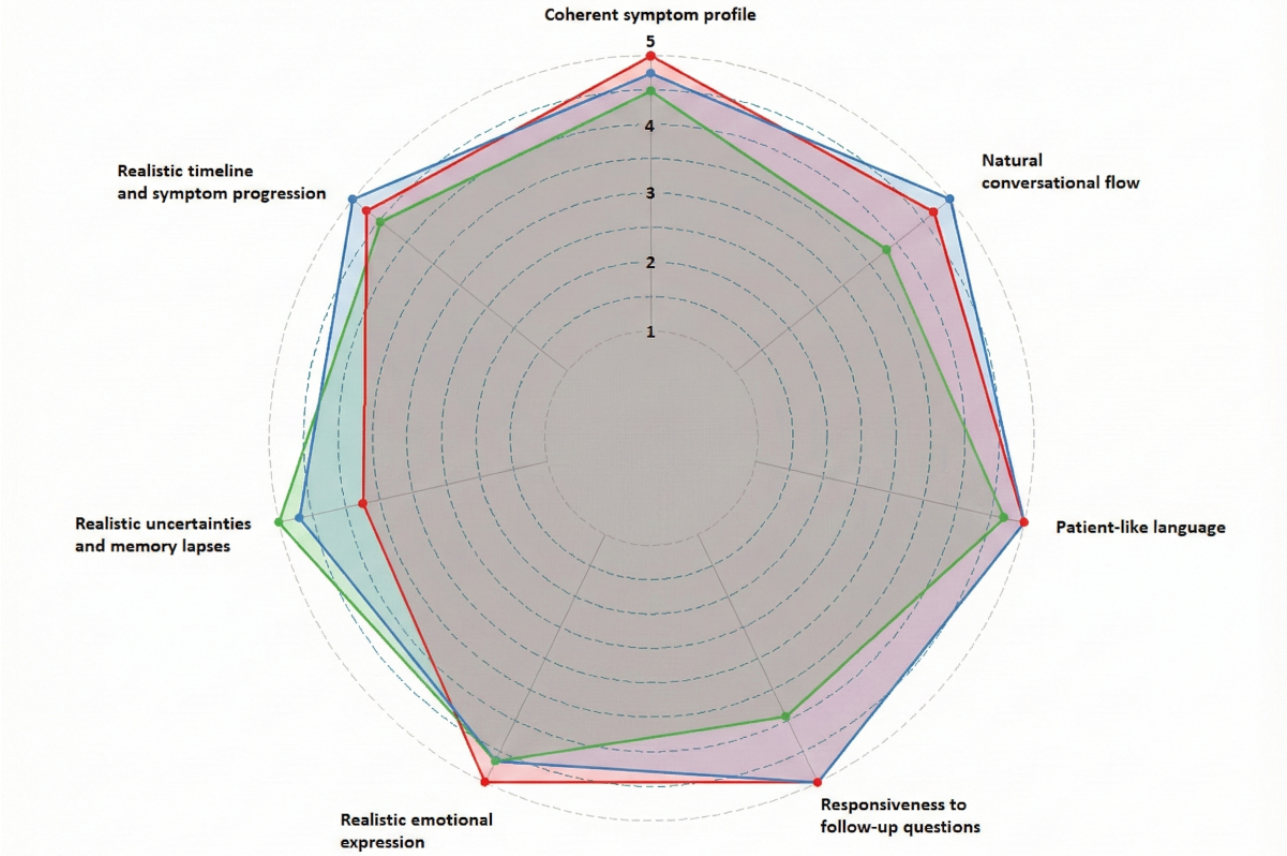
Table 1. Comparison of language models' quality assessed by clinical experts (n=4; scale 1 "totally disagree" to 5 "totally agree").

Evaluation criterion	Grok 3, mean (minimum-maximum)	GPT-4, mean (minimum-maximum)	Claude 3 Opus, mean (minimum-maximum)
The symptom profile described by the simulated patient is coherent.	4.5 (4-5)	4.75 (4-5)	4.25 (3-5)
The simulated patient describes the symptom presentation in language typical of real patients.	4.25 (4-5)	4.25 (4-5)	4.0 (2-5)
The simulated patient can describe realistic timelines and symptom progression.	4.75 (4-5)	4.5 (4-5)	4.25 (4-5)
The simulated patient displays realistic uncertainties and memory lapses.	3.75 (3-5)	3.0 (2-4)	4.0 (3-5)
The emotions portrayed appear realistic in the context of the simulated situation (medical history taking).	3.5 (3-4)	3.75 (3-4)	3.5 (2-5)
The flow and pace of the conversation appear natural.	4.25 (3-5)	4.0 (3-5)	3.25 (2-4)
The simulated patient responds appropriately to targeted follow-up questions.	4.75 (4-5)	4.75 (4-5)	3.75 (2-5)

Grok 3 demonstrated strengths in depicting realistic timelines and symptom progression, as well as in providing appropriate responses to follow-up questions (both mean scores 4.75, SD 0.5). Its lowest score was observed in the portrayal of realistic emotions (mean 3.5, SD 0.58).

GPT-4's primary strength was the coherence of the symptom presentation (mean 4.75, SD 0.5), outperforming the other 2 models in this category. It also matched Grok 3 in its ability to respond appropriately to follow-up questions (mean 4.75, SD 0.5). However, GPT-4 showed the weakest performance in conveying realistic uncertainties and memory lapses (mean 3.0, SD 0.82).

Figure 4. Strengths and weaknesses of 3 large language models (LLMs) for simulating patients. Colors refer to the different LLMs: green=Grok 3, red=GPT-4, and blue=Claude 3 Opus.



Claude 3 Opus exhibited the greatest variance in its ratings. The lowest scores of this model were found in the categories of natural conversation flow (mean 3.25, SD 0.96) and the portrayal of realistic emotions (mean 3.5, SD 1.0). Notably, the use of language typical of real patients (mean 4, SD 1.41) and the ability to respond appropriately to follow-up questions (mean 3.75, SD 1.26) showed the highest variability in responses. Figure 4 shows the comparison of the 7 evaluation criteria and the 3 language models in a spider chart.

A Friedman test was conducted to evaluate differences in ratings across repeated measures (raters and questions) among the 3 models (Table 2). None achieved statistical significance at the α=.05 level (all *P* values of >.14). The analysis revealed no statistically significant differences in performance between the models (χ^2_2=2.086; *P*=.45). Kendall coefficients of concordance (W values) are small (≈0-0.19), and the CIs are wide, so the degree of agreement on model rankings is very uncertain, suggesting that there are minimal practical differences in the performance of the models.

Table 2. Model comparison using Friedman tests per question: chi-square value, *P* value, and Kendall coefficient of concordance (W) with a bootstrap 95% CI to indicate agreement on model ranking.

Question	Chi-square (*df*)	*P* value	Kendall W (95% CI)
Q1	3.00 (2)	.22	0.14 (0.00-0.61)
Q2	0.00 (2)	>.99	0.00 (0.00-0.42)
Q3	3.00 (2)	.22	0.14 (0.00-0.61)
Q4	3.00 (2)	.22	0.14 (0.00-0.61)
Q5	0.67 (2)	.72	0.05 (0.00-0.61)
Q6	0.93 (2)	.63	0.11 (0.00-0.81)
Q7	4.00 (2)	.14	0.19 (0.00-0.75)

Interrater Comparison

Examination of rater-specific statistics revealed considerable heterogeneity in rating patterns. R1 exhibited the most lenient rating tendency (mean 4.524, SD 0.794), while R2 demonstrated the strictest ratings (mean 3.762, SD 1.064), exhibiting the greatest variability. Notably, R2's mean rating for Claude 3 Opus (mean 2.714, SD 0.881) was substantially lower than the ratings given by the other evaluators, which suggests either rater bias or a divergent interpretation of the evaluation criteria. R4

displayed the most consistent rating pattern, only rating with 3 or 4; however, this consistency may reflect restricted range use rather than genuine reliability.

Interrater reliability was assessed using the Kendall coefficient of concordance (W) across raters. Specifically, we thereby assessed how consistently they rank the 3 models. We used a bootstrap over raters (resampling columns with replacement) for percentile 95% CIs. Results are shown in Table 3. Values for Kendall W are all very low (0-0.19), showing at best a weak agreement. CIs start at 0 and extend up to 0.6, reflecting extreme imprecision.

Table 3. Interrater agreement per question, reported as the Kendall coefficient of concordance (W) across raters.

Question	Kendall W (95% CI)
Q1	0.141 (0-0.609)
Q2	0.000 (0-0.422)
Q3	0.141 (0-0.609)
Q4	0.141 (0-0.609)
Q5	0.047 (0-0.609)
Q6	0.109 (0-0.812)
Q7	0.188 (0-0.750)

Effect Size Analysis

Table 4 presents the results of the post hoc pairwise comparisons. We treated the data as within-rater and used a cluster bootstrap over raters to get a 95% CI percentile for Cliff δ. We calculated Cliff δ pooled across all 7 questions. None of the 3 pairwise comparisons achieved statistical significance after Holm-Bonferroni correction. The comparison between Claude 3 Opus and Grok 3 approached but did not reach significance (W=30.00; *P*=.07; Holm α=0.017), with a medium effect size (*r*=0.322). The comparison between Claude 3 Opus and GPT-4 showed a small effect (W=31.50; *P*=.18; *r*=0.249), while the comparison between GPT-4 and Grok 3 demonstrated a negligible effect (W=22.50; *P*=.59; *r*=0.096). In practice, the signed-rank statistics are very small because there are only 4 paired observations per contrast. With 4 raters, there is no robust evidence that any model consistently outperforms another on any single criterion. The data are descriptive rather than confirmatory.

Table 4. Results for post hoc Wilcoxon signed-rank tests, Holm-corrected for multiple comparisons.

Comparison	n	W statistic	*P* value	*z* score	Effect size (*r*)	Mean difference	Holm α	Cliff's δ (95% CI)
Claude 3 Opus vs Grok 3	28	30	.07	–1.704	0.322	–0.393	0.017	–0.21 (–0.591 to 0.059)
Claude 3 Opus vs GPT-4	28	31.5	.18	–1.318	0.249	–0.286	0.025	–0.15 (–0.128 to 0.556)
GPT-4 vs Grok 3	28	22.5	.59	–0.509	0.096	–0.108	0.05	–0.07 (–0.184 to 0.128)

Discussion

Principal Findings

This paper presents a training system for medical students to train interaction with diverse patient populations using LLMs and generating vignettes for case-based learning. The evaluation shows good usability and provides first insights into the differences of LLMs when simulating patients and their characteristics.

Usability and User Experience

The SUS results from 5 participants provide an initial indication that the tool may fulfill its core requirements for simplicity, efficiency, and appeal. The technical implementation can thus be deemed a success and forms a robust foundation for further development. However, good usability alone does not suffice to create an effective medical learning instrument. Qualitative responses highlighted a crucial shortcoming, that is, the repeated wish for a "didactic loop"—a cycle of history taking, diagnostic feedback, and case summary—was identified as an important finding in the qualitative evaluation. Without this step, the application remains primarily a tool for communication training, leaving its full potential for fostering clinical reasoning and diagnostic skills untapped. Such a feedback mechanism is considered critical for sustainable learning and repeated use. Traditionally, simulated patients are not only tasked with portraying a case but also with providing structured feedback to learners. This complex requirement—involving both role-play and performance evaluation—necessitates extensions to our simulator. An additional extension would be to consider the guidelines and communication strategies taught in medical education and to assess whether the students apply them in their interactions.

The feasibility and validity of an AI-based feedback system were demonstrated by Holderried et al [32], who showed that a GPT-4–powered chatbot could generate patient simulations and provide automated feedback, which closely matched expert human evaluations (Cohen κ=0.832). Their approach—using a dedicated "feedback prompt" to analyze the dialogue—directly aligns with the development pathway proposed in this work. This feedback capability of LLMs was also demonstrated by Cook et al [16] using GPT-3.5-Turbo.

Quality of AI-Generated Simulations and LLM Comparison

This analysis provides no evidence for statistically significant differences in performance among the 3 evaluated LLMs. Our findings should be interpreted in light of the limited discriminative signal in the rating data, reflected by very low interrater reliability and a strong ceiling effect (67/84, 81% of all ratings at 4-5). This reflects clearly the small rater panel. Accordingly, the comparative results are exploratory and underpowered and are offered as preliminary indications rather than definitive effects.

Overall, results show the general suitability of the tool, as all tested models were able to generate plausible, realistic patient roles useful for practicing history taking, with each statement rated between "Neutral" (3 points) and "Strongly agree" (5 points), and no ratings of "Strongly disagree" (1 point). These findings echo previous studies, which have also found that current LLMs can create clinical vignettes with high linguistic and medical accuracy, though expert revision is often required. For instance, Yanagita et al [23] found 97% of LLM-generated vignettes usable with minor modifications, while Takahashi et al [22] noted considerable potential for ChatGPT-4–generated cases in medical education, but with room for improvement in realism. Our system differs from such approaches, as we are suggesting a human-in-the-loop approach for vignette generation, providing 2 options—specifying all characteristics manually using the vignette template or using AI-support.

All models scored highest for symptom coherence and realistic timelines. Statements assessing the realism of the dialogue itself, however, received lower scores. This suggests that the models were better at maintaining factual accuracy than at simulating natural role play. The most significant difference between Grok 3 and GPT-4 was in their portrayal of realistic uncertainty (Figure 4). Although GPT-4 performed strongly overall and achieved the highest scores for symptom coherence and emotional realism, it struggled to depict realistic uncertainty—a key component of authentic medical interviews. By contrast, Claude 3 Opus received the lowest scores in most categories but performed best in conveying uncertainty. Grok 3 performed particularly well in core medical content, coherent symptom simulation, and realistic timeline generation. GPT-4 and Grok 3 demonstrated similar abilities when addressing follow-up questions, whereas Claude 3 Opus fell behind, particularly in terms of maintaining conversational flow and using language typical of patients. Claude 3 Opus also exhibited the greatest variability in expert ratings, particularly with regard to patient language, follow-up responses, and emotional realism. Furthermore, it was the only model to insert stage directions into the dialogue, which some users found distracting while others valued the added nonverbal context. These differences have to be considered with care, as each model was tested with a different case example, and the differences may also be due to the different sociodemographic characteristics of the simulated patients. Because of the small expert panel to judge the models, we cannot recommend any model as the primary model for medical patient simulations. Evaluations involving a larger number of raters are required to confirm these initial findings. Future improvements should also focus on prompting the models to simulate authentic human emotional expression and memory lapses, as these aspects are essential for the realism of clinical interviews.

In this study, we did not assess whether the answers from the simulated patient are realistic. A previous study showed that the answers of the simulated patient are correct, but the study was limited in size [9]. In principle, having an incomplete or even wrong answer from a simulated patient could be considered a realistic case, as in the real world, patients will not disclose everything and can share wrong information as well. This adds an additional dimension to the training scenario. In future work, we will study whether the correctness can be ensured and whether specific learning cases can be developed that allow training skills of diagnosing and medical history taking when patients are not honest or are incomplete in their answers.

Educational Use Cases and Implications

As exemplified, we describe 2 educational use cases of the patient simulator. The first use case concerns the dynamic and authentic patient simulation integrating psychosocial complexity. In this use case, the simulator supports learning the patient-doctor interaction, given the psychosocial complexity of patients. Besides medical symptoms, the simulator uses a variety of contextual factors in the simulated interaction, including language barriers, family-related stressors, cultural sensitivities, and varying levels of trust or mistrust in the health care system. By integrating these dimensions, learners can experience and learn to overcome communication difficulties that often arise when treating diverse patient populations. For example, a patient with limited health literacy may misinterpret medical terminology, while a patient from a different cultural background may express distress indirectly or be reluctant to share personal information due to systemic mistrust.

A second use case focuses on training young surgeons to hold informed consent discussions with patients before surgical procedures. Such discussions are crucial for ensuring ethical and legally sound medical practice, as well as fostering patient trust and facilitating shared decision-making. However, novice clinicians often feel uncertain about communicating complex medical information, addressing patient fears, or managing emotional reactions under time pressure. For this use case, the simulator provides an authentic, dialogic environment in which students or young surgeons can practice and refine their communication strategies during simulated preoperative consultations. The AI-driven patient responds dynamically based on the learner's explanations, empathy, and ability to balance technical accuracy with clarity and reassurance. This approach aligns with the principles of deliberate practice and experiential learning, emphasizing repeated, feedback-oriented engagement with realistic scenarios to achieve competence.

By interacting with the AI-based patient simulator, students will acquire domain-specific communication skills and develop a deeper understanding of the opportunities and limitations of AI in professional contexts. The simulated encounters provide an environment in which learners must navigate complex psychosocial dynamics, such as language barriers, family stressors, and mistrust of the health care system, requiring a high degree of empathy, cultural sensitivity, and adaptive

communication strategies. Although promising, the use of a simulator within student education raises some ethical challenges. Among them is a risk for overreliance on scripted behavior, as LLM behavior might be predictable, and students miss learning to deal with unpredictable behavior. Biases can be amplified when the personality modeling is not diverse. Further, students need to understand that the simulator is a training aid, not a real patient interaction.

For educators, the system opens up new possibilities for integrating innovative AI tools into medical education in a meaningful way from a pedagogical perspective. It supports the expansion of teaching practice through technology-enhanced learning approaches, enabling instructors to design more personalized, reflective, and data-driven learning experiences.

At a broader societal level, the use of such a simulator addresses the urgent need to improve health care by facilitating more effective communication between physicians and patients from diverse backgrounds. By explicitly incorporating psychosocial factors such as language barriers, cultural differences, and institutional mistrust, the system helps to foster equity and inclusion in medical education, and ultimately in patient care.

Limitations

Several methodological limitations must be considered when interpreting the results. The LLM evaluation was conducted with only 4 infectious disease experts, with low interrater reliability, limiting the statistical generalizability of the findings. In addition to these constraints, the study is statistically underpowered for detecting small or moderate differences between models. We used a frequentist analytic framework to generate CIs for the expert ratings; however, given the limited sample size, these intervals should be understood as descriptive indicators of uncertainty rather than tools for confirmatory inference. In other words, the frequentist estimates do not support strong population-level claims but instead provide a bounded range of plausible effect sizes based on the observed data. This aligns with the formative nature of the study. The statistical outputs offer preliminary signals to guide future, adequately powered evaluations rather than serving as evidence of definitive model differences. Because each model was tested on a single vignette and models were not counterbalanced across cases, the observed differences may also be due to the cases rather than model performance.

Only 2 students in their 6th semester contributed to the requirement collection process. Although attempted, it was not possible to gain the interest of additional students for this phase. The requirements were extended by information from scientific literature. It would be beneficial to get additional input from medical educators. Similarly, usability testing involved only 5 medical students, excluding the perspective of teachers in medical education who would use the tool for creating vignettes for teaching purposes. The raters were not specifically trained, which could have led to varying interpretations of the items of the Likert scale.

Each LLM was tested with a single case vignette, chosen to minimize expert workload but potentially limiting the assessment's breadth. The analysis was also confined to 3 proprietary, closed-source models, without including open-source alternatives—reflecting a gap that future research should address. Further, the case and LLM were not blinded to the expert raters. This may have introduced bias, as they were able to see the name of the model they were interacting with.

The survey instrument used for expert evaluation was newly developed for this study and had not been previously validated or piloted, representing a challenge common to this research field, as described by Yanagita et al [23]. They therefore analyzed subjective expert opinions descriptively without mean comparisons. The usability test focused on learners' perspectives without structured instructional units.

Subsequent studies should include a comprehensive evaluation with medical educators to validate the system's suitability as a patient simulation tool. Larger numbers of raters and randomized model assignments to different vignettes would address current methodological limitations. A longitudinal study measuring actual learning outcomes in students who regularly train with the system, compared to a control group, would be necessary to empirically demonstrate didactic effectiveness.

Future Work

In summary, this work successfully addresses the challenge of developing a highly usable simulation frontend and demonstrates the potential of LLMs as patient simulation tools. At the same time, it clearly outlines the necessary steps required to transform the tool into a comprehensive educational instrument. The most urgent improvement is the integration of an automated feedback mechanism that analyzes the conducted conversation and provides feedback to the student on what to improve. Building on validated approaches, another AI agent could analyze the dialogue and provide structured feedback on history-taking categories. This would address students' expressed need for a formal case closure and clearer learning outcomes. The addition of digital note-taking tools and integration of structured data (eg, laboratory values, vital signs) would increase both complexity and realism, better meeting the needs of advanced learners. To have a more realistic training situation, the simulation tool should be extended by a voice user interface that allows a speech-based interaction.

Emerging LLMs such as GPT-4o, Gemini, Grok 3, and Claude 3.5 Sonnet now offer advanced multimodal capabilities. Future iterations could leverage these features to enrich vignettes with relevant visual findings, further enhancing didactic value.

Acknowledgments

GPT-4.1 was used for translating the text written in German into English and for improving the writing style. The authors acknowledge the use of Genspark AI [33] for assistance with statistical data analysis, code generation, and visualization. All analytical approaches were independently validated, and all interpretations and conclusions are solely those of the authors.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

Access to the system can be provided upon request. Example prompts are available as multimedia appendices. Additional data sharing is not applicable to this article as no data sets were generated or analyzed during this study.

Authors' Contributions

Conceptualization: AN, AE, DR
Methodology: AN, AE
Investigation: AE
Visualization: AE
Software: AE
Writing – original draft: AN, AE, DR, KD
Writing – review & editing: AN, AE, DR, KD
Supervision: KD
Project administration: KD

Conflicts of Interest

None declared.

Multimedia Appendix 1
Example prompts and vignettes.
[DOCX File , 2159 KB - mededu_v11i1e81271_app1.docx]

Multimedia Appendix 2
Generation of a structured case vignette based on a prompt.
[DOCX File , 19 KB - mededu_v11i1e81271_app2.docx]

References

1. Faller H. Patientenorientierte Kommunikation in der Arzt-Patient-Beziehung [Patient-centered communication in the physician-patient relationship]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2012;55(9):1106-1112. [doi: 10.1007/s00103-012-1528-x] [Medline: 22936477]
2. Schweizerische Akademie der Medizinischen Wissenschaften. Kommunikation im medizinischen Alltag. Ein Leitfaden für die Praxis. Zenodo. 2023. URL: https://zenodo.org/records/8224985 [accessed 2025-12-10]
3. Thistlethwaite JE, Davies D, Ekeocha S, Kidd JM, MacDougall C, Matthews P, et al. The effectiveness of case-based learning in health professional education. A BEME systematic review: BEME Guide No. 23. Med Teach 2012;34(6):e421-e444. [doi: 10.3109/0142159X.2012.680939] [Medline: 22578051]
4. Schmidt HG, Rotgans JI, Yew EHJ. The process of problem-based learning: what works and why. Med Educ 2011;45(8):792-806. [doi: 10.1111/j.1365-2923.2011.04035.x] [Medline: 21752076]
5. Barrows HS. An overview of the uses of standardized patients for teaching and evaluating clinical skills. AAMC. Acad Med 1993;68(6):443-51; discussion 451. [doi: 10.1097/00001888-199306000-00002] [Medline: 8507309]
6. Issenberg SB, McGaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. Med Teach 2005;27(1):10-28. [doi: 10.1080/01421590500046924] [Medline: 16147767]
7. Schoenberg NE, Ravdal H. Using vignettes in awareness and attitudinal research. Int J Soc Res Methodol 2000;3(1):63-74. [doi: 10.1080/136455700294932]
8. Evans SC, Roberts MC, Keeley JW, Blossom JB, Amaro CM, Garcia AM, et al. Vignette methodologies for studying clinicians' decision-making: validity, utility, and application in ICD-11 field studies. Int J Clin Health Psychol 2015;15(2):160-170 [FREE Full text] [doi: 10.1016/j.ijchp.2014.12.001] [Medline: 30487833]
9. Reichenpfader D, Denecke K. Simulating diverse patient populations using patient vignettes and large language models. : ELRA and ICCL; 2024 Presented at: Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024; 2025 November 30; Torino, Italia p. 20-25 URL: https://aclanthology.org/2024.cl4health-1.3/ [doi: 10.18653/v1/2025.cl4health-1.0]
10. Flanagan OL, Cummings KM. Standardized patients in medical education: a review of the literature. Cureus 2023;15(7):e42027 [FREE Full text] [doi: 10.7759/cureus.42027] [Medline: 37593270]

11. Svyntozelska O, Suarez NRE, Demers J, Dugas M, LeBlanc A. Socioeconomic and demographic factors influencing interpersonal communication between patients with chronic conditions and family physicians: a systematic review. Patient Educ Couns 2025;131:108548 [FREE Full text] [doi: 10.1016/j.pec.2024.108548] [Medline: 39657391]
12. Theodosopoulos L, Fradelos EC, Panagiotou A, Dreliozi A, Tzavella F. Delivering culturally competent care to migrants by healthcare personnel: a crucial aspect of delivering culturally sensitive care. Soci Sci 2024;13(10):530. [doi: 10.3390/socsci13100530]
13. Shukla M, Schilt-Solberg M, Gibson-Scipio W. Medical mistrust: a concept analysis. Nurs Rep 2025;15(3):103 [FREE Full text] [doi: 10.3390/nursrep15030103] [Medline: 40137676]
14. Fiala MA. Discrimination, medical mistrust, and delaying cancer screenings and other medical care. JCO Oncol Pract 2023;19(11_suppl):159-159. [doi: 10.1200/op.2023.19.11_suppl.159]
15. Kaczynski MA, Benitez G, Shehadeh F, Mylonakis E, Fiala MA. Perceived discrimination in the healthcare setting and medical mistrust: findings from the health information national trends survey, 2022. J Gen Intern Med 2025;40(11):2491-2498. [doi: 10.1007/s11606-025-09369-x] [Medline: 39838250]
16. Cook DA, Overgaard J, Pankratz VS, Del Fiol G, Aakre CA. Virtual patients using large language models: scalable, contextualized simulation of clinician-patient dialogue with feedback. J Med Internet Res 2025;27:e68486 [FREE Full text] [doi: 10.2196/68486] [Medline: 39854611]
17. Cook DA. Creating virtual patients using large language models: scalable, global, and low cost. Med Teach 2025;47(1):40-42. [doi: 10.1080/0142159X.2024.2376879] [Medline: 38992981]
18. Öncü S, Torun F, Ülkü HH. AI-powered standardised patients: evaluating ChatGPT-4o's impact on clinical case management in intern physicians. BMC Med Educ 2025;25(1):278 [FREE Full text] [doi: 10.1186/s12909-025-06877-6] [Medline: 39979969]
19. Borg A, Georg C, Jobs B, Huss V, Waldenlind K, Ruiz M, et al. Virtual patient simulations using social robotics combined with large language models for clinical reasoning training in medical education: mixed methods study. J Med Internet Res 2025;27:e63312 [FREE Full text] [doi: 10.2196/63312] [Medline: 40053778]
20. Benfatah M, Marfak A, Saad E, Hilali A, Nejjari C, Youlyouz-Marfak I. Assessing the efficacy of ChatGPT as a virtual patient in nursing simulation training: a study on nursing students' experience. Teach Learn Nurs 2024;19(3):e486-e493. [doi: 10.1016/j.teln.2024.02.005]
21. Coşkun, Kıyak YS, Budakoğlu. ChatGPT to generate clinical vignettes for teaching and multiple-choice questions for assessment: a randomized controlled experiment. Med Teach 2025;47(2):268-274. [doi: 10.1080/0142159X.2024.2327477] [Medline: 38478902]
22. Takahashi H, Shikino K, Kondo T, Komori A, Yamada Y, Saita M, et al. Educational utility of clinical vignettes generated in japanese by ChatGPT-4: mixed methods study. JMIR Med Educ 2024;10:e59133 [FREE Full text] [doi: 10.2196/59133] [Medline: 39137031]
23. Yanagita Y, Yokokawa D, Uchida S, Li Y, Uehara T, Ikusaka M. Can AI-generated clinical vignettes in Japanese be used medically and linguistically? J Gen Intern Med 2024;39(16):3282-3289. [doi: 10.1007/s11606-024-09031-y] [Medline: 39313665]
24. Grafstein E, Unger B, Bullard M, Innes G. Canadian emergency department information system (CEDIS) presenting complaint list (version 1.0). CJEM 2003;5(1):27-34. [doi: 10.1017/s1481803500008071] [Medline: 17659149]
25. Vercel. URL: https://vercel.com [accessed 2025-12-04]
26. Ban Y, Wang R, Zhou T, Cheng M, Gong B, Hsieh CJ. Understanding the impact of negative pompts: when and how do they take effect? arXiv Preprint posted online on June 5, 2024. [doi: 10.48550/arXiv.2406.02965]
27. Min S, Lyu X, Holtzman A, Artetxe M, Lewis M, Hajishirzi H, et al. Rethinking the role of demonstrations: what makes in-context learning work? arXiv Preprint posted online on February 25, 2022. [doi: 10.48550/arXiv.2202.12837]
28. Rosenbloom L, Xu J, Zhang LL, Chen Q, Feng X, Chen Y, et al. Beyond prompt content: enhancing LLM performance via content-format integrated prompt optimization. arXiv Preprint posted online on February 6, 2025. [doi: 10.5260/chara.21.2.8]
29. Brooke J. SUS: a quick and dirty usability scale. In: Usability Evaluation in Industry. London: Taylor and Francis; 1996.
30. Schweizerische Ethikkommissionen für die Forschung am Menschen. Quality assurance or research requiring approval? Guideline from swissethics to support the differentiation between quality assurance (quality assurance studies, quality control studies) and research projects requiring approval.. Swiss Ethics. 2020. URL: https://swissethics.ch/themen/qualitaetssicherung-oder-bewilligte-forschung [accessed 2025-12-11]
31. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the System Usability Scale. Int J Hum-Comput Interact 2008;24(6):574-594. [doi: 10.1080/10447310802205776]
32. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, Festl-Wietek T, Holderried M, Eickhoff C, et al. A language model-powered simulated patient with automated feedback for history taking: prospective study. JMIR Med Educ 2024;10:e59213 [FREE Full text] [doi: 10.2196/59213] [Medline: 39150749]
33. Genspark AI. URL: https://www.genspark.ai/ [accessed 2025-12-04]

Abbreviations

AI: artificial intelligence
CEDIS: Canadian Emergency Department Information System
LLM: large language model
SUS: System Usability Scale

Edited by D Chartash; submitted 25.07.25; peer-reviewed by PF Chen, A Munoz-Zavala, M Elbattah; comments to author 27.09.25; revised version received 30.11.25; accepted 30.11.25; published 12.12.25.

Please cite as:
Elhilali A, Ngo ASH, Reichenpfader D, Denecke K
Large Language Model–Based Patient Simulation to Foster Communication Skills in Health Care Professionals: User-Centered Development and Usability Study
JMIR Med Educ 2025;11:e81271
URL: https://mededu.jmir.org/2025/1/e81271
doi:10.2196/81271
PMID:

Correction: Comparing the Perceived Realism and Adequacy of Venipuncture Training on an in-House Developed 3D-Printed Arm With a Commercially Available Arm: Randomized, Single-Blind, Cross-Over Study

Susan Gijsbertje Brouwer de Koning[1], PhD; Amy Hofman[2], PhD; Sonja Gerber[3]; Vera Lagerburg[4,5], PhD; Michelle van den Boorn[1], MSc

[1]3D Lab, Department of Computerization, Automation and Medical Technology (iMED), OLVG Hospital, 9 Oosterpark, Amsterdam, The Netherlands

[2]Department of Research and Epidemiology, OLVG Hospital, Amsterdam, The Netherlands

[3]Skillslab, OLVG Hospital, Amsterdam, The Netherlands

[4]Department of Medical Physics, OLVG Hospital, Amsterdam, The Netherlands

[5]Department of Medical Physics and Instrumentation, St. Antonius Ziekenhuis, Nieuwegein, The Netherlands

Corresponding Author:
Susan Gijsbertje Brouwer de Koning, PhD
3D Lab, Department of Computerization, Automation and Medical Technology (iMED), OLVG Hospital, 9 Oosterpark, Amsterdam, The Netherlands

Related Article:

Correction of: https://mededu.jmir.org/2025/1/e71139

Abstract

(JMIR Med Educ 2025;11:e89670) doi:10.2196/89670

In "Comparing the Perceived Realism and Adequacy of Venipuncture Training on an in-House Developed 3D-Printed Arm With a Commercially Available Arm: Randomized, Single-Blind, Cross-Over Study" [1], the authors made one addition.

The following affiliation has been added as affiliation 5 and attached to author VL:

Department of Medical Physics and Instrumentation, St. Antonius Ziekenhuis, Nieuwegein, The Netherlands

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Brouwer de Koning SG, Hofman A, Gerber S, Lagerburg V, van den Boorn M. Comparing the perceived realism and adequacy of venipuncture training on an in-house developed 3D-printed arm with a commercially available arm: randomized, single-blind, cross-over study. JMIR Med Educ 2025 Nov 4;11:e71139. [doi: 10.2196/71139] [Medline: 41187260]

Submitted 16.12.25; this is a non–peer-reviewed article;accepted 16.12.25; published 22.12.25.

Please cite as:
Brouwer de Koning SG, Hofman A, Gerber S, Lagerburg V, van den Boorn M
Correction: Comparing the Perceived Realism and Adequacy of Venipuncture Training on an in-House Developed 3D-Printed Arm With a Commercially Available Arm: Randomized, Single-Blind, Cross-Over Study
JMIR Med Educ 2025;11:e89670
URL: https://mededu.jmir.org/2025/1/e89670
doi:10.2196/89670

Advantages of a Virtual Collaborative Research Dermatology Laboratory

Natasha E Barton[1], BS, BA; Kenny Ta[2*], BA; Angela R Loczi-Storm[3*], BS; Cory A Dunnick[4*], MD; Robert P Dellavalle[5*], MD, PhD, MSPH

[1]School of Medicine, University of Colorado, 13001 E 17th Pl, Aurora, CO, United States

[2]School of Medicine, University of Minnesota, Minneapolis, MN, United States

[3]College of Osteopathic Medicine of the Pacific-Northwest, Western University of Health Sciences, Lebanon, OR, United States

[4]Department of Dermatology, University of Colorado, Aurora, CO, United States

[5]Department of Dermatology, University of Minnesota, Minneapolis, MN, United States

[*]these authors contributed equally

Corresponding Author:
Natasha E Barton, BS, BA
School of Medicine, University of Colorado, 13001 E 17th Pl, Aurora, CO, United States

Abstract

The Dellavalle/Dunnick Dermato-Epidemiology Lab transitioned from a single campus to a dual-campus collaboration between the University of Colorado and the University of Minnesota in 2024. Since the 2020 COVID-19 pandemic, the laboratory has been operating on Zoom and allows medical students from any institution to join. This innovative laboratory structure offers students and other researchers unique opportunities to engage in dermatological research and develop professional networks across two large academic institutions. The laboratory's model embraces a virtual collaborative approach, promotes inclusivity, encourages student-led inquiry, and provides a structured environment for professional development and academic output. Through its commitment to diverse student perspectives and interdisciplinary cooperation, the Dellavalle/Dunnick Dermato-Epidemiology Lab creates a new, equitable, nationwide model for research and mentorship in dermatology, supporting medical students, residents, and fellows to navigate future careers in dermatology.

(JMIR Med Educ 2025;11:e65697) doi:10.2196/65697

KEYWORDS

medical students; collaborative research; virtual collaboration; dermatology research; tutorial; editorial

Introduction

The Dellavalle/Dunnick Dermato-Epidemiology Lab represents an innovative collaboration between the University of Colorado (CU) and the University of Minnesota (UMN). This dual-campus model offers a unique platform for medical students, residents, and fellows, particularly those from institutions without dedicated dermatology programs, to engage in high-impact dermatological research. The laboratory's goals are multifaceted: to foster professional networks, promote a culture of scientific curiosity and inquiry within the dermatology field, and enhance students' research credentials. This editorial will summarize the key components and operational structure contributing to the laboratory's sustained engagement and productivity across institutions, providing a practical framework for medical educators, researchers, or administrators looking to replicate this scalable model in other medical specialties. By focusing on strategies for creating an inclusive, collaborative research environment, we will explore how mentorship strategies, logistical planning, digital tools, and virtual collaboration can bridge gaps in medical education and support the next generation of physician-scientists, particularly in settings where formal dermatology programs or research infrastructure may be limited.

Laboratory Background

The Dellavalle/Dunnick Dermato-Epidemiology Lab was founded as an in-person research group at CU. It operated under a traditional model of weekly on-site meetings, primarily with CU medical students. This structure had been in place since the laboratory's creation in the early 2000s.

During the COVID-19 pandemic, the laboratory transitioned to a fully virtual format using platforms such as Zoom (Zoom Communications) and Google Workspace. This transition significantly expanded the laboratory's geographic reach and accessibility, allowing medical students from over 30 institutions nationwide to participate. Participation is open and nonselective; students are not required to apply but are vetted and invited to join meetings and become involved as their interest develops.

While the laboratory has an accessible website, its growth has largely been driven by its strong reputation, word-of-mouth referrals, and visibility at academic conferences. The research

credentials of the laboratory's principal investigators, authors RPD and CAD, further enhance its appeal, attracting motivated trainees eager to contribute to impactful dermatological projects. RPD has an h-index of 70 and more than 145,000 citations, and CAD has an h-index of 36 and more than 5,000 citations [1,2].

When RPD transitioned to UMN, the lab's commitment to virtual collaboration ensured continuity. Weekly Zoom meetings provide a platform for updates, allowing for the seamless integration of students and residents from multiple locations. This virtual approach supported the laboratory's ongoing success, allowing it to thrive as a dual-campus operation.

The laboratory's focus on collaboration has led to contributions from students at a range of institutions, including Rocky Vista University College of Osteopathic Medicine, Case Western Reserve University, Kansas City University, Texas Tech El Paso, A.T. Still University, SUNY Upstate University, Texas A&M, The Ohio State University, and many others. These efforts have resulted in numerous publications and sustained high-impact research projects, further cementing the laboratory's role in advancing dermatological science through interdisciplinary collaboration.

Laboratory Operations

Joining the Laboratory

The Dellavalle/Dunnick Dermato-Epidemiology Lab uses an open-access, inclusive model for student participation. When students learn about the laboratory through the website, a conference presentation, or word of mouth, they email CAD and/or RPD. One of the unique aspects of our laboratory is that students are not required to apply or be selected to join. Instead, interested students are provided the Zoom link and the shared laboratory Google Doc (Textbox 1 and Multimedia Appendix 1). They are invited to attend weekly meetings without expectations of prior research experience.

Textbox 1. Fictitious lab executive summary. Details included were created for the purposes of this tutorial and are not accurate or representative. Formatting is the same as the real shared document.

Dellavalle/Dunnick Dermato-Epidemiology Weekly Research Lab Meeting

Time: Tuesdays 12:15 pm Mountain time zone (Denver)

Zoom: [Zoom Link]

Zoom meeting ID: 111 222 3333

Google doc: [Google Document Link]

Lab website: derm-epi.com

Dermatology Faculty:

Robert Dellavalle, MD, PhD, MSPH, 720-111-3333

Cory A. Dunnick, MD, 303-111-3333,

Clinical Research Fellow, Lab Coordinators, Residents, Fellows:

John Meisenheimer, MD

U. Minnesota Medical Students:

(2025)

(2026)

(2027): Kenny Ta

(2028)

CU Medical Students:

(2025)

(2026)

(2027): Natasha Barton

(2028)

Medical Students at Other Institutions:

Angela Loca-Storm (WesternU-COMPNW)

Manuscripts Submitted:

1. Advantages of a Virtual Collaborative Research Dermatology Laboratory - Natasha, Kenny, Angela, tutorial to *JMIR Med Education*

Manuscripts Needing Revision:

1. Advantages of a Virtual Collaborative Research Dermatology Laboratory - Natasha, Kenny, Angela, tutorial to *JMIR Med Education*

Dormant Projects:

1. Advantages of a Virtual Collaborative Research Dermatology Laboratory - Natasha, Kenny, Angela, tutorial to *JMIR Med Education*

Active IRBs:

1. COMIRB 111: Survey on Advantages of a Virtual Collaborative Research Dermatology Laboratory - Natasha, Kenny, Angela

Active Grants:

1. Survey on Advantages of a Virtual Collaborative Research Dermatology Laboratory - Natasha, Kenny, Angela

Grants Available:

1. Sulzberger Education AAD innovation grant (LOI opens June 2025 – $5K/30K)

Upcoming Events:

1. May 7-10, 2025, San Diego, SID
2. July 10-13, 2025, Chicago, AAD Innovation Academy
3. March 27-31, 2026 Denver, AAD
4. May 13-16, 2026 Chicago, SID

Recommended:

1. Brief Faculty Development Videos
2. Listen to Dermasphere blog (Spotify)

New participants are encouraged to observe their first few meetings to gain familiarity with the laboratory's structure and active projects. As they become more comfortable, they are encouraged to contact established laboratory members to join projects that interest them. This approach reduces barriers to participation and allows self-directed engagement based on availability, interest, and experience.

Weekly Operations

Consistent, weekly Zoom meetings serve as the backbone of laboratory operations. The meetings follow a structured format, including a "popcorn-style" check-in, where each member introduces themselves, their role, medical school, and current location. After the check-in, we engage in personal development activities, such as discussing key insights from recent conferences or watching brief faculty development videos from the Association of Professors of Dermatology [3].

Following the professional development segment, we review the current list of ongoing projects. This includes updates on submitted projects awaiting decisions, projects needing revisions, and new ideas. Laboratory members provide updates on their respective projects, and if papers or posters need review, the members will share their screen to go through the material with the group. We then go over any upcoming deadlines for conferences or grant applications. The meeting ends with a final "popcorn-style" check-out with the opportunity to provide any last-minute updates.

Shared Laboratory Document

The shared laboratory document is key to our operations and is essentially the "holy grail" of the laboratory. This Google Doc is a centralized hub for all essential information related to laboratory activities. It includes a comprehensive list of laboratory members' contact details, ensuring easy communication across our collaborative network.

Key sections of the document include the following:

- Lab member information
- Projects submitted awaiting decision (this section tracks all projects that have been submitted for review and are waiting for feedback or approval)
- Projects with revisions (this lists projects that have received reviewer comments and are currently undergoing revisions)
- Active projects (this section highlights ongoing research projects)
- Institutional review board (IRB) protocols (a list of all active IRB protocols currently in effect for our research projects)
- Active grants (this section includes information on approved grants and details on who is leading or contributing to each)
- Open conferences with submission deadlines (this section provides an updated list of conferences open for submissions, along with their deadlines)
- Grant submission deadlines (a list of upcoming grant applications with specific submission dates)
- Upcoming dermatology events (a calendar of dermatology-related events that may interest lab members for networking or continuing education)

The shared laboratory document lists the names of the members currently working on each active project, IRB protocol, and grant. This ensures that all laboratory members can stay updated on who is involved in each initiative and fosters a collaborative, transparent research environment. The document is updated weekly to ensure all members can access current information and deadlines.

Project Development

Participation in the laboratory is self-motivated, and each member is encouraged to develop their own unique research ideas. Once an idea is proposed, our principal investigators, along with other laboratory members, offer suggestions and support to help get the project started. In addition to individual project ideas, our principal investigators often provide project concepts and assemble teams to work on them. After a project is initiated and a team is formed, smaller, more focused meetings take place outside of the weekly laboratory meetings to continue project work.

Funding

The laboratory operates without dedicated, laboratory-specific funding. Instead, students are encouraged to seek financial support for their research projects through grants available at their home institutions or from national sources. While many projects do not require funding beyond literature access and student time, grant opportunities are essential for supporting conference travel, publication fees, and community outreach efforts. Students who secure funding are encouraged to share their grant proposals with other laboratory members as templates, creating a cycle of shared learning and resource building. The shared Google Doc also lists all active grants that laboratory members have received, as well as open grant applications, providing students with up-to-date opportunities to apply for funding. By using this collaborative approach to funding, the laboratory fosters a culture of self-sufficiency and resourcefulness, allowing students to secure financial support for their community-based research endeavors.

Steps for Creating a Successful Collaborative Laboratory

Step 1: Setting Clear Goals and Objectives for Research Programs

Every successful research program begins with a clear set of objectives that guide its mission and outcomes. The Dellavalle/Dunnick Dermato-Epidemiology Lab operates with well-defined goals rooted in the desire to offer students

substantial research opportunities and mentorship while promoting the exploration of novel scientific questions. Key goals include providing professional mentorship, fostering a collaborative research environment that encourages scientific curiosity, and enhancing student research opportunities. These objectives are not only important for the success of the laboratory but are also critical for helping medical students navigate the competitive landscape of dermatology residency applications, which became more difficult with the transition of the United States Medical Licensing Examination Step I exam to pass/fail grading in 2022 [4,5].

To replicate this model, laboratory groups must first identify their core objectives. These could include goals such as improving students' exposure to field-specific research, encouraging scientific curiosity, providing mentorship opportunities, and offering access to high-impact, publishable projects. Once these objectives are established, they will serve as a framework for the program's success and provide clarity on how to best serve the students involved.

Step 2: Structuring the Program for Success

The success of the Dellavalle/Dunnick Dermato-Epidemiology Lab is heavily dependent on its strong organizational structure and clear communication channels. To maintain the flow of projects and ensure consistent progress, the laboratory holds weekly meetings where students and faculty members provide updates on ongoing research, discuss any challenges, and plan the next steps. These regular check-ins foster a sense of continuity and accountability, which is especially important when managing research teams spread across multiple campuses. One of the key benefits of this structure is that it allows students from various institutions to remain engaged, regardless of their geographic location.

Weekly meetings and centralized documentation keep research organized and moving forward. Using tools like Google Docs supports team transparency and accountability. Replicating this model involves consistent communication, flexibility, and logistical coordination across institutions. In addition, it's important to maintain a flexible, adaptable approach to account for the varying schedules and needs of participants from different institutions.

Step 3: Creating an Inclusive and Collaborative Research Environment

The laboratory's success can be attributed to its focus on creating a collaborative environment where students feel valued and empowered to contribute to meaningful research, regardless of their academic background. This ethos is reflected in the laboratory's recruitment policy, which invites students from institutions with and without established dermatology programs to participate in research. This inclusive approach fosters camaraderie between MD and DO students, enhancing professional networks and building connections early in their careers.

Replicating such an inclusive environment requires institutions to recruit and include participants consciously and to create a culture that values all contributions, regardless of a student's institutional affiliation. Offering mentorship and guidance from experienced researchers and ensuring that students from various backgrounds have equal access to resources and support are essential steps in creating a thriving research community.

Step 4: Ensuring Program Sustainability and Resources

Long-term sustainability of a virtual collaborative research program requires strategic planning, flexible leadership, and efficient resource management. Many collaborative research programs rely on extramural grants and institutional support to finance operations. Program leaders should prioritize securing long-term funding sources, including government grants, philanthropic support, or institutional backing, to ensure the sustainability of the research program. Since most academic institutions already provide access to platforms like Zoom and Microsoft 365, investing in a solid technology infrastructure can significantly streamline communication and enhance collaboration.

Step 5: Replicating the Model: Key Insights

To replicate the success of the Dellavalle/Dunnick Dermato-Epidemiology Lab, it's important to focus on several core principles, outlined in Textbox 2.

Textbox 2. Core principles.

Defining clear objectives

Establish clear and measurable goals for the program, ensuring they align with students' goals and the broader goals of your institution or research community.

Structuring for organization

Implement a consistent structure for regular meetings, project tracking, and task management.

Promoting inclusivity

Create an inclusive research environment that invites participation from students across multiple institutions and backgrounds and encourages collaboration.

Securing resources

Plan for the sustainability of the program by securing funding and investing in technology platforms that facilitate virtual collaboration.

Performing ongoing evaluation

Continuously assess the program's impact and refine based on participant feedback.

Conclusion

The Dellavalle/Dunnick Dermato-Epidemiology Lab offers a practical model for building scalable, inclusive, and collaborative multicampus research programs. Its structured virtual environment lowers access barriers, supports consistent mentorship, and engages students from schools without dermatology departments, addressing common gaps in academic medical training.

As medical education adapts to more decentralized and technology-driven formats, this model illustrates how virtual infrastructure and intentional design can sustain student scholarship and faculty engagement. By applying the principles outlined in this editorial—clear objectives, open access, consistent workflows, and distributed leadership—other institutions can replicate the success of the Dellavalle/Dunnick Dermato-Epidemiology Lab and contribute to the next generation of physician-scientists who are prepared to excel in an increasingly competitive medical landscape.

Authors' Contributions

NEB: investigation, methodology, project administration, writing—original draft, writing—review and editing
KT: investigation, writing—original draft, writing—review and editing
ARL-S: writing—original draft, writing—review and editing
CAD: methodology, writing—review and editing
RPD: investigation, methodology, writing—review and editing

Conflicts of Interest

RPD is the editor-in-chief of *JMIR Dermatology*. CAD reports receiving royalties from UpToDate and is a speaker for Pfizer.

Multimedia Appendix 1
Fictitious laboratory executive summary. Details included were created for the purposes of this tutorial and are not accurate or representative. Formatting is the same as the real shared document.
[DOCX File, 8 KB - mededu_v11i1e65697_app1.docx]

References

1. Robert Dellavalle profile. Google Scholar. URL: https://scholar.google.ca/citations?user=Qo_2B2cAAAAJ&hl=en&oi=ao [accessed 2025-04-18]
2. Cory A Dunnick profile. Google Scholar. URL: https://scholar.google.ca/citations?hl=en&user=bBC9v24AAAAJ [accessed 2025-04-18]
3. Resources: faculty development. Association of Professors of Dermatology. URL: https://www.dermatologyprofessors.org/bfd_chronological.php [accessed 2025-04-06]
4. Yeh C, Desai AD, Wilson BN, et al. Cross-sectional analysis of scholarly work and mentor relationships in matched dermatology residency applicants. J Am Acad Dermatol 2022 Jun;86(6):1437-1439. [doi: 10.1016/j.jaad.2021.06.861] [Medline: 34214622]
5. Burgess A, Oates K, Goulston K. Role modelling in medical education: the importance of teaching skills. Clin Teach 2016 Apr;13(2):134-137. [doi: 10.1111/tct.12397] [Medline: 26119778]

Abbreviations

CU: University of Colorado
IRB: institutional review board
UMN: University of Minnesota

Edited by B Lesselroth; submitted 23.08.24; this is a non–peer-reviewed article;accepted 23.09.25; published 30.10.25.

Please cite as:
Barton NE, Ta K, Loczi-Storm AR, Dunnick CA, Dellavalle RP
Advantages of a Virtual Collaborative Research Dermatology Laboratory
JMIR Med Educ 2025;11:e65697
URL: https://mededu.jmir.org/2025/1/e65697
doi:10.2196/65697

Leveraging Datathons to Teach AI in Undergraduate Medical Education: Case Study

Michael Steven Yao[1,2,3], BS; Lawrence Huang[3,4*], BS; Emily Leventhal[3,5*], BA; Clara Sun[3,6], BS; Steve J Stephen[3,7,8], MBA; Lathan Liou[3,5], MPhil

[1]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States
[2]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA, United States
[3]MDplus, New York, NY, United States
[4]Warren Alpert Medical School, Brown University, Providence, RI, United States
[5]Icahn School of Medicine at Mount Sinai, New York, NY, United States
[6]School of Medicine, Case Western Reserve University, Cleveland, OH, United States
[7]School of Medicine and Dentistry, University of Rochester, Rochester, NY, United States
[8]Simon Business School, University of Rochester, Rochester, NY, United States
[*]these authors contributed equally

Corresponding Author:
Lathan Liou, MPhil
MDplus, New York, NY, United States

Abstract

Background: As artificial intelligence and machine learning become increasingly influential in clinical practice, it is critical for future physicians to understand how such novel technologies will impact the delivery of patient care.

Objective: We describe 2 trainee-led, multi-institutional datathons as an effective means of teaching key data science and machine learning skills to medical trainees. We offer key insights on the practical implementation of such datathons and analyze experiences gained and lessons learned for future datathon initiatives.

Methods: We detail 2 recent datathons organized by MDplus, a national trainee-led nonprofit organization. To assess the efficacy of the datathon as an educational experience, an opt-in postdatathon survey was sent to all registered participants. Survey responses were deidentified and anonymized before downstream analysis to assess the quality of datathon experiences and areas for future work.

Results: Our digital datathons between 2023 and 2024 were attended by approximately 200 medical trainees across the United States. A diverse array of medical specialty interests was represented among participants, with 43% (21/49) of survey participants expressing an interest in internal medicine, 35% (17/49) in surgery, and 22% (11/49) in radiology. Participant skills in leveraging Python for analyzing medical datasets improved after the datathon, and survey respondents enjoyed participating in the datathon.

Conclusions: The datathon proved to be an effective and cost-effective means of providing medical trainees the opportunity to collaborate on data-driven projects in health care. Participants agreed that datathons improved their ability to generate clinically meaningful insights from data. Our results suggest that datathons can serve as valuable and effective educational experiences for medical trainees to become better skilled in leveraging data science and artificial intelligence for patient care.

(JMIR Med Educ 2025;11:e63602) doi:10.2196/63602

KEYWORDS

data science education; datathon; machine learning; artificial intelligence; undergraduate medical education

Introduction

The exploration of machine learning (ML), artificial intelligence (AI), and other data science-driven technologies is becoming increasingly popular within clinical medicine [1-5]. Given the rapidly growing presence of ML in health care innovation, it is important for both current and future physicians to understand the fundamentals of ML technology and how they may help inform clinical decision-making.

However, data science and AI education in current medical school curricula are lacking. Despite recent efforts to integrate AI learning objectives into medical education [6-10], few US medical schools have formally integrated AI-based topics into their curricula. Pupic et al [11] and Civaner et al [12] report studies of small self-selected groups of medical students and

residents participating in both student- and faculty-led electives covering the fundamental theory behind AI applications for medicine. However, opportunities facilitating real-world experience remain limited [13,14].

One potential method for hands-on AI education popular across many fields of science and engineering is the "datathon," which is a short competition where teams of students work together to create new solutions to domain-specific challenges through leveraging real-world data and algorithms. Following Daneshvar et al [15], we also make the important distinction between datathons and hackathons. Traditionally, hackathons are product-orientated initiatives where team projects are primarily focused on programming novel products and applications. By contrast, the primary learning objectives for our datathons were to (1) teach student participants how to analyze complex datasets to support clinical insights, and (2) leverage ML models to derive these clinical insights from data. Oyetade et al [16] offer a scoping review of datathons and found that such events help students learn both technical and soft skills and argue that datathon-based pedagogies be incorporated in classroom environments. Silver et al [17] describe a hackathon event for current attendings in clinical practice and found that study participants were better equipped to accelerate specialty-focused innovation after the hackathon. However, similar events specifically designed for medical students and other undergraduate trainees are not well described in the literature.

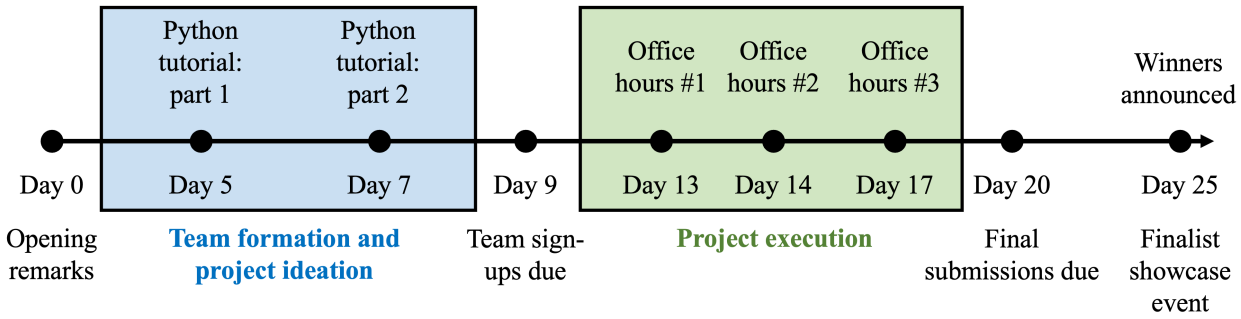
In this work, we hypothesize that datathons can be an effective training initiative to teach skills in AI to medical trainee participants. To evaluate this hypothesis, we describe 2 datathons hosted by MDplus, a 501(c)3 national student-run nonprofit whose mission is to support and empower future physician-innovators. We describe the structure of the events, present data on educational outcomes, and offer resources and recommendations for putting together similar events in the future. Our results suggest that datathons and similar events may be an effective means for AI education for medical students.

Methods

Overview

In this section, we detail the logistics of organizing and executing 2 trainee-led datathon events. A number of features distinguish our events from prior work. First, the datathons are trainee-led; all members of the organizing committee were undergraduate medical trainees at the time of the event. Second, the datathons were held digitally over the course of approximately 3 weeks (Figure 1). Finally, the target participant audience of our datathons included current undergraduate medical trainees at institutions granting doctoral degrees. These features of our datathons substantially differentiate them from prior work [15-17], and also affected our design and organization of the events that we detail below.

Figure 1. Overview of the datathon event. The MDplus datathon ran for approximately 4 weeks and was loosely divided into two parts: (1) Team formation and project ideation and (2) project execution.



Timeline and Participant Recruitment

We, the datathon organizing team, detail 2 datathon events organized by MDplus between 2023 and 2024, herein referred to as "the datathons." Each datathon ran for approximately 3 weeks (Figure 1), and was organized by the medical trainee-led executive team of MDplus, consisting of a core datathon planning team of 8 medical trainees. To accommodate the participation of medical trainees from across the United States, the entirety of both datathons was held digitally. The MDplus' Slack community, monthly newsletter, and social media pages (LinkedIn, Instagram, and Twitter) were used to advertise the datathon. Over the span of 3 months prior to the start of the datathon, 2 organizing team members were tasked with recruiting sponsors, mentors, knowledge experts, and judges through the MDplus and personal networks, while 3 organizing team members—all with prior experience working as software engineers prior to medical school—crafted and iterated the educational material and dataset curation for the event. One team member helped publicize the event on social media. Registration for the event was limited to current trainees (ie, medical students, residents, and graduate students) in the United States. To provide a fair learning environment for trainees, our organizing team opted to exclude attending physicians, industry professionals, and individuals with extensive technical backgrounds in software engineering from participation. Participants were asked to form their own teams of 3-5 individuals.

Datathon Theme and Dataset

Each of the datathons focused on a specific theme to help participants contextualize their projects within a specific application relevant to health care. The theme of the 2024 datathon was responsible generative AI and that of the 2023 datathon was value-based care (VBC). Generative AI is an area of ML that uses technologies such as large language models (LLMs) to create new content by learning patterns from existing human-generated examples [18-20]. While such technologies

have the potential to improve health care delivery, recent work has highlighted a growing need to better evaluate how clinicians can use these tools responsibly before real-world integration is possible [21-23]. Separately, VBC refers to a health care delivery model in which providers are held accountable for improving patient outcomes. In a VBC system, providers are often rewarded with incentivized payments based on quality of care, provider performance, and the patient experience [24].

To enable participants to explore projects related to each of these themes, a medical dataset was made available for participants to use in each of our datathons. All datasets were made available via Hugging Face (Hugging Face, Inc), a public repository to facilitate the sharing of ML data and models. In our 2023 VBC datathon, participants were required to use the Medical Information Mart for Intensive Care (MIMIC)-IV dataset [25], a single-site dataset of patient records and admission details. Briefly, the MIMIC-IV dataset contains anonymized patient data aggregated from over 500,000 patients at the Beth Israel Deaconess Medical Center between 2008 and 2019. Variables from this rich dataset include electrocardiograms, medical imaging studies, health records, and patient laboratory values and outcomes, among others. We chose to use this dataset specifically for the datathon because of the following factors:

Public Availability

In similar prior events organized by the authors, we found that procuring a real-world dataset of health care data can often be prohibitively expensive or constraining, especially for trainee-led initiatives with limited budgets. To circumvent this problem, we used the MIMIC-IV dataset, which is made publicly available by Johnson et al [25].

Real Patient Data and Outcomes

The primary learning objective of our datathons is to teach participants how to derive data-driven insights to affect and ultimately improve patient care. We therefore sought to provide real patient data for participants to explore and use for their projects in alignment with this goal.

Prevalence of Prior Work

The vast majority of our participants have minimal (if any) prior experience with programming and data analysis techniques. For this reason, the abundance of prior literature and publicly available coding resources for interacting with the MIMIC-IV dataset helped lower the barrier to participating in the datathon.

Multiple Modalities of Data

Many participants have individual academic and personal interests in medicine, and we sought to encourage participants to craft and work on projects that were interesting to them. The abundance of textual, image, biomedical signal, and laboratory data available in the MIMIC-IV dataset was important to make this possible.

All participants in the datathon were required to sign a data use agreement and complete responsible data handling training in order to gain access to the MIMIC-IV dataset. Participating teams were tasked with thinking critically about quantitative methods, conducting appropriate analyses (eg visualization, statistics, and other computational tools), and contextualizing clinical insights into actionable proposals that solve a problem related to VBC for relevant stakeholders.

While organizing for the 2024 generative AI datathon, we found that one limitation of the MIMIC-IV dataset was its size and complexity, making it unwieldy for some participants to work with for their projects. To overcome this challenge while simultaneously retaining the desirable features listed above, our 2024 datathon introduced the concept of datathon “tracks”: teams were able to choose to participate in 1 of 3 tracks within the broader theme of responsible generative AI. Each track was associated with its own dataset: (1) Clinical Documentation track participants used the MTS-Dialog dataset of patient-physician conversation transcripts from Abacha et al [26]; (2) Medical Education track participants used the MedQA dataset of practice medical board examination questions from Jin et al [27]; and (3) Mental Health track participants used the SuicideWatch and Mental Health Collection dataset of tagged social media posts from Ji et al [28]. Participants were allowed to participate in at most 1 of the 3 tracks.

Resources and Support

An official datathon page [29] was created for participants as a central hub with instructions, registration, and materials for the event. Links to the datathon’s Github Repository were provided with written tutorials and example code, including (1) downloading and overview of the datasets; (2) introduction to Python (Python Software Foundation; offered in both the 2023 and 2024 datathons; see Multimedia Appendix 1); and (3) an introduction to R (R Foundation; offered only in the 2023 datathon). Optional workshops and private Zoom (Zoom Communications, Inc) events with experienced data scientists were offered to participating trainees, including Python and R bootcamps, oral presentation workshops, and a prerecorded Zoom talk with physician experts. The scope of the projects was largely left up to the discretion of individual team members; participant teams were encouraged to leverage the optional workshop sessions and public discussion channels on Slack if they would benefit from discussing potential project ideas with others, although no explicit guidance on project ideation or constraints was given other than all teams had to (1) use the official datathon dataset and (2) work on a project under the broad datathon theme (ie, VBC in 2023 and responsible generative AI in 2024) and track. No tutorials or structured datathon programming were provided for teaching participants how to use GitHub, GitLab, Microsoft Excel, or other computing tools. Communication and announcements throughout the datathon were conducted through Slack.

Submission Requirements and Judging Criteria

In the 2023 VBC datathon, teams were asked to submit a written technical report of their work without restrictions on the word count and were asked to record a 5-minute-long oral presentation highlighting key contributions and findings. Participants were free to use any programming language or software to perform their analysis. In the 2024 generative AI datathon, teams were asked to submit a 1-page extended abstract with at most 1 figure and unlimited references and a written technical report without word count restrictions. Judging criteria in both datathons

included statistical rigor, relevance to the datathon theme (VBC), creativity of visualization and analysis, and team diversity (Multimedia Appendix 2).

Final Showcase Event

In the 2023 VBC datathon, an internal set of 4 blinded judges composed of members of the MDplus datathon organizing committee evaluated the initial anonymized submissions and selected 7 finalist teams to present at the finalist datathon showcase event. Each team played their recorded 5-minute oral presentations and were allotted 2 minutes immediately after for responding to judge questions. A panel of 5 judges—recruited for their diverse range of expertise in the VBC space—evaluated the finalists' submissions. In total, 3 of the judges are health care executives, 4 are practicing clinicians, and 1 is a product manager.

In the 2024 generative AI datathon, an internal set of 3 blinded judges composed of members of the organizing committee evaluated the 14 initial anonymized team submissions and selected 8 finalist teams to present at the final datathon showcase. Finalist teams were invited to a 2-hour finalist showcase event where they were each allotted 8 minutes for a live oral presentation followed by 2 minutes of question answering with the judges. We recruited a panel of 4 judges to evaluate the finalist submissions: 1 judge is a software engineer at a health care company, 1 judge is a postdoctoral fellow in a health care AI lab, and 2 judges are practicing physicians in the United States. In general, we found the live oral presentations to be better received by the judges and audience members than playing prerecorded presentations.

Postdatathon Survey

Upon the conclusion of each datathon, an anonymous 16-question open survey (Multimedia Appendix 3) was electronically sent to all registered participants that submitted a final project via both Slack and email; this survey study was exempted by the University of Pennsylvania Institutional Review Board (protocol #856530). The survey was created in close collaboration with an attending physician at a US academic medical institution with expertise in medical education and assessing educational outcomes and was piloted within the datathon organizing team prior to the public release of the survey. Participants were requested to complete the survey within the 2 weeks immediately following the conclusion of the respective datathon, and the survey remained open for 3 weeks. Participant emails were collected to ensure that no individual filled out the survey multiple times but were removed prior to analysis. The optional, opt-in survey asked respondents questions pertaining to team demographics, medical education status, medical specialty interest, familiarity with technical and computational tools, and subjective datathon quality. The questions were divided between 4 survey pages, each taking approximately 1 minute to complete; no partial survey responses were submitted. Participants were asked to rate their familiarity with quantitative tools before and after the datathon on a 4-point scale (1=no familiarity, 4=a lot of familiarity) and were also informed that the study results would be anonymized and deidentified prior to analysis. To assess the efficacy of the aforementioned technical Python and R tutorials for datathon participants, we compared them against participant subjective familiarity with quantitative tools—namely, GitHub and Microsoft Excel—that were not taught explicitly as a part of the datathon. Data were analyzed using the Fisher exact test in Python 3. To better characterize participant experiences during the datathon, survey respondents also rated their agreement with a set of 5 standardized statements regarding (1) overall enjoyment of the datathon, (2) VBC topic understanding, (3) ability to identify problems in health care, (4) ability to generate insights from data, and (5) likelihood of future datathon participation. Participant sentiment was quantified using a 5-point Likert scale (1=strongly disagree, 5=strongly agree) [30].

Ethical Considerations

This study was exempted by the University of Pennsylvania Institutional Review Board (protocol #856530). All opt-in participants provided informed consent prior to data collection and were not compensated for participating in our optional, opt-in survey as a part of our study. Confidentiality and privacy were maintained during data acquisition and analysis, and participants had the right to withdraw their data from the study at any time without any consequences.

Results

Datathon Logistics

In the 2023 datathon, 28 teams consisting of a total of 109 participants registered for the datathon, of which 13 of the initial registered teams submitted a final project, while in the 2024 datathon, 25 teams consisting of 110 participants registered for the datathon, of which 14 of the initial registered teams submitted a final project. Among the submitted projects, 7 and 8 were chosen as finalists to present at the synchronous digital showcase in 2023 and 2024, respectively. In the 2023 VBC datathon, the 7 projects addressed a variety of topics related to VBC, including chronic kidney disease underdiagnosis, the efficacy of social work referrals, and readmission rates for alcohol-related conditions, among others. Similarly, the 2024 responsible generative AI datathon featured 2 clinical documentation track teams, 3 medical education track teams, and 3 mental health track teams. We include brief descriptions of each of the finalist projects in Table 1. The final showcase was followed by the announcement of the 3 winning projects; we announced the winning teams at the end of the showcase in the 2023 datathon and 48 hours after the end of the showcase in the 2024 datathon.

Table . Sample datathon project descriptions. Descriptions of finalist datathon projects for the 2023 and 2024 MDplus datathons are shown to illustrate the diversity of project submissions from participating teams.

Theme or track	Project description
Value-based care	• Minimizing chronic kidney disease (CKD) underdiagnosis using machine learning • Significant association of social work referral and 30-day unplanned hospital readmission for patients with alcohol-related disorders using MIMIC[a]-IV data • Can we curb frequent emergency department (ED) visits due to alcohol-related conditions? • Automatic knowledge graph extraction from medical discharge notes for clinical decision support • Contrast overuse in patients with renal disease: a targeted analysis • Analyzing acuity as a tool for value-based care • Machine learning-driven forecasting and characterization of the intensive care unit (ICU)-admitted heart failure patient population in the MIMIC-IV (version 04) database
Responsible generative AI: clinical documentation	• Cost-benefit analysis of non-artificial intelligence (AI) and AI models implemented for predicting chief complaints • Bridging speech documentation and clinical support through LLM[b] automation • Automating trust in AI-generated clinical notes: developing a look-up tool for real-time verification
Responsible generative AI: medical education	• Using a LLM for USMLE[c] preparation via generative AI • Use of LLMs in assessing how age and gender affect model accuracy in clinical reasoning
Responsible generative AI: mental health	• Reassessing specialist models: risks in fine-tuning LLMs for mental health tasks • Robust text classification and grounded LLM integration for personalized mental health support • Characterizing suicidal ideation subtypes in social media posts via unsupervised contrastive feature identification

[a]MIMIC: Medical Information Mart for Intensive Care.

[b]LLM: large language model.

[c]USMLE: United States Medical Licensing Examination

The organization-accrued cost of organizing and running the datathons was US $28 per participant, averaged over the number of participants who individually registered for the datathon regardless of whether they ultimately submitted a final project. The majority of expenses supported prize money, computing resources for participants, technical skill-based workshops, and other resources that were provided during the datathon. In our experience, most of the costs accrued were for (1) the prize money of the datathon winners and (2) honorariums for the guest judges in the finalist showcase events. We primarily relied on sponsorships from industry partners to provide computing resources for participants, and MDplus community members readily volunteered to help lead technical skill-based workshops and offer pro-bono mentorship to participating teams.

Survey Results

Out of the 219 registered participants (summed over both datathons), 61 (28%) completed the postdatathon survey (Table 2). A majority who completed the survey identified as male (71%, 43/61) and were under the age of 25 years (61%, 37/61). Survey respondents self-reported as Asian (69%, 42/61), White (20%, 12/61), Middle Eastern or North African (3.3%, 2/61), Hispanic or Latinx (3.3%, 2/61), or Black or African American (1.6%, 1/61); and 3.3% (2/61) preferred not to say. In total, 49/61 (80%) of survey respondents were medical students (Table 3); there was a wide range of medical specialty interests amongst the medical trainee survey respondents, with internal medicine (21/49), surgery (17/49), and radiology (11/49) being the most popular specialties.

Table . Demographic information of participants who completed the postdatathon survey (N=61).

Characteristics	Value, n (%)
Age, years	
<25	37 (61)
25–30	22 (36)
30–35	1 (1.6)
≥35	1 (1.6)
Self-reported race and ethnicity	
Asian	42 (69)
Hispanic or Latinx	2 (3.3)
Middle Eastern or North African	2 (3.3)
White	12 (20)
Black or African American	1 (1.6)
Prefer not to say	2 (3.3)
Gender	
Male	43 (71)
Female	18 (29)
Sexual orientation	
Heterosexual or straight	57 (93)
Bisexual, gay, lesbian, or other	4 (6.6)
Disability status	
Does not identify as a person with a disability	54 (89)
Does identify as a person with a disability	5 (8.2)
Prefer not to answer	2 (3.3)
Current education status	
Medical student or resident physician	49 (80)
Other	12 (20)

Table . Datathon participant analysis. Current medical education status and medical specialty interest information for participants who completed the postdatathon survey (N=49) filtered by medical student and resident physician status. Note that respondents were allowed to select multiple medical specialties.

Characteristics	Value
Current medical education status, n (%)	
First-year medical student	14 (29)
Second-year medical student	19 (39)
Third-year medical student	6 (12)
Year-out medical student	4 (8.2)
Fourth-year medical student	5 (10)
Resident physician	1 (2.0)
Medical specialty interests, n (%)	
Anesthesia or critical care	9 (18)
Cardiology	1 (2.0)
Dermatology	5 (10)
Emergency medicine (EM)	4 (8.2)
Family medicine (FM)	1 (2.0)
Internal medicine	21 (43)
Mental health counseling and therapy	2 (4.1)
Neurology	9 (18)
Obstetrics and gynecology (OB/GYN)	3 (6.1)
Ophthalmology	5 (10)
Pediatrics	5 (10)
Physical medicine and rehabilitation (PM&R)	1 (2.0)
Plastic surgery	1 (2.0)
Psychiatry	8 (16)
Radiology	11 (22)
Surgery (general or unspecified)	17 (35)
Orthopedic surgery	2 (4.1)
Not currently exploring a medical specialty	1 (2.0)

Familiarity with quantitative tools, Python, R, Github/Gitlab, and Microsoft Excel before and after participating in the datathon was assessed (Figure 2). As a reminder, a core component of the programming of both our VBC and generative AI datathons was the educational workshops and tutorials on data analysis and ML skills using Python. Workshops on the programming language R were only offered in the 2023 VBC datathon. As our negative controls, we also asked participants to rate their skills with Github/Gitlab and Microsoft Excel; neither of these software were primary educational components of the datathons. As expected, participant familiarity with GitHub/Gitlab and Microsoft Excel did not significantly change before and after the datathon (GitHub/Gitlab: P=.92; Microsoft Excel: P=1.00; pairwise Fisher exact test). In contrast, subjective participant familiarity with Python significantly improved through participation in the datathons (P=.04; pairwise Fisher exact test); familiarity with R showed some evidence of improvement (P=.83; pairwise Fisher exact test), although it did not reach the traditional threshold for statistical significance likely due to the limited sample size of the study. Our reports support that targeted educational tutorials during the datathon event can empower participants with improved technical skills relevant to data science applications in medicine.

Figure 2. Bar plot visualizing participant self-assessment of technical skills before and after participating in the datathon for all 61 survey responses. Python was the only skill out of the 4 above that was an educational component in both the 2023 VBC and 2024 generative AI datathons. Participant scores correspond to the following: (1) no familiarity; (2) a little familiarity; (3) some familiarity; (4) a lot of familiarity. * Indicates a statistically significant difference in the distribution of scores before and after participating in the datathon (Python: *P*=.041; pairwise Fisher exact test). n.s. indicates no statistically significant difference in the distribution of scores. (R: *P*=.83; GitHub/Gitlab: *P*=.92; Microsoft Excel: *P*=1.00; pairwise Fisher exact test).

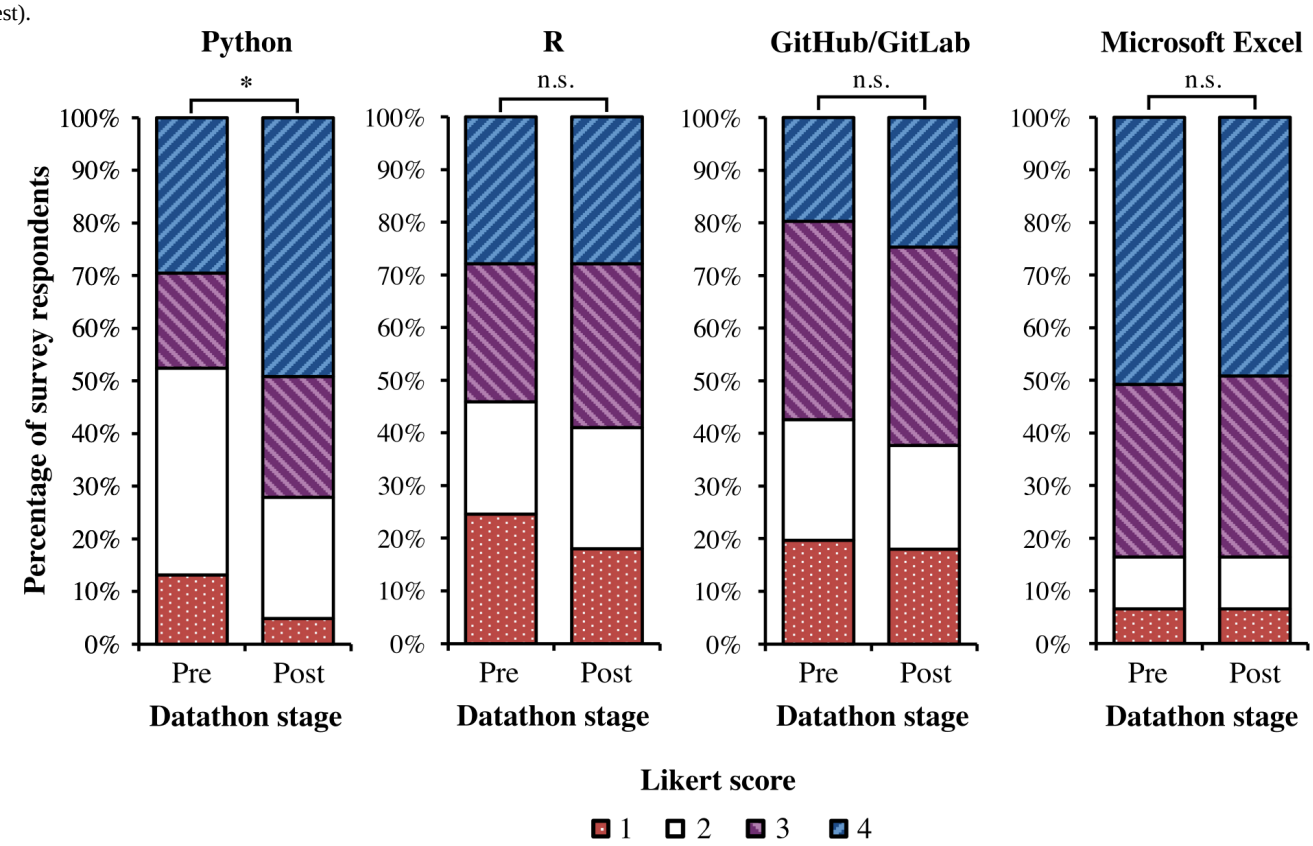
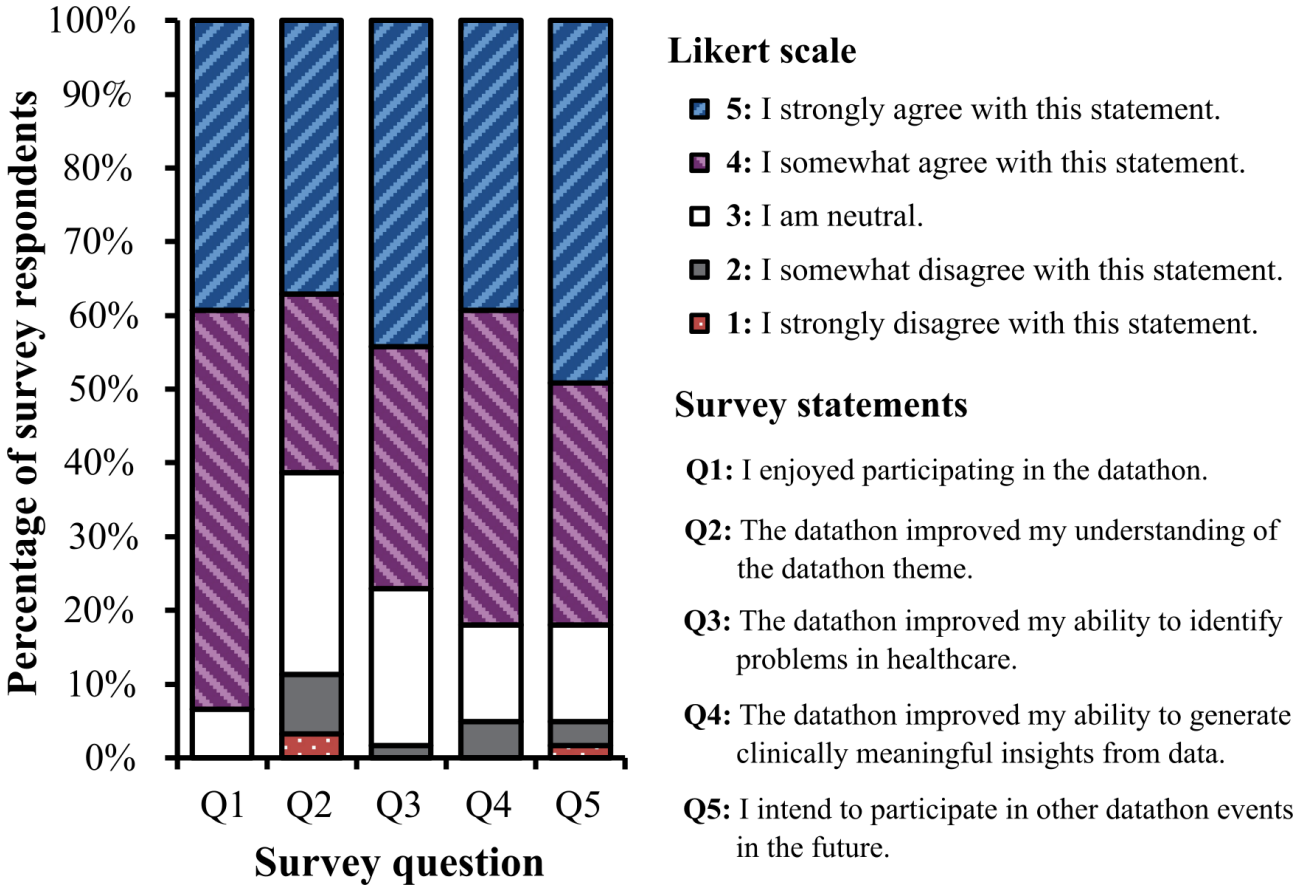


Figure 3 examines the participant experience quantified by participant agreement with a set of standardized statements. Overall, 57/61 (93%) survey respondents enjoyed participating in the datathon, and 38/61 (62%) respondents affirmed that the datathon improved their understanding of the VBC or responsible generative AI theme (ie, Likert score of 4 or 5). We also found that 47/61 (77%) respondents stated that their ability to identify problems in health care improved, and 50/61 (82%) respondents agreed that they were better equipped to generate meaningful insights from data. Of the 61 participants, 50 participants (82%) also expressed interest in participating in similar datathon events in the future. For each of these statements, an “agreeable sentiment” was determined by indicating a Likert scale value of either 4 (“I somewhat agree with the statement”) or 5 (“I strongly agree with the statement”) on a 5-point scale in the participant survey response.

Figure 3. Bar plot visualizing survey results assessing for subjective datathon quality. Participant scores correspond to the following: (1) Strongly disagree; (2) Disagree; (3) Neither agree nor disagree; (4) Agree; and (5) Strongly agree.



Qualitative Survey Results

The survey also included an open-ended response option for participants to provide any additional comments. There was a mix of short, positive comments and comments that offered suggestions for future events. Based on our qualitative analysis, key areas for improvement to consider for future datathon iterations include (1) ensuring a balanced distribution of technical skills between participating teams; (2) expediting the team creation process; and (3) offering additional technical workshops and tutorials to participants. Representative example unedited participant comments are shown below:

> *I think in the future, it'd be more effective to make sure each team at least has a "senior" tech lead (someone with 3-5+ years of tech experience) and a "junior" tech lead (1-2 years) to ensure there is great education for all parties involved, as well as greater quality of work. This is of course for folks seeking out teams and not those who already have a team formed that they are comfortable with.*

> *…I feel like the team creation process could've been a little faster and I was only able to join a team around halfway into the datathon which didn't give us enough time to work on our idea. But overall, I really appreciate the effort and time put in by everyone involved and I definitely hope to be involved in this again!*

Discussion

Principal Findings

In this work, we describe an instance of a trainee-led datathon to teach medical trainees how to effectively leverage modern computational tools to solve real-world problems in medicine. We show preliminary evidence that trainees become more familiar with foundational skills such as reading and writing computer programs in Python and R, are satisfied with their participation, and are eager to participate in similar initiatives in the future. To our knowledge, our national, trainee-led datathons were the first to bring together teams of medical students, residents, and graduate students to propose data-driven solutions within VBC. Our study ultimately supports that datathons can be effective platforms to teach medical trainees how to leverage AI to advance clinical medicine.

Logistical Insights and Best Practice Recommendations

In this section, we offer additional discussion on subjective design choices and lessons learned from the MDplus datathon organizing team. We hope that our experiences and takeaways can serve as a foundation for which future datathon educational initiatives can build upon.

Perhaps one of the most notable logistical details that distinguish our datathons from related hackathons that are traditionally organized by computer science students outside of medicine is that our datathons each spanned the course of multiple weeks asynchronously, whereas hackathons are often held over the

course of a few days in a single physical location. While we recognize that there are likely untapped benefits with this alternative strategy, we chose to run an extended digital datathon due to two primary reasons: (1) to support participation from MDplus members spanning multiple countries and timezones; and (2) to minimize potential time conflicts with concurrent medical school curricula for participants. In our work, these 2 constraints together necessarily precluded an in-person datathon; in situations where either one or both constraints are not limiting, future work may warrant exploring similar datathon initiatives spanning a few days hosted in a single physical location.

Separately, we also emphasize the importance of carefully choosing the datasets used in the datathon. In our 2023 VBC datathon using the MIMIC-IV dataset, we retrospectively observed that some participants initially struggled with the technical implementation details of working with the MIMIC-IV dataset due to the sheer volume of data available and the preprocessing steps before any ML modeling could be done. This consideration was especially important as the majority of participants registered in the datathon with little or no prior experience with computer programming (Figure 2). At the same time, participants also voiced enthusiasm for the diversity of data available in the MIMIC-IV dataset—making multiple modalities of data available, such as medical imaging, textual clinical documentation, biometric signals, and tabular data, allowed for participating teams to design and execute projects tailored to their specific interests. In our 2024 generative AI datathon, we found that the introduction of datathon "tracks" enabled us to offer 3 diverse dataset options while simultaneously removing the extra data processing steps outside the scope of the datathon learning objectives.

We also evaluated the utility of unstructured "office-hour" sessions where participating teams could ask experienced members of the community for assistance with their projects. Despite holding multiple office-hour sessions at different times of the day throughout the datathon, we found that only 1 team attended any of the office-hour sessions in the 2023 VBC datathon. Because of this low attendance, we opted to remove synchronous office hours from the 2024 datathon programming and instead implemented a custom anonymous discussion forum via the datathon Slack communication channel where participants could ask questions anonymously that could be viewed and answered by anyone. Subjectively, we found that this asynchronous mode of communication made it easier for participants to seek help with their projects and observed greater engagement in public discussions after this feature was implemented. Future work is warranted to more rigorously evaluate the utility of such interventions.

Finally, we acknowledge that disciplines such as medicine and computer science have historically seen disproportionate participation from trainees of certain racial, socioeconomic, and gender backgrounds. These systemic trends well described in prior work [31,32] are reproduced in our datathons as well (Table 2); as topics such as data science, VBC, and generative AI become increasingly important components of modern health care, it is crucial that all future clinicians from all backgrounds can interact meaningfully with these concepts and their applications. We hope that future work will explore how to reduce barriers to participation for historically marginalized groups of trainees.

Related Work

The majority of prior work published in related literature details short datathons lasting a few days at a single physical location with a different target participant group. Hochheiser et al [33] describe a 2-day datathon consisting of 5 participating teams of clinicians and informaticians working on elucidating potential sources of bias within health care ML models. While their synchronous datathon model may be suitable for participants at a single physical site, such a model was intractable for our purposes as participating trainees were distributed across multiple institutions and time zones. Sobel et al [15] detail a similar datathon at a single physical location, but their study was primarily conducted with undergraduate and graduate students with pre-existing computational backgrounds, as opposed to undergraduate medical trainees from institutions granting postdoctoral fellowships as in our case. Anecdotally, we found evidence of similar initiatives held at the institutional level, such as the Digital Critical Care Datathon [34], the New York University Health Tech Datathon [35], and the Society of Critical Care Medicine Datathon [36]; each of these were single-institution initiatives with different datathon design constraints. To our knowledge, we are the first to describe a trainee-led, multi-institutional, asynchronous datathon effort and demonstrate preliminary evidence of its efficacy and potential role in the future of medical education.

Limitations

There are also limitations associated of our study. Firstly, our datathon was coordinated digitally with participants joining from across the United States. While we acknowledge there are both benefits and drawbacks to a datathon (as opposed to their in-person counterparts), we leave a rigorous comparison between their utilities in modern medical education paradigms for future work. Furthermore, both participating in the datathon and completing the postparticipation survey were opt-in processes, and so it is unclear how our findings would translate to undergraduate medical trainees who might have systematically chosen to not participate in the datathons—for example, potential participants who were more hesitant in learning about AI and data science practices in medicine and those whose medical school coursework made concurrent participation in the datathon unfeasible. Our survey results also exclude individuals who initially signed up to express interest in participating in the datathon but ultimately decided not to submit a final project. Given the opt-in design of our survey study, we were unable to assess the efficacy of our datathons for these individuals. Future work might evaluate how similar initiatives could scale across more diverse participant profiles and foster participation from student trainees of all backgrounds and perspectives. Finally, our postparticipation survey makes use of retrospective questions that ask participants to subjectively reflect on their skill development, rather than an objective evaluation of participant skills through a standardized programming examination. We chose this study design for two primary reasons: (1) because of the diverse array of participant projects, the skill sets that they developed through their

participation in the datathon are likely equally diverse, making a single standardized examination challenging to construct; and (2) in our initial efforts in organizing the datathon, we hypothesized that the survey response rate would be too low to adequately power our study if we asked participants to complete opt-in programming examinations. We leave the exploration of using more standardized assessments of programming skill competencies attained through datathon initiatives as future work.

Conclusions

Ultimately, the goal of this datathon was to provide opportunities for trainees—especially medical students—to improve their data skills and to identify data-driven solutions to problems in health care. Participants practiced using hands-on data science and artificial intelligence to explore meaningful clinical problems and voiced a collective interest in continuing to participate in similar initiatives in the future. Overall, our results and collective experiences suggest that datathons can be valuable within undergraduate medical education.

Acknowledgments

The authors thank the teams at Hugging Face, Doximity, Merck, Conduce Health, and AvoMD for their generous support and sponsorship of the 2023 and 2024 MDplus datathon events, and the datathon judges (listed in no particular order) Reza Alavi, MD, MHS, MBA; Caroline Berchuck, MD, MPH; Amit Phull, MD; Sid Salvi; Kathryn Teng, MD, MBA, FACP; Andrea Green; David Dupee, MD; Marc Triola, MD; and Jannes Jegminat, PhD for lending their time and expertise in helping run a successful datathon event. The authors also thank Eric Shan, Katie Link, Julia Bondar, Vamsi Chodisetty, and other members of the MDplus community and executive team for their help in organizing and running the datathon event. MSY was supported by NIH F30 MD020264, and ELL was supported by NIH T32 training grant GM146636; the content in this manuscript is solely the responsibility of the authors and does not necessarily represent the official view of the NIH.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Introduction to Python Datathon Tutorial.
[PDF File, 150 KB - mededu_v11i1e63602_app1.pdf]

Multimedia Appendix 2
Judging Rubric for Datathon Finalist Showcase Event.
[XLSX File, 13 KB - mededu_v11i1e63602_app2.xlsx]

Multimedia Appendix 3
Post-Datathon Survey.
[DOCX File, 18 KB - mededu_v11i1e63602_app3.docx]

References

1. Hege I, Kononowicz AA, Adler M. A clinical reasoning tool for virtual patients: design-based research study. JMIR Med Educ 2017 Nov 2;3(2):e21. [doi: 10.2196/mededu.8100] [Medline: 29097355]
2. Pongdee T, Larson NB, Divekar R, Bielinski SJ, Liu H, Moon S. Automated identification of aspirin-exacerbated respiratory disease using natural language processing and machine learning: algorithm development and evaluation study. JMIR AI 2023 Jun 12;2:e44191. [doi: 10.2196/44191] [Medline: 39105270]
3. Chae A, Yao MS, Sagreiya H, et al. Strategies for implementing machine learning algorithms in the clinical practice of radiology. Radiology 2024 Jan;310(1):e223170. [doi: 10.1148/radiol.223170] [Medline: 38259208]
4. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. N Engl J Med 2023 Mar 30;388(13):1201-1208. [doi: 10.1056/NEJMra2302038] [Medline: 36988595]
5. Kendale S, Bishara A, Burns M, Solomon S, Corriere M, Mathis M. Machine learning for the prediction of procedural case durations developed using a large multicenter database: algorithm development and validation study. JMIR AI 2023 Sep 8;2:e44909. [doi: 10.2196/44909] [Medline: 38875567]
6. Seth P, Hueppchen N, Miller SD, et al. Data science as a core competency in undergraduate medical education in the age of artificial intelligence in health care. JMIR Med Educ 2023 Jul 11;9:e46344. [doi: 10.2196/46344] [Medline: 37432728]
7. Arango-Ibanez JP, Posso-Nuñez JA, Díaz-Solórzano JP, Cruz-Suárez G. Evidence-based learning strategies in medicine using AI. JMIR Med Educ 2024 May 24;10:e54507. [doi: 10.2196/54507] [Medline: 38801706]
8. Shimizu I, Kasai H, Shikino K, et al. Developing medical education curriculum reform strategies to address the impact of generative AI: qualitative study. JMIR Med Educ 2023 Nov 30;9:e53466. [doi: 10.2196/53466] [Medline: 38032695]

9. Furlan R, Gatti M, Mene R, et al. Learning analytics applied to clinical diagnostic reasoning using a natural language processing-based virtual patient simulator: case study. JMIR Med Educ 2022 Mar 3;8(1):e24372. [doi: 10.2196/24372] [Medline: 35238786]
10. Sun L, Yin C, Xu Q, Zhao W. Artificial intelligence for healthcare and medical education: A systematic review. Am J Transl Res 2023;15(7):4820-4828. [Medline: 37560249]
11. Pupic N, Ghaffari-zadeh A, Hu R, et al. An evidence-based approach to artificial intelligence education for medical students: A systematic review. PLOS Digit Health 2023;2(11):e0000255. [doi: 10.1371/journal.pdig.0000255]
12. Civaner MM, Uncu Y, Bulut F, Chalil EG, Tatli A. Artificial intelligence in medical education: A cross-sectional needs assessment. BMC Med Educ 2022 Nov 9;22(1):772. [doi: 10.1186/s12909-022-03852-3] [Medline: 36352431]
13. Gray K, Slavotinek J, Dimaguila GL, Choo D. Artificial intelligence education for the health workforce: expert survey of approaches and needs. JMIR Med Educ 2022 Apr 4;8(2):e35223. [doi: 10.2196/35223] [Medline: 35249885]
14. Chan KS, Zary N. Applications and Challenges of Implementing Artificial Intelligence in Medical Education: Integrative Review. JMIR Med Educ 2019 Jun 15;5(1):e13930. [doi: 10.2196/13930] [Medline: 31199295]
15. Sobel J, Almog R, Celi L, Yablowitz M, Eytan D, Behar J. How to organise a datathon for bridging between data science and healthcare? Insights from the Technion-Rambam machine learning in healthcare datathon event. BMJ Health Care Inform 2023 Sep;30(1):e100736. [doi: 10.1136/bmjhci-2023-100736] [Medline: 37696642]
16. Oyetade K, Zuva T, Harmse A. Educational benefits of hackathon: a systematic literature review. WJET 2022;14(6):1668-1684 [FREE Full text] [doi: 10.18844/wjet.v14i6.7131]
17. Silver JK, Binder DS, Zubcevik N, Zafonte RD. Healthcare hackathons provide educational and innovation opportunities: a case study and best practice recommendations. J Med Syst 2016 Jul;40(7):177. [doi: 10.1007/s10916-016-0532-3] [Medline: 27277278]
18. Barabucci G, Shia V, Chu E, Harack B, Laskowski K, Fu N. Combining multiple large language models improves diagnostic accuracy. NEJM AI 2024 Oct 24;1(11). [doi: 10.1056/AIcs2400502]
19. Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open 2024 Oct 1;7(10):e2440969. [doi: 10.1001/jamanetworkopen.2024.40969] [Medline: 39466245]
20. Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. NPJ Digit Med 2024 Jan 24;7(1):20. [doi: 10.1038/s41746-024-01010-1] [Medline: 38267608]
21. Ong JCL, Chang SYH, William W, et al. Ethical and regulatory challenges of large language models in medicine. Lancet Digit Health 2024 Jun;6(6):e428-e432. [doi: 10.1016/S2589-7500(24)00061-X] [Medline: 38658283]
22. Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large language models propagate race-based medicine. NPJ Digit Med 2024;6(1):195. [doi: 10.1038/s41746-023-00939-z]
23. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med 2024 Sep;30(9):2613-2622. [doi: 10.1038/s41591-024-03097-1] [Medline: 38965432]
24. Teisberg E, Wallace S, O'Hara S. Defining and implementing value-based health care: a strategic framework. Acad Med 2020 May;95(5):682-685. [doi: 10.1097/ACM.0000000000003122] [Medline: 31833857]
25. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 2023 Jan 3;10(1):1. [doi: 10.1038/s41597-022-01899-x] [Medline: 36596836]
26. Ben Abacha A, Yim WW, Fan Y, Lin T. An empirical study of clinical note generation from doctor-patient encounters. Presented at: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics; 2023; Dubrovnik, Croatia p. 2291-2302. [doi: 10.18653/v1/2023.eacl-main.168]
27. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Appl Sci (Basel) 2021;11(14):6421. [doi: 10.3390/app11146421]
28. Ji S, Li X, Huang Z, Cambria E. Suicidal ideation and mental disorder detection with attentive relation networks. Neural Comput & Applic 2022 Jul;34(13):10309-10319. [doi: 10.1007/s00521-021-06208-y] [Medline: 33746365]
29. MDplus. MDplus DS & AI - datathon. URL: http://ai.mdplus.community/datathon/2023 [accessed 2024-01-18]
30. Likert R. A technique for the measurement of attitudes. Arch Psychol 1932;140:1-55.
31. Hendricks-Sturrup R, Simmons M, Anders S, et al. Developing ethics and equity principles, terms, and engagement tools to advance health equity and researcher diversity in AI and machine learning: modified Delphi approach. JMIR AI 2023 Dec 6;2:e52888. [doi: 10.2196/52888] [Medline: 38875540]
32. Aggarwal R, Farag S, Martin G, Ashrafian H, Darzi A. Patient perceptions on data sharing and applying artificial intelligence to health care data: cross-sectional survey. J Med Internet Res 2021 Aug 26;23(8):e26162. [doi: 10.2196/26162] [Medline: 34236994]
33. Hochheiser H, Klug J, Mathie T, et al. Raising awareness of potential biases in medical machine learning: Experience from a Datathon. medRxiv 2024 Nov 2:2024.10.21.24315543. [doi: 10.1101/2024.10.21.24315543] [Medline: 39502657]
34. Elbers P, Thoral P, Bos LDJ, Greco M, Wendel-Garcia PD, Ercole A. The ESICM datathon and the ESICM and ICMx data science strategy. Intensive Care Med Exp 2024 Mar 12;12(1):29. [doi: 10.1186/s40635-024-00615-w] [Medline: 38472595]
35. Health Tech Datathon. NYU grossman school of medicine. URL: https://med.nyu.edu/our-community/health-technology/events/health-tech-datathon [accessed 2025-04-07]

36. Datathon. Society of Critical Care Medicine (SCCM): The Intensive Care Professionals. URL: https://sccm.org/research/discovery-research-network/datascience/datathon [accessed 2024-12-01]

Abbreviations

LLM: large language model
MIMIC: Medical Information Mart for Intensive Care
ML: machine learning
VBC: value-based care

Edited by B Lesselroth; submitted 24.06.24; peer-reviewed by M Knopp, PM Naliyatthaliyazchayil, S Purkayastha; revised version received 02.12.24; accepted 25.02.25; published 16.04.25.

Please cite as:
Yao MS, Huang L, Leventhal E, Sun C, Stephen SJ, Liou L
Leveraging Datathons to Teach AI in Undergraduate Medical Education: Case Study
JMIR Med Educ 2025;11:e63602
URL: https://mededu.jmir.org/2025/1/e63602
doi:10.2196/63602

Effect of Immersive Virtual Reality Teamwork Training on Safety Behaviors During Surgical Cases: Nonrandomized Intervention Versus Controlled Pilot Study

Lukasz Mazur[1,2], PhD; Logan Butler[1], MD; Cody Mitchell[1], BS; Shaian Lashani[1], BS; Shawna Buchanan[1], BSN; Christi Fenison[1], MA; Karthik Adapa[1], MD, PhD; Xianming Tan[3], PhD; Selina An[4], MD; Jin Ra[4], MD

[1]Department of Radiation Oncology, Division of Healthcare Engineering, School of Medicine, University of North Carolina, Campus Box 7512, Chapel Hill, NC, United States

[2]School of Information and Library Science, University of North Carolina, Chapel Hil, NC, United States

[3]Gillings School of Global Public Health, University of North Carolina, Biostatistics, Chapel Hill, NC, United States

[4]Department of Surgery, School of Medicine, University of North Carolina, Chapel Hill, NC, United States

Corresponding Author:
Lukasz Mazur, PhD
Department of Radiation Oncology, Division of Healthcare Engineering, School of Medicine, University of North Carolina, Campus Box 7512, Chapel Hill, NC, United States

Abstract

Background: Approximately 4000 preventable surgical errors occur per year in the US operating rooms, many due to suboptimal teamwork and safety behaviors. Such errors can result in temporary or permanent harm to patients, including physical injury, emotional distress, or even death, and can also adversely affect care providers, often referred to as the "second victim."

Objective: Given the persistence of adverse events in the operating rooms, the objective of this study was to quantify the effect of an innovative and immersive virtual reality (VR)–based educational intervention on (1) safety behaviors of surgeons in the operating rooms and (2) sense-making regarding the overall training experience.

Methods: This mixed methods pre- versus postintervention pilot study was conducted in a large academic medical center with 55 operating rooms. Safety behaviors were observed and quantified using validated Teamwork Evaluation of Non-Technical Skills instrument during surgical cases at baseline (101 observations; 83 surgeons) and postimmersive VR based intervention (postintervention: 24 observations within each group; intervention group [with VR training; 10 surgeons] and control [no VR training; 10 surgeons]). VR intervention included a 45-minute immersive VR-based training incorporating a pre- and postdebriefing based on Team Strategies and Tools to Enhance Performance and Patient Safety (TeamSTEPPS) principles to improve safety behaviors. A 2-tailed, 2-sample *t*-test with adjustments for multiplicity of the tests was used to test for significance in observable safety behaviors between the groupings. The debriefing data underwent analysis through the phenomenological analysis method to gain insights into how participants interpreted the training.

Results: Preintervention, all safety behaviors averaged slightly above "acceptable" scores, with an overall average of 2.2 (range 2 - 2.3; 0 - 3 scale). The 10 surgeons that underwent our intervention showed statistically significant (P<.05) improvements in 90% (18/20) of safety behaviors when compared to the 10 surgeons that did not receive the intervention (overall average 2.5, range 2.3 - 2.7 vs overall average 2.1, range 1.9 - 2.2). Our qualitative analysis based on 492 quotes from participants suggests that the observed behavioral changes are a result of an immersive experience and sense-making of key TeamSTEPPS training concepts.

Conclusions: VR-based immersive training intervention focused on TeamSTEPPS principles seems effective in improving safety behaviors in the operating rooms as quantified via observations using the Teamwork Evaluation of Non-Technical Skills instrument. Further research with larger, more diverse sample sizes is needed to confirm the generalizability of these findings.

International Registered Report Identifier (IRRID): RR2-10.2196/40445.

(JMIR Med Educ 2025;11:e66186) doi:10.2196/66186

KEYWORDS

Teamwork Evaluation of Non-Technical Skills; TENTS; Team Strategies and Tools to Enhance Performance and Patient Safety; TeamSTEPPS; immersive virtual reality; virtual reality; VR; safety behavior; surgical error; operating room; OR; training intervention; training; pilot study; nontechnical skills; surgery; surgical; patient safety; medical training; medical education

Introduction

High-quality health care necessitates ongoing efforts to reduce the occurrence of medical errors [1]. Surgical patients face heightened risks of adverse outcomes related to errors due to the invasive nature of surgical procedures [2]. It is estimated that more than 4000 preventable surgical errors occur annually on a national scale [1,2]. Such errors can result in temporary or permanent harm to patients, including physical injury, emotional distress, or even death, and can also adversely affect care providers, often referred to as the "second victim." A notable example is the unintended retention of foreign objects, which is believed to happen at least once in every 5500 surgeries [3]. This can lead to the need for reoperation, extended hospital stays, and complications such as sepsis. Furthermore, the average additional cost associated with each incident of unintended retention is estimated to exceed US $200,000 [4]. Common underlying causes of surgical errors identified by the Joint Commission include the lack of established policies and procedures, issues related to hierarchy and intimidation, ineffective communication among care team, and the failure of staff to relay pertinent patient information [5]. Additionally, factors such as excessive workload, time constraints, and burnout are linked to increased error rates [5]. Addressing these root causes has proven challenging, as complex health care delivery systems tend to evolve over time, leading to the emergence of new failure sources and pathways. Teamwork skills are often essential for preventing such errors that could lead to patient harm [6]. [7], To address these issues, the Team Strategies and Tools to Enhance Performance and Patient Safety (TeamSTEPPS) framework was specifically designed as a resource to help health care providers improve patient safety behaviors through effective communication, leadership, situation monitoring, and mutual support [8]. [9,10] By using the TeamSTEPPS framework, it is possible to assess the use and quality of patient safety education and behaviors among operating room staff and establish a baseline for improvement.

Virtual reality (VR) is a digital technology that enables a virtual manifestation of the real world [11]. VR provides a more captivating experience compared to viewing a conventional video, as it fully envelopes the viewer within the narrative [12]. In VR, the audience becomes an integral part of the story rather than merely an onlooker. The advantages of immersive VR primarily include: (1) viewers are placed within a 360-degree environment, where each movement of the head unveils new dimensions of the scene. Conversely, traditional video confines viewers to a fixed perspective on a flat screen. (2) Participants in VR actively engage with the narrative, rather than remaining passive spectators. (3) A profound experiential learning opportunity, allowing participants to engage with the contextual realities of a surgical error in the surgical environment.

While live mock simulations with standardized patients are used for certain health care scenarios, the logistical demands of accurately recreating a surgical environment are substantial. These include the coordination of a full surgical team, time spent in an operating room, and the creation of a realistic setting, along with patient representation and various special effects that must align with the error, necessitating cleanup and reset after each simulation. The extensive resources required make this approach neither cost-effective nor scalable for providing multiple realistic experiences for individual learners. Furthermore, a live mock simulation often fails to address the comprehensive needs of learners, as it does not provide insights into the broader contributing factors and repercussions that extend beyond the confines of the operating room. While this disconnect may not significantly affect certain types of learning, such as factual recall, the deeper cognitive processing of theoretical scenarios, particularly in the context of complex situations, could be enhanced through more experiential learning methods such as VR [13]. Experiential learning fosters a personalized and cognitive interaction with educational content, highlighting the relationship between learning and its practical application in the real world. If health care workers are to engage in sense-making regarding the intricate realities associated with adverse events, VR may provide distinct advantages over 2D content and live mock simulations by offering an immersive perspective of events as they would unfold in real life.

In recent years, VR has seen increasing use for safety training across industries [14-17]. VR safety training has great potential, as it allows trainees to experience complex, challenging situations that are difficult to replicate in the real world due to ethical, cost, and time constraints. However, there is a lack of research examining the efficacy of immersive VR-based interventions focused on safety education and behaviors as proposed by the TeamSTEPPS framework. Thus, given the persistence of adverse events in the operating rooms, the objective of this pilot study was to quantify the effect of an innovative and immersive VR-based educational intervention focused on TeamSTEPPS on (1) safety behaviors of surgeons in the operating rooms and (2) sense-making regarding the overall training experience and contributing factors associated with the surgical error.

Methods

Ethical Considerations

This study obtained ethics approval via the Institutional Review Board of the University of North Carolina at Chapel Hill (22 - 1150). Participants were provided with and signed an informed consent form before engaging with activities related to this study. Participants in the intervention group were provided with a small token of appreciation ($25) for compleating the VR intervention process. Patients or the public were not involved in the design, conduct, reporting, or dissemination plans of our research. All results are reported in an aggregate manner to ensure the privacy and confidentiality of participants. The protocol for the full mixed methods pre-versus postintervention study design was published in *JMIR Research Protocols* [18].

Recruitment

A scripted email with a flyer was sent through a listserv to inform prospective participants about this study. Our research team was on standby to answer any questions from prospective participants, and the principal investigators' contact information was available on the flyer. First, for the baseline measurement, a volunteer sample of 83 surgeons (35 attendings, 41 residents,

and 7 fellows; 36 female and 47 male) was enrolled and participated. Surgeons volunteered by providing verbal consent to be observed and scored for safety behaviors before each surgical case. One surgeon refused to participate (no reasons stated). For the pilot study, 10 out of 83 (12%) surgeons (7 attendings and 3 residents; 5 female and 5 male) volunteered to undergo the immersive VR-based educational intervention and to undergo the follow-up observations and scoring. An additional 10 of 83 (12%) surgeons (10 attendings and 0 residents; 3 female and 7 male) volunteered to undergo follow-up observation and scoring without being exposed to the intervention.

Study Design

This mixed methods pre- versus postintervention study with a baseline and intervention or control groups was conducted at a large academic medical center with 55 operating rooms. For the baseline measurement, safety behaviors were quantified while observing 101 surgical cases from October 31, 2022, to February 21, 2023, with 83 surgeons. For the pilot study, 24 observations within each group were conducted from April 17, 2023, to November 2, 2023, with 10 surgeons in the intervention and control group, respectively. Data was collected using the Teamwork Evaluation of Non-Technical Skills (TENTS) instrument. We opted for the TENTS instrument instead of alternatives such as the TeamSTEPPS Team Performance Observation Tool due to our requirement for a more focused evaluation of nontechnical teamwork aspects. TENTS is specifically designed to assess nontechnical dimensions of teamwork, encompassing team member interactions, task delegation, and information management, which makes it particularly relevant in situations where technical skills are not the primary concern. Additionally, TENTS provides a more detailed observational framework, allowing for a comprehensive evaluation of specific teamwork behaviors, in contrast to the broader categories covered by the TeamSTEPPS Team Performance Observation Tool. In our study, we used TENTS items 1A to 4D to assess individual behaviors of surgeons in independent settings (except for the 3 residents who were part of the intervention group), while items 5 and 6 were used to evaluate the overall team functioning and leadership. To train the observers for scoring items 5 and 6, we assembled a team of 17 medical student volunteers who were tasked with conducting observations using the TENTS instrument. They used a simplified scoring scale ranging from 0 to 3, where 0 indicates expected behavior not observed, 1 signifies observed behavior that was poorly executed or counterproductive, 2 denotes acceptable performance, and 3 represents excellent performance. The students received 1.5 hours of instruction on TeamSTEPPS patient safety behaviors and were trained to apply the validated TENTS instrument consistently. Following this, they participated in facility tours and were guided through the observation protocol, which included a practical demonstration of the TENTS scoring process. The students were assigned to surgical cases by aligning their weekly schedules with the relevant cases, while being blinded to the treatment status of the surgeons during postintervention observations. TENTS scores were recorded during real-time observations of surgical procedures using paper forms, and students were required to provide written justifications for behaviors rated as 1 (poor) or 3 (excellent) to offer context for their evaluations.

Interventions or Exposures

We used state-of-the-art filming equipment to capture a 360° view of the event and the perspective of those involved in the error event and contributing events. We built the scripts for the scenes, recruited actors (attendings, residents, students, and administrators) with lived experiences in health care to help with the filming, identified filming locations, rehearsed all the scenes, and filmed our scenes. This training was delivered to the participants using a VR head-mounted display to ensure an immersive environment (see Multimedia Appendix 1 for a 1-minute summary clip of the training). Specifically, we used the Pico Neo 3 Pro Eye headset (PICO Technology Co Ltd) with 6DoF VR hardware or software to administer the training.

Participants were exposed to a 45-minute immersive VR-based training based on TeamSTEPPS principles to improve safety behaviors. Overall, our overarching training story is focused on a human error that occurs at the operating table but is caused by many factors, as explained by the James Reason Swiss Cheese Model [19] and Human Factors Analysis and Classification System (HFACS) [20]. Training involved a standardized pre- and postbriefing aimed at comprehending and identifying potential behavioral enhancements within the training narrative, grounded in TeamSTEPPS principles. Specifically, we sought to collect participants' perspectives on the overall training experience and their interpretation of the TENTS and HFACS-related factors that contributed to the patient safety incident illustrated in the VR training. The debriefing data underwent analysis through the phenomenological analysis method [21] to gain insights into how participants interpreted the training and the patient safety incidents they encountered, thereby refining their understanding, behavior, and commitment to TeamSTEPPS practices. This qualitative analysis was performed by 2 researchers, a senior surgical attending and a fourth-year surgical resident, using the data frame theory of sense-making [22]. The primary objective of this qualitative effort was to enhance our understanding of the main outcomes and measures outlined below.

Main Outcomes and Measures

The primary outcome measures were the pooled average and 95% CI of observed TeamSTEPPS related behaviors quantified using the validated TENTS instrument (including 20 types of safety behaviors across 4 domains [communication, leadership, situation monitoring, and mutual support] scored from 0=expected but not observed, 1=observed but poorly performed or counterproductive, 2=observed and acceptable, and 3=observed and excellent).

Statistical Analysis

A 2-tailed, ANOVA was used to test for significance between the groupings. For given comparators (eg, [1] vs [2] or [2] vs [3]), we used a false discovery rate (FDR) control (the Benjamini-Hochberg procedure) to adjust the resulting *P* values (from the 2-sample *t*-tests) to account for the multiplicity of the tests. We claimed a result to be statistically significant if the adjusted *P* value is less than .05. All analyses were performed

with R statistical software. Data analysis was conducted from August 10 to December 1, 2023.

Results

Preintervention, all safety behaviors averaged slightly above 2 (“acceptable”), with an overall average of 2.2 (range 2 - 2.3). There was no significant difference between the intervention and the control group at the preintervention stage. The results indicated that the 10 surgeons that underwent our intervention showed statistically significant (P<.05) improvements in 90% (18/20) of safety behaviors when compared to the 10 surgeons that did not receive the intervention (overall average 2.5, range 2.3 - 2.7 vs overall average 2.1, range 1.9 - 2.2; Table 1). Our qualitative analysis revealed 492 individual quotes. The results suggest that the observed behavioral changes are a result of a sense-making emerging from 5 specific themes as discussed below.

Table . Summary of results (baseline vs intervention).

	TENTS[a] behavior	Baseline, mean (SD)	Intervention, mean (SD)	*P* value
1a	Communicates and receives information appropriately	2.227 (0.53)	2.708 (0.455)	.001
1b	Comfortable speaking up and asking questions	2.237 (0.452)	2.625 (0.484)	.001
1c	Responses to feedback between team members	2.135 (0.504)	2.458 (0.498)	.02
1d	Communicates and receives information to or from patient	2.097 (0.433)	2.261 (0.439)	.14
1e	Uses language in urgent situations appropriately	2.133 (0.505)	2.444 (0.497)	.04
1f	Uses teamwork tools	2.26 (0.464)	2.542 (0.498)	.03
1g	Learns together, focuses on improvement following a problem	2.222 (0.451)	2.55 (0.497)	.02
2a	Leaders effectively manage team during their roles	2.274 (0.493)	2.708 (0.455)	.001
2b	Verbalizes plan: intentions, recommendations, or timeframes	2.247 (0.501)	2.583 (0.571)	.02
2c	Delegates tasks appropriately	2.104 (0.369)	2.458 (0.498)	.007
2d	Instructs as appropriate to the situation	2.281 (0.475)	2.542 (0.498)	.03
3a	Pays attention to surroundings or environment	2.11 (0.526)	2.5 (0.5)	.005
3b	Aware of each other, contributions, strengths, and weaknesses	2.208 (0.433)	2.5 (0.5)	.02
3c	Verbalizes adjustments in plan as changes occur	2.236 (0.428)	2.524 (0.499)	.03
4a	Willingness to ask for help or additional resources	2.26 (0.464)	2.708 (0.455)	.001
4b	Willingness to support others across different roles	2.253 (0.437)	2.417 (0.493)	.15
4c	Accomplishes and prioritizes tasks appropriately	2.103 (0.338)	2.333 (0.471)	.04
4d	Employs conflict resolution	2.038 (0.341)	2.316 (0.465)	0.03
5	Rating of how well the team functioned as a whole	2.258 (0.44)	2.708 (0.455)	0.001
6	Rate how well leaders functioned and how the team responded	2.247 (0.434)	2.708 (0.455)	0.001

[a]TENTS: Teamwork Evaluation of Non-Technical Skills.

Discussion

Principal Results

Overview

The results show that participants exposed to our intervention displayed improved levels of safety behaviors, as quantified by the TENTS instrument, in 90% (18/20) of the safety behaviors measured. Specifically, quantitative data suggests that surgical teams were more effective in developing and maintaining a dynamic awareness of the situation in the operating room. This was achieved by assembling and understanding data from various sources (eg, patient, team, time, displays, and equipment), and using strong communication and leadership skills to think ahead and provide clear direction while being

considerate of individual team members' needs. Importantly, these improvements were observed not only at the individual level, as shown by the TENTS instrument, but also in the overall team functioning and leadership, as indicated by aggregate measures of "how well the team functioned as a whole," "how well the leaders functioned," and "how the team responded."

The 2 behaviors that did not reach statistical significance, despite trending positively, were 1d (communicating and receiving information with patients), and 4b (willingness to support others across different roles) (Table 1). For the communication behavior, very few such interactions were observed during the surgical cases, limiting our assessment of this behavior. Regarding the willingness to support others across roles, this was the highest-scoring behavior at baseline, suggesting it may have been challenging to improve further. This implies that in the dynamic and complex surgical environment, team members may struggle to step outside their designated roles to provide support to one another, as they focus intently on delivering excellent patient care and ensuring safety within their specific responsibilities.

Our qualitative analysis suggests that these behavioral improvements materialized from the enhanced understanding of skills needed in the operating room by reinforcing critical behaviors related to the sense-making of themes presented below.

Need for Effective Teamwork and Communication

The VR training modeled effective versus ineffective team communication, demonstrating how dismissing or ignoring concerns can lead to errors and decrease team trust. Participants may have practiced assertive communication, such as how a resident can escalate a concern when an anesthesiologist is distracted or how surgical technology can improve instrument handling. VR also emphasized calling out errors in a constructive way, rather than scolding or ignoring them.

The most consistent topics addressed by the subjects during the poststudy interview were teamwork and communication, and how these traits are essential for a well-functioning operating room. Participants noted how the VR training vividly illustrated both the consequences of poor communication and the benefits of a cohesive team. The training emphasized that errors often arise not solely from technical mistakes but from an inability to effectively relay and escalate concerns. The lack of teamwork and communication in the ineffective VR scenario was particularly alarming to participants. One individual highlighted the disconnect among team members: "There was clearly not a deep relationship between the surgeon, the resident, the scrub, the circulator, and the anesthesiologist; they just seemed completely disconnected [and] in their own worlds." Another key issue was the absence of assertive communication when concerns arose. One participant recalled: "The resident did pick up on the change in the heart rate tone, questioned the anesthesiologist about it, who blew her off." The absence of closed-loop communication was another prevalent theme. Participants remarked on how essential feedback loops were often missing in the ineffective scenarios, which contributed to preventable errors: "No closed-loop communication. Just a lot of people suggesting things but the other person either wasn't listening or just kind of ignored it and moved along. Even when it was something that could have prevented safety issues." Overall, the poststudy interviews reinforced that teamwork and communication are foundational to operating room effectiveness. The VR training provided a powerful demonstration of how dismissing concerns, failing to engage in open dialogue, and neglecting structured communication can significantly hinder patient safety. By recognizing these pitfalls and emphasizing assertive, closed-loop communication, participants reported a newfound appreciation for fostering a more cohesive and communicative operating room team.

Emphasis on Empathetic Workplace Culture and Psychological Safety

VR heightened participants' awareness of how fatigue, dismissive communication, and team dynamics impact psychological safety in the operating room. By simulating real-world scenarios, it demonstrated how exhaustion and distractions compromise decision-making, how seemingly minor quips or sarcasm can erode team cohesion, and how validating trainee concerns fosters a culture of safety and respect. This reinforced the importance of clear, professional communication and proactive leadership in creating a supportive and effective surgical intervention.

The VR training underscored how the psychological safe environment of the operating room directly impacts team effectiveness, patient safety, and overall workplace culture. Participants became more empathetic and attuned to how fatigue, dismissive behavior, and team dynamics can create a toxic versus supportive surgical setting. By immersing participants in scenarios where exhaustion led to oversight, sarcasm eroded trust, and concerns were either validated or dismissed, the training highlighted the importance of maintaining a psychologically safe and healthy work environment. One of the most striking realizations was how fatigue and distraction—often seen as inevitable in surgical practice—could significantly impair decision-making and communication. As one participant noted, "Exhaustion, lack of sleep, and lack of focus on the attending's part... distraction from the anesthesiologist... these are indicators that their wellness score is probably not stellar." Additionally, the training revealed how subtle, seemingly harmless behaviors can undermine psychological safety. Participants observed how sarcastic remarks or casual quips, even when meant humorously, created an environment where individuals felt less comfortable speaking up. "There were a lot of quips...I don't think they contributed too much. And they can be detrimental." Another crucial takeaway was the need to legitimize concerns when raised, rather than allowing the pressures of a high caseload to override safety. One participant reflected, "I think we can always do better legitimizing people raising concerns...the pressure to feel rushed and move quickly ... can obviously be counterproductive." Perhaps most telling was the recognition that team dynamics set the tone for the entire operating room. When interpersonal relationships are strained, it affects everyone, from the attending surgeon to the anesthesia technician. One participant encapsulated this sentiment: "Because I think we've all been in rooms where [if] the staff doesn't get along …it makes it miserable for everyone." The VR training effectively demonstrated how the psychological

environment of the operating room shapes both patient safety and team dynamics, making participants more aware of the subtle but powerful ways fatigue, communication styles, and validation of concerns impact surgical outcomes. By immersing participants in realistic scenarios, the training seemed to inspire an appreciation for operating rooms, where they do have a psychologically healthy environment.

Need for Leadership With Personal Responsibility and Accountability

VR simulation highlighted the ripple effect of individual actions—how fatigue, inattention, or lack of engagement from a team member impacts the whole operating room. It also reinforced that leaders (attendings, residents, or nurses) set the tone for safety culture, whether by ensuring protocols are followed or by fostering an environment where concerns can be raised. The VR training may have helped participants recognize their personal accountability in maintaining operating room safety and identifying behaviors that contribute to or undermine team effectiveness.

The training reinforced the critical role that leadership plays in shaping operating room culture, particularly in fostering accountability and ensuring adherence to safety protocols. Participants observed how leadership—or the lack thereof—had a cascading effect on communication, decision-making, and overall team cohesion. By placing participants in scenarios where leadership failures led to errors or unsafe practices, the training emphasized the responsibility of every operating room member to contribute to a culture of accountability. A key takeaway was the attending surgeon's responsibility in setting the tone for the operating room environment. As 1 participant remarked, "The attending surgeon sort of sets the tone in many ways, and by not looking into concerns...[and] cutting corners in order to be able to increase throughput...—that's just not good leadership." Beyond the attending, participants recognized how personal accountability extends to every team member. The training exposed moments where concerns were voiced quietly but never formally escalated, leading to missed opportunities to address potential safety issues. One participant reflected, "...everyone was making little comments, but no one was, again, like really saying them. They were all kind of talking to themselves...." The VR also made participants more aware of how the pressure to move quickly can lead to cutting corners, potentially compromising patient outcomes. One participant acknowledged, "We all get caught up in rush, rush, rush…and I've done that before, where you look through the labs, like, it's probably fine. It usually is. But what if it's not?" Ultimately, the training reinforced the idea that each case demands responsibility and accountability. The VR experience demonstrated how a disengaged or inattentive leader could undermine these traits by dismissing concerns or prioritizing efficiency over protocol.

Need for Stability

The VR training elicited participants' desires and personal experiences to have surgical teams that are consistent and connected. Understandably, those who are working in a high-stakes, high-stress environment would want to mitigate other factors that could lead to negative patient outcomes or contribute to workplace burnout and distress. The VR training showed participants what can happen with a more discordant or unstable working environment, which can mimic reality, and many participants were quick to point out how deleterious that can be to the dynamics of an operating room.

The training highlighted the critical role that stability plays in fostering an effective and supportive surgical environment. Participants emphasized the need for consistency in team composition, resource availability, and leadership presence—elements that are often taken for granted but can significantly impact patient safety and staff well-being. In the high-stakes, high-stress environment of the operating room, an unstable or discordant team structure can lead to inefficiencies, miscommunication, and increased burnout, all of which were vividly demonstrated in the VR scenarios. Many participants noted how instability, whether due to staffing shortages, systemic pressures, or administrative constraints, can disrupt operating room dynamics. One participant reflected on the broader hospital structure, stating, "It didn't seem like there was support from a majority of people...No one person can do it all." A major recurring theme was the strain placed on attendings who were expected to be in multiple places at once, highlighting the impact of systemic pressures on operating room stability. As 1 participant observed, "I mean, there were clearly systemic pressures for the attending to be in multiple places at one time, I would say primarily pressure for throughput, short staffing." This speaks to the broader challenges of balancing efficiency with quality care. This particularly resonated with our interviewees as it is something that a vast majority of health care workers can relate to. By experiencing the challenges of an unstable operating room environment with the constant, relatable pressure to do more, participants gained a deeper appreciation for the structures and policies needed to foster a more reliable and effective surgical setting.

Emphasis on Outcome and Attention to Detail

This VR module showed participants what can happen when means do not at all justify ends. In a system where outcomes, whether it be several cases completed or the speed of the operation, are prioritized over how those end points are achieved, it can lead to consequential errors. While operating rooms need to maintain a high level of efficiency, the consequences of doing so can come at the expense of the patient. It can be challenging for surgeons to balance their commitment to good patient care with intense pressures for increased efficiency and decreased case turnaround time.

The VR training underscored the risks of prioritizing efficiency and case volume over patient safety and procedural integrity. Participants recognized how a results-driven culture, where speed and throughput are emphasized over safe surgical practice, can lead to critical errors and compromise team effectiveness. While efficiency is a necessary component of modern surgical workflows, the VR scenarios illustrated the consequences of allowing productivity pressures to override fundamental principles of patient care. One participant stated, "many other factors—fatigue [and]the pressure from hospital administration for revenue and productivity," when reflecting on the systemic pressures driving an outcome-based mindset. The VR module

demonstrated how this tunnel vision manifests at different levels of the operating room team. One participant observed, "The resident's main mission was 'I gotta close so I can go to the other room.' The surgeon was like 'I gotta get these 12 cases done because that's what administration says.' The circulator was like 'We gotta get these cases done so we can all go home.'" Participants also reflected on the human cost of this approach, not just for the patients but for the surgical teams themselves. One particularly striking insight was, " ...it's fine to be efficient. It's not fine to be in a hurry. And I think that it's really, really important for all of us to think about...." The training allowed participants to experience the tension between efficiency and quality care, a concept that is ubiquitous among health care settings. By highlighting the potential dangers of an outcome-driven approach, the VR module reinforced the need for deliberate, methodical teamwork, where safety is never sacrificed in the name of speed or the bottom line.

The VR training also reinforced the critical importance of attention to detail in the operating room, highlighting how even small lapses in accuracy can lead to significant consequences. By immersing participants in real-world scenarios, the module demonstrated how factors such as fatigue, communication breakdowns, and time pressures can contribute to oversights. It emphasized that attention to detail is not just an individual responsibility but a collective effort, where every team member plays a role in maintaining surgical precision.

The training reinforced the critical role of accuracy in the operating room, emphasizing how seemingly minor lapses in attention to detail can have significant consequences for patient outcomes. Participants recognized that surgical safety is not solely dependent on technical skill but also on the thoroughness of preoperative preparation, intraoperative vigilance, and collective situational awareness. Several interviewees pointed out the dangers of neglecting critical details in the operating room. For example, many participants highlighted that the failure to properly assess a patient's anticoagulation status before surgery leads to cancellation of cases and delays in care. Another key theme was the lack of a shared mental model among the surgical team, leading to fragmented awareness of the patient's condition. As one participant described, "I'm not sure that there was a complete understanding of the entire situation... that global shared mental model of where everybody was, the implications of the decisions, and how things were happening was kind of missing." The scenarios also illustrated that once the ball was set in motion, no one had the care or will to change it. One participant noted, "It seemed like both the surgeon and the resident didn't have a good sense of who the patient was, what the case was, if they were on anticoagulation. And then even when the resident realized and brought it up to the attending, the attending was like, 'It's fine, it's too late, moving on.'" Ultimately, the VR experience reinforced the necessity of meticulous preparation, comprehensive team awareness, and an environment where concerns are acknowledged rather than dismissed. It demonstrated that attention to detail is not merely an individual responsibility but a collective effort, where every member of the surgical team plays a vital role in ensuring safe and effective patient care.

Comparison With Prior Work

This pilot study is the first to quantify the effects of an immersive VR-based educational intervention focused on improving TeamSTEPPS-related behaviors among surgeons in the operating room. Overall, the findings align with previous non–VR-based research highlighting the importance of teamwork training for enhancing soft skills critical to patient safety [23-27]. Our findings also align with the conclusions of the work by Abelson et al [11] that supports the motion that VR is a feasible solution for team-based training, and Gasteiger et al [12] that postulate that technical and nontechnical skills training programs using VR for health care staff may trigger perceptions of realism and deep immersion and enable easier visualization, interactivity, enhanced skills, and repeated practice in a safe environment, which in turn may improve skills and increase learning, knowledge, and learner satisfaction. Notably, prior work using VR to teach TeamSTEPPS for cesarean section surgery showed that the VR-based content improved teamwork competencies in interprofessional surgical teams [28]. By addressing the need for teamwork training, while using the TeamSTEPPS framework, and incorporating innovative educational technologies such as VR, this study demonstrated how collaboration among surgical team members can be enhanced [28]. Finally, in a randomized trial, Liaw et al [13] showed that learning outcomes did not show an inferiority of team training using VR when compared with live simulations, which supports the potential use of VR to substitute conventional simulations for communication team training.

However, many of these initiatives concentrate on skills pertinent to the immediate context of errors, such as communication and teamwork in the operating room, as well as technical competencies, while often neglecting training related to systemic flaws, the culture of patient safety, and the unreliable thought processes and behaviors that can lead to mistakes or hinder their prevention. We propose that trainees' comprehension of the factors leading to patient safety incidents as highlighted by the HFACS framework, their sense-making of safety behaviors as outlined by the TENTS tool, and their understanding of TeamSTEPPS principles in the context of the surgical error can lead to improvements in patient safety.

Limitations

While the findings of this study offer valuable insights, there are several important limitations to consider. First, the results are based on a single experiment with a relatively small sample size from 1 academic medical center. To address the small sample size, we used a 2-tailed, ANOVA for significance between the groupings using the FDR to adjust the resulting *P* values (from the 2-sample *t*-tests) to account for the multiplicity of the tests. For a 2-tailed, ANOVA for significance between the groupings, without the FDR control, the analysis would require approximately 68 participants to obtain a medium and to large effect size, the power level of 0.8, and an α of .05. Larger-scale studies could also take into account various confounding factors, such as training levels, gender, and race. Additionally, the possibility that more motivated individuals enrolled in the intervention group may have skewed the results. Second, the participants' awareness of being observed may have

influenced their performance, potentially biasing the results toward better patient safety practices. To mitigate this effect, all participants were allowed to withdraw from the study at any time and were assured that their individual results would remain confidential. Third, the TENTS instrument and scoring could be imperfect. We address this by conducting robust training and practice with the TENTS tool and by blinding students to the treatment status of the surgeons during postintervention observations. Despite the limitations associated with this approach, the involvement of medical student volunteers was essential for executing this extensive observational study without incurring financial costs. Future iterations of this research could leverage such a program, benefiting both the institution through enhanced understanding of behaviors in operating rooms and the students through valuable operating room exposure and hospital experience. In summary, while the findings offer valuable insights, the limitations of this single-site study with a small sample size and potential participant bias must be considered. Larger, more robust studies will be needed to validate and expand upon these preliminary results. There is also a need for longitudinal studies to assess effects over time.

Conclusions

A VR-based immersive training program focused on TeamSTEPPS principles appears effective at improving safety behaviors, as measured by the TENTS tool. Our qualitative analysis based on 492 quotes from participants, suggest that the observed behavioral changes are a result of an immersive experience and sense-making of key training concepts where participants could see the consequences of suboptimal teamwork (eg, poor leadership and communication, lack of attention to detail, failure to take responsibility, low psychological safety to speak up, etc). Given the persistent patient safety issues in operating rooms nationwide, such innovative and immersive patient safety education programs could provide a scalable intervention to help reduce patient harm in the long run. However, further research with larger, more diverse sample sizes is needed to confirm the generalizability of these findings.

Acknowledgments

The authors want to acknowledge the contribution of the following medical student volunteers in collecting the data used for this study: Kwadwo Ansah, Cameron Kurz, Julian Mobley, SL, Ricardo Crespo, Mary Kaufman, Kainat Aslam, Dakota Perez, Grace Fuller, Grace Tabor, Lily Bender, Alexander Requarth, Hannah Black, Brenderia Cameron, Annie Bright, Zhanè Washington, Sierra Parkinson, Collin Shick, and Mandi. We also want to send special thanks to the following medical students for their contribution to the qualitative data analysis: Anu Chaparala, Nathan Adams, and SL.

Conflicts of Interest

LM is an advisor and equity holder in Communify.us LLC and the founder of MaiaZura LLC. SB, JR, and CF are equity holders and co-owners of MaiaZura LLC.

Multimedia Appendix 1
One-minute overview clip of our VR training used in this study. VR: virtual reality.
[MP4 File, 8549 KB - mededu_v11i1e66186_app1.mp4]

References

1. Rodziewicz TL, Houseman B, Vaqar S, Hipskind JE. Medical error reduction and prevention. In: StatPearls: StatPearls Publishing; 2024. URL: https://www.ncbi.nlm.nih.gov/books/NBK499956/ [accessed 2024-02-12]
2. Panagioti M, Khan K, Keers RN, et al. Prevalence, severity, and nature of preventable patient harm across medical care settings: systematic review and meta-analysis. BMJ 2019 Jul 17;366:l4185. [doi: 10.1136/bmj.l4185] [Medline: 31315828]
3. Cima RR, Kollengode A, Garnatz J, Storsveen AS, Weisbrod CA, Deschamps C. Incidence and characteristics of potential and actual retained foreign object events in surgical patients. J Am Coll Surg 2008 Jul;207(1):80-87. [doi: 10.1016/j.jamcollsurg.2007.12.047] [Medline: 18589366]
4. Williams TL, Tung DK, Steelman VM, Chang PK, Szekendi MK. Retained surgical sponges: findings from incident reports and a cost-benefit analysis of radiofrequency technology. J Am Coll Surg 2014 Sep;219(3):354-364. [doi: 10.1016/j.jamcollsurg.2014.03.052] [Medline: 25081938]
5. Sentinel Event Alert 51: preventing unintended retained foreign objects. The Joint Commission. 2022 May 1. URL: https://tinyurl.com/3k8tn89r [accessed 2025-04-23]
6. Keebler JR, Dietz AS, Lazzara EH, et al. Validation of a teamwork perceptions measure to increase patient safety. BMJ Qual Saf 2014 Sep;23(9):718-726. [doi: 10.1136/bmjqs-2013-001942]
7. Costar DM, Hall KK. Improving team performance and patient safety on the job through team training and performance support tools: a systematic review. J Patient Saf 2020 Sep;16(3S Suppl 1):S48-S56. [doi: 10.1097/PTS.0000000000000746] [Medline: 32810001]
8. TeamSTEPPS 30. Agency for Healthcare Research and Quality. URL: https://www.ahrq.gov/teamstepps-program/index.html [accessed 2025-04-23]

9. Skaret MM, Weaver TD, Humes RJ, Carbone TV, Grasso IA, Kumar H. Automation of the I-PASS tool to improve transitions of care. J Healthc Qual 2019;41(5):274-280. [doi: 10.1097/JHQ.0000000000000174] [Medline: 31483392]
10. Lin WT, Mayer C, Lee BO. Validity and reliability of the teamwork evaluation of non-technical skills tool. Aust J Adv Nurs 2019;36(3):67-74 [FREE Full text] [doi: 10.37464/2019.363.1460]
11. Abelson JS, Silverman E, Banfelder J, Naides A, Costa R, Dakin G. Virtual operating room for team training in surgery. Am J Surg 2015 Sep;210(3):585-590. [doi: 10.1016/j.amjsurg.2015.01.024] [Medline: 26054660]
12. Gasteiger N, van der Veer SN, Wilson P, Dowding D. How, for whom, and in which contexts or conditions augmented and virtual reality training works in upskilling health care workers: realist synthesis. JMIR Serious Games 2022 Feb 14;10(1):e31644. [doi: 10.2196/31644] [Medline: 35156931]
13. Liaw SY, Ooi SW, Rusli KDB, Lau TC, Tam WWS, Chua WL. Nurse-physician communication team training in virtual reality versus live simulations: randomized controlled trial on team communication and teamwork attitudes. J Med Internet Res 2020 Apr 8;22(4):e17279. [doi: 10.2196/17279] [Medline: 32267235]
14. Zhao J, Xu X, Jiang H, Ding Y. The effectiveness of virtual reality-based technology on anatomy teaching: a meta-analysis of randomized controlled studies. BMC Med Educ 2020 Apr 25;20(1):127. [doi: 10.1186/s12909-020-1994-z] [Medline: 32334594]
15. Kyaw BM, Saxena N, Posadzki P, et al. Virtual reality for health professions education: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019 Jan 22;21(1):e12959. [doi: 10.2196/12959] [Medline: 30668519]
16. Wang P, Wu P, Wang J, Chi HL, Wang X. A critical review of the use of virtual reality in construction engineering education and training. Int J Environ Res Public Health 2018 Jun 8;15(6):1204. [doi: 10.3390/ijerph15061204] [Medline: 29890627]
17. Makransky G, Klingenberg S. Virtual reality enhances safety training in the maritime industry: an organizational training experiment with a non‐WEIRD sample. Computer Assisted Learning 2022 Aug;38(4):1127-1140 [FREE Full text] [doi: 10.1111/jcal.12670]
18. Mazur LM, Khasawneh A, Fenison C, et al. A novel theory-based virtual reality training to improve patient safety culture in the department of surgery of a large academic medical center: protocol for a mixed methods study. JMIR Res Protoc 2022 Aug 24;11(8):e40445. [doi: 10.2196/40445] [Medline: 36001370]
19. Reason J. Human error: models and management. BMJ 2000 Mar 18;320(7237):768-770. [doi: 10.1136/bmj.320.7237.768] [Medline: 10720363]
20. Wiegmann DA, Shappell SA. A Human Error Approach to Aviation Accident Analysis: The Human Factors Analysis and Classification System, 1st edition: Routledge; 2017:45-71. [doi: 10.4324/9781315263878]
21. Alase A. The Interpretative Phenomenological Analysis (IPA): a guide to a good qualitative research approach. IJELS 2017;5(2):9. [doi: 10.7575/aiac.ijels.v.5n.2p.9]
22. Klein G, Phillips JK, Rall EL, Peluso DA. A data-frame theory of sensemaking. In: Hoffman RR, editor. Expertise Out of Context: Proceedings of the Sixth International Conference on Naturalistic Decision Making: Routledge; 2007:45-71. [doi: 10.4324/9780203810088]
23. McEwan D, Ruissen GR, Eys MA, Zumbo BD, Beauchamp MR. The effectiveness of teamwork training on teamwork behaviors and team performance: a systematic review and meta-analysis of controlled interventions. PLoS ONE 2017;12(1):e0169604. [doi: 10.1371/journal.pone.0169604] [Medline: 28085922]
24. Todsen T, Melchiors J, Wennerwaldt K. Use of virtual reality to teach teamwork and patient safety in surgical education. 2018 Presented at: 2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR); Mar 18-22, 2018; Reutlingen, Germany. [doi: 10.1109/VR.2018.8446469]
25. Aaberg OR, Hall-Lord ML, Husebø SIE, Ballangrud R. A human factors intervention in a hospital - evaluating the outcome of a TeamSTEPPS program in a surgical ward. BMC Health Serv Res 2021 Feb 3;21(1):114. [doi: 10.1186/s12913-021-06071-6] [Medline: 33536014]
26. Shi R, Marin-Nevarez P, Hasty B, et al. Operating room in situ interprofessional simulation for improving communication and teamwork. J Surg Res 2021 Apr;260:237-244. [doi: 10.1016/j.jss.2020.11.051] [Medline: 33360307]
27. Kaplan HJ, Spiera ZC, Feldman DL, et al. Risk reduction strategy to decrease incidence of retained surgical items. J Am Coll Surg 2022 Sep 1;235(3):494-499. [doi: 10.1097/XCS.0000000000000264] [Medline: 35972170]
28. Khoshnoodifar M, Emadi N, Mosalanejad L, Maghsoodzadeh S, Shokrpour N. A new practical approach using TeamSTEPPS strategies and tools: - an educational design. BMC Med Educ 2024 Jan 4;24(1):22. [doi: 10.1186/s12909-023-04803-2] [Medline: 38178071]

Abbreviations

FDR: false discovery rate
HFACS: Human Factors Analysis and Classification System
TeamSTEPPS: Team Strategies and Tools to Enhance Performance and Patient Safety
TENTS: Teamwork Evaluation of Non-Technical Skills
VR: virtual reality

Edited by B Lesselroth; submitted 05.09.24; peer-reviewed by KP Wong, P Greilich; revised version received 04.03.25; accepted 06.04.25; published 01.05.25.

Please cite as:
Mazur L, Butler L, Mitchell C, Lashani S, Buchanan S, Fenison C, Adapa K, Tan X, An S, Ra J
Effect of Immersive Virtual Reality Teamwork Training on Safety Behaviors During Surgical Cases: Nonrandomized Intervention Versus Controlled Pilot Study
JMIR Med Educ 2025;11:e66186
URL: https://mededu.jmir.org/2025/1/e66186
doi:10.2196/66186

Assessment of Large Language Model Performance on Medical School Essay-Style Concept Appraisal Questions: Exploratory Study

Seysha Mehta[1*], BA; Eliot N Haddad[1*], BS; Indira Bhavsar Burke[2], MHPE, MD; Alana K Majors[1], PhD; Rie Maeda[1], BA; Sean M Burke[2], MD; Abhishek Deshpande[1], MD, PhD; Amy S Nowacki[1], PhD; Christina C Lindenmeyer[1], MD; Neil Mehta[1], MBBS

[1]Cleveland Clinic Lerner College of Medicine, School of Medicine, Case Western Reserve University, 9500 Euclid Ave, G10, Cleveland, OH, United States

[2]Department of Internal Medicine, The University of Texas Southwestern Medical Center, Dallas, TX, United States

[*]these authors contributed equally

Corresponding Author:
Neil Mehta, MBBS
Cleveland Clinic Lerner College of Medicine, School of Medicine, Case Western Reserve University, 9500 Euclid Ave, G10, Cleveland, OH, United States

Abstract

Bing Chat (subsequently renamed Microsoft Copilot)—a ChatGPT 4.0–based large language model—demonstrated comparable performance to medical students in answering essay-style concept appraisals, while assessors struggled to differentiate artificial intelligence (AI) responses from human responses. These results highlight the need to prepare students and educators for a future world of AI by fostering reflective learning practices and critical thinking.

(JMIR Med Educ 2025;11:e72034) doi:10.2196/72034

KEYWORDS

essay-type questions; large language models; generative AI; Microsoft Copilot; artificial intelligence

Introduction

Large language models (LLMs) are of growing interest in medical education. LLMs have demonstrated passing scores on the United States Medical Licensing Examination (USMLE), raising questions about their impact on assessment frameworks [1], including whether artificial intelligence (AI) can successfully answer essay-style, reasoning-based questions and whether assessors can distinguish AI-generated and student-written responses. Our medical school's preclinical students complete application-level, essay-type questions—concept appraisals (CAPPs)—every week (Multimedia Appendix 1) [2]. We evaluated LLMs' performance on CAPPs and examined assessors' ability to distinguish AI-generated and human responses.

Methods

Study Design

Ten retired CAPP questions were selected, ensuring representation from multiple preclinical organ-system blocks, including gastroenterology, endocrinology, musculoskeletal science, cardiorespiratory medicine, hematology, renal biology, and immunology. Retired CAPPs were used, so that currently used ones were not exposed to students. Answering these required literature review and application of knowledge to clinical scenarios.

Five student responses from previous classes (before availability of LLMs) were randomly selected and deidentified. Individuals at various medical training levels generated AI responses via Bing Chat (subsequently renamed Microsoft Copilot; Multimedia Appendix 1), which used GPT-4 algorithms and had similar performance on medical tasks as ChatGPT 4.0—the most advanced LLM at the time of study [3,4]. Users first prompted Bing Chat by using the original CAPP text and then iteratively refined prompts to generate more comprehensive answers and match institutional standards without manual editing (Multimedia Appendix 1).

Ten expert assessors graded responses to 1 CAPP question each. While unaware that any responses had been AI-generated, they graded 5 deidentified student responses and 2 AI-generated responses (presented in random order) for their CAPP question, using a standard rubric (Multimedia Appendix 1). For 2 CAPPs, 4 student responses were used instead of 5 due to lack of consent for inclusion in the registry. Grading each CAPP took approximately 30 minutes; thus, a larger sample size was infeasible for this exploratory study. Afterward, assessors

identified whether responses were AI- or student-generated and provided their rationales.

Scoring differences between human- and AI-generated responses and identification accuracy were evaluated, using descriptive statistics. Thematic analysis was conducted on assessors' classification rationales; 2 team members independently analyzed reasons to identify themes, compared findings, and reconciled differences (Multimedia Appendix 1).

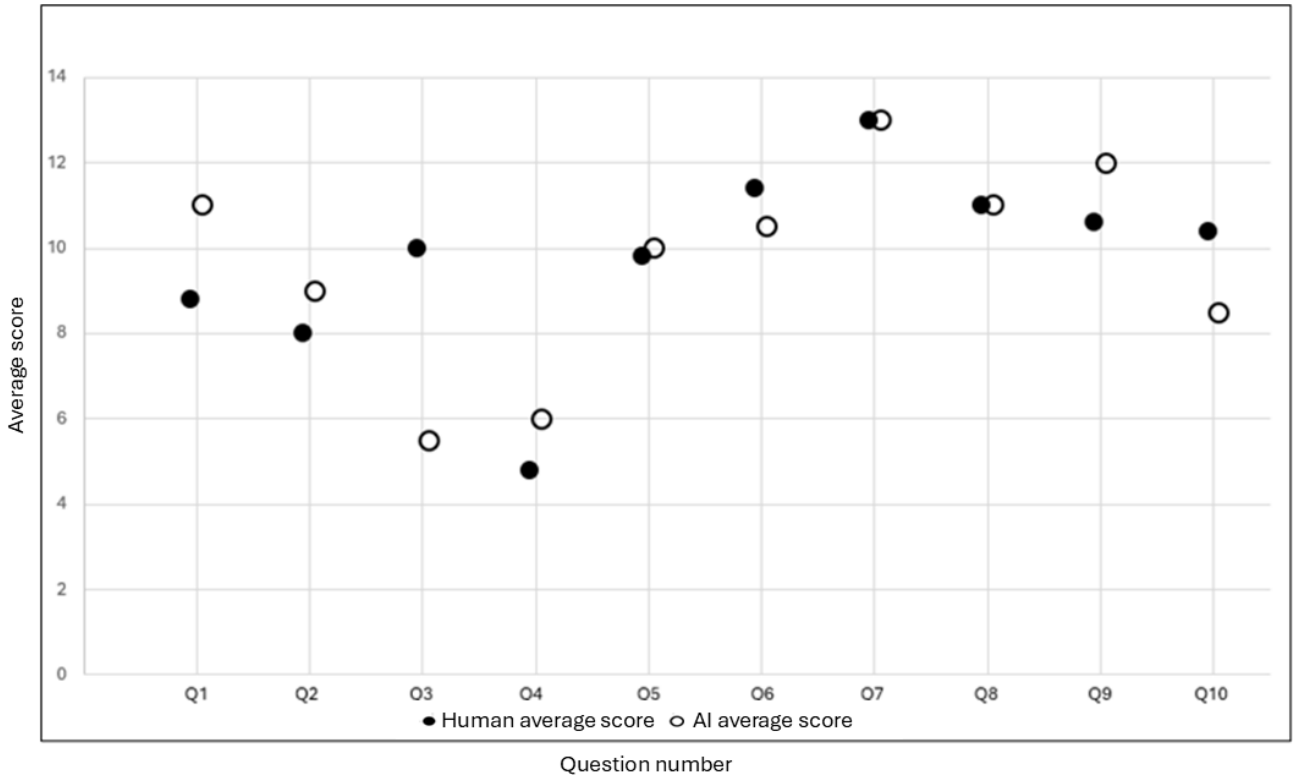
Ethical Considerations

This study used deidentified data from the Cleveland Clinic Institutional Review Board–approved registry #6600. Since this was a registry for which students had already provided informed consent, separate informed consent was not required. Each CAPP reviewer was paid US $100.

Results

AI responses received scores higher than or equal to those for human responses for most questions, with substantial performance variability; AI scored better than, equivalent to, or worse than humans, depending on the CAPP question (Figure 1).

Figure 1. Average of human vs AI scores for each question. CAPP questions were answered either by students (human) or by prompting Microsoft Copilot (AI). Expert graders scored the CAPP questions based on a rubric. The average scores received by humans and AI are shown by question (colored vs open circles, respectively). AI responses received scores higher than or equal to those for human responses for most questions. Each question had a unique maximum score. This figure illustrates the relative scores of humans vs AI. AI: artificial intelligence; CAPP: concept appraisal.



Assessors correctly identified response sources 53% (36/68) of the time (student responses: 27/48, 56%; AI-generated responses: 9/20, 45%). Only 1 assessor correctly classified all responses. Consistent with other studies, 1 assessor who used AI detection tools did not have much success [5] (Table 1).

Table . Percentage of responses correctly identified as human or artificial intelligence (AI) responses for each critical appraisal (CAPP) question.[a]

Question number	Correctly identified responses, n/N (%)
Q1	3/6 (50)
Q2	3/7 (43)
Q3	3/7 (43)
Q4	6/7 (86)
Q5	3/6 (50)
Q6	2/7 (29)
Q7[b]	0/7 (0)
Q8	5/7 (71)
Q9	4/7 (58)
Q10[c]	7/7 (100)

[a]Responses for each question were graded by 1 expert. Expert graders were blinded and were not told which responses were generated by humans vs AI.

[b]Despite utilization of AI detection tools, 1 assessor did not correctly classify any of the responses (Q7).

[c]Only 1 assessor correctly classified all responses for their CAPP question (Q10).

Thematic analysis showed that the most cited reason for identification was the perceived "writing style," though many assessors noted an inability to distinguish categories (Multimedia Appendix 1).

Discussion

We demonstrate that AI can provide high-quality answers to essay-style medical education questions requiring detailed research and knowledge application. Content experts struggled to distinguish AI-generated and human-written responses, underscoring the challenges of identifying academic misuse of generative AI.

Iterative prompting of Microsoft Copilot was essential for generating acceptable responses. This process mirrors students' typical workflow for refining drafts through edits; thus, iterative prompting does not necessarily disadvantage AI. Our findings highlight concerns about potential overreliance on AI and its implications for assessment validity, especially as recent survey data suggest that 89% of students use ChatGPT during self-study [6,7].

Given AI responses' similarity to human responses, institutions must consider frameworks for integrating AI into assessments without compromising academic integrity [8]. Potential strategies include structured classroom use of AI during collaborative group work (eg, requiring students to assess AI responses and cite primary evidence to support or refute them) [7,9].

Study limitations include a small sample of AI-generated responses and the research's exploratory nature. Expanding the sample size and including additional questions could provide insights on AI's performance (relative to humans) for specific question types (Multimedia Appendix 1). Additionally, the findings prompt further discussions on ethically integrating generative AI into medical curricula while ensuring students develop critical appraisal and independent reasoning skills [7,10].

AI's performance suggests its potential as a learning enhancement tool. However, medical educators must implement strategies for preventing overreliance on AI, fostering reflective learning practices and critical thinking, and maintaining assessment integrity.

Acknowledgments

The authors would like to thank the following individuals for serving as concept appraisal (CAPP) graders: William Albabish, William Cantrell, Thomas Crilley, Ryan Ellis, Andrew Ford, Emily Frisch, Jeffrey Schwartz, Michael Smith, Mohammad Sohail, and Anirudh Yalamanchali. Financial support was received from The Jones Day Endowment Fund.

Authors' Contributions

IBB and NM contributed to the literature review. NM, AKM, and CCL contributed to the conceptual design. SM, NM, ASN, and AD contributed to data analysis and visualization. IBB and SMB contributed to thematic analysis. SM, ENH, and NM contributed to manuscript writing. All authors contributed to the critical revision of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials regarding concept appraisal questions and grading, Bing Chat (subsequently renamed Microsoft Copilot), the iterative prompting used in this study, and the thematic analysis.
[DOCX File, 148 KB - mededu_v11i1e72034_app1.docx]

References

1. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. JMIR Med Educ 2023 Oct 20;9:e48785. [doi: 10.2196/48785] [Medline: 37862079]
2. Bierer SB, Dannefer EF, Taylor C, Hall P, Hull AL. Methods to assess students' acquisition, application and integration of basic science knowledge in an innovative competency-based curriculum. Med Teach 2008;30(7):e171-e177. [doi: 10.1080/01421590802139740] [Medline: 18777415]
3. Cai LZ, Shaheen A, Jin A, et al. Performance of generative large language models on ophthalmology board-style questions. Am J Ophthalmol 2023 Oct;254:141-149. [doi: 10.1016/j.ajo.2023.05.024] [Medline: 37339728]
4. Amin KS, Davis MA, Doshi R, Haims AH, Khosla P, Forman HP. Accuracy of ChatGPT, Google Bard, and Microsoft Bing for simplifying radiology reports. Radiology 2023 Nov;309(2):e232561. [doi: 10.1148/radiol.232561] [Medline: 37987662]
5. Elkhatat AM, Elsaid K, Almeer S. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. Int J Educ Integr 2023 Sep 1;19(1):17. [doi: 10.1007/s40979-023-00140-5]
6. Westfall C. Educators battle plagiarism as 89% of students admit to using OpenAI's ChatGPT for homework. Forbes. 2023 Jan 28. URL: https://www.forbes.com/sites/chriswestfall/2023/01/28/educators-battle-plagiarism-as-89-of-students-admit-to-using-open-ais-chatgpt-for-homework/ [accessed 2025-04-01]
7. Mehta S, Mehta N. Embracing the illusion of explanatory depth: a strategic framework for using iterative prompting for integrating large language models in healthcare education. Med Teach 2025 Feb;47(2):208-211. [doi: 10.1080/0142159X.2024.2382863] [Medline: 39058399]
8. Silverman JA, Ali SA, Rybak A, van Goudoever JB, Leleiko NS. Generative AI: potential and pitfalls in academic publishing. JPGN Rep 2023 Nov 8;4(4):e387. [doi: 10.1097/PG9.0000000000000387] [Medline: 38034432]
9. Jowsey T, Stokes-Parish J, Singleton R, Todorovic M. Medical education empowered by generative artificial intelligence large language models. Trends Mol Med 2023 Dec;29(12):971-973. [doi: 10.1016/j.molmed.2023.08.012] [Medline: 37718142]
10. Halkiopoulos C, Gkintoni E. Leveraging AI in e-learning: personalized learning and adaptive assessment through cognitive neuropsychology—a systematic analysis. Electronics (Basel) 2024 Sep 22;13(18):3762. [doi: 10.3390/electronics13183762]

Abbreviations

AI: artificial intelligence
CAPP: concept appraisal
LLM: large language model
USMLE: United States Medical Licensing Examination

Edited by LT Car; submitted 02.02.25; peer-reviewed by D Chartash, R Yang; revised version received 11.05.25; accepted 16.05.25; published 16.06.25.

Please cite as:
Mehta S, Haddad EN, Burke IB, Majors AK, Maeda R, Burke SM, Deshpande A, Nowacki AS, Lindenmeyer CC, Mehta N
Assessment of Large Language Model Performance on Medical School Essay-Style Concept Appraisal Questions: Exploratory Study
JMIR Med Educ 2025;11:e72034
URL: https://mededu.jmir.org/2025/1/e72034
doi:10.2196/72034

Perceptions and Intentions to Use Generative AI Among First-Year Medical Students in Japan: Cross-Sectional Survey Study

Hiroshi Tajima[1], MD, PhD; Hajime Kasai[1,2], MD, PhD; Kiyoshi Shikino[3], MD, MHPE, PhD; Ikuo Shimizu[1,2], MD, MHPE, PhD; Shoichi Ito[1,2,3], MD, PhD

[1]Department of Medical Education, Graduate School of Medicine, Chiba University, 1-8-1 Inohana, Chuo-ku, Chiba, Japan

[2]Health Professional Development Center, Chiba University Hospital, Chiba, Japan

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

Corresponding Author:
Hiroshi Tajima, MD, PhD
Department of Medical Education, Graduate School of Medicine, Chiba University, 1-8-1 Inohana, Chuo-ku, Chiba, Japan

Abstract

An April 2025 survey of 118 first-year Japanese medical students found high use of generative artificial intelligence (84.7%) but limited formal learning (49.2%), with strong learning interest yet neutral assignment use, indicating a need for structured literacy in generative artificial intelligence.

(JMIR Med Educ 2025;11:e77552) doi:10.2196/77552

KEYWORDS

generative artificial intelligence; medical students; digital literacy; perceptions; learning behavior; Japan

Introduction

Generative artificial intelligence (GenAI), particularly large language models like ChatGPT (OpenAI), has emerged as a tool for brainstorming, information gathering, proofreading, translating, and other academic tasks [1]. The rapid evolution of ChatGPT has made it essential for medical students to understand and use these tools [2]. As digital technologies become more embedded in health care education, understanding how future physicians will engage GenAI is vital. Many students demonstrate familiarity with digital tools and are exposed to them during childhood. One study showed that medical students had positive attitudes toward GenAI, albeit with low use rates [3]. In Japan, although some educators have adopted GenAI, information on how incoming medical students educated during the GenAI boom perceive these tools remains limited [4]. Thus, we explored GenAI use, learning behaviors, and perceptions of first-year medical students in Japan.

Methods

Overview

An anonymous online survey targeting first-year students immediately after high school graduation was conducted at the Chiba University School of Medicine in April 2025; 120 students were invited to participate and 118 valid responses were received (response rate 98.3%).

The questionnaire comprised four sections: (1) demographics; (2) prior GenAI exposure; (3) willingness to learn and perceptions (16 items rated on a 5-point Likert scale [1=strongly disagree to 5=strongly agree], based on prior studies [5,6]); and (4) an open-ended item on academic use. The items were pilot-tested for clarity and reliability.

Group differences were examined using independent 2-tailed *t* tests and chi-square tests. Spearman correlations were used to identify factors associated with learning motivation. Internal consistency (Cronbach α) was 0.81 for risk awareness (6 items), 0.81 for benefit recognition (6 items), and 0.53 for concern items (4 items; reverse-coded). Statistical significance was set at $P<.05$. Two independent coders performed content analysis of open-ended responses, with substantial agreement (κ=0.67). Analyses were conducted using JMP Pro (version 18; JMP Statistical Discovery LLC).

Ethical Considerations

This study was approved by the ethics committee of Chiba University Graduate School of Medicine (approval 3425). Electronic informed consent was obtained prior to participation. The study database was anonymized. Study data are stored on an offline, password-protected computer with full-disk encryption. Participants received no compensation for this study. All methods were in accordance with the relevant guidelines and regulations.

Results

Among 118 responses (n=79, 66.9% male; mean age 18.5, SD 0.7 years), 84.7% (n=100) reported GenAI use, mainly for language learning (n=53, 44.9%) and information searches (n=46, 39.0%). Only 49.2% (n=58) of the participants had

learned about GenAI, primarily through browsing (n=34, 28.8%) or through peers (n=24, 20.3%).

Students demonstrated high willingness to learn (mean score 4.3, SD 0.9), with good understanding (mean scores 3.9 - 4.6) and positive attitudes, especially regarding the relevance of GenAI (mean score 4.3, SD 0.8) (Table 1). Perceived individualization was lower (mean score 3.5, SD 1.0), with concerns over educational impact (mean score 3.6, SD 0.9).

Willingness to learn correlated positively with expectations of GenAI's benefits, such as enhancing digital literacy and career preparedness (both $P<.001$).

Students with learning experience were slightly younger (mean age 18.4, SD 0.7 vs 18.7, SD 0.8 years; $P=.02$) and demonstrated greater awareness of potential bias in GenAI outputs ($P=.045$).

Table . Medical students' responses regarding demographics, knowledge, willingness to use, and concerns about generative artificial intelligence (GenAI) technologies.

Item		All (n=118)	Experienced[a] group (n=58)	Inexperienced group (n=60)	*P* value
Sex, n (%)					.13
	Male	79	35	44	
	Female	39	23	16	
Age (years), mean (SD)		18.5 (0.7)	18.4 (0.7)	18.7 (0.8)	.02
Willingness to learn about GenAI (1=strongly negative to 5=strongly positive), mean score (SD)		4.3 (0.9)	4.3 (0.8)	4.2 (0.9)	.32
Attitude toward using GenAI for assignments (1=strongly negative to 5=strongly positive), mean score (SD)		3.0 (1.0)	2.9 (1.0)	3.0 (1.0)	.65
Knowledge of GenAI technologies, mean score (SD)					
	I understand that GenAI has limitations in handling complex tasks.	3.9 (0.9)	4.1 (0.9)	3.9 (0.9)	.21
	I understand that GenAI may produce outputs that are factually incorrect.	4.6 (0.7)	4.6 (0.6)	4.6 (0.7)	.67
	I understand that GenAI may generate outputs that are contextually inappropriate.	4.4 (0.7)	4.4 (0.7)	4.4 (0.6)	.9
	I understand that GenAI outputs may sometimes reflect bias or discrimination.	4.0 (0.9)	4.2 (0.9)	3.9 (0.9)	.04
	I understand that GenAI is pattern-based and may have limited applicability in specific contexts.	4.0 (1.0)	4.1 (1.0)	3.8 (0.9)	.11
	I understand that that GenAI lacks emotional intelligence and may produce insensitive content.	4.1 (1.0)	4.2 (0.9)	3.9 (1.0)	.11
Willingness to use GenAI technologies, mean score (SD)					
	I am considering incorporating GenAI into my learning and practice in the future.	3.9 (0.8)	3.9 (0.8)	3.8 (0.9)	.53
	Students need to learn how to effectively leverage GenAI for their career paths.	4.3 (0.8)	4.3 (0.7)	4.3 (0.9)	.59
	I believe that GenAI can enhance my digital literacy skills.	3.7 (1.0)	3.8 (0.9)	3.6 (1.1)	.33
	I believe GenAI can offer novel insights beyond my own thinking.	3.8 (0.9)	3.8 (0.8)	3.9 (0.9)	.73

Item		All (n=118)	Experienced[a] group (n=58)	Inexperienced group (n=60)	*P* value
	I believe that GenAI can instantly offer personalized suggestions for my assignments.	3.5 (1.0)	3.5 (1.0)	3.5 (1.0)	.99
	I think GenAI is a great tool because it is available 24-7.	4.0 (0.9)	4.0 (0.8)	3.9 (1.0)	.36
Concerns about GenAI technologies, mean score (SD)					
	Completing assignments using GenAI undermines the value of a university education.	3.6 (0.9)	3.6 (0.9)	3.5 (0.9)	.53
	GenAI limits opportunities for students to interact during face-to-face lectures and group work.	3.2 (1.0)	3.3 (0.9)	3.1 (1.0)	.25
	GenAI hinders the development of portable skills, such as teamwork, problem-solving, and leadership.	3.5 (0.9)	3.6 (0.8)	3.4 (1.0)	.12
	I feel that I am becoming overly dependent on GenAI.	2.1 (1.1)	2.2 (1.2)	2.0 (1.1)	.25

[a]"Experienced" refers to students who reported any learning about GenAI, including formal instruction and learning from informal sources, such as web browsing or peers.

The attitude toward using GenAI for assignments was neutral (mean score 3.0, SD 1.0). Open-ended responses were categorized into 3 attitudinal groups, "positive," "cautious," and "negative," reflecting enthusiasm, balanced concerns, or skepticism (Table 2).

Table . Keyword categories extracted from free-text responses regarding the academic use of generative artificial intelligence (GenAI).

Category		Keywords	Responses, n	Quotes
Positive				
	Convenience and efficiency	Efficient, convenient, easy	25	GenAI improves work efficiency because it can quickly search for information.
	Information handling and use	Information gathering, use, referencing, critical thinking	23	Using GenAI as a mock interviewer to question my responses was helpful.
	Quality and outcomes	Quality improvement, attractive, useful	9	High-quality information can be obtained.
	Necessity and attitude	Necessary, available resources	6	We should develop the ability to use GenAI effectively.
	Creativity	Original, idea	4	By integrating it with my own ideas, I can come up with more original concepts.
Cautious				
	Quantity and scope	Excessive, limited range, moderate, balance	27	I believe using GenAI within a limited range, such as proofreading my own writing, is acceptable.
	Use and operation	Assistance, proper use, permission	11	I think using GenAI as a learning assistant is good, but relying on it entirely is not advisable.
	Perception and emotion	Unclear, anxiety	5	I still don't fully understand GenAI, so I have some anxiety, but I also have a desire to master its use.
Negative				
	Thinking and creativity	Think independently, cessation of thinking, boring	40	Our ability to think independently will be lost.
	Ethics and norms	Dishonesty, fairness, rules	9	Presenting GenAI-generated work as one's own is, in my view, equivalent to plagiarism.
	Uncertainty and constraints	Ambiguity, limitations	3	Accuracy of GenAI is not guaranteed.

Discussion

Most medical students had used GenAI before university; however, structured learning was limited. Minor differences in perceptions between those with and without learning experience suggest that casual exposure alone may be insufficient to develop a critical understanding of GenAI, highlighting the need for foundational GenAI education in universities [7].

Although the students were interested in learning about GenAI, they remained cautious about using it for assignments, reflecting varied levels of familiarity and trust. While most viewed GenAI positively, some expressed concern that overreliance could hinder creativity and critical thinking.

GenAI dependence may hinder originality and decision-making [8], whereas trust in GenAI may foster motivation and proactive learning [9]. Students' neutral views suggest uncertainty or low confidence regarding appropriate use, underscoring the need for a balanced education that addresses concerns and fosters critical evaluation skills. This aligns with previous studies highlighting the importance of reforming curricula to integrate GenAI-related competencies, including ethical reasoning, clinical relevance, and communication skills [10].

A limitation of this study was that only first-year students at a single institution were included. Further, the "concern" subscale had relatively low internal consistency, and qualitative analysis relied on a single open-ended item. Finally, multiple comparisons were performed without correction, increasing the risk of type I errors. Nevertheless, this exploratory study aimed to identify trends rather than test specific hypotheses.

In conclusion, our findings support the inclusion of structured GenAI curricula in higher education, suggesting that such programs go beyond technical training to address students' expectations, values, and concerns.

Conflicts of Interest

None declared.

References

1. Regarding the academic handling of generative AI at universities and technical colleges. Ministry of Education, Culture, Sports, Science and Technology (Japan). 2023 Jul 14. URL: https://www.mext.go.jp/content/20230714-mxt_senmon01-000030762_1.pdf [accessed 2025-05-03]
2. Lee H. The rise of ChatGPT: exploring its potential in medical education. Anat Sci Educ 2024;17(5):926-931. [doi: 10.1002/ase.2270] [Medline: 36916887]
3. Tao W, Yang J, Qu X. Utilization of, perceptions on, and intention to use ai chatbots among medical students in China: national cross-sectional study. JMIR Med Educ 2024 Oct 28;10:e57132. [doi: 10.2196/57132] [Medline: 39466038]
4. Final report on utilization of advanced technologies and educational data in future schools and educational settings. Ministry of Education, Culture, Sports, Science and Technology (Japan). 2024 Jul 14. URL: https://www.mext.go.jp/content/2025414-mxt_shuukyo01_000033776_03.pdf [accessed 2025-07-28]
5. Chan CKY, Hu W. Students' voices on generative AI: perceptions, benefits, and challenges in higher education. Int J Educ Technol High Educ 2023;20(1):43. [doi: 10.1186/s41239-023-00411-8]
6. Deschenes A, McMahon M. A survey on student use of generative AI chatbots for academic research. Evid Based Libr Inf Pract 2024;19(2):2-22. [doi: 10.18438/eblip30512]
7. Zhai C, Wibowo S, Li LD. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. Smart Learn Environ 2024;11(1):28. [doi: 10.1186/s40561-024-00316-7]
8. Doshi AR, Hauser OP. Generative AI enhances individual creativity but reduces the collective diversity of novel content. Sci Adv 2024 Jul 12;10(28):eadn5290. [doi: 10.1126/sciadv.adn5290] [Medline: 38996021]
9. Huang J, Mizumoto A. Examining the effect of generative AI on students' motivation and writing self-efficacy. Digit Appl Linguist 2024;1:102324. [doi: 10.29140/dal.v1.102324]
10. Shimizu I, Kasai H, Shikino K, et al. Developing medical education curriculum reform strategies to address the impact of generative AI: qualitative study. JMIR Med Educ 2023 Nov 30;9:e53466. [doi: 10.2196/53466] [Medline: 38032695]

Abbreviations

GenAI: generative artificial intelligence

Edited by A Hasan Sapci; submitted 16.05.25; peer-reviewed by C Yi, VT Hoang; revised version received 18.08.25; accepted 12.10.25; published 13.11.25.

Please cite as:
Tajima H, Kasai H, Shikino K, Shimizu I, Ito S
Perceptions and Intentions to Use Generative AI Among First-Year Medical Students in Japan: Cross-Sectional Survey Study
JMIR Med Educ 2025;11:e77552
URL: https://mededu.jmir.org/2025/1/e77552
doi:10.2196/77552

Impact of Clinical Decision Support Systems on Medical Students' Case-Solving Performance: Comparison Study with a Focus Group

Marco Montagna[1*], MD; Filippo Chiabrando[1*], MD; Rebecca De Lorenzo[1], MD; Patrizia Rovere Querini[1,2], MD, PhD; Medical Students[3]

[1]School of Medicine, Vita-Salute San Raffaele University, Via Olgettina 58, Milan, Italy

[2]Unit of Medical Specialties and Healthcare Continuity, IRCCS San Raffaele Scientific Institute, Milan, Italy

[3]

[*]these authors contributed equally

Corresponding Author:
Marco Montagna, MD
School of Medicine, Vita-Salute San Raffaele University, Via Olgettina 58, Milan, Italy

Abstract

Background: Health care practitioners use clinical decision support systems (CDSS) as an aid in the crucial task of clinical reasoning and decision-making. Traditional CDSS are online repositories (ORs) and clinical practice guidelines (CPG). Recently, large language models (LLMs) such as ChatGPT have emerged as potential alternatives. They have proven to be powerful, innovative tools, yet they are not devoid of worrisome risks.

Objective: This study aims to explore how medical students perform in an evaluated clinical case through the use of different CDSS tools.

Methods: The authors randomly divided medical students into 3 groups, CPG, n=6 (38%); OR, n=5 (31%); and ChatGPT, n=5 (31%); and assigned each group a different type of CDSS for guidance in answering prespecified questions, assessing how students' speed and ability at resolving the same clinical case varied accordingly. External reviewers evaluated all answers based on accuracy and completeness metrics (score: 1 - 5). The authors analyzed and categorized group scores according to the skill investigated: differential diagnosis, diagnostic workup, and clinical decision-making.

Results: Answering time showed a trend for the ChatGPT group to be the fastest. The mean scores for completeness were as follows: CPG 4.0, OR 3.7, and ChatGPT 3.8 (P=.49). The mean scores for accuracy were as follows: CPG 4.0, OR 3.3, and ChatGPT 3.7 (P=.02). Aggregating scores according to the 3 students' skill domains, trends in differences among the groups emerge more clearly, with the CPG group that performed best in nearly all domains and maintained almost perfect alignment between its completeness and accuracy.

Conclusions: This hands-on session provided valuable insights into the potential perks and associated pitfalls of LLMs in medical education and practice. It suggested the critical need to include teachings in medical degree courses on how to properly take advantage of LLMs, as the potential for misuse is evident and real.

(JMIR Med Educ 2025;11:e55709) doi:10.2196/55709

KEYWORDS

chatGPT; chatbot; machine learning; ML; artificial intelligence; AI; algorithm; predictive model; predictive analytics; predictive system; practical model; deep learning; large language models; LLMs; medical education; medical teaching; teaching environment; clinical decision support systems; CDSS; decision support; decision support tool; clinical decision-making; innovative teaching

Introduction

Clinical reasoning and decision-making are at the core of the medical workflow. If they are accurate and grounded on solid and updated evidence, they help ensure the best health outcomes for patients. Clinical decision support systems (CDSS) have been implemented to aid practitioners in this duty [1-3]. Clinical practice guidelines (CPG) serve as the prototype for CDSS. They are published and updated at varying frequencies by scientific societies and policy makers, covering virtually every medical field or disorder. Over time, the number and complexity of CPG have increased, resulting in more detailed and robust recommendations. However, this has also led to reduced immediacy and ease of access and comprehension for medical professionals. Additionally, there may be multiple CPG for a single pathological condition, sometimes with conflicting recommendations. As a potential solution, emerging technologies on the internet have given rise to new CDSS options known as online repositories (ORs). These repositories, like encyclopedias, consolidate and synthesize knowledge

related to various medical disorders. They draw from current practices, available CPG, and the latest published evidence, making this information easily accessible to physicians. Typically provided by publishing groups, ORs often require subscription-based access. Two of the most popular options are "UpToDate" (by Wolters Kluwer [4]) and "BMJ Best Practice" (by BMJ Publishing Group [5]), both available as websites and mobile apps. The recent introduction of large language models (LLMs) for public use has generated both excitement and debate. Their adoption has rapidly grown across various human activities [6]. Many foresee the immense potential benefits of applying such technology to medical practice, while others harbor concerns about the dangers it might pose if left unregulated and misaligned [7-12].

Without a doubt, LLMs like ChatGPT represent a new generation of CDSS with unparalleled assistance capabilities. They can engage in active interactions with users and directly interpret medical information, extending far beyond simple guideline consultation. They can suggest possible diagnostic workups (DWs) or treatment algorithms [6]. In such cases, physicians would no longer need to navigate extensive datasets of clinical information, distill practical advice from lengthy text pages, or grapple with uncertainty about consulting the correct or sufficient sources. On the flip side, it is evident that LLMs also carry the potential for misuse, which could lead to significant harm to patients [7-9,11,13]. There's a risk of guiding clinicians down erroneous thought processes, potentially resulting in wasted time and the unintentional complication of cases. When the alternatives being evaluated are either incorrect or become excessively numerous, the complexity of a case may inevitably worsen. As a result, there is legitimate concern regarding how the indiscriminate use of LLMs might inadvertently drive-up health care costs. This underscores the importance of integrating LLMs into clinical decision-making (CDM) processes with caution and judiciousness.

However, although the adoption of innovative CDSS tools is steadily rising, the lack of dedicated training in their proper utilization undermines their full potential as valuable aids [9,10].

The aim of the present study was to investigate how senior medical students employ CDSS in the resolution of a clinical case with the ultimate intention of designing specific educational programs. Specifically, we conducted a hands-on session to compare CPG, ORs, and an LLM (ChatGPT) in terms of speed and accuracy of the clinical decisions proposed after consultancy with the CDSS.

Methods

Study Design

The present is a report of a hands-on practical session taking place during the Course of Internal Medicine at our university. The subject of the analysis was the quality of students' answers to a number of open-ended questions related to clinical reasoning and problem-solving, as a proxy for their capacity to employ different CDSS. A fictional clinical case was designed by the authors to control for complexity. Additionally, ChatGPT (version 3.5) generative capabilities were used to fabricate vital parameters, physical examination, and laboratory results for the fictional patient. ChatGPT was asked to include confounding factors in the answers provided. The authors revised generated elements to make sure they met the study requirements. The complete clinical case, open-ended questions and conversation with ChatGPT are available in Multimedia Appendix 1 and Multimedia Appendix 2.

Participants, Recruitment Strategies, and Sampling Method

Students attending the last lesson in the academic year 2022/23 Course of Internal Medicine of the International MD Program degree at Vita-Salute San Raffaele University, Milan (IT), were all included in the study, with convenience sampling. No exclusion criteria were applied.

Experimental Groups, Randomization, and Blinding

In total, 3 groups with comparable numbers of students were defined at the beginning of the lesson (Multimedia Appendix 3). Starting from the first seating rows, students were randomly assigned a number from 1 to 3 and consequently formed the 3 groups. Each group nominated a delegate who randomly picked an envelope containing the indication of the type of CDSS to be used by his or her group, either (i) CPG, (ii) OR, or (iii) ChatGPT. For each group, only the delegate was allowed web-based access to the CDSS. Group assignments were open label. The group assigned to CPG was allowed to use the internet to search for and consult CPG deemed useful to solve the case. The group assigned to OR was allowed to use the internet to access UpToDate and look for articles and algorithms or tables deemed useful to solve the case. The group assigned to ChatGPT was allowed to log into the LLM and use it to ask and gather information deemed useful to solve the case.

The delegate was also in charge of sending his or her group's clinical decision to the researchers via a mobile phone SMS text message. An inspector of the research staff was assigned to each group to guarantee that only the assigned CDSS was used. The questions were shown along with the presentation of the clinical case in a Microsoft PowerPoint (Microsoft Corporation) slideshow. For each question, a countdown timer was shown on the projector screen, and the start time was recorded by the researchers. Time required by each group for each answer was calculated by subtracting the start time from the mobile message arrival time.

Outcomes/Assessment

In total, 1 junior resident in internal medicine, 1 senior resident in internal medicine, and 2 internal medicine junior consultants were asked to perform blind external assessment of the answers. They were provided with a form containing the same clinical case and questions shown to the students together with the answers given by each group, with no details on CDSS used. Answers were graded from 1 to 5 in terms, respectively, of completeness and accuracy. Completeness was described as: "Is the answer complete or does it miss anything?". Accuracy was described as: "Is the answer precise and adherent to clinical practice, or too vague, too wide, too superficial?"

Sample Size

Sample size was not defined a priori. Sample size was determined by the number of students that attended the lesson on that day.

Statistical Analysis

Scores are reported as mean. For further analysis, the 8 questions were grouped into 3 domains according to the students' skill investigated: (1) differential diagnosis (DD, Q 1-5-6), (2) DW (Q 2 - 7), and (3) CDM (Q 3-4-8). The Kruskal-Wallis test was performed to establish whether there was a significant difference among the 3 groups in the times and scores overall and in each student skill domain. Microsoft Excel (Microsoft Corporation) and GraphPad Prism (version 9.0; GraphPad Software) were used as software tools for the analysis.

Ethical Considerations

This study does not require ethical approval as this is a report of data collected for monitoring and reporting purposes of innovative teaching activities taking place at Vita-Salute San Raffaele University, according to the Self-assessment-Evaluation-Accreditation system (AVA) of ANVUR, to which our institution is subject [14]. The data were generated during a hands-on session taking place during a usual lesson and in a teaching, non-experimental environment, with no risks for the participants. Data represent the output of each student group; they therefore collect aggregated information, with no individual identity linked to them. No confidential data were collected. No compensation was provided to participants, and they were able to opt out at any time during the lesson. AVA aims to improve the quality of teaching and research carried out in the Italian universities through the application of a quality assurance model based on internal procedures for planning, management, self-assessment, and improvement of training and scientific activities and on an external verification carried out in a clear and transparent manner. The requirements of the new AVA3 model underline the importance for the universities to promote, support, and monitor the participation of teachers in training and teaching refresher initiatives in the various disciplines, including those relating to the use of innovative teaching methodologies, also through the use of online tools and the provision of multimedia teaching materials [14]. The presented data were collected in this context, and, accordingly, no ethics approval was applied for (Page 3, Art.5, Clause 2 of [15]).

Results

A total of 16 students were included: 6 allocated to the CPG group (*F*=5, 83%), 5 to the OR group (*F*=2, 40%), and 5 to the ChatGPT group (*F*=3, 60%).

During the presentation of the clinical case, all 3 groups were presented with questions, and students were required to provide their responses as quickly as possible, within predefined time limits. The time taken to answer each question was recorded for all groups. Except for one response, all answers were given within the allocated time (see Table 1). Of the 49 total allocated minutes, the CPG group took 41 minutes to complete the clinical case, the OR group 45 minutes, and the ChatGPT group 38 minutes. The total time taken to answer, expressed as a percentage of the allocated time, was not significantly different among groups (*P*=.69).

Table . Time (min) required for answers and mean score received at the external assessment in terms of completeness and accuracy for each answer given by each group. Overall time for completion and median score across all answers for each group are also reported. Allotted time was exceeded only in Q4 by OR group.

	Clinical practice guidelines			Online repositories			ChatGPT		
	Time (min)	Quality		Time (min)	Quality		Time (min)	Quality	
		Completeness, mean (SD)	Accuracy, mean (SD)		Completeness, mean (SD)	Accuracy, mean (SD)		Completeness, mean (SD)	Accuracy, mean (SD)
Q1. Rank the possible differential diagnoses in terms of probability.	8	4.0 (0.8)	3.8 (1.3)	8	3.5 (0.6)	3.0 (0.8)	8	4.0 (0.8)	3.8 (0.5)
8 min[a]									
Q2. Based on the previous list, which diagnostic workup would you set up?	7	4.0 (0.8)	3.8 (0.5)	7	4.8 (0.8)	2.5 (1.0)	6	3.8 (0.5)	3.8 (1.3)
8 min[b]									
Q3. Which values are altered?	3	4.0 (0.8)	4.3 (1.0)	5	3.8 (1.0)	4.0 (1.2)	2	4.0 (0.8)	2.3(1.3)
5 min[c]									
Q4. Which treatment do you start?	5	4.3 (1.5)	3.8 (1.5)	6	4.0 (0.8)	4.3 (1.0)	4	4.0 (0.8)	4.3 (1.0)
5 min[c]									
Q5. Which are the possible causes of hypercalcemia?	5	3.8 (1.0)	3.8 (1.0)	4	3.8 (0.5)	3.8 (1.0)	6	3.0 (0)	3.3 (1.3)
8 min[a]									
Q6. Can you narrow down the previous list based on these findings?	3	4.0 (1.4)	4.3 (1.5)	5	3.0 (0.8)	2.0 (0.8)	5	4.3 (1.0)	4.0 (0.8)
5 min[a]									
Q7. Which are the primary diagnostic tests that you order?	5	4.3 (1.0)	4.5 (1.0)	5	3.8 (1.5)	3.4 (0.7)	2	4.0 (0.8)	4.3 (0.9)
5 min[b]									
Q8. Which therapeutic choice do you offer to the patient?	5	3.8 (0.5)	4.3 (1.0)	5	3.5 (1.3)	3.4 (1.3)	5	3.8 (1.0)	4.0 (1.1)
5 min[c]									

	Clinical practice guidelines			Online repositories			ChatGPT		
	Time (min)	Quality		Time (min)	Quality		Time (min)	Quality	
		Completeness, mean (SD)	Accuracy, mean (SD)		Completeness, mean (SD)	Accuracy, mean (SD)		Completeness, mean (SD)	Accuracy, mean (SD)
TOT/mean[a]	41	4.0 (0.2)	4.0[d] (0.3)	45	3.8 (0.5)	3.3[d] (0.8)	38	3.8 (0.4)	3.7[d] (0.7)

[a]Differential diagnosis

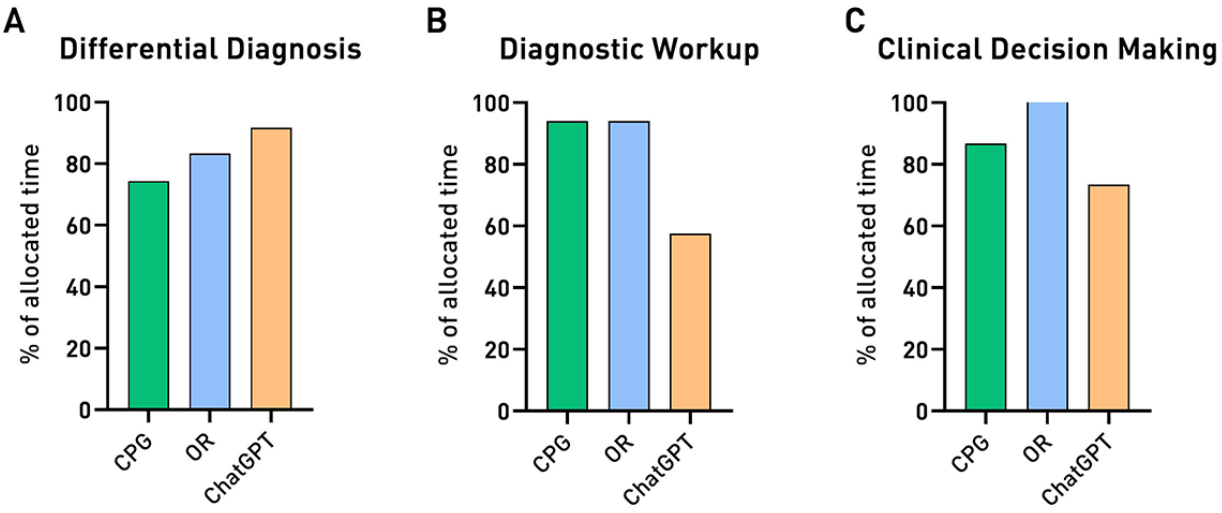
[b]Diagnostic workup

[c]Clinical decision-making

[d]*P*=.02.

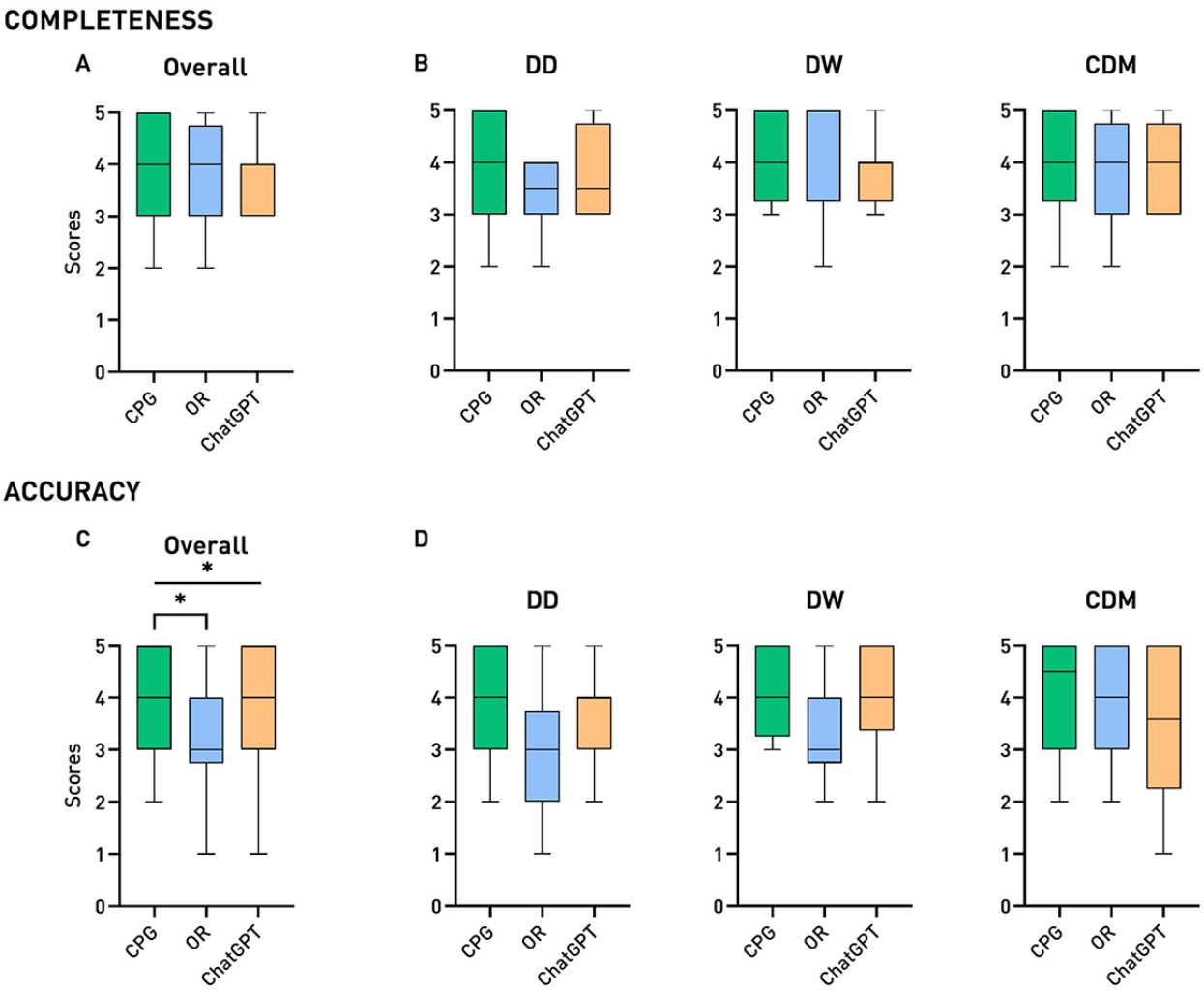
The questions were then categorized into 3 major domains: DD (Q 1-5-6), DW (Q 2 - 7), and CDM (Q 3-4-8). The time taken to answer, as a percentage of the allocated time, by each group of students according to the provided domains is shown in Figure 1(A, B and C). While no statistically significant differences were observed, it is worth noting that the ChatGPT group tended to respond more quickly to questions related to DW and CDM.

Figure 1. Sum of the time taken by the 3 groups of students to answer questions in the 3 domains. Results are shown as the percentage of the total allocated time for that domain. CPG: clinical practice guidelines; OR: online repositories.



Answers were assessed for completeness, which considered the depth of information provided, and for accuracy, which assessed adherence to clinical practice versus an excess or superficiality of information. These evaluations were conducted by 4 external reviewers who were blinded to the group assignment. Scores are reported in Table 1 and Figure 2.

Figure 2. Box plots of scores obtained by the 3 groups of students for (A)overall completeness, (C)overall accuracy, (B)completeness in the 3 domains, (D)accuracy in the 3 domains. DD: differential diagnosis; DW: diagnostic workup; CDM: clinical decision-making (*P<.05).



Overall, the CPG group performed best, reaching the highest mean scores for completeness (4.0) and accuracy (4.0) (Figure 2A and C). The ChatGPT group comes in second place, with equal completeness (3.8 vs 3.8) compared to the OR group but higher accuracy (3.7 vs 3.3). The Kruskal-Wallis test showed non-significant difference among groups for overall completeness (P=.49), and a significant difference in overall accuracy between the 3 groups (P=.02), particularly—at post hoc analysis—between the CPG and the OR group (P=.02).

Aggregating scores according to the 3 students' skill domains, trends in differences among the groups emerge more clearly (Figure 2B and C). When it comes to generating differential diagnoses, the CPG group was the most complete (3.9) and accurate (3.9) among the 3, whereas the OR group has the worst scores for both completeness and accuracy categories, with mean scores of 3.4 and 2.9, respectively, among the lowest registered. On the contrary, whenever students were asked to provide a DW for the patient, the OR group appeared as the most complete (4.3), although the least accurate (3.0), considering the source of the competition. In this domain, the other 2 groups come quite close (4.1 and 4.1 for CPG; 3.9 and 4.0 for ChatGPT) and maintain coherent scores in between their own completeness and accuracy. Lastly, when knowledge had to be applied to drive clinical decisions, as tested in CDM questions, the 3 groups obtained similar scores in terms of completeness, while the CPG group succeeded as the most accurate (4.1), followed by the OR group (3.9) and the ChatGPT group (3.5).

Of note, it appears evident how the CPG group performed the best in nearly all domains and maintained almost perfect alignment between its completeness and accuracy. The ChatGPT group maintained an overall mediocre yet stable performance, without ever achieving the best scores. On the other hand, the OR group showed mixed features, with notable peaks of performance–with scores equal to or higher than the ChatGPT contender–but some other dramatic drops in answer quality in terms of accuracy in more than one skill domain.

Discussion

In recent years, the availability of tools to support clinicians in the diagnostic and therapeutic processes has grown considerably. Although the use of CDSS is widespread, individuals often use them without specific education and pay little attention to their

inherent limitations, especially in the case of their newest potential counterparts, such as ChatGPT [1].

To assess how final-year medical school students make use of the available CDSS and to begin considering an instructional approach for the use of such tools, we designed the experiment outlined in this study. Observing the students' interaction with the assigned CDSS during the resolution of a clinical case and analyzing their answers, we recorded specific criticalities regarding the use of each CDSS.

In terms of rapidity of use, ChatGPT seems to represent a significant breakthrough in the world of CDSS. If traditional encyclopedic or textbook-like written resources call the reader to go through the entire material to properly understand a topic and retrieve correct clinical answers or guidance, an instantaneous chat environment allows for both a quick overview of disciplines and—at the same time—deep vertical dives into specific details. Students using ChatGPT arrived at the required answer almost always faster than their colleagues aided by CPG or OR, especially in questions regarding DW and CDM. It seems that the role of ChatGPT in the chat dialogue more closely resembles the attitude of a human counterpart (for example, a senior teaching physician asked for guidance), whose responses are fast, direct, and usually finely targeted. Such answers follow the students' line of reasoning and indirectly encourage them to choose a unique path of solution out of many possible scenarios. This dynamic emerges as brilliantly effective whenever the students embark on the correct clinical thought process but can lead to disastrous consequences whenever students feed cognitive biases or overt errors into their chat conversation. On the contrary, CPG and OR offer vast amounts of information, such as long lists of items and in-depth descriptions, whose digestion is neither easy nor fast. Therefore, whenever consulted correctly and for enough time, CPG and OR—especially the former, our results seem to suggest—generally help students give answers of higher quality, both in terms of completeness and accuracy. No real-time interaction is present; therefore, they appear almost immune to reader-introduced bias and misinterpretation.

In the modern era, velocity is a precious commodity, especially in the fast-paced clinical context, where less and less time is available for extensive reference consultation. This might influence current and future generations of medical school students to prefer chatbot-based guidance over preformed texts as routine help throughout their study [1,8].

As said, in terms of highest answer quality, there seems to be no rival to CPG. Old-fashioned guidelines might be slower to consult but grant far greater quality information to students, helping them to be complete and accurate in key tasks, such as generating differentials and deploying clinical decisions. Possibly, CPG might also enhance the student's comprehension of the analyzed topic, given the broader context and deeper description always provided. Nonetheless, in a continuously evolving context of increasing number and complexity of CPG, it would be presumptuous to expect students to rely on their use only [4,5,9].

ORs offered unexpected and ambiguous results, as the performance of these students was not stable across questions, and their answers could reach both extremes of the score spectrum. ORs are intrinsically designed and thought of as an evolution of CPG, more accessible and applicable to practice. This aspect may emerge in the excellent quality of answers given by this group in the DW domain, where students were able to follow the detailed workup algorithms offered by OR, which can graphically synthesize even complex clinical scenarios. An interesting experience across teaching hospitals in Japan evidenced a somewhat significant positive correlation between the use of OR and score performance in a national general medicine test, in a numerous population of over 3000 residents. According to the authors, frequent logging into and consultation of UpToDate might have contributed to improved clinical reasoning skills, specifically in the tested domains of DD generation, and clinical decision deployment [16].

Specific Observations Around the Students' Use of Each Type of CDSS

CPG Group

For this group's ability to reach a correct solution, the crucial step seemed to be selecting the proper guideline to consult. Once correctly identified, suitable CPG contain virtually everything a physician should know about a disease. During the initial process of elaboration of a clinical scenario, though, it is not immediately clear which disease entity, often more than one, is going to be selected as a candidate diagnosis for the patient. Choosing the right CPG can therefore be quite challenging. Additionally, CPG deliver their content in the form of plain text and interspersed summary tables and charts. The students had some difficulty in focusing on the right chart.

OR Group

Likely due to their lack of experience with the tool, the students failed to search for symptoms within the query. One hypothesis could be that they relied on their prior knowledge to make decisions rather than on the results obtained from each query, preventing them from breaking free from their preconceptions.

ChatGPT Group

Our results revealed students to lack substantial background on how to properly approach an LLM chatbot. *Masterprompting*, referred to as the assignment of the role and behavior expected from the chatbot, was not provided by the students.

Using an LLM as a CDSS inevitably introduces a "prompt bias," for which the human subjective way of reasoning and choice of questions to be asked directly influence how the chatbot perceives the information and how it transforms its responses accordingly. Along such a general trend, the students' prompts were not properly designed and translated into confused and misleading answers by ChatGPT. For example, clinical details were provided without any structure or further clinical context, triggering diverging suggestions by the chatbot. Accordingly, hints were given by the instructor on asking the LLM directly how to convey information for it to understand and elaborate on such information at its best, but consistent results did not follow.

On other instances, it was ChatGPT itself that derailed students' reasoning. For example, a chatbot answer about laboratory

values, lacking a relative unit of measurement and normal reference ranges, leads to misinterpretation of hypercalcemia as hypocalcemia, a critical mistake.

ChatGPT does not provide any literature citation and guideline reference to support its own line of reasoning, as other CDSS do in the form of easily accessible links to further explanations and deeper dives (eg differential diagnoses, varying DWs, tables of available first-line therapies, etc) [6]. Such an absence seemed to be associated with a significantly lower propensity of students to question either their own knowledge or the answers formulated by ChatGPT, blindly accepting the information provided. Whenever such indulgence was noted and pointed out by the instructors, students confessed how certain answers provided by ChatGPT were not completely clear and understandable; nonetheless, they willingly accepted them as valid.

Limitations

Some limitations of this report must be underlined. First, the restricted number of students who took part in the experiment. This may have led to uneven distribution of differently ranked students in the groups, despite the random strategy used for group definition. Second, our methodological constraints lead to insufficient statistical power to draw sound conclusions. Lastly, the experiment has not yet been repeated, and the results have been further confirmed or discarded with other clinical cases.

Conclusions

As in many other disciplines, the adoption of LLMs in medical practice and in the medical school curriculum is inevitable [10]. Our hands-on session suggests the critical need to include in medical degree courses teachings on how to properly take advantage of LLMs such as ChatGPT, as we verified that the potential for misuse is evident and real.

Our experience suggests the need for medical students to be acquainted with LLMs in their learning process and future profession. ChatGPT does not provide nor teach a reasoning method to approach a medical case resolution, a relevant issue for it to be recognized as part of the armamentarium in formal medical education. CPG and OR, on the contrary, most often provide step-by-step guidance on how to behave in each clinical scenario, how to approach diagnosis, and how to address treatment of diseases. References and recommendation strength form the cornerstone of these tools and help the student get progressively acquainted with ever-updating medical knowledge [4,5].

In conclusion, regarding the upcoming future, we suggest medical educators to:

- start to increasingly incorporate and refer to LLMs in their teachings, also by building tailored case studies [17-20];
- favor practice-based learning by using LLMs as a help to navigate guidelines and repositories with more ease and speed;
- exploit the very limitations of LLMs—such as the lack of an explicit reasoning method or unsure reliance on the latest published literature—to prompt students to consciously provide them themselves to the chatbot, turning the CDSS consultation process into a bidirectional teaching environment, possibly uncovering biases and misconceptions on both sides;
- help students in focusing on their accountability: they should be pushed to continuously look for evidence and validation of their own clinical reasoning, avoiding relying completely on that of LLMs.

Acknowledgments

The authors thank the Medical Students as Collaborators attending the hands-on session: ChatGPT group - Mariam Datukishvili, Roberto Leone, Mariagiulia Giugliano, Marifrancesca Forquet, Ino De Martino; UpToDate group - Lisa Marie Pereira, Pietro Felisatti, Francesco Paleari, Carlotta Dell'Anna Misurale, Angelo Manfredi; Guidelines Group - Giulia Pacini, Francesca Crippa, Ivan Shashkin, Carola Alberoni, Mazvita Mungwadzi, Sara Coacci. The authors thank Prof. Angelo Andrea Maria Manfredi for his valuable insights on the matter.

Authors' Contributions

FC and MM involved in conceptualization, methodology, investigation, data curation, formal analysis, and writing: original draft; RDL contributed to methodology, investigation, data curation, and writing: review & editing; MD students contributed to investigation and resources; PRQ involved in conceptualization, methodology, supervision, and writing: review & editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
ChatGPT conversation history for clinical case generation.
[DOCX File, 35 KB - mededu_v11i1e55709_app1.docx]

Multimedia Appendix 2
Clinical Case and questions for students.
[PDF File, 56 KB - mededu_v11i1e55709_app2.pdf]

Multimedia Appendix 3
Workflow Diagram.
[PDF File, 20 KB - mededu_v11i1e55709_app3.pdf]

References

1. Goodman KE, Rodman AM, Morgan DJ. Preparing physicians for the clinical algorithm era. N Engl J Med 2023 Aug 10;389(6):483-487. [doi: 10.1056/NEJMp2304839] [Medline: 37548320]
2. Huang S, Liang Y, Li J, Li X. Applications of clinical decision support systems in diabetes care: scoping review. J Med Internet Res 2023 Dec 8;25(1):e51024. [doi: 10.2196/51024] [Medline: 38064249]
3. Wiwatkunupakarn N, Aramrat C, Pliannuom S, et al. The integration of clinical decision support systems into telemedicine for patients with multimorbidity in primary care settings: scoping review. J Med Internet Res 2023 Jun 28;25(1):e45944. [doi: 10.2196/45944] [Medline: 37379066]
4. Evidence-Based Clinical Decision Support System| UpToDate. : Wolters Kluwer URL: https://www.wolterskluwer.com/en/solutions/uptodate [accessed 2023-07-11]
5. BMJ best practice. URL: https://bestpractice.bmj.com/ [accessed 2023-07-11]
6. Introducing ChatGPT. URL: https://openai.com/blog/chatgpt [accessed 2023-07-19]
7. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023 Jul 6;6(1):120. [doi: 10.1038/s41746-023-00873-0] [Medline: 37414860]
8. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst 2023 Mar 4;47(1):33. [doi: 10.1007/s10916-023-01925-4] [Medline: 36869927]
9. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023;6:1169595. [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
10. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887. [doi: 10.3390/healthcare11060887] [Medline: 36981544]
11. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
12. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ 2023 Mar 6;9:e46885. [doi: 10.2196/46885] [Medline: 36863937]
13. The Lancet Digital Health. ChatGPT: friend or foe? Lancet Digit Health 2023 Mar;5(3):e102. [doi: 10.1016/S2589-7500(23)00023-7] [Medline: 36754723]
14. II cycle (AVA3) – ANVUR – agenzia nazionale di valutazione del sistema universitario e della ricerca. URL: https://www.anvur.it/en/activities/ava/periodic-accreditation/ii-cycle-ava3/ [accessed 2024-10-01]
15. President of Italian Republic. Presidential decree february 1st, 2010, n. 76. rome. 2010 URL: https://www.normattiva.it/uri-res/N2Ls?urn:nir:presidente.repubblica:decreto:2010-02-01;76~art12-com4-letd [accessed 2024-10-25]
16. Kataoka K, Nishizaki Y, Shimizu T, et al. Hospital se of a web-based clinical knowledge support system and in-training examination performance among postgraduate resident physicians in Japan: nationwide observational study. JMIR Med Educ 2024 May 30;10(1):e52207. [doi: 10.2196/52207]
17. Lee J, Wu AS, Li D, Kulasegaram KM. Artificial intelligence in undergraduate medical education: a scoping review. Acad Med 2021 Nov 1;96(11S):S62-S70. [doi: 10.1097/ACM.0000000000004291] [Medline: 34348374]
18. Tsopra R, Peiffer-Smadja N, Charlier C, et al. Putting undergraduate medical students in AI-CDSS designers' shoes: an innovative teaching method to develop digital health critical thinking. Int J Med Inform 2023 Mar;171:104980. [doi: 10.1016/j.ijmedinf.2022.104980]
19. Feng S, Shen Y. ChatGPT and the future of medical education. Acad Med 2023 Aug 1;98(8):867-868. [doi: 10.1097/ACM.0000000000005242] [Medline: 37162219]
20. Jacobs SM, Lundy NN, Issenberg SB, Chandran L. Reimagining core entrustable professional activities for undergraduate medical education in the era of artificial intelligence. JMIR Med Educ 2023 Dec 19;9(1):e50903. [doi: 10.2196/50903] [Medline: 38052721]

Abbreviations

CDM: clinical decision-making
CDSS: clinical decision support systems
CPG: clinical practice guidelines
DD: differential diagnosis
DW: diagnostic workup
LLM: large language model
OR: online repositories

Edited by TDA Cardoso, T Leung; submitted 21.12.23; peer-reviewed by E Ogut, JJ Ríos Blanco; revised version received 25.10.24; accepted 26.10.24; published 18.03.25.

Please cite as:
Montagna M, Chiabrando F, De Lorenzo R, Rovere Querini P, Medical Students
Impact of Clinical Decision Support Systems on Medical Students' Case-Solving Performance: Comparison Study with a Focus Group
JMIR Med Educ 2025;11:e55709
URL: https://mededu.jmir.org/2025/1/e55709
doi:10.2196/55709

Original Paper

Comparing the Effectiveness of Multimodal Learning Using Computer-Based and Immersive Virtual Reality Simulation–Based Interprofessional Education With Co-Debriefing, Medical Movies, and Massive Online Open Courses for Mitigating Stress and Long-Term Burnout in Medical Training: Quasi-Experimental Study

Sirikanyawan Srikasem[1], MSc; Sunisa Seephom[1], PhD; Atthaphon Viriyopase[2], PhD; Phanupong Phutrakool[3], PhD; Sirhavich Khowinthaseth[2,4], MD; Khuansiri Narajeenron[2,4], MD, MSc, CHSE, MHPE; ER-VIPE Study Group[5]

[1]Adult and Gerontological Nursing Department, Srisavarindhira Thai Red Cross Institute of Nursing, Bangkok, Thailand

[2]Department of Emergency Medicine, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand

[3]Chula Data Management Center, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand

[4]Department of Emergency Medicine, King Chulalongkorn Memorial Hospital, The Thai Red Cross Society, Bangkok, Thailand

[5]See Acknowledgments

Corresponding Author:
Khuansiri Narajeenron, MD, MSc, CHSE, MHPE
Department of Emergency Medicine
Faculty of Medicine
Chulalongkorn University
1873 MDCU Faculty
Pathumwan
Bangkok, 10330
Thailand
Phone: 66 0855054209
Email: khuansiri.n@chula.ac.th

Abstract

Background: Burnout among emergency room health care workers (HCWs) has reached critical levels, affecting up to 43% of HCWs and 35% of emergency medicine personnel during the COVID-19 pandemic. Nurses were most affected, followed by physicians, leading to absenteeism, reduced care quality, and turnover rates as high as 78% in some settings such as Thailand. Beyond workforce instability, burnout compromises patient safety. Each 1-unit increase in emotional exhaustion has been linked to a 2.63-fold rise in reports of poor care quality, 30% increase in patient falls, 47% increase in medication errors, and 32% increase in health care–associated infections. Burnout is also associated with lower job satisfaction, worsening mental health, and increased intent to leave the profession. These findings underscore the urgent need for effective strategies to reduce stress and burnout in emergency care.

Objective: This study aimed to evaluate the effectiveness and effect size of a multimodal learning approach—Emergency Room Virtual Simulation Interprofessional Education (ER-VIPE)—that integrates medical movies, massive online open courses (MOOCs), and computer- or virtual reality (VR)–based simulations with co-debriefing for reducing burnout and stress among future health care professionals compared with approaches lacking co-debriefing or using only movies and MOOCs.

Methods: A single-blind, quasi-experimental study was conducted at a university hospital from August 2022 to September 2023 using a 3-group treatment design. Group A (control) participated in a 3D computer-based, simulation-based interprofessional education (SIMBIE) without debriefing. Group B received the ER-VIPE intervention. Group C received the same as Group B, but the computer-based SIMBIE was replaced with 3D VR-SIMBIE. SIMBIE activities simulated a COVID-19 pneumonia crisis. Outcomes included the Dundee Stress State Questionnaire (DSSQ) and the Copenhagen Burnout Inventory, with trait anxiety as a behavioral control. Stress and burnout were measured at baseline, pre-intervention, postintervention, and 1-month follow-up. Generalized estimating equations were used to analyze group differences, with statistical significance set at $P<.05$.

Results: We randomized 87 undergraduate students from various health programs into the 3 groups (n=29 each). Participants' mean age was 22 years, with 71% (62/87) as women. After the 1-month post-SIMBIE follow-up, adjusted analyses revealed positive trends in DSSQ-engagement across all groups, with Group B showing a significant increase compared with Group A (mean difference=3.93; *P*=.001). DSSQ-worry and DSSQ-distress scores decreased nonsignificantly across all groups. Burnout scores also improved across groups, with Group B having a significantly lower score than Group A (mean difference=–2.02; *P*=.02). No significant burnout differences were found between Group C and Groups A or B.

Conclusions: A multimodal learning approach combining medical movies, MOOCs, and 3D computer-based SIMBIE with co-debriefing effectively improved engagement, reduced stress, and lowered burnout among future health care professionals. This scalable educational framework may help enhance well-being and resilience in high-pressure clinical environments.

(JMIR Med Educ 2025;11:e70726) doi:10.2196/70726

KEYWORDS

emergency medicine; stress; anxiety; burnout; interprofessional education; virtual reality; simulation; medical movies; MOOCs; simulation-based interprofessional education; emergency room virtual interprofessional education

Introduction

Background

Prior to the onset of the COVID-19 pandemic, burnout was already a significant concern among emergency health care professionals. For example, 60% of emergency medicine (EM) physicians [1] and 26% of emergency nurses [2] reported experiencing burnout on a regular basis. Although no quantitative, emergency department–specific studies from Thailand exist for the pre-COVID-19 period, broader national surveys have documented substantial levels of burnout among Thai nurses. In a study of more than 2000 nurses working in community hospitals, 32% reported high emotional exhaustion, 18% reported high depersonalization, and 35% experienced low personal accomplishment [3]. Notably, Nantsupawat et al [3] found that each 1-unit increase in emotional exhaustion score was associated with a 2.63-fold increase in the likelihood of reporting fair or poor quality of care, 30% increase in patient falls, 47% increase in medication errors, and 32% increase in infections. Supplementing this quantitative evidence, a qualitative study by Yuwanich et al [4] identified unique emergency department–specific stressors, such as patients' and their relatives' behaviors as well as power imbalances, affecting nurses in Thailand. Together, these baseline findings underscore that emergency health care workers (HCWs) were already highly vulnerable to burnout prior to the COVID-19 pandemic.

Working in an emergency room during crises or pandemics under tight time constraints, limited experience, and with newly formed teams is highly stressful. HCWs face heavy workloads, interpersonal conflicts, limited social support, and high-acuity patients [5]. The need for rapid, complex decision-making in this unpredictable environment intensifies stress, contributing to high burnout rates, especially during COVID-19 [6-8]. A systematic review reported that the prevalence of overall burnout among HCWs during the COVID-19 pandemic was high, reaching 43%, with 35% of EM HCWs identified as being at high risk for burnout [9]. The profession type—such as nurse, physician, or resident—was a significant factor influencing the rate of burnout and its domains, whereas gender did not show a consistent association. Among EM professionals, nurses were the most affected, followed by physicians [9]. The psychological impact of the COVID-19 pandemic extended beyond burnout, encompassing a wide range of adverse mental health outcomes, including stress, anxiety, depression, inadequate sleep, posttraumatic stress disorder symptoms, and secondary trauma [10]. Anxiety, depression, stress, and burnout levels were notably higher among physicians and nurses than in other health care roles, consistent with findings from other systematic reviews [9,11]. Gualano et al [11] further identified several pandemic-related factors associated with increased burnout risk, such as resource shortages, fear of COVID-19 infection, and social stigma. Consequently, the prevalence of turnover intention among emergency nurses has been reported at 45% globally [12], with a notably higher rate of 78% observed in Thailand [13]. Thailand is currently facing a critical nursing workforce shortage, reflected in a nurse-to-population ratio of approximately 1:400. Projections estimate an annual attrition of around 7000 nurses [13].

The COVID-19 health care crisis has profoundly impacted frontline health care professionals, exposing them to cumulative traumatic experiences that elevate stress and burnout levels. Frontline workers consistently report higher emotional exhaustion and depersonalization compared with those in non-COVID-19 units. Resilience and adaptive defense mechanisms are crucial for mitigating these effects, while younger age, female gender, increased COVID-19 exposure, and less resilient coping strategies predict greater vulnerability to stress and burnout. These findings highlight the urgent need for support programs focused on resilience-building and stress management for HCWs directly involved in COVID-19 patient care [14].

Health care workforce retention is a growing global concern, particularly in the postpandemic context. Depression and anxiety also positively correlate with higher absenteeism rates [15], reduced job commitment [16], lower job satisfaction [17], more medical leave [18], and poorer quality of life [19]. A 2023 study in Thailand reported that the revised model fit the data and accounted for 45% of the variance in nurses' intention to leave [20]. Burnout was the strongest factor, affecting intention both directly and indirectly via job satisfaction and professional commitment. Work-family conflict and the nursing practice environment influenced intention indirectly through these pathways. Thailand also faces a physician shortage, with only 0.5 to 0.8 doctors per 1000 people—below the World Health

Organization (WHO) standard of 1:1000. The pandemic exacerbated these challenges, leaving many frontline providers with burnout and limited mental health support [21]. These findings highlight the urgent need for targeted retention strategies. Thailand's experience offers insights for global health systems aiming to reduce burnout and strengthen workforce resilience.

Prolonged stress and burnout among HCWs have serious consequences for both individual well-being and the health care system. These conditions contribute to a range of physical and mental health issues, including fatigue, anxiety, social isolation, depression, and an elevated risk of suicide [22,23]. They also adversely affect patient care, leading to decreased quality, increased medical errors, patient dissatisfaction, and reduced job satisfaction [22,24-26]. Furthermore, stress and burnout are linked to higher turnover intentions and career disengagement [22,27]. Consistent with these findings, a study in Thailand reported that job stress significantly influences physicians' intentions to resign from their roles in health care [28]. These findings underscore the critical need for effective strategies to protect HCWs' well-being and maintain high-quality patient care. The COVID-19 pandemic underscored the critical importance of strengthening preparedness for future pandemics and health crises. A key lesson involves the need to enhance the effectiveness and efficiency of HCWs, which aligns with the first objective of the *Global Strategy on Human Resources for Health: Workforce 2030*, issued by WHO in 2016 [29].

Prior Work

Overview

A comprehensive literature review identified and analyzed existing studies on burnout and its prevention strategies, characterizing burnout as a chronic response to prolonged stress in work or academic settings. The review emphasized online learning as a key approach to mitigating burnout, highlighting the importance of providing emotional support; enhancing educator training; and using diverse, interactive tools to foster student engagement and motivation [30]. In addition, several strategies aimed at enhancing collaboration to combat stress and burnout in health care, including interprofessional education (IPE), medical movies, massive open online courses (MOOCs), and simulation-based interprofessional education (SIMBIE) incorporating co-debriefing and psychosocial support, will be discussed in detail in the following sections.

IPE to Enhance Collaboration as a Strategy to Combat Stress and Burnout in Health Care

Among various methodologies, such as mentorship and supervision [31,32], improving collaborative practice has demonstrated success for boosting HCW effectiveness and efficiency. Collaborative practice is when "multiple health workers from different professional backgrounds provide comprehensive services by working with patients, their families, careers and communities to deliver the highest quality of care across settings" [33]. It enhances effectiveness and efficiency by focusing on 4 key competencies: values and ethics, roles and responsibilities, communication, and teams and teamwork [34]. These competencies foster effective communication, satisfaction with high-quality teamwork, interprofessionalism, and a positive attitude with mutual respect for diverse health care disciplines [35,36]. Therefore, collaborative practice is crucial for enhancing patient safety in emergency departments [37-40]. Studies suggest that collaborative practice skills should be developed early in professional education [41-44].

IPE, where "two or more professions learn about, from and with each other to enable effective collaboration and improve health outcomes" [33], is more effective for teaching collaborative practices than other methods, such as traditional lectures [45], meeting discussions [42], video-based education [46], or simulation-based education involving only one profession [46]. IPE fosters positive perceptions of interprofessionalism through experiential learning, helping HCWs develop the 4 key competencies essential for collaborative practices. For over a decade, IPE has been integrated into university studies to promote teamwork [47,48], enhance understanding of professional roles and responsibilities [49], strengthen communication skills [50,51], and expand interprofessional knowledge [52,53].

A meta-analysis review study by Sezgin and Bektas [54] found IPE significantly improves communication competency (n=7; 95% CI 0.26 to 0.82; $P<.001$) and teams-and-teamwork competency (n=4; 95% CI 0.25 to 0.56; $P<.001$) among HCWs. However, the analysis included only 8 randomized controlled trials (RCTs)—too few for subgroup analysis of key intervention characteristics. Similarly, the meta-analysis by Marion-Martins and Pinho [55] showed IPE enhances values-and-ethics competency by fostering cross-professional mutual respect ($P=.007$; n=1) and improves roles-and-responsibilities competency (n=2; $P=.004$), yet the limited number of studies in each analysis restricts the generalizability of these findings. Overall, meta-analyses have shown that IPE interventions significantly improve health care systems (12 studies: standardized mean difference [SMD]=1.37, 95% CI 0.92 to 1.82 [56]; 6 studies: mean difference 7.19, 95% CI 2.61 to 11.77 [55]). A recent scoping review also highlighted IPE's positive impact on organizational culture, climate, and staff attachment [57]. Additionally, IPE initiatives improve health care professionals' work environments and strengthen multidisciplinary team effectiveness [58].

Based on previous evidence-based studies, we can conclude that IPE is an effective methodology for fostering collaborative practices, indirectly alleviating workplace stress, reducing burnout, and decreasing turnover intentions while enhancing HCWs' well-being, career satisfaction, and perceived service quality. Despite these benefits, Frenk et al, for The Lancet Commissions [59], criticized health education curricula for being fragmented and focused on a single profession, failing to prepare students for team-based clinical practice [59]. Traditional siloed education persists, hindering collaboration and patient safety [60].

Medical Movies to Enhance Collaboration as a Strategy to Combat Stress and Burnout in Health Care

The integration of films and television series into medical education, a practice known as cinemeducation [61], has been demonstrated to effectively enhance learning outcomes [62,63].

Cinemeducation has been widely implemented across various disciplines, including medical diagnostics [64], nursing [65], pharmacology [66,67], psychiatry [68-70], and psychology [71,72]. Importantly, cinemeducation extends beyond the mere screening of films in classrooms; it is grounded in a structured pedagogical framework that involves a sequence of carefully designed steps before, during, and after the educational activity [73,74]. Surprisingly, to the best of our knowledge, no study has examined the direct relationship between cinemeducation and stress or burnout among HCWs. Most studies investigated the effectiveness of cinemeducation for alleviating stigma, one of the greatest barriers to mental health treatment [75-79] that indirectly contribute to levels of stress [80,81] and burnout [82,83].

The study by Zeppegno et al [84] indicated that cinemeducation has the potential to reduce stigma, foster positive attitudes toward psychiatry, and enhance students' ability to manage anxiety when confronted with others' distress, with effects lasting up to 6 months. The pilot study by Rehl et al [85] demonstrated the effectiveness of cinematic virtual reality (cine-VR) training—a combination of cinemeducation and VR—for reducing stigmatizing attitudes among osteopathic medical students toward patients with opioid use disorder. The study also reported increased empathy, aligning with Vygotsky's theory of learning, which emphasizes learning through collaboration with others using reflection and authentic activities in real-life situations [86]. Similarly, the study by Kontos et al [87] used a filmed version of a research-based theatrical production to reduce caregivers' stigmatizing toward dementia. The study by Hawke et al [88] demonstrated that film-based interventions could reduce stigma among health care service providers toward individuals with bipolar disorder, with effects persisting for 1 month. Additionally, the pilot study by Linton et al [89] found that the film "Wounded Healer" significantly reduced stigma toward mental illness among health care students. Based on the reviewed studies, there is no doubt that cinemeducation is an effective approach for reducing stigma among HCWs and students.

MOOCs to Enhance Collaboration as a Strategy to Combat Stress and Burnout in Health Care

MOOCs are characterized as (1) "massive," enabling access to thousands of learners; (2) "open," with no enrollment fees; (3) "online," delivered via the web; and (4) "courses," offering structured content aligned with specific learning objectives [90]. They were first introduced in 2008 by Stephen Downes and George Siemens [91]. MOOCs have since revolutionized distance learning, driven by growing demand for flexible, accessible education [92,93]. The COVID-19 pandemic further accelerated their adoption, making MOOCs a central focus in education, with over 16,300 courses offered by 950 universities and more than 180 million enrollments globally [94] on platforms such as Coursera, edX, FutureLearn, Thai MOOCs, and Udacity [95,96].

Pedagogically, MOOCs fall into 3 categories: cMOOCs (connectivist), xMOOCs (extension of something else), and iMOOCs (integrated) [97,98]. cMOOCs, based on Downes' connectivist principles, emphasize networked learning; peer interaction; and evolving, learner-driven content [99-101]. In contrast, xMOOCs prioritize instructor-led, content-centered learning with limited interaction [102]. iMOOCs, developed by Universidade Aberta, blend the collaborative, flexible features of cMOOCs with the structured, assessed nature of xMOOCs, encouraging self-directed learning, peer engagement, and artifact-based assessment (eg, presentations, videos, mind maps) [98]. Due to their scalability and accessibility, MOOCs are highly effective for educational interventions across diverse global audiences [103]. In health care education, MOOCs serve various roles—from promoting health literacy among the public (eg, dementia education [104]) to supporting just-in-time training during health crises [105] and enhancing HCWs' well-being and resilience against stress and burnout [106].

MOOCs are an effective platform for delivering evidence-based interventions—such as stress and burnout education and mindfulness training—to HCWs, improving resilience while reducing stress and burnout. Recently, Ricker et al [107] conducted a single-group cohort study evaluating the Physician Well-being Course, a 4.5-hour online program covering well-being fundamentals (sleep, nutrition, exercise, resilience, and mindfulness), followed by a 2-week self-selected resiliency activity (10 minutes daily). The course was offered to postgraduate year-1 residents who could join voluntarily. Among 87 enrollees, 53 (61%) completed the course, with meditation being the most frequently selected resiliency activity (36/53, 68%) and sleep being the most frequently reported wellness behavior (22/36, 61%). Postintervention assessments showed statistically significant improvements in emotional exhaustion, depersonalization, and resilience (P<.05, paired t test). The authors suggested a key strategy to obtain a relatively high completion rate was attributed to offering the course during the preresidency timing, when residents had additional time and energy. Limitations of the study were the lack of follow-up assessing long-term effects and a control group.

Peterson et al [106] conducted a single-group cohort study to evaluate a pilot program supporting nursing students with coping with stress and burnout during the COVID-19 pandemic. The intervention consisted of an 8-hour self-paced online course and a 1-hour Zoom support group addressing 7 objectives including stress recognition, crisis response, self-care planning, and psychological safety. Psychoeducational resources such as mindfulness and breathing exercises were embedded, and participants were encouraged to choose a "battle buddy" for peer connection. Of 360 enrollees, 224 completed both pre- and postcourse surveys. Significant improvements were observed in calmness, connectedness, coping capacity, and hopefulness (P<.001, paired t test), reflecting gains in psychological flexibility [108], a protective factor against mental health deterioration [109-111]. The program also enhanced a sense of agency [65] through shared experiences and personalized resilience strategies, contributing to reduced acute stress [112]. Notably, burnout risk decreased significantly (OR 0.58, 95% CI 0.4-0.9; P<.006), which the authors linked to reduced isolation and increased social connectedness—consistent with theoretical models that identify social support and community belonging as key protective factors [113]. However, the study

lacked a control group and follow-up data to assess long-term effects.

The asynchronous nature of MOOCs enables participants to engage with content at their own convenience and pace, thereby accommodating demanding clinical schedules. Additionally, MOOCs can support virtual peer groups, which promote shared coping strategies, social connectedness, and a sense of community. These elements are critical mechanisms for fostering psychological resilience and enhancing HCWs' sense of agency, ultimately contributing to a reduction in acute stress—even under challenging circumstances [112].

SIMBIE With Co-Debriefing and Emotional Psychosocial Support to Enhance Collaboration as a Strategy to Combat Stress and Burnout in Health Care

As elaborated in the IPE section, IPE is an effective method for fostering the 4 competencies essential for collaborative practices, resulting in the mitigation of stress and burnout among health care professionals. Among other educational strategies [114-118], a recent scoping review and related studies indicate that SIMBIE, defined as "when participants and facilitators from two or more professions are engaged in a simulated health care experience to achieve shared or linked objectives and outcomes" [119], is an effective method for teaching the 4 interprofessional competencies in the emergency department [120-123]. In addition to improving technical skills [118,124,125], evidence-based studies have also demonstrated that SIMBIE enhances nontechnical skills such as communication and teamwork [54,118,124-130], with effects sustained for up to 6 months [124]. This is particularly significant given the increasing recognition that nontechnical skills are critical determinants of patient safety and quality of care [131-133].

Additionally, studies demonstrated unique advantages of SIMBIE over other educational strategies for learning IPE. First, SIMBIE offers a safe [134,135], controlled [127,136], and realistic [121,137] environment where learners from various health care professions can practice collaboration and teamwork without posing any risk to real patients [44,127]. Second, SIMBIE enables repetitive and deliberate practice, allowing learners to refine their skills by learning from mistakes in a risk-free setting—something not always feasible in real clinical environments [138-143]. Third, SIMBIE facilitates active authentic experiential learning by immersing learners in realistic clinical situations that prepare them for real-world practice [137,144,145]. Unlike passive formats, substantial research in adult education arguably emphasizes that active participation significantly enhances learning effectiveness [118,146-152]. Authentic experiential learning of SIMBIE benefits its uses in stress inoculation (SIT) training. Couarraze et al [153] reported positive effects on stress, anxiety, and burnout of anesthesia and critical care workers after attending simulation training based on critical situation exposure: The effects lasted for at least 1 week. Fourth, SIMBIE can expose learners to rare and complex medical conditions that they may not encounter during typical clinical rotations [145,154], thereby equipping them to manage a broader range of clinical scenarios more effectively. Studies have shown that using simulation as a teaching strategy significantly reduces State-Trait Anxiety Inventory (STAI) scores, with this reduction persisting at a 1-week posttraining follow-up. These findings underscore the benefits of simulation-based education, particularly for residents in anesthesia and intensive care [153]. Similarly, Shamputa et al [155] highlighted the effectiveness of virtual IPE initiatives for fostering interprofessional collaboration (IPC), particularly during the COVID-19 pandemic.

Beyond economical [156], logistical [157], scalability [158], and resilience-related [159] benefits, virtual SIMBIE has shown potential for enhancing educational outcomes. A recent meta-analysis reported that virtual SIMBIE significantly improves students' clinical reasoning and performance [39]. A systematic review and meta-analysis by Jiang et al [151] found no significant differences between virtual and real-world SIMBIE in terms of improving knowledge, procedural skills, clinical reasoning, or communication skills. Liaw et al [160] reported that desktop VR–induced stress levels of medical and nursing students, measured using blood pressure and heart rate, were comparable to the level induced by face-to-face simulation with during simulated rapidly deteriorating patient situations. Similarly, Shamputa et al [155] highlighted the effectiveness of virtual IPE initiatives at fostering IPC, particularly during the COVID-19 pandemic. Although virtual SIMBIE offers standardized and immersive practice scenarios, development of tacit knowledge for clinical practice is facilitated through debriefing [161-164]. Students consistently rated debriefing as essential to their learning experience, a finding supported by prior research [165-167]. Therefore, it is imperative that facilitators of virtual SIMBIE receive appropriate debriefing training to effectively guide these critical reflective processes [93].

Debriefing in the context of simulation-based health care education refers to "the facilitated discussion between two or more individuals in order to guide reflection and review performance, with the intent of gaining insight and understanding such that future performance is improved" [168]. Debriefing is regarded as an interactive, bidirectional, and reflective conversation that involves some degree of facilitation—whether by a facilitator, multiple facilitators (co-debriefing [169]), or the learners themselves (self-guided debriefing [170])—to support the reflective learning process [168,171,172]. When conducted effectively, debriefing thus serves as a critical component of learning in SIMBIE [163,173-175]. Its importance lies in its role as a process of reflection-on-action [176,177], which is a central element of Kolb's [178] experiential learning cycle. Experience alone during simulation is insufficient to produce learning; rather, it is the intentional reflection on that experience that facilitates deeper understanding [164,173,174,179-182].

Attending debriefing sessions has been studied as a potential strategy for reducing stress and preventing burnout. Although a small study reported no significant effect on burnout scores, most participants appreciated the emotional and social support offered through these sessions and recommended debriefing as a beneficial approach for junior medical residents [183]. Attending debriefing sessions has been shown to significantly reduce the risk of burnout and alleviate work-related

posttraumatic stress among intensive care staff, even after accounting for resilience and other contributing factors [184].

Several studies demonstrated that team debriefing after critical events provides various aspects of job and personal resources [185]: for example, (1) psychosocial support [186-193], (2) improvement in teamwork and interprofessional relationships [190,191,194-197], (3) learning and performance improvement [191,193,196-198], and (4) team-based cultural enhancement [189,190,199]. The Job Demands-Resources (JD-R) model [185] hypothesizes that the resources—both job-related (eg, knowledge, skills, abilities, social support) and personal (eg, self-efficacy)—function as a protective "buffer" that mitigates the adverse impact of job demands on individual strain, thereby reducing the risk of burnout [185,200,201].

This hypothesis is supported by both qualitative and quantitative evidence. Qualitative studies highlight the benefits of SIMBIE and debriefing interventions for enhancing emotional support and team reflection [140,186,187,190,191,193,195,202-206]. Quantitative studies further support these findings. For example, although Gunasingam et al [183] found no statistically significant reduction in burnout scores, participants valued the emotional and social support gained from debriefing sessions. Similarly, Colville et al [184] reported that debriefing was significantly associated with reduced burnout and work-related posttraumatic stress among intensive care staff, even after adjusting for resilience and other factors. These findings underscore debriefing as a time-efficient, evidence-based strategy for mitigating burnout in both simulation-based and real-world clinical environments.

Theoretical Frameworks

Stress and Burnout

Stress experienced within teams engaged in collaborative practice can be broadly categorized into 2 types, based on the differential impact of stressors: individual stress and team stress [207-209]. One of the most widely accepted definitions of stress is provided by Lazarus and Folkman [210], who conceptualize individual stress as "a particular relationship between the person and the environment that is appraised by the person as taxing or exceeding his or her resources and endangering his or her well-being" [211]. Weaver et al [208] extended the definition to the team context, defining team stress as "a particular relationship between the team and its environment, including other team members, that is appraised by the team members as taxing or exceeding their resources and/or endangering their well-being." Fundamentally, team stress represents a distinct collective psychological construct, whereas individual stress reflects a psychophysiological phenomenon experienced at the individual level [207].

Although individual stress and team stress occur at different levels, they can exert reciprocal and cross-level influences on performance at both levels [209]. Cross-level studies have revealed how specific team-level stressors—such as role ambiguity [212], team climate [213], and team conflict [214]—affect individual team members' attitudes, behaviors, and emotional states in ways that are relevant to their performance. Conversely, studies have also demonstrated that individual-level stressors, including workload [215] and time pressure [216], can significantly impact team performance. For instance, Savelsbergh et al [217] found that excessive team-level quantitative workload negatively affects both individual and team performance by inhibiting team learning behaviors and further impairs individual performance indirectly through elevated individual workload [217]. Additionally, Kamphuis et al [218] reported that team-level intervention strategies can modulate the effectiveness of individual-level interventions aimed at enhancing individual performance. Collectively, these findings underscore that both individual performance and team performance are shaped by complex interactions between team-level stressors—reflecting characteristics of the team environment—and individual-level stressors—reflecting the attributes and experiences of individual team members.

Several studies have reported that prolonged exposure to occupational distress can ultimately lead to burnout [187,219-229]. This burnout induction process can be explained by the JD-R model [185], which is 1 of the 2 leading theoretical frameworks on burnout [230]. According to the JD-R model, burnout occurs when the availability of resources—both job-related (eg, knowledge, skills, abilities, social support) and personal (eg, self-efficacy)—is insufficient over time to buffer the negative effects of job demands on individual strain [185,200,201]. The model hypothesizes that increasing access to resources—physical (eg, assistance with wearing personal protective equipment [PPE]), psychological (eg, positive feedback from peers or supervisors), social (eg, peer support), or organizational (eg, stress-coping programs)—can help mitigate the risk of burnout.

Cinemeducation

Our cinemeducation was created on the basis of vicarious learning, which is the modification of an observer's behavior that is similar to that of a model by watching the model's behavior be reinforced or punished [231]. The effectiveness of cinemeducation stems from its ability to simplify the comprehension of complex situations by presenting real-life role-playing scenarios that illustrate both effective and ineffective practices. This approach enhances the understanding of human behaviors and their impact on colleagues and patients. Movies in cinemeducation serve as a unique and engaging learning tool [232,233], capable of stimulating debate [234] and providing insights into students' perspectives on topics that might otherwise remain unexpressed [235-237] or be overshadowed by technical considerations [238]. Cinemeducation thus transforms implicit aspects of human behavior, teamwork, and professional interactions into explicit, tangible learning experiences, fostering deeper comprehension, critical reflection, and meaningful discussions.

Furthermore, movies effectively evoke emotions that foster active participation and deeper learning [239]. They leverage vicarious learning by allowing viewers to emotionally connect with characters [240]. Emotional experiences tend to be more vividly remembered [241-243], facilitating vicarious learning among students, a process supported by the Yerkes-Dodson law [244]. Our arguments on the effectiveness of cinemeducation have been further supported by empirical evidence suggesting

that movies are effective at reducing stigma by fostering active learning through emotional engagement [245-247] and by imparting values centered on lived experiences [88]. Moreover, movies could represent a highly complex form of symbolic content that conveys nuanced and richly layered audiovisual information to students through storytelling. Beyond merely delivering content, cinema functions as both a tool and a reflective space, akin to psychotherapeutic sessions, allowing viewers to project aspects of their own psyche onto movies [248]. This process enables viewers to immerse themselves in the narrative, fostering self-reflection and deeper engagement. Additionally, cinema is believed to provide a contemporary experiential medium that momentarily detaches viewers from their daily lives, engaging their unconscious mind in a manner comparable to hypnosis and dreaming [249]. These immersive experiences contribute to the formation of subjective world views, a concept theoretically grounded in Jungian [250] and Hillmanian [251] perspectives on the relationship between images and archetypes. According to this framework, archetypal experiences—facilitated through engagement with cinematic imagery—support learning from both cognitive and emotional perspectives [252]. Such experiences encourage students to engage in discovery and self-reflection, aligning with the principles of inductive learning [253].

Cinemeducation for teaching professionalism as well as stress and burnout education was delivered to HCWs using xMOOCs. xMOOCs represent a practical and scalable solution to deliver interventions that mitigate stress and prevent burnout to health care professionals. The asynchronous nature of xMOOCs enables participants to engage with content at their own convenience and pace, thereby accommodating demanding clinical schedules.

SIMBIE and Co-Debriefing

Experiential and Phase Learning

In this study, the learning facilitated by SIMBIE and co-debriefing is grounded in experiential learning theory [178]. The experiential learning theory conceptualizes learning as an ongoing, cyclical process comprising 4 stages "whereby knowledge is created through the transformation of experience" [178]. The cycle begins with a concrete experience, which, in this context, is generated through SIMBIE, which provides a safe and controlled environment for repetitive and deliberate practice. This is followed by reflective observation and abstract conceptualization, both of which primarily happen during co-debriefing sessions. In the reflective observation stage, facilitators guide learners to examine their experiences from multiple perspectives, fostering deeper insight into the significance of their actions and decisions. These insights then serve as the foundation for abstract conceptualization, where learners attempt to synthesize and generalize their experiences into broader theories or hypotheses. The final phase, active experimentation, occurs when learners are encouraged to test their emerging understandings by engaging in subsequent simulated scenarios in SIMBIE. At this stage, learners apply their knowledge of "good" and "bad" responses, decisions, and actions to analogous situations. These new experiences then initiate another cycle of learning. Experiential learning theory underscores that meaningful learning does not arise from experience alone but rather from deliberate reflection on those experiences.

Additionally, the intervention program was developed based on the 3 sequential stages of phase learning [254], which are rooted in experiential learning theory [178] and principles of instructional scaffolding [255]: (1) prebriefing, (2) participation, and (3) co-debriefing. The prebriefing phase is defined as "a time when the facilitator illustrates the purpose of the simulation, the learning objectives, the process of debriefing, and what it entails" [172]. This phase aims to enhance learners' engagement and involvement [256]. Effective prebriefing must communicate to learners that they are entering a controlled and purposeful environment for reflective practice, where making errors is not only accepted but expected as part of the learning process [256]. Critically, psychological safety should be established from the outset—ideally during the very first interaction between facilitators and learners [135,168,257]. In the participation phase, the virtual SIMBIE environment enables learners to acquire concrete experiences by enacting prior knowledge about social and medical skills, clinical roles, and IPC. This occurs within a setting that is safe [134,135], controlled [127,136], and realistic [121,137]. These conditions allow for repetitive and deliberate practice, which supports skill refinement through learning from mistakes in a low-risk environment—an opportunity often not feasible in actual clinical practice [138-143]. Following participation, the co-debriefing phase facilitates learners' reflective learning through a bidirectional, interactive dialogue, often cofacilitated by multiple instructors [168,171,172]. When implemented effectively, co-debriefing serves as a critical component of learning through SIMBIE [163,173-175]. The concrete experiences gained during simulation, while necessary, are not sufficient for learning; it is the intentional reflection on those experiences that enables learners to derive deeper understanding [173]. Through this reflective process, learners engage in observation and abstract conceptualization, promoting active experimentation and eventual behavioral change—core mechanisms described in experiential learning theory [164,174,179-182].

Theoretical Framework for SIMBIE

In addition to constructivism, experiential learning theory, situated learning theory, and andragogy, as outlined in [258], our SIMBIE approach is also grounded in resilience theory. The term "resilience" is derived from the Latin word resilire, meaning "to leap back" [259]. Resilience has been explored across a wide range of scientific disciplines—including engineering [260], ecology [261], psychology [262], and health care [263]—resulting in varied definitions and conceptualizations [259,262]. Even within the field of psychology, perspectives differ: Some researchers define resilience as the capacity for positive adaptation in the face of significant adversity [264], whereas others view it as the ability to maintain a stable equilibrium under stress [265]. Despite these variations, most definitions converge on 2 core elements: positive adaptation and the presence of adversity [266]. Given the conceptual ambiguity, our study does not aim to redefine resilience but instead focuses on evaluating whether proposed resilience factors (such as the 4 competencies of collaborative

practice) demonstrably confer resilience. To guide this inquiry, we adopted the bidimensional framework for resilience research [267], which categorizes influencing factors into 2 domains: intrinsic factors (resilience and risk factors internal to the individual or team) and external factors (protective and environmental risk factors).

One of the most compelling findings in resilience research, which informs our SIMBIE design, is the concept of the steeling effect—the idea that resilience can be cultivated through exposure to manageable risk rather than through avoidance of all adversity. There is increasing recognition that an appropriate level of exposure may be essential for organizing, calibrating, or "tuning" the adaptive systems of a unit, such as an individual [268,269] or a team [270], in preparation for future, more intense, unpredictable adversity. Crucially, the type, degree, and duration of the exposure must remain within a range that is manageable for the unit. If the exposure exceeds the unit's adaptive capacity, the resulting impact differs depending on the nature of the unit. In individuals, unmanageable exposure may lead to the sensitizing effect, increasing vulnerability to future stressors [271]. In contrast, teams subjected to overwhelming adversity may experience a breakdown in collective identity, with members becoming increasingly individualistic and self-protective, thereby eroding team cohesion and effectiveness [272]. The steeling effect is consistent with the broader understanding that coping with, engaging in, and confronting challenges—rather than avoiding them—are fundamental to adult growth, learning, and development [273]. Persistent avoidance of adversity in individuals can result in reduced self-efficacy, low adaptive flexibility, and the stagnation of higher-order cognitive and emotional development, ultimately limiting self-actualization [273]. At the team level, avoidance may lead to increased brittleness and decreased tolerance to disruption, which can compromise safety and elevate the risk of harm during future crises [274]. Furthermore, such avoidance can hinder a team's ability to identify and revise latent systemic threats, which, when accumulated, can significantly impair team performance [275].

Theoretical Framework for Co-Debriefing

Learners' engagement in co-debriefing within this study was grounded in 3 interrelated theoretical frameworks: (1) social constructivism [276], (2) zone of proximal development [277], and (3) transformative learning theory [278]. The theory of social constructivism emphasizes the centrality of social interaction in collaborative learning. It conceptualizes learning as an active, meaning-making process shaped by learners' engagement with lived experiences [276,279]. Meaning, in this view, is coconstructed through dialogue and shared reflection with peers. Learning is thereby characterized as active, constructive, self-controlled, social, and situational [280]. Complementing this perspective, Vygotsky's [277] zone of proximal development further explains that learning is optimized when learners are supported in tasks slightly beyond their independent abilities, guided by a more knowledgeable other such as a facilitator. In co-debriefing, facilitators assume this role by guiding reflective discussions that help learners bridge knowledge gaps and deepen clinical reasoning and team collaboration.

Transformative learning theory emphasizes the transformation of learners' "frames of reference"—the existing ready-made meanings for interpreting experiences [281]. These frames consist of deeply held assumptions that guide how learners perceive, understand, and respond to situations. Frenk et al [59] identified transformative learning as the most advanced of 3 successive learning levels—informative, formative, and transformative—with the latter aiming to fundamentally shift perspectives, rather than merely convey knowledge or instill professional behaviors [59]. According to Mezirow [278], transformation occurs when learners revise their frames of reference to become "more inclusive, discriminating, open, emotionally capable of change, and reflective so that they may generate beliefs and opinions that will prove more true or justified to guide action." This transformation unfolds through 3 key processes. First, disorienting dilemmas, such as challenging simulated scenarios in SIMBIE, disrupt learners' assumptions [278]. Second, critical reflection enables them to question and evaluate these assumptions [282], going beyond technical methods to examine underlying reasoning. Third, reflective discourse involves dialogic engagement, allowing learners to explore alternative perspectives and collaboratively seek mutual understanding in a psychologically safe environment [278]. In co-debriefing, facilitators foster this safety, encouraging open dialogue that supports both individual and collective transformation through shared reflection of their peers, making the transformation both individual and collective.

Psychological Safety

Co-debriefing in this study was guided by the principle of psychological safety to enhance its effectiveness and efficiency [283]. Psychological safety is recognized as a foundational—and even essential—condition for effective debriefing [120,171,284]. It fosters open dialogue by creating an environment where participants feel safe to share experiences and emotions without fear of judgment, embarrassment, or punishment [173,256,285,286]. This environment supports creativity, engagement, speaking up, and learning [283,287,288] while reducing face-saving behaviors such as withdrawal or avoidance of critique [164,284].

To foster psychological safety, facilitators use both explicit strategies, such as setting clear objectives, ensuring confidentiality, and choosing appropriate settings [289], and implicit cues, including respectful tone and open body language [284]. Quiet, private simulation spaces and thoughtful physical arrangements further support this environment [164,290]. However, co-debriefing interprofessional groups poses unique challenges due to differences in participants' backgrounds, experiences, professional identities, and learning goals [164,291,292]. These complexities extend to co-debriefers themselves, who may also differ in training, experience, and facilitation styles [163,169,293]. Maintaining emotional and psychological safety amid such diversity can be difficult, especially when power imbalances are present [284,286].

Power dynamics—how authority and influence affect interpersonal interactions [294-296]—exist in both clinical settings [297,298] and simulations like SIMBIE [120]. If unaddressed, power imbalances may suppress participation and

hinder the development of interprofessional competencies [292,299]. Despite their importance, power dynamics are often avoided in co-debriefings, possibly because the topic is perceived as taboo [164,289]. Given SIMBIE's goal of enhancing real-world collaborative practice [33], explicitly addressing power imbalances during debriefing may strengthen psychological safety and improve learning outcomes [299]. Power imbalances may also arise among co-debriefers. Poorly managed dynamics can lead to tension, miscommunication, and perceived hierarchies [169,299]. These issues may stem from positional power (based on role) or expert power (based on knowledge) [300,301]. In practice, one debriefer may dominate discussions, interrupt others, or address only one profession, undermining the value of co-debriefing and diminishing diverse perspectives [163,302]. To fulfill SIMBIE's interprofessional goals, co-debriefers must be mindful of these dynamics and intentionally foster equitable, respectful collaboration that models the competencies they aim to teach.

Goal of This Study

Although prior research supports the individual benefits of cinemeducation, SIMBIE with co-debriefing, and MOOCs for emergency health care professionals, these modalities have primarily been studied in isolation. A few studies have examined integrated approaches—such as combining simulation with cinemeducation—but focused on different contexts and outcomes [74,85,303]. To the best of our knowledge, no study has integrated medical movies, MOOCs, and virtual SIMBIE with co-debriefing as a combined strategy. This study aimed to explore their integration as an IPE approach to enhance collaborative practice, mitigate stress, and prevent burnout among health care professionals. This study was intended for medical educators and other health care education professionals—including nurses, pharmacies, and allied health educators—as well as policymakers involved in health professions education and curriculum development. However, a quasi-experimental design was adopted due to practical constraints in randomizing participants within a clinical setting during the COVID-19 pandemic.

Therefore, to address these gaps in the literature and build upon the promising effects of individual strategies, the goal of this study was to evaluate the effectiveness and effect size of a multimodal learning approach—grounded in the aforementioned theoretical frameworks—that integrates medical movies, MOOCs, and either computer-based or VR-based simulation with co-debriefing, collectively referred to as Emergency Room Virtual Simulation Interprofessional Education (ER-VIPE), at improving self-reported stress levels and reducing burnout among future health care professionals. Stress and burnout were measured using the Dundee Stress State Questionnaire (DSSQ) and the Copenhagen Burnout Inventory (CBI), respectively. This quasi-experimental study compared the outcomes of the ER-VIPE intervention with alternative approaches, including (1) computer-based simulation without co-debriefing, (2) VR-based simulation without co-debriefing, and (3) medical movies and MOOCs alone. We hypothesized that the ER-VIPE multimodal learning model would be more effective at reducing stress and preventing burnout than any single-component or nondebriefing method.

Methods

Research Design and Setting

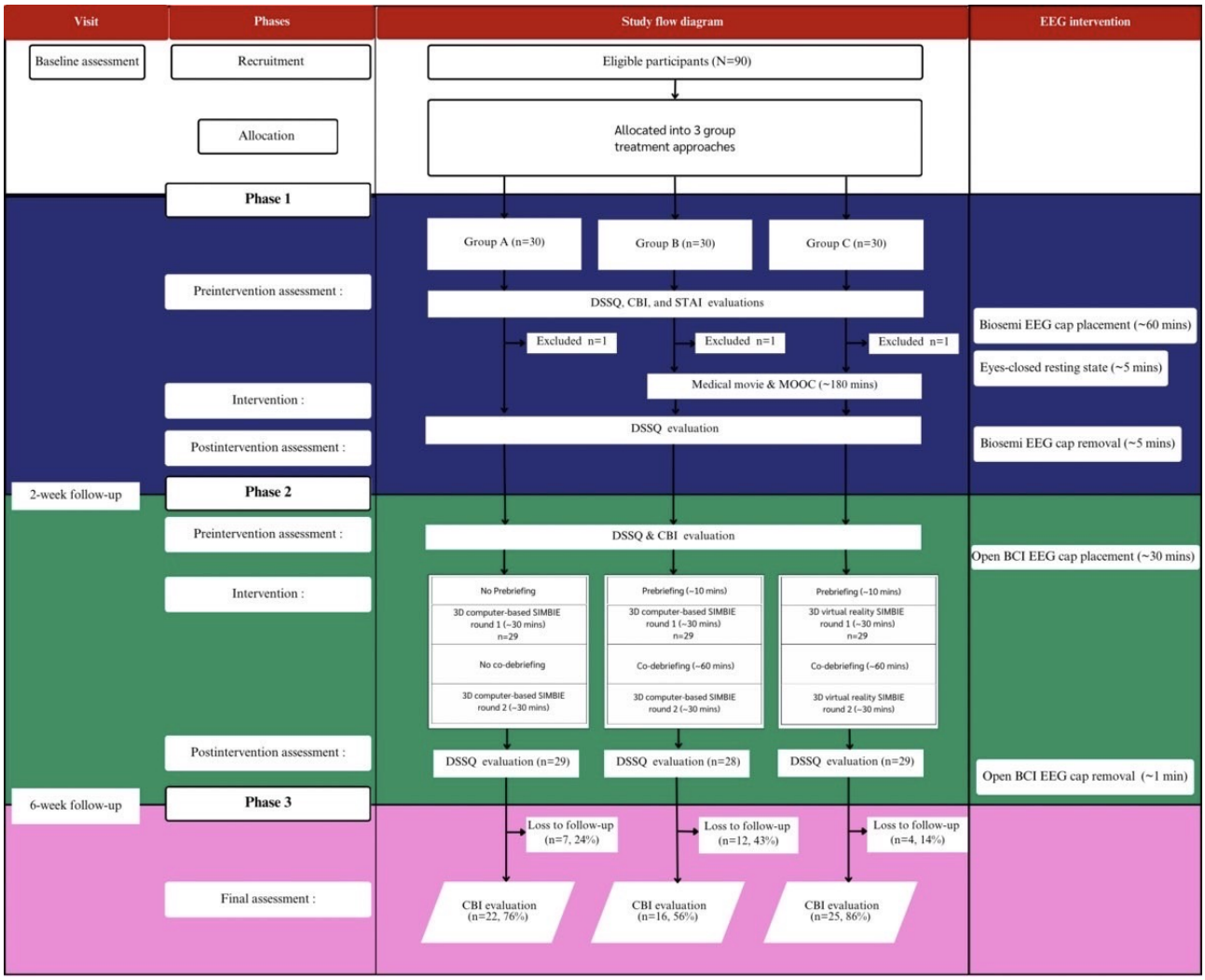
A quasi-experimental study with a single-blinded statistician was conducted to compare stress and burnout levels among health care professional students who were divided into 3 groups: 2 groups received novel educational interventions, and 1 group served as the control group with traditional learning methods. The study was conducted at a 1500-bed university-affiliated hospital in Bangkok, Thailand, which serves as a training site for multiple health care professional programs. A quasi-experimental design was used due to the practical challenges of randomization in a clinical setting during the COVID-19 pandemic.

Participants and Sampling

After institutional review board approval, undergraduate clinical students from 5 health care disciplines (medicine, nursing, pharmacy, radiologic technology, and medical technology) were recruited via announcements, Line, and posters and provided informed consent. Interested students enrolled through a QR-linked Google Form. The principal investigator's contact was provided for inquiries. Participation was voluntary and scheduled outside of regular academic activities to avoid disruption.

Eligibility criteria included healthy individuals aged 18 years to 25 years who were enrolled as 5th-year medical students, 5th- or 6th-year pharmacy students, 4th-year medical technologist students, or 3rd- or 4th-year nursing students or radiological technologist students. The population size included approximately 533 undergraduate clinical students across 5 health care professions: medicine, nursing, pharmacy, medical technology, and radiological technology. A total of 147 students expressed interest and registered to participate in the study. Exclusion criteria included substance use (eg, smoking), a history of neurological or psychiatric disorders, Patient Health Questionnaire-9 (PHQ-9) score ≥9 [304], regular use of antidepressants, and belonging to vulnerable groups, such as pregnant individuals or students with severe illnesses. Moreover, since the potential impact of SIMBIE-induced stress was uncertain during this initial phase of the study conducted amid the COVID-19 pandemic, we prioritized participant safety. There was a concern that SIMBIE could induce stress levels exceeding the optimal arousal threshold for learning, as described by the Yerkes-Dodson law [244]. Participants were also excluded if they missed 2 scheduled appointments or failed to comply with study preparation requirements on 2 occasions (see Figure 1).

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) diagram [305] and participant flow throughout the study, which used a 3-group treatment design. BCI: brain-computer interface; CBI: Copenhagen Burnout Inventory; DSSQ: Dundee Stress State Questionnaire; EEG: electroencephalography; MOOC: massive open online course; PHQ-9: Patient Health Questionnaire-9; RT: radiological technology student SIMBIE: Simulation-Based Interprofessional Education; STAI: State-Trait Anxiety Inventory.



The study recruited undergraduate clinical-level students from the faculty of medicine, the institute of nursing, the faculty of pharmaceutical sciences, and the faculty of allied health sciences. Of the 90 students who initially met the inclusion criteria, 3 radiological technology students were excluded due to PHQ-9 scores exceeding the clinical threshold for depression, resulting in a final sample of 87 participants. The sample included 15 medical students, 30 nursing students, 15 pharmacy students, 15 medical technologist students, and 12 radiological technologist students. These exclusions were made to minimize potential confounding effects and ensure that observed outcomes could be more accurately attributed to the intervention.

Participants were allocated into 3 groups (A, B, and C) using a stratified convenience sampling method to ensure balance in baseline characteristics such as age, education level, and prior clinical experience. Each group was further divided into 5 interprofessional subgroups, with each subgroup consisting of 1 medical student, 2 nursing students, 1 pharmacy student, 1 medical technologist student, and 1 radiological technologist student.

Group A (control) participated in a 3D computer-based SIMBIE without oral debriefing. Group B received a medical movie, a MOOC, a 3D computer-based SIMBIE, and an oral co-debriefing session. Group C received a medical movie, a MOOC, a 3D VR SIMBIE, and an oral co-debriefing session. Stress levels were assessed using the DSSQ at 4 time points: before and after the medical movie and MOOC sessions (intervention phase 1) and before and after the SIMBIE sessions (intervention phase 2). Burnout levels were measured using the CBI at 3 time points: prior to the medical movie and MOOC sessions (intervention phase 1), prior to the SIMBIE sessions (intervention phase 2), and during the follow-up assessment conducted 4 weeks after completing intervention phase 2 (phase 3). The flow of electroencephalogram procedures is also shown in Figure 1 (also see Multimedia Appendix 1), explaining how this potential confounding factor varied across groups. All times reported were approximate.

Data Collection

Data were collected between August 2022 and September 2023. Baseline assessments included demographic information, the DSSQ to measure stress, and the CBI to evaluate burnout as the

primary outcomes. Trait anxiety was assessed once at baseline using the STAI. All questionnaires were administered online via Qualtrics. Data were collected in 3 phases. Phase 1 included DSSQ assessments before and after participants watched a medical movie and completed a MOOC, with only a pre-intervention CBI assessment conducted. Phase 2 involved DSSQ assessments before and after the SIMBIE activity, conducted 2 weeks later, along with a pre-activity CBI assessment. Phase 3 was a follow-up burnout assessment using the CBI, conducted 4 weeks after the completion of all activities. To minimize bias, data collection and analysis were performed by a single-blinded statistician. The timeline and details of the study are summarized in Figure 1.

In addition, the online survey has been reported in accordance with CHERRIES (Checklist for Reporting Results of Internet E-Surveys), which is provided as Multimedia Appendix 2. Although this study is not an RCT and thus does not require a trial registration number, we completed the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth; V1.6.1) submission/publication form to ensure transparency and completeness in reporting. This form is included as Multimedia Appendix 3.

Traditional Learning Approaches

The traditional learning approach involved online, lecture-based instruction provided to health care students during the COVID-19 pandemic within a uniprofessional educational framework. These methods did not include any curriculum nor intervention specifically designed to reduce stress, build resilience, or promote interprofessional team training. Group A, serving as the control group, received no IPE exposure nor preparatory learning activities beyond the traditional lecture format. Specifically, they did not participate in lectures or cinemeducation on interprofessional collaboration, role-play activities, or pair-and-share exercises. Additionally, Group A did not undergo any preparatory learning before the simulation. Instead, they virtually engaged in a 3D computer-based SIMBIE scenario with other participants in group A and without oral co-debriefing, which served as the traditional learning condition in this study.

Experimental Groups and Interventions

This study used a 2-arm design with Groups B and C as the experimental groups. Both groups underwent a series of interventions, including a combination of medical movie, MOOC, and 3D SIMBIE simulation, followed by a co-debriefing session, collectively referred to as ER-VIPE. Group B used a computer-based 3D SIMBIE for approximately 30 minutes, while Group C used a 3D VR SIMBIE for the same duration. Both groups participated in a 60-minute oral co-debriefing session (see Figure 1). The medical movies used in the simulations were standardized across all participants to ensure consistency in content. Similarly, the MOOC materials were delivered and maintained uniformly throughout the study to ensure consistent learning experiences. In addition, the simulation scenarios were standardized so that all participants were exposed to the same conditions and challenges, thereby minimizing potential variability.

Multimodal IPE Design Based on the SIT Framework

For the experimental groups, we implemented SIT [194] to induce the steeling effect through the integration of cinemeducation, coping-with-stress strategies via MOOC, SIMBIE, and co-debriefing. SIT is a structured cognitive behavioral intervention aimed at fostering resilience by equipping individuals with the tools needed to effectively manage stress. The approach consists of 3 interrelated and overlapping phases: conceptualization, skills acquisition and consolidation, and application and follow-through.

In the conceptualization phase, participants were introduced to the cognitive, emotional, and behavioral manifestations of stress using cinemeducation through medical movies and coping-with-stress strategies. A flipped classroom approach via cinemeducation and MOOC were used to optimize preparedness for the coming SIMBIE sessions [306]. Presession engagement with medical content via xMOOC platforms supported the development of interprofessional knowledge and encouraged positive attitudes IPE, Team Strategies and Tools to Enhance Performance and Patient Safety (TeamSTEPPS), and stress coping strategies [307]. This model enhanced learner engagement and emotional investment, thereby maximizing the effectiveness of in-session SIMBIE activities [308]. The integration of stress management into clinical education has been widely recognized as essential to the development of professional competencies [309].

In the skills acquisition and consolidation phase, learners developed a repertoire of coping mechanisms—including relaxation techniques; cognitive restructuring; Identify, Situation, Background, Assessment, and Recommendation (ISBAR) [49]; closed-loop communication; and collaborative communication strategies—within SIMBIE sessions that approximate real-world clinical stressors, such as time constraints and multisensory demands (visual and auditory stimuli). Skills consolidation was reinforced through co-debriefing, which provided structured opportunities for critical reflection and feedback.

In the final application and follow-through phase, learners applied these skills during a second round of SIMBIE and were progressively intensified to balance cognitive load and emotional challenge while maintaining learners within their zone of proximal development. SIT not only supports the development of self-regulation and problem-solving capabilities but also enhances self-efficacy by helping participants reframe stressors as challenges rather than threats. Moreover, it fosters a psychologically safe learning environment in which learners can engage with high-stakes situations without fear of failure—ultimately promoting more adaptive coping strategies, reduced anxiety, and improved performance under pressure.

Medical Movie (Cinemeducation)

Participants viewed a 75-minute medical movie developed in-house that depicted an interprofessional emergency team managing high-stress scenarios. The film emphasized communication, shared mental models, team reasoning, team support, and collaborative problem-solving for stress mitigation. Viewing was conducted individually and asynchronously, with

no postfilm group discussion. Although cinemeducation is evidence-based for reducing stigma [84,85,88], its impact on stress or burnout has not been previously studied.

MOOCs

The MOOCs comprised 7 lessons covering essential topics, including Interprofessional Education Collaborative core competencies for IPC practice, TeamSTEPPS principles, team-based clinical reasoning for diagnosis, stress management and coping strategies, IPE for patient safety, and ethical principles for collaboration. A 15-minute segment focused specifically on stress management and coping strategies, addressing significant stressors faced by emergency department personnel, such as disease outbreaks, heavy workloads, and interpersonal challenges. These stressors were categorized as external (eg, workplace pressures and unexpected events) or internal (eg, personality traits like perfectionism and difficulties with work-life balance). The module emphasized problem-solving approaches over emotional reactions to challenges, aiming to mitigate the long-term effects of stress, including burnout symptoms like emotional exhaustion, depersonalization, and reduced personal accomplishment. All participants were required to complete both the medical movie and MOOCs as well as pass the pre- and posttest exams to receive a certificate and meet the eligibility criteria for the flipped classroom before attending the SIMBIE session.

Prebriefing Design

To promote psychological safety, we implemented structured prebriefing strategies based on best practices [135,168,257]. Facilitators and learners introduced themselves, shared prior experiences, and established ground rules emphasizing confidentiality, active participation, and a focus on performance improvement rather than individual critique [162,171,172,284]. Effective prebriefing must communicate to learners that they are entering a unique, controlled environment for reflective practice where making errors is acceptable and expected as part of the learning process [256].

A “fiction contract” was cocreated to encourage engagement and suspend disbelief despite simulation limitations [310-312]. Learners were oriented to the simulation space, equipment, and environment to reduce anxiety and enhance participation [182,313-315].

For co-debriefing, facilitators conducted prebriefing discussions to align objectives, align areas of expertise, clarify roles, agree on debriefing methods, clarify who will lead different elements, and address potential challenges [168,169]. This precoordination ensured a smooth, cohesive, and collaborative debriefing process and helped establish a shared mental model between co-debriefers [169]. Facilitators also attended to implicit cues—such as tone, expressions, and body language—to foster trust and a positive psychological climate [316-318].

3D SIMBIE Design

The 3D SIMBIE simulations, available in both computer-based and VR formats, were designed to progressively increase in complexity and stress, replicating high-acuity emergency scenarios reflective of real-world challenges during the COVID-19 pandemic. Each simulation integrated both technical and TeamSTEPPS-based nontechnical learning objectives in the context of a multidisciplinary emergency department encounter. Participants worked as a 6-member health care team to diagnose and manage a complex case involving a 70-year-old male patient with COVID-19, chronic obstructive pulmonary disease, hypertension, diabetes mellitus, and a documented allergy to ceftriaxone. The presence of a distressed spouse added an emotionally charged layer to the scenario, requiring participants to apply both clinical and interpersonal skills.

Technical training included time-sensitive interventions such as intubation, ventilator management, laryngeal mask airway insertion, cricothyroidotomy, hyperkalemic crisis, use of PPE, and coordination of portable chest imaging. Clinical reasoning was guided toward formulating an accurate diagnosis and implementing a safe treatment plan under pressure, closely mirroring the demands of real-world emergency care clinical environments and emphasizing decision-making under pressure, as illustrated by the multiple stressor events shown in Figures 2 and 3. Nontechnical training focused explicitly on the 5 TeamSTEPPS domains—team structure, communication, leadership, situation monitoring with the STEP framework, and mutual support—using strategies such as ISBAR, closed-loop communication, and real-time psychosocial support. Communication occurred via open microphones to simulate authentic interprofessional interaction, and learners engaged with the patient’s spouse through a branching dialogue system that enabled them to select empathetic responses in real time.

An interactive game-based interface further reinforced TeamSTEPPS competencies as our main learning objectives. Each participant was equipped with a HP bar, a symbolic “health point” indicator representing emotional resilience under pressure. The HP bar depletes over time, particularly during high-stakes decision points or delays in clinical action such as managing a “can’t intubate, can’t ventilate” situation, video laryngoscope battery failure, or sudden hypotension**.** Mutual support was emphasized as teammates could replenish each other’s HP by pressing a “plus” button—an interactive metaphor for peer support and collaborative coping. The HP bar thus served both as a visual representation of stress load and as a positive reinforcement mechanism for TeamSTEPPS-aligned behaviors. These features were designed to enhance psychological fidelity and simulate cognitive load realism by integrating both clinical complexity and emotional dynamics.

This study was conducted during the COVID-19 pandemic, during which all learners were limited to online lectures without clinical exposure. Given these constraints, along with the logistical challenges of assembling multiprofessional student teams and facilitators, we designed the study to ensure both research feasibility and standardization. To address safety concerns for both participants and researchers, a control group (Group A) was exposed to a 3D computer-based SIMBIE simulation without debriefing. All participants provided informed consent after being fully briefed on the study’s potential risks and benefits. Inclusion criteria required that participants have a PHQ-9 score within the normal range, to minimize psychological risk. Trained research assistants were present throughout the simulation to observe and provide support

as needed. Participants were also informed of their right to withdraw from the study at any time without penalty. These measures were implemented to uphold ethical standards and ensure fairness, psychological safety, and relevance to the context of the pandemic.

The 3D SIMBIE and 3D VR SIMBIE platforms were developed by our ER-VIPE team, which comprises professionals from various disciplines including emergency physicians, nurses, pharmacists, medical technologists, radiologic technologists, communication arts specialists, instructional designers, architects, psychologists, and experts in the humanities and education. This interdisciplinary team collaborated with a specialized group of engineers in immersive learning technologies for health care education. Both platforms use advanced simulation software to create realistic, interactive environments that enable participants to engage in high-stakes medical scenarios.

The key distinction lay in the level of immersion: The 3D version used standard computer input devices (eg, screen, keyboard, and mouse), while VR SIMBIE used headsets and controllers to deepen emotional engagement and presence. This allowed participants to physically move and interact within the environment in a fully immersive and realistic manner. The VR format enhances the sense of presence, making participants feel as though they are "inside" the clinical scenario, which may result in greater cognitive and emotional stress compared with the 3D desktop version. Figures 2 and 3 illustrate the user interface, visual comparisons, and experiential differences between both platforms. In addition, they show the health point (HP) bar, which can reach a maximum of 50 points, with the ability to restore a teammate's energy up to 5 times, adding 10 points each time. Each simulation concluded with a structured co-debriefing session, enabling participants to reflect on clinical decision-making and TeamSTEPPS performance, consolidating targeted nontechnical teamwork competencies.

Figure 2. Health point (HP) bar's features, functions, and examples of stressful situations encountered by health care professionals while practicing stress coping strategies in 3D computer-based simulation-based interprofessional education (SIMBIE): (A) triage nursing student entering for patient assessment, walking in, putting on a mask, assessing the patient, and measuring vital signs; (B) medical student struggling with unsuccessful endotracheal tube insertion due to a difficult airway; (C) nursing student connecting a patient to a ventilator; (D) a pharmacy student advising a patient's family against taking photos, emphasizing patient privacy and ethical confidentiality; (E) radiological technologist and nursing students collaborating to reposition the patient for a chest X-ray while monitoring for accidental dislodgment of the endotracheal tube; and (F) medical technologist student improperly removing PPE after working in a lab where COVID-19 was detected, triggering a system warning for the wrong sequence.

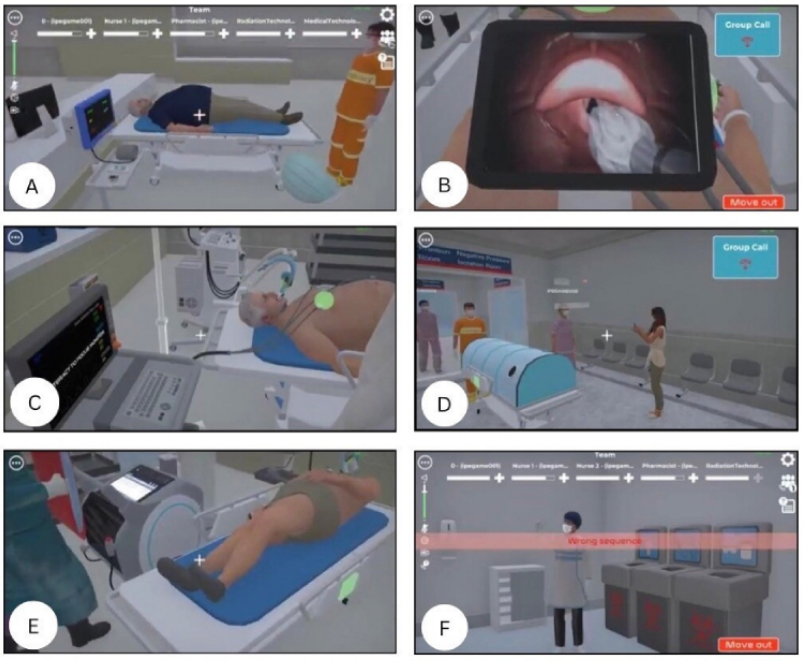
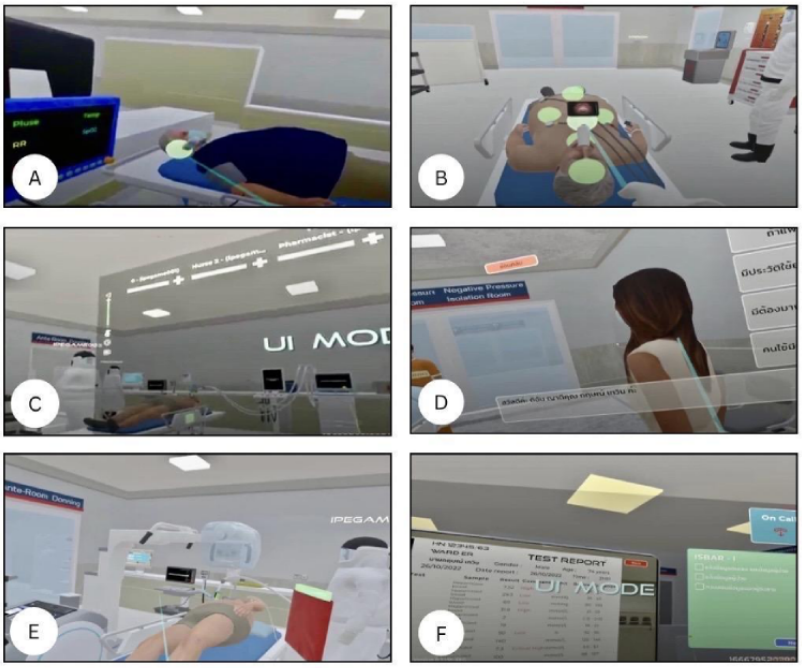


Figure 3. Health point (HP) bar's features and functions and examples of stressful situations and professional perspectives encountered by 6 health care professionals while practicing Team Strategies and Tools to Enhance Performance and Patient Safety (TeamSTEPPS) and stress-coping strategies in a 3D virtual reality (VR)–based simulation-based interprofessional education (SIMBIE) scenario: (A) nursing student perspective during initial triage, including vital sign measurements, such as temperature using a green laser pointing at the ear; (B) medical student perspective while performing endotracheal intubation using a video laryngoscope; (C) nursing student perspective while receiving a phone call from a pharmacy student alerting the team to a drug allergy and each profession's HP bar; (D) pharmacy student perspective while politely warning a patient's relative taking a photo of the patient and clinical staff; (E) radiologic technologist student view while assisting the nursing student with repositioning the patient to safely remove a lead X-ray plate, ensuring proper side rail placement to prevent accidental extubation or falls; and (F) medical technologist student perspective while urgently communicating a critical potassium level using the Identify, Situation, Background, Assessment, and Recommendation (ISBAR) communication framework to the team via phone.



Co-Debriefing Design

Following the initial 30-minute SIMBIE session, participants engaged in a 60-minute oral co-debriefing session, guided by TeamSTEPPS as the primary learning framework. This session provided a safe, nonjudgmental environment for team-based problem-solving and crisis management. Facilitated by instructors, it aimed to enhance experiential learning, reinforce team dynamics, and foster a no-blame culture. After the co-debriefing, participants in Group B replayed the 3D computer-based SIMBIE for a final 30-minute session, while participants in Group C replayed the 3D VR SIMBIE for their concluding 30-minute session.

In this study, co-debriefing was conducted using two complementary conceptual frameworks designed to optimize critically reflective learning: (1) the PEARLS (Promoting Excellence and Reflective Learning in Simulation) framework [169] and (2) the "divide and conquer" approach [169]. The PEARLS framework offers a straightforward and structured method for facilitating consistent, reliable exchanges of information during co-debriefing. It involves 4 sequential phases: reactions, during which learners can voice emotionally salient concerns or choose discussion points; description, during which facilitators and learners establish a shared, factual account of what happened; analysis, which explores performance gaps and encourages critical discussion; and summary, which consolidates the discussion into key learning points.

The reaction phase helps to establish a shared and psychologically safe foundation for learning and reflection by allowing learners to express their immediate thoughts and emotional responses [315]. For instance, facilitators began with questions such as, "What emotions are you experiencing right now?" to validate participants' initial reactions and set a supportive tone for reflection. The inclusion of this phase not only supports psychosocial well-being but also may enhance learner engagement by increasing the relevance and resonance of subsequent discussions. In the description phase, debriefers introduce observations coupled with evaluative judgments and may shorten the phase if learners appear to have a shared understanding of the case. This is followed by the analysis phase, in which facilitators guide group reflection using structured prompts. Examples of such prompts included: "Can you describe what was on your mind during the most intense moments?" "In what ways did your team interact and support each other under pressure?" "What specific actions did you take to address the issue, and how successful were they?" and "Are there any other methods you think might have worked in that situation?" Facilitators also applied the plus/delta technique, asking questions such as "What aspects of the scenario went smoothly?" and "What elements could be improved next time?" In addition, the advocacy-inquiry model was used to promote deeper insight into decision-making. For example, a facilitator might say "I'd like to talk about mutual support in TeamSTEPPS. I observed three unsuccessful intubation attempts. I was concerned about the effect of prolonged hypoxia.

What was going through your mind as your team worked to establish the airway?"—encouraging participants to share their thought processes and mental models. Finally, the summary phase concluded the session with a review of critical insights. Facilitators asked learners to identify key lessons they planned to apply in future clinical practice, such as, "What are the main takeaways from this experience that you would carry into real-life situations?"

Additionally, we implemented the "divide and conquer" approach during co-debriefing, assigning a lead role—typically the medical debriefer—to guide the session [59,163,319]. Under this model, co-debriefers conveniently divide thematic responsibilities based on the TeamSTEPPS framework, thereby minimizing redundant cognitive processing. This division of labor helps reduce the intrinsic and extraneous cognitive load experienced by co-debriefers—a key consideration in accordance with cognitive load theory, which posits that reducing unnecessary mental burden enhances co-debriefers' performance and consequently learning outcomes of learners [311,320].

To prevent power imbalances among co-debriefers, we followed strategies suggested by Oriot et al [293] on agreeing in advance about the techniques, approaches, and educational strategies we will use—including our use of nonverbal communication. This is important because, as Holmes and Mellanby [164] noted, "debriefers debrief slightly differently, everybody has a slightly different focus, different speeds and different process for how you do it." Establishing nonverbal communication allows co-debriefers to "authorize" each other to speak using pre-agreed signals. This approach helps prevent interruptions or conflicts, ensuring that one debriefer does not undermine the other's process. Additionally, we also conducted a "post-debriefing huddle," which includes peer coaching and provides an opportunity for co-debriefers to review challenges, resolve misunderstandings, and collaboratively develop strategies for future sessions [168,290].

Outcome Assessments

To assess outcomes, DSSQ and CBI were administered. Additionally, Anxiety Trait scores from the STAI were measured as a control variable. Anxiety Trait scores were assessed prior to each intervention. Participants self-reported on the DSSQ before and after each intervention (4 total measurements per participant). Burnout assessments were conducted before each intervention and followed up 1 month later (3 measurements per participant). The e-survey was developed and administered using the Qualtrics platform (see Multimedia Appendix 4).

DSSQ and CBI were selected based on their strong psychometric properties and suitability for use with health care professionals in short-term interventions. The DSSQ provides a multidimensional assessment of psychological stress, capturing emotional, cognitive, and motivational states, and is particularly suited for evaluating acute stress responses. We selected the CBI because it is more applicable to the context of our study population—health care students undergoing both clinical and academic training. The CBI is particularly relevant for assessing 3 core dimensions of burnout: work-related, personal, and client-related. Unlike the Maslach Burnout Inventory, which was originally developed for use with long-term professionals in workplace settings, the CBI offers a broader conceptualization of burnout that includes personal exhaustion—an important consideration for student populations. Designed specifically for health care contexts, the CBI has demonstrated high reliability and validity across diverse clinical settings. Furthermore, it has shown greater sensitivity than the Maslach Burnout Inventory, with studies reporting higher burnout detection rates (53% vs 35%) [321]. This suggests that the CBI may be more effective at identifying early or less overt symptoms of burnout, aligning well with our goal of preventive intervention among students. It is considered a reliable tool across cultures and has been associated with future risks such as absenteeism, sleep issues, painkiller use, and intention to leave. Its subscales also reflect meaningful changes in burnout over time [322].

These instruments were chosen over other tools due to their brevity, sensitivity to short-term changes, and practical applicability within the study's timeframe. Both scales are concise, enabling efficient administration without overburdening participants—an important consideration during pandemic-related disruptions and online learning contexts. Although both tools are self-reported and may be subject to response bias, they offer a balanced trade-off between feasibility and diagnostic utility. The DSSQ may not fully capture stress under extreme crisis conditions, and the CBI may not reflect long-term burnout trajectories; however, their capacity to assess immediate changes in stress and burnout levels makes them appropriate for evaluating the short-term impact of the ER-VIPE intervention.

A Measure of Anxiety Trait

The subscale of trait anxiety from the short version of the Spielberger STAI [323] was ordered and paid for with the necessary permissions obtained. It is used to assess an individual propensity to anxiety. Trait anxiety data were collected and included as a control variable in the analysis. There are 5 items, and the items use a 4-point Likert scale (1=not at all to 4=very much so). An example is, "In general, I feel that difficulties are piling up so that I cannot overcome them." The Cronbach α of the original scale was found to be acceptable (α=0.91). This measure was translated into Thai using a back-translation method by J Chavanovanich, PhD (email, August 21, 2024; see Multimedia Appendix 4).

A Measure of State Stress

The short-version DSSQ [324] was used to assess the level of subjective stress state. Professor Helton granted us permission for its use. The DSSQ is a 24-item multidimensional measurement comprising task engagement, distress, and worry. The items use a 5-point Likert scale (1=not at all to 5=extremely). An example item is, "I was motivated to do the task." DSSQ is a well-established tool for assessing acute stress, recognized for its reliability and validity. It consistently achieves high internal consistency, with Cronbach α for the original scale found to be acceptable, exceeding 0.80 across all dimensions. This measure was translated into Thai using a back-translation method by J Chavanovanich, PhD (email, August 21, 2024). We chose psychological tools, such as the DSSQ, instead of physiological measures like heart rate and heart rate variability because they allow for more detailed differentiation between

engagement (items: 2, 5, 11, 12, 13, 17, 21, and 22), distress (items: 1, 3, 4, 6, 7, 8, 9, and 10), and worry (items: 14, 15, 16, 18, 19, 20, 23, and 24), offering a comprehensive understanding of stress states (see Multimedia Appendix 4).

A Measure of Burnout

A 6-item personal burnout subscale from the CBI [322] was used to assess the level of prolonged physical and psychological exhaustion. The items use a 5-point Likert scale (1=never/almost never to 5=always). An example item is, "How often are you physically exhausted?" The measure was translated into Thai using a back-translation method by J Chavanovanich, PhD (email, August 21, 2024). Permission to translate and use the CBI for this research was granted by the National Research Centre for the Working Environment by T Clausen, PhD (email, June 28, 2022). The CBI has demonstrated strong psychometric properties across various studies. It exhibits high internal consistency, with Cronbach α coefficients typically exceeding 0.90 for both the overall scale and its subscales (personal, work, and patient burnout). Construct validity was supported by confirmatory factor analysis, and convergent validity was demonstrated through strong correlations with a previously validated measurement [322,325-327]. We focused exclusively on personal burnout, as the participants were students who were neither working nor patients. Concentrating on personal burnout was expected to provide a more direct response to the research question and increase the likelihood of successful study completion, although this may reduce validity (see Multimedia Appendix 4).

Statistical Analysis

Differences and Changes Over Time

Descriptive statistics were summarized using means (SDs) and medians (IQRs). Frequencies and percentages were used for categorical data. Comparisons between groups in demographic characteristics at baseline were conducted using the Wilcoxon Mann-Whitney *U* test, Kruskal-Wallis *H* test with the Dunn post hoc test, chi-square test, or Fisher exact test, as appropriate. Generalized estimating equations (GEEs) were used to analyze changes over time and assess differences in improvement scores for burnout and DSSQ between groups, with adjustments for anxiety. DSSQ engagement, distress, and worry scores were measured using specific items from the DSSQ. Engagement scores among Groups A, B, and C were obtained from items 2, 5, 11, 12, 13, 17, 21, and 22. Distress scores were derived from items 1, 3, 4, 6, 7, 8, 9, and 10, while worry scores were obtained from items 14, 15, 16, 18, 19, 20, 23, and 24. To assess changes over time and compare improvements in burnout and DSSQ scores between groups, GEEs were used, with adjustments for anxiety as a covariate. Both an intention-to-treat (ITT) analysis and a per-protocol (PP) analysis were performed. Missing data were imputed using the last observation carried forward method. Statistical significance was set at a 2-tailed *P* value of <.05 for all analyses. Stata version 15 (Stata Corp) [328] was used for data analyses.

Effect Size Calculation for Burnout Reduction

To assess the magnitude of intervention effects on burnout reduction, Cohen *d* was calculated for each of the 5 intervention conditions using mean differences and pooled SDs. Five effect sizes were computed according to the structure of the study. A negative effect size indicated a reduction in burnout compared with baseline, while a positive effect size indicated an increase in burnout.

I. Effect Size of Medical Movies and MOOCs

To evaluate the effect of medical movies and MOOCs on burnout after 2 weeks, we calculated the mean difference in burnout scores between phase 2 (pre-SIMBIE) and phase 1 (before medical movies and MOOCs) within Group B and C, using pooled SDs. An alternative comparison using Group A as a control group was also considered to assess whether burnout reduction occurred over 2 weeks without movie and MOOC exposure, accounting for potential external factors.

II. Effect Size of Computer-Based SIMBIE

To evaluate the effect of the computer-based SIMBIE simulation after 4 weeks, we calculated the mean difference in burnout scores between phase 3 (4 weeks post-SIMBIE) and phase 2 (pre-SIMBIE) within Group A using pooled SDs.

III. Effect Size of VR-Based SIMBIE

To evaluate the effect of VR-based SIMBIE simulation after 4 weeks, we calculated the mean difference in burnout scores between phase 3 and phase 2 within Group C (after follow-up losses) using pooled SDs.

IV. Effect Size of ER-VIPE (Computer-Based Version)

To evaluate the full ER-VIPE program (medical movies + MOOCs + computer-based SIMBIE with co-debriefing) after 6 weeks, we calculated the mean difference in burnout scores between phase 3 (6 weeks post-SIMBIE) and phase 1 (before medical movies and MOOCs) within Group B using pooled SDs.

V. Effect Size of ER-VIPE (VR-Based Version)

To evaluate the VR-based version of ER-VIPE (medical movies + MOOCs + VR-based SIMBIE with co-debriefing), the same approach was used. Burnout scores at phase 3 and phase 1 were compared within Group C using pooled SDs.

Sample Size Calculation

A power analysis for a repeated measures multivariate analysis of variance (MANOVA) conducted in G*Power (version 3.1.9.4) with a within-between interaction indicated a required sample size of 82 participants to detect an effect size of 0.35, with an α level of .05 and a power of .80. To account for potential dropout, 10% was added, resulting in a total sample size of 87 participants.

Ethical Considerations

The study protocol was approved by the Ethics Committee for Research in Human Subjects of the Faculty of Medicine, Chulalongkorn University, Thailand (IRB number 0366/65). Participants were informed about the study's objectives, procedures, potential risks, and benefits. Each participant received US $30 in recognition of their time and contribution, in accordance with institutional review board approval. This information was provided both orally and in writing before obtaining informed consent. Participants were assured of their

right to make voluntary decisions and withdraw from the study at any time. Data were anonymized to maintain confidentiality.

Results

Demographics

A total of 87 undergraduate clinical students from various professional programs participated in the study, with 29 students in each group (A, B, and C). The sample was predominantly female (62/87, 71%) with a mean age of 21.87 (SD 1.13) years. No significant differences were found in demographic characteristics between the groups. However, analysis of the debriefing duration revealed significant differences between Group B (mean 50.93, SD 12.61) and Group C (mean 60.33, SD 9.75), as shown in Table 1.

Table 1. Demographics by treatment group.

Factor	Total sample	Group A[a]	Group B[b]	Group C[c]	*P* value
Gender, n (%)					.80[d]
Female	62 (71)	22 (76)	20 (69)	20 (69)	
Male	25 (29)	7 (24)	9 (31)	9 (31)	
Age (years), mean (SD)	21.87 (1.13)	21.83 (0.93)	21.86 (1.03)	21.93 (1.41)	—[e]
Age (years), median (IQR)	22 (21-22)	22 (21-22)	22 (21-22)	21 (21-22)	.90[f]
Academic year, n (%)					.63[g]
Third year	12 (14)	3 (10)	7 (24)	2 (7)	
Fourth year	47 (54)	17 (59)	13 (45)	17 (59)	
Fifth year	18 (21)	6 (21)	5 (17)	7 (24)	
Sixth year	10 (11)	3 (10)	4 (14)	3 (10)	
Academic grade, mean (SD)	3.23 (0.38)	3.24 (0.40)	3.23 (0.43)	3.22 (0.31)	—
Academic grade, median (IQR)	3.24 (3-3.50)	3.25 (3-3.53)	3.23 (3.03-3.50)	3.24 (3-3.48)	.92[f]
Debriefing duration (minutes), mean (SD)	54.96 (12.29)	—	50.93 (12.61)	60.33 (9.75)	—
Debriefing duration (minutes), median (IQR)	57 (45-63)	—	49 (45-58.50)	63 (60-69)	.01[h]
Debriefing staff, mean (SD)	5.33 (1.75)	—	5.69 (1.85)	4.97 (1.59)	—
Debriefing staff, median (IQR)	6 (4-6)	—	6 (4-7)	6 (5-6)	.06[h]

[a]Group A (control) participated in 3D computer-based simulation-based interprofessional education (SIMBIE) without oral debriefing.

[b]Group B received a medical movie, a massive open online course (MOOC), a 3D computer-based SIMBIE, and an oral co-debriefing session.

[c]Group C received a medical movie, a MOOC, a 3D virtual reality SIMBIE, and an oral co-debriefing session.

[d]Chi-square test.

[e]Not applicable.

[f]Kruskal-Wallis *H* test.

[g]Fisher exact test.

[h]Wilcoxon Mann-Whitney *U* test.

Baseline Outcomes and Time Interval Assessments Across Groups

Baseline assessments of burnout and DSSQ measures (engagement, distress, and worry) showed no significant differences among Groups A, B, and C (see Table S1 in Multimedia Appendix 5). Similarly, the mean interval between the postintervention DSSQ assessment in phase 1 and the pre-intervention DSSQ assessment in phase 2 did not differ significantly across groups. However, a significant difference emerged in the mean burnout assessment interval between the postintervention assessment in phase 2 and the final assessment in phase 3: Group A (mean 33.82, SD 12.90 days) differed significantly from Group B (mean 48.20, SD 23.48 days) and Group C (mean 37.20, SD 6.52 days). See Table S2 in Multimedia Appendix 6.

Overall Assessment Outcomes Among Groups

The study aimed to assess individual state stress, as measured using the mean DSSQ scores, and burnout outcomes, as measured using the CBI, across Groups A, B, and C. Group A (control) participated in a 3D computer-based SIMBIE without oral debriefing. Group B received a medical movie, MOOC, 3D computer-based SIMBIE, and oral co-debriefing session. Group C received a medical movie, MOOC, 3D VR SIMBIE, and oral co-debriefing session. Statistical analysis was conducted

using a GEE approach, with adjustments for confounding factors such as anxiety trait. Measurements were taken during 3 phases, as illustrated in Figure 1: phase 1 (movie and MOOC intervention), phase 2 (SIMBIE intervention), and phase 3 (6-week follow-up). These results were obtained through an ITT analysis (see Figure 4, Figure 5, and Table S3 in Multimedia Appendix 7). A PP analysis further corroborated the findings from the ITT analysis, demonstrating consistent results and interpretations for the mean change in DSSQ scores (see Table S4 in Multimedia Appendix 8, Figure S1 in Multimedia Appendix 9, and Figure S2 in Multimedia Appendix 10).

Figure 4. Based on intention-to-treat analysis, changes in (A) Dundee Stress State Questionnaire (DSSQ)-engagement scores, (B) DSSQ-distress scores, and (C) DSSQ-worry scores with the movie and massive open online course (MOOC) intervention as well as changes in (D) DSSQ-engagement scores, (E) DSSQ-distress scores, and (F) DSSQ-worry scores with the simulation-based interprofessional education (SIMBIE) intervention. VR: virtual reality.

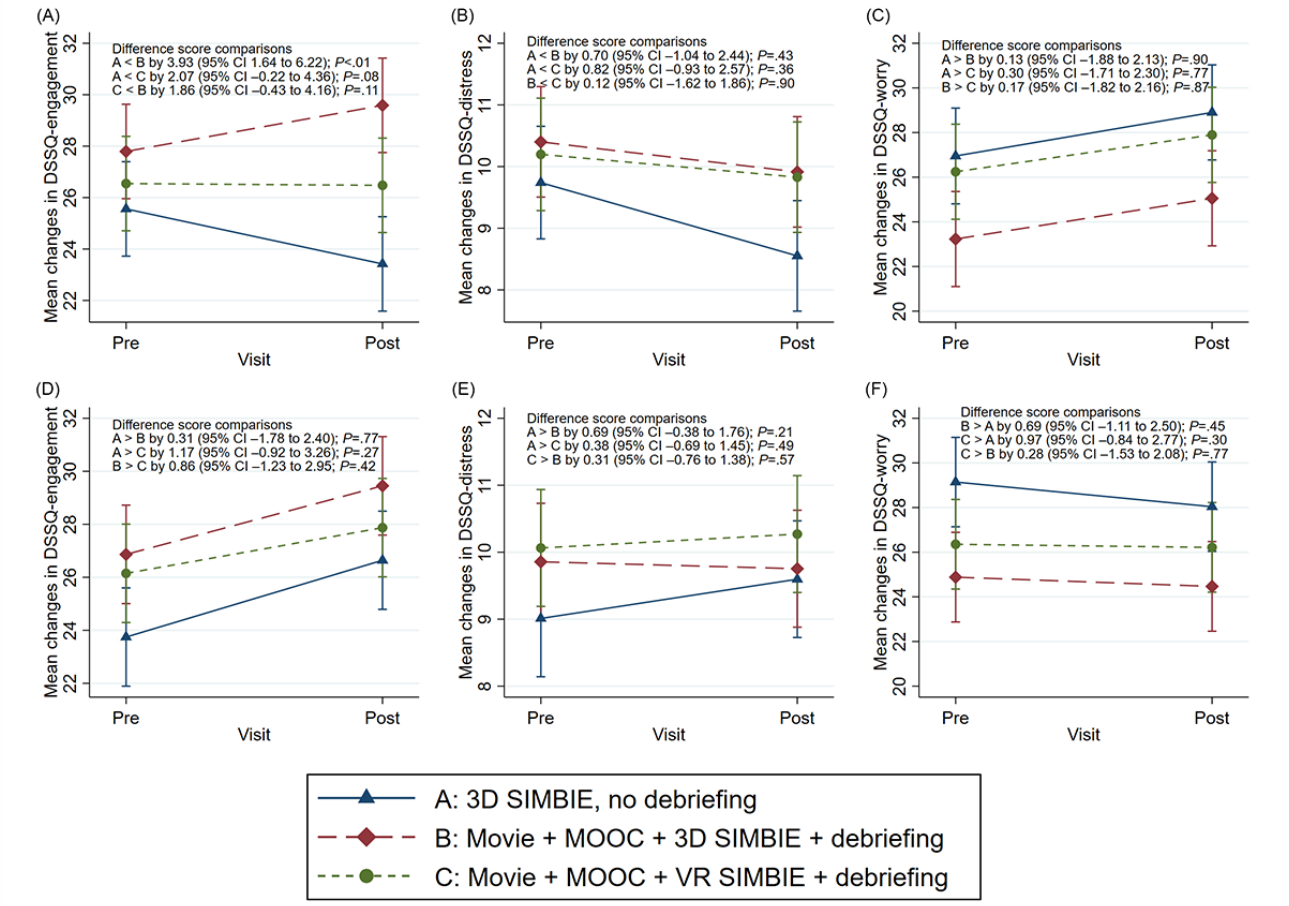
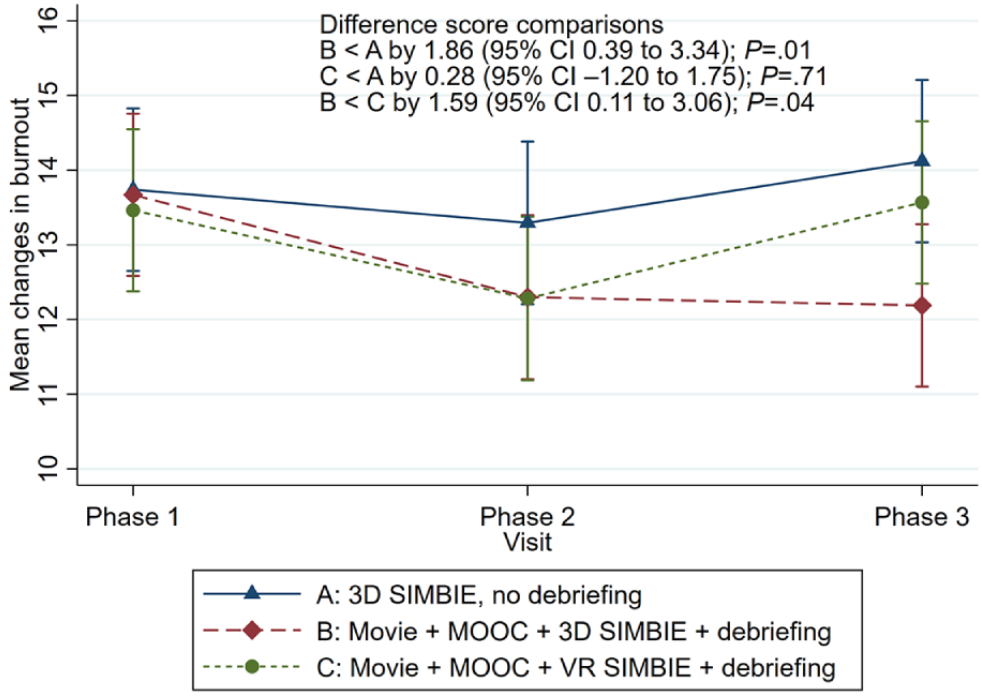


Figure 5. Changes in Copenhagen Burnout Inventory (CBI) scores based on intention-to-treat analysis using generalized estimating equations (GEEs), adjusted for anxiety traits as a control variable. MOOC: massive open online course; SIMBIE: simulation-based interprofessional education; VR: virtual reality.



Measuring State Stress With the DSSQ

Overview

The DSSQ [329] was selected as the primary tool in this study due to its ability to assess momentary stress states across 3 dimensions: task engagement, distress, and worry. Designed for real-time application, the DSSQ is particularly well-suited for simulation-based educational research [330]. It enabled the detection of subtle, immediate changes in learners' stress responses before and after each ER-VIPE session. In contrast, the Perceived Stress Scale-10 [331] is a globally recognized instrument widely used in health care studies. It provides a reliable measure of general perceived stress over the past month and is appropriate for large-scale screening.

DSSQ-Engagement Scores Among Groups

Following exposure to a medical movie and MOOCs, Group B had significantly higher DSSQ engagement scores, adjusted for confounding factors, than Group A (mean difference 3.93; *P*<.001). Group C had nonsignificantly higher scores than Group A (mean difference 2.07; *P*=.08), and nonsignificantly lower scores than Group B (mean difference –1.86; *P*=.11). After participating in the SIMBIE process all 3 groups demonstrated positive trends in DSSQ engagement scores, adjusted for confounding factors. Group B scored higher than Group A, though the difference was not significant (mean difference –0.31; *P*=.77). Similarly, Group C had nonsignificantly higher scores than Group A (mean difference –1.17; *P*=.27) and nonsignificantly lower engagement scores than Group B (mean difference –0.86; *P*=.42). See Figure 4, Table S3 in Multimedia Appendix 7, Table S4 in Multimedia Appendix 8, and Figure S1 in Multimedia Appendix 9.

DSSQ-Distress Scores Among Groups

Focusing on DSSQ-distress following exposure to a medical movie and MOOCs, the results indicated negative trends across all 3 groups. Group B exhibited nonsignificantly higher distress scores than Group A, with a mean difference of 0.70 (*P*=.43). Similarly, Group C had nonsignificantly higher distress scores than Group A, with a mean difference of 0.82 (*P*=.36) and nonsignificantly lower distress scores than Group B, with a mean difference of 0.12 (*P*=.90). After participating in the SIMBIE process and adjusting for confounding factors, only Group B exhibited negative trends in DSSQ distress scores. However, the differences for all 3 groups were not significant. Group B scored higher than Group A, but the mean difference was not statistically significant (mean difference –0.69; *P*=.21). Similarly, Group C had nonsignificantly higher scores than Group A (mean difference –0.38; *P*=.49) and nonsignificantly lower scores than Group B (mean difference 0.31; *P*=.57). See Figure 4, Table S3 in Multimedia Appendix 7, Table S4 in Multimedia Appendix 8, and Figure S1 in Multimedia Appendix 9.

DSSQ-Worry Scores Among Groups

Regarding DSSQ-worry following exposure to a medical movie and MOOCs, all 3 groups exhibited positive trends. Group B had nonsignificantly lower worry scores than Group A, with a mean difference of –0.13 (*P*=.90). Group C also had nonsignificantly lower worry scores than Group A (mean difference –0.30; *P*=.77) and nonsignificantly higher scores than Group B (mean difference –0.17, *P*=.87). After participating in the SIMBIE process and adjusting for confounding factors, all 3 groups had negative trends in DSSQ-worry, though no significant differences were detected.

Group B had nonsignificantly lower worry scores than Group A (mean difference 0.69; *P*=.45). Similarly, Group C had nonsignificantly lower worry scores than Group A (mean difference 0.97; *P*=.30) and nonsignificantly higher scores than Group B (mean difference 0.28; *P*=.77). See Figure 4, Table S3 in Multimedia Appendix 7, Table S4 in Multimedia Appendix 8, and Figure S1 in Multimedia Appendix 9.

Measuring Burnout: Comparative Outcomes by Group

Our analysis of the CBI scores revealed discrepancies in results and interpretations between the ITT and PP analyses. In the ITT analysis comparing burnout outcomes between the pre-intervention assessment (phase 1) and the final assessment (phase 3), Group B had significantly lower burnout scores than Group A (mean difference –1.86, 95% CI 0.39 to 3.34; *P*=.01). Similarly, Group B had significantly lower burnout scores than Group C, with a mean difference of 1.59 (95% CI 0.11 to 3.06; *P*=.04). No statistically significant differences were observed between Group C and Group A, with a mean difference of –0.28 (95% CI –1.20 to 1.75; *P*=.07). See Figure 5, Table S3 in Multimedia Appendix 7, Table S4 in Multimedia Appendix 8, and Figure S2 in Multimedia Appendix 10.

In the PP analysis comparing burnout outcomes between the pre-intervention assessment (phase 1) and the final assessment (phase 3), Group B again had significantly lower burnout scores than Group A, with a mean difference of –2.02 (*P*=.02). However, the mean difference between Group B and Group C was not significant (mean difference 1.6; *P*=.07). Similarly, no statistically significant differences were observed between Group C and Group A, with a mean difference of –0.42 (*P*=.61). See Table S3 in Multimedia Appendix 7, Table S4 in Multimedia Appendix 8, and Figure S2 in Multimedia Appendix 10.

Effect Sizes for Burnout Change Across the Multimodal Interventions

The ER-VIPE computer-based intervention (Group B) at 8 weeks demonstrated the largest reduction, with a small-to-moderate effect size (*d* =–0.31, 95% CI –0.78 to 0.15). Similarly, the combined medical movie and MOOC intervention (Groups B and C, pre-SIMBIE) at 2 weeks yielded a small-to-moderate effect (*d*=–0.30, 95% CI –0.63 to 0.03). In contrast, the computer-based SIMBIE with co-debriefing intervention (Group B) at 4 weeks showed a slight increase in burnout, with a small effect size (*d*=0.09, 95% CI –0.50 to 0.67). The ER-VIPE VR-based intervention (Group C) at 8 weeks was associated with an increase in burnout, with a moderate-to-large effect size (*d*=0.45, 95% CI –0.25 to 1.15). Notably, the VR-based SIMBIE with co-debriefing (Group C) at 4 weeks resulted in a substantial increase in burnout, demonstrating a large effect size (*d*=0.83, 95% CI 0.22 to 1.44), as shown in Table 2.

Table 2. Effect sizes for changes in burnout across multimodal interventions.

Group(s)	Intervention				Phase comparison	Duration (weeks)	Effect size, Cohen *d* (95% CI)[a]
	Medical movie + MOOCs[b]	Computer-based SIMBIE[c]	VR[d]-based SIMBIE	Co-debriefing			
B[e]	Yes	Yes	No	Yes	1[f] vs 3[g]	8	–0.31 (–0.75 to 0.15)
B, C[h]	Yes	No	No	No	1 vs 2[i]	2	–0.30 (–0.63 to 0.03)
A[j]	No	Yes	No	No	2 vs 3	6	–0.19 (–0.75 to 0.36)
B	No	Yes	No	Yes	2 vs 3	6	0.09 (–0.50 to 0.67)
C	Yes	No	Yes	Yes	1 vs 3	8	0.45 (–0.25 to 1.15)
C	No	No	Yes	Yes	2 vs 3	6	0.83 (0.22 to 1.14)

[a]Negative values indicate reductions in burnout: positive values indicate increases in burnout.

[b]MOOCs: massive open online courses.

[c]SIMBIE: simulation-based interprofessional education.

[d]VR: virtual reality.

[e]Group B received the movie + MOOCs, 3D computer-based SIMBIE, and co-debriefing.

[f]Phase 1: baseline; before movie + MOOCs.

[g]Phase 3: 8-week follow-up from baseline.

[h]Group C received the movie + MOOCs, VR-based SIMBIE, and co-debriefing.

[i]Phase 2: before SIMBIE; 2-week follow-up.

[j]Group A (control) received only 3D computer-based SIMBIE without debriefing.

Discussion

ER-VIPE: An Innovative Approach to Reducing Stress and Burnout

This study evaluated the effectiveness and quantified the effect sizes of multimodal educational strategies—medical movies, MOOCs, and computer- or VR-based SIMBIE with co-debriefing (collectively termed ER-VIPE)—for reducing burnout among future health care professionals. The ER-VIPE computer-based intervention (Group B) demonstrated the most notable reduction in burnout, with a small-to-moderate effect size (d=–0.31), followed closely by the combined medical movie and MOOC intervention after 2 weeks (d=–0.30).

In contrast, the computer-based SIMBIE with co-debriefing intervention (Group B) was associated with a minimal increase in burnout, reflected by a small effect size (d=0.09). The ER-VIPE VR-based intervention (Group C) showed a moderate-to-large increase in burnout (d=0.45), while the VR-based SIMBIE after 4 weeks resulted in the largest increase (d=0.83).

These findings suggest that the ER-VIPE computer-based multimodal approach, particularly when combined with co-debriefing, is more effective at mitigating burnout than the VR-based ER-VIPE intervention or single-component SIMBIE interventions. Additionally, medical movies and MOOCs alone contributed to a modest reduction in burnout after 2 weeks.

Discussion: On the Meaningfulness of Effect Sizes in Burnout Prevention

The ER-VIPE computer-based intervention (Group B) demonstrated a small-to-moderate effect on burnout reduction, with an SMD of 0.31 (95% CI –0.15 to 0.78). Although the CI includes 0, the upper bound suggests a potentially meaningful impact, warranting further investigation in larger samples.

The ER-VIPE VR-based intervention (Group C) demonstrated a small-to-moderate effect for increasing burnout, with an SMD of 0.45 (95% CI –0.25 to 1.15). Although the CI includes 0, the upper bound suggests a potentially meaningful impact, warranting further investigation in larger samples.

Group B (ER-VIPE computer-based) demonstrated significantly lower burnout scores than Group A (computer-based SIMBIE without co-debriefing), with a mean difference of –1.86 (95% CI 0.39 to 3.34; P=.01). Although the beta coefficient (β=1.8, 95% CI 0.39 to 3.34) indicates a statistically significant effect, these findings should be interpreted with caution. Sole reliance on P values can be misleading, as statistical significance does not necessarily imply clinical or practical relevance. To address this, we calculated effect sizes (Cohen d) to better understand the magnitude of the observed effects. The ER-VIPE computer-based intervention yielded a small-to-medium effect size (d=–0.31), supporting its potential educational and clinical value.

From a practical standpoint, even small to modest effect sizes may have meaningful implications, especially in the context of early burnout prevention. The CBI (maximum 30 points), which was used in this study, is a culturally validated and widely adopted tool for assessing prolonged physical and psychological exhaustion. Previous studies have linked CBI scores to future risks such as absenteeism, sleep disturbances, increased use of painkillers, and intention to leave the profession [322]. In our study, a maximum reduction of approximately 3.34 points across 6 items (an average change of 0.56 per item on a 5-point Likert scale for 6 items) may reflect a meaningful shift in participants' experiences—for example, from reporting they "always" feel overwhelmed to "seldom" feeling that they "can't take it anymore." Such changes, particularly in at-risk populations, could represent significant psychological relief.

Nonetheless, these findings should be interpreted cautiously. Self-reported burnout scores may be influenced by short-term emotional states or response-shift bias. Although the observed trend in burnout reduction is promising, especially for the ER-VIPE computer-based group, we consider the results preliminary and recommend further exploration through longitudinal research and real-world implementation studies.

Effectiveness of Preparatory Tools and Instructional Design Considerations

Medical movies and MOOCs were used as preparatory tools prior to SIMBIE participation. Although these online modalities demonstrated some positive effects, their impact was variable and appeared to depend on factors such as learner readiness, the presence of an instructor [332], and the incorporation of structured guidance to direct learner attention toward key instructional scenes [333]. Consistent with our findings, the integration of cinemeducation and simulation has been demonstrated to be both feasible and effective, producing synergistic and convergent benefits from the perspectives of preclinical medical students. This combined approach also helps address some limitations of simulation alone in replicating complex professional scenarios, as supported by survey responses [334]. These findings underscore the critical role of intentional instructional design for optimizing the effectiveness of online educational interventions. Among all the interventions examined, the ER-VIPE computer-based model produced the most pronounced reduction in stress and burnout. This outcome may be attributed to its congruence with evidence-based learning principles; the integration of thoughtful SIMBIE design elements; and the use of accessible, user-friendly technology.

Applying Experiential, Flipped, and SIT Frameworks in Learning Design

The learning design in this study was grounded in 3 complementary educational frameworks: Kolb's experiential learning cycle, the flipped classroom model, and SIT. According to [335], learning is most effective when it progresses through 4 stages: concrete experience, reflective observation, abstract conceptualization, and active experimentation. To support this cycle, MOOCs were used to deliver theoretical foundations, while medical movies promoted critical thinking and contextual understanding by translating abstract concepts into realistic scenarios. Mayer's cognitive theory of multimedia learning suggests that such video-based materials enhance knowledge retention by presenting information in a more concrete and engaging format [336]. Furthermore, analyzing film content through the lens of learned theories supports both cognitive and

emotional development [337], while audiovisual elements foster realism and a more immersive learning experience [338-340].

Learners initially engaged in a SIMBIE session to establish a baseline experience, followed by co-debriefing guided by trained facilitators. This reflective process emphasized TeamSTEPPS concepts and aligned with Kolb's stages of reflective observation and abstract conceptualization. Learners then applied newly acquired insights in a subsequent SIMBIE session, fulfilling the active experimentation phase.

To enhance preparedness, a flipped classroom approach was used. Presession exposure to medical movies and MOOCs helped build interprofessional knowledge and promoted positive attitudes toward IPE, TeamSTEPPS, and stress coping strategies [307]. This model improved learner engagement and emotional investment while maximizing the effectiveness of in-session SIMBIE activities [308]. Importantly, integrating stress management into real-world clinical education has been recognized as essential to professional education [309].

SIT further supported the design by providing a structured approach to building stress resilience. SIT involves 3 phases: conceptualization (understanding stress and personal reactions) via cinemeducation and MOOCs, skill acquisition (developing coping strategies such as communication) via SIMBIE, and skills consolidation reinforced through co-debriefing and application (practicing these skills in simulated or realistic scenarios as second round). This approach enabled learners to build psychological readiness for high-stress clinical environments. Similarly, Liaw et al [160] demonstrated that computer-based virtual SIMBIE could positively influence stress responses, measured via blood pressure and heart rate, and enhance team rescue performance, showing no significant difference compared with physical simulations. These findings were particularly valuable during the COVID-19 pandemic, as desktop VR offered a safe and effective alternative. Building on this, our study contributes new evidence that desktop-based SIMBIE within the ER-VIPE framework not only supports immediate stress management but also leads to sustained reductions in burnout observed at the 8-week follow-up.

In contrast, an RCT by Blanchard et al [268] found that a tutorial module on stress and stress management, followed by repeated non-IPE VR simulations under moderate to high stress conditions, was a feasible instructional approach. Participants in the intervention group reported lower perceived stress; reduced electrodermal activity, a physiological marker of stress measured through changes in the electrical conductivity of sweat on the hands or feet; and greater perceived competence after completing the test module compared with the control group. The training also appeared to facilitate desensitization to stress in future simulated scenarios. The differing outcomes between our study and the RCT by Blanchard et al [268] may be attributed to key methodological and contextual differences. Although the study by Blanchard et al used a controlled design with repeated VR exposure under moderate stress, induced by ecologically salient auditory stimuli, our quasi-experimental study featured more complex IPE scenarios involving a greater number of multiprofessional participants and team-based stress dynamics. These high-stakes, emotionally charged simulations may have exceeded the optimal arousal level described by the Yerkes-Dodson law [244], potentially leading to heightened stress.

Interestingly, a multicenter prospective randomized trial involving EM residents examined the effects of brief mental skills training delivered 1 month prior to simulation. The study reported no significant differences in subjective or objective stress responses, measured using heart rate variability and the STAI, during a non-IPE, manikin-based resuscitation simulation. This lack of measurable impact may be attributed to the extended gap between the conceptual (didactic) phase and the skill acquisition (simulation) phase, potentially hindering effective learning transfer. Moreover, the high-stress experiences of real-life clinical pressure may have reduced the simulation's ability to elicit authentic stress responses. Notably, the study did not describe the debriefing process (consolidation phase) and did not include a second simulation for deliberate practice (application phase), which may have further limited the effectiveness of the SIT framework.

SIMBIE Design Elements Supporting Safe, Realistic, and Collaborative Learning

The SIMBIE platform in this study was intentionally designed to replicate realistic clinical scenarios involving diverse health care students, with a focus on psychological safety, structured learning, and interprofessional collaboration. The learning sequence began with a prebriefing to outline objectives, facilitate ice-breaking activities, foster familiarity among participants, and establish a safe learning climate. A unique HP bar feature was incorporated to provide peer support during gameplay, and the session concluded with structured debriefing led by trained facilitators using a no-blame, psychologically safe approach—all conducted within a clearly defined time frame.

These design features align with evidence emphasizing that high-quality simulation-based education should include structured scenario progression, adaptability to learner actions, and the identification of critical performance points. Time-conscious design and high-fidelity environments further enhance learner engagement and realism. Equally important is the role of facilitators, who not only guide learning but also ensure that cultural diversity, psychological safety, and shared understanding of team roles are respected. Their guidance reinforces learning outcomes, encourages reflection, and supports the development of interprofessional competencies [341,342].

Findings from this study align with those of previous literature suggesting that interprofessional simulation-based approaches, when combined with skilled debriefing, enhance active engagement, self-efficacy, realistic role enactment, and collaborative team-based learning [340]. These experiences help students build essential interprofessional skills prior to clinical practice [343], gain insight into other professional roles [344,345], and foster deeper participation, particularly when stress is acknowledged and validated by peers or facilitators [346].

Enhancing Accessibility and Reducing Stress Through User-Friendly Technology

The use of a computer-based platform in SIMBIE significantly enhanced accessibility and ease of use compared with VR headsets, which often require more complex operation. Consistent with previous studies, computer-based simulations have been shown to reduce negative emotional responses during learning [124,347]. Moreover, cine-VR training—delivered through head-mounted displays or 360-degree video—has been shown to enhance empathy among health care professional students, with no reported technological issues nor adverse effects [348]. In contrast, VR-based simulations have been associated with higher levels of distress due to technostress (eg, ergonomics, cybersickness, visual fatigue), or technological barriers, including user unfamiliarity, discomfort, and reluctance to adopt new tools [349-351]. Poorly designed VR experiences may further limit accessibility and user comfort [352,353]. In light of these challenges, computer-based simulation emerges as a more practical, scalable, and cost-effective educational approach [334], particularly in contexts aiming to reduce stress and promote learner engagement.

Limitations

This study has several limitations. Conducted in a single university-based setting, the findings may have limited generalizability to other settings that differ in context and available facilities. This study was conducted during the middle of the COVID-19 pandemic, limiting its comparability to conventional clinical teaching, which is primarily conducted in clinical settings. The exclusion of participants with pre-existing mental health conditions may limit the generalizability of findings to broader populations. Although anxiety traits were controlled as a confounding factor, other potential influences on stress and burnout, such as individual stressors, baseline coping mechanisms, and levels of social support, were not assessed in this study and may have influenced participants' responses. As a result, the findings should be interpreted with caution, as unmeasured psychosocial factors may have moderated participants' responses to the simulation-based intervention. Although the DSSQ specifically measures stress during assigned tasks and the CBI captures broader aspects of exhaustion, these tools may not fully reflect the complexity of stress experiences in all contexts.

Additionally, dropout rates of 24%, 44%, and 14% in Groups A, B, and C, respectively, during the transition from phase 2 to phase 3 were addressed. We did not know the exact reasons for participant dropout. However, potential contributing factors may include exam schedules, personal or academic stress, mental health concerns, and minor technical issues related to the online survey platform sent via email. Additionally, the delayed delivery of monetary incentives—provided only after the second simulation, several weeks before the final assessment—may have reduced participants' motivation to complete the study. We used GEEs to minimize sensitivity to missing data. An ITT analysis with imputed data was also performed to validate the PP findings. This indicates that, even though the dropout rate was low, the results were likely not significantly different from the original findings.

Future Study

Future research should investigate the effectiveness of movies, MOOCs and virtual SIMBIE for reducing stress and burnout among diverse groups, such as postgraduate professionals and those in intensive care or prehospital settings. Future studies should consider including participants with mild or well-managed mental health comorbidities to enhance the generalizability and applicability of the intervention to more diverse, real-world clinical settings. We also recommend conducting an RCT comparing the intervention with traditional educational approaches in a nonpandemic context, such as in-person clinical teaching. To strengthen the rigor of future research, objective measures of stress, such as heart rate variability, electrodermal activity, or electroencephalography, should be incorporated. Additionally, potential confounding factors, including individual stressors, baseline coping mechanisms, and levels of social support, should be carefully assessed and controlled. A mixed methods approach is suggested to explore the mechanisms of impact and identify areas for improvement. Furthermore, longitudinal studies using multiple stress and burnout assessment methods are necessary to comprehensively evaluate long-term outcomes.

Conclusion

This study highlights the effectiveness of a novel multimodal learning approach that integrates movie-based education, MOOCs, and a virtual 3D computer-based SIMBIE with co-debriefing for reducing burnout and improving self-reported stress levels by enhancing engagement and reducing worry and distress among undergraduate clinical students. These innovative IPE strategies are designed to help multiprofessional students manage stress, reduce burnout, and develop collaborative problem-solving skills within authentic simulation environments. By alleviating the pressures and risks typically associated with clinical practice, the interactive and scalable 3D computer-based ER-VIPE platform supports 21st-century health care learners, where patient safety is paramount. Integrating these tools and strategies into IPE programs offers significant potential to enhance well-being and resilience while preparing health care students and early-career professionals for a smooth transition into clinical practice.

Acknowledgments

This research was supported by the Second Century Fund, Chulalongkorn University. The authors thank the Chulalongkorn Healthcare Advanced Multi-Profession Simulation Center (CHAMPS) for their invaluable assistance. The authors are also grateful to Surachai Pianpetchlert, Thepwinphan Theppitak, and the research assistants—Kitnipat Boonydhammakul, Sutasinee Chaidej,

Chayanit Trakulpipat, Sirisopha Suwanchinda, Thanyared Sangsawad, Chaiwat Takkanat, Thanes Yunirundorn, and Nuttarin Panswad—for their contributions to data collection, graphic design, and IT support.

We gratefully acknowledge the *Emergency Room Virtual Simulation Interprofessional Education (ER-VIPE)* study group for their collaboration in developing the educational movies, massive open online courses (MOOCs), and the simulation-based interprofessional education (SIMBIE) platform. The contributors associated with the ER-VIPE Study Group are as follows: Chanya Thanomlikhit (Department of Nursing, King Chulalongkorn Memorial Hospital, The Thai Red Cross Society); Jennifer Chavanovanich (Faculty of Psychology, Chulalongkorn University); Jiraphan Ritsamdang (Faculty of Pharmaceutical Sciences, Chulalongkorn University); Kawin Dhanakoses (Faculty of Architecture, Chulalongkorn University); Kittisak Potisartra (Faculty of Engineering, Chulalongkorn University); Krittin Bunditanukul (Faculty of Pharmaceutical Sciences, Chulalongkorn University); Narisorn Kongruttanachok (Faculty of Medicine, Chulalongkorn University); Nattanun Chanchaochai (Faculty of Arts, Chulalongkorn University); Nattawit Tanjapatkul (Faculty of Engineering, Chulalongkorn University); Navaporn Worasilchai (Faculty of Allied Health Sciences, Chulalongkorn University); Nhabhat Chaimongkol (Faculty of Education, Chulalongkorn University); Pataraporn Kheawwan (Faculty of Nursing, Chulalongkorn University); Porntiwa Sunpawut (Srisavarindhira Thai Red Cross Institute of Nursing); Sararas Khongwirotphan (Faculty of Allied Health Sciences, Chulalongkorn University); Sawitree Suayod (Faculty of Allied Health Sciences, Chulalongkorn University); Somchit Eiam-Ong (Faculty of Medicine, Chulalongkorn University); Sopon Jakdetchai (Faculty of Engineering, Chulalongkorn University); Sujinat Jitwiriyanont (Faculty of Arts, Chulalongkorn University); Suwimon Rojnawee (Faculty of Nursing, Chulalongkorn University); Supachai Chuenjitwongsa (Faculty of Dentistry, Chulalongkorn University); Thititip Tippayamontri (Faculty of Allied Health Sciences, Chulalongkorn University); Tippayaporn Pavavimol (Faculty of Communication Arts, Chulalongkorn University); Vishnu Kotrajaras (Faculty of Engineering, Chulalongkorn University).

Authors' Contributions

S Srikasem, S Seephom, AV, PP, SK, and KN contributed to writing the original draft. All authors participated in reviewing and editing the manuscript and approved the final version for submission. Conceptualization was led by KN, while methodology was developed collaboratively by KN and PP. Software development was managed by KN. Validation and formal analysis were performed by KN and PP. Investigation was conducted by KN and AV. Resources were provided by KN, AV, and PP. Data curation was handled by KN, AV, and PP. Visualization was contributed by KN, PP, S Srikasem, AV, and SK. Project supervision and administration were managed by KN, and funding acquisition was overseen by KN.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Electroencephalogram (EEG) Procedure.
[DOCX File , 11 KB - mededu_v11i1e70726_app1.docx]

Multimedia Appendix 2
CHERRIES (Checklist for Reporting Results of Internet E-Surveys).
[DOCX File , 19 KB - mededu_v11i1e70726_app2.docx]

Multimedia Appendix 3
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 3419 KB - mededu_v11i1e70726_app3.pdf]

Multimedia Appendix 4
Measurement instruments for state stress, burnout, and anxiety trait.
[DOCX File , 525 KB - mededu_v11i1e70726_app4.docx]

Multimedia Appendix 5
Baseline burnout and Dundee Stress State Questionnaire (DSSQ) scores.
[DOCX File , 20 KB - mededu_v11i1e70726_app5.docx]

Multimedia Appendix 6
Time interval assessments.
[DOCX File , 23 KB - mededu_v11i1e70726_app6.docx]

Multimedia Appendix 7
Improvement in burnout and Dundee Stress State Questionnaire (DSSQ) scores: intention-to-treat analysis.
[DOC File , 25 KB - mededu_v11i1e70726_app7.doc]

Multimedia Appendix 8
Per-protocol analysis for burnout and Dundee Stress State Questionnaire (DSSQ) improvements.
[DOCX File , 28 KB - mededu_v11i1e70726_app8.docx]

Multimedia Appendix 9
Dundee Stress State Questionnaire (DSSQ) changes based on per-protocol analysis.
[DOCX File , 198 KB - mededu_v11i1e70726_app9.docx]

Multimedia Appendix 10
Copenhagen Burnout Inventory (CBI) changes based on per-protocol analysis.
[DOCX File , 135 KB - mededu_v11i1e70726_app10.docx]

References

1. Arora M, Asha S, Chinnappa J, Diwan AD. Review article: burnout in emergency medicine physicians. Emerg Med Australas 2013 Dec 09;25(6):491-495. [doi: 10.1111/1742-6723.12135] [Medline: 24118838]
2. Adriaenssens J, De Gucht V, Maes S. Determinants and prevalence of burnout in emergency nurses: a systematic review of 25 years of research. Int J Nurs Stud 2015 Feb;52(2):649-661. [doi: 10.1016/j.ijnurstu.2014.11.004] [Medline: 25468279]
3. Nantsupawat A, Nantsupawat R, Kunaviktikul W, Turale S, Poghosyan L. Nurse burnout, nurse-reported quality of care, and patient outcomes in Thai hospitals. J Nurs Scholarsh 2016 Jan;48(1):83-90. [doi: 10.1111/jnu.12187] [Medline: 26650339]
4. Yuwanich N, Akhavan S, Nantsupawat W, Martin L. Experiences of occupational stress among emergency nurses at private hospitals in Bangkok, Thailand. OJN 2017;07(06):657-670. [doi: 10.4236/ojn.2017.76049]
5. Dixon E, Murphy M, Wynne R. A multidisciplinary, cross-sectional survey of burnout and wellbeing in emergency department staff during COVID-19. Australas Emerg Care 2022 Sep;25(3):247-252 [FREE Full text] [doi: 10.1016/j.auec.2021.12.001] [Medline: 34906441]
6. Chor WPD, Ng WM, Cheng L, Situ W, Chong JW, Ng LYA, et al. Burnout amongst emergency healthcare workers during the COVID-19 pandemic: a multi-center study. Am J Emerg Med 2021 Aug;46:700-702 [FREE Full text] [doi: 10.1016/j.ajem.2020.10.040] [Medline: 33129643]
7. Zhang Q, Mu M, He Y, Cai Z, Li Z. Burnout in emergency medicine physicians: a meta-analysis and systematic review. Medicine (Baltimore) 2020 Aug 07;99(32):e21462 [FREE Full text] [doi: 10.1097/MD.0000000000021462] [Medline: 32769876]
8. Mong M, Noguchi K. Emergency room physicians' levels of anxiety, depression, burnout, and coping methods during the COVID-19 pandemic. Journal of Loss and Trauma 2021 Jun 08;27(3):212-228. [doi: 10.1080/15325024.2021.1932127]
9. Alanazy ARM, Alruwaili A. The global prevalence and associated factors of burnout among emergency department healthcare workers and the impact of the COVID-19 pandemic: a systematic review and meta-analysis. Healthcare (Basel) 2023 Aug 07;11(15):2220 [FREE Full text] [doi: 10.3390/healthcare11152220] [Medline: 37570460]
10. Alanazi TNM, McKenna L, Buck M, Alharbi RJ. Reported effects of the COVID-19 pandemic on the psychological status of emergency healthcare workers: a scoping review. Australas Emerg Care 2022 Sep;25(3):197-212 [FREE Full text] [doi: 10.1016/j.auec.2021.10.002] [Medline: 34802977]
11. Gualano MR, Sinigaglia T, Lo Moro G, Rousset S, Cremona A, Bert F, et al. The burden of burnout among healthcare professionals of intensive care units and emergency departments during the COVID-19 pandemic: a systematic review. Int J Environ Res Public Health 2021 Aug 02;18(15):8172 [FREE Full text] [doi: 10.3390/ijerph18158172] [Medline: 34360465]
12. Ren H, Xue Y, Li P, Yin X, Xin W, Li H. Prevalence of turnover intention among emergency nurses worldwide: a meta-analysis. BMC Nurs 2024 Sep 11;23(1):645 [FREE Full text] [doi: 10.1186/s12912-024-02284-2] [Medline: 39261866]
13. Sungbun S, Naknoi S, Somboon P, Thosingha O. Impact of the COVID-19 pandemic crisis on turnover intention among nurses in emergency departments in Thailand: a cross sectional study. BMC Nurs 2023 Sep 27;22(1):337 [FREE Full text] [doi: 10.1186/s12912-023-01495-3] [Medline: 37759190]
14. Oyat FWD, Oloya JN, Atim P, Ikoona EN, Aloyo J, Kitara DL. The psychological impact, risk factors and coping strategies to COVID-19 pandemic on healthcare workers in the sub-Saharan Africa: a narrative review of existing literature. BMC Psychol 2022 Dec 01;10(1):284 [FREE Full text] [doi: 10.1186/s40359-022-00998-z] [Medline: 36457038]
15. Deady M, Collins D, Johnston D, Glozier N, Calvo R, Christensen H, et al. The impact of depression, anxiety and comorbidity on occupational outcomes. Occup Med (Lond) 2022 Jan 13;72(1):17-24. [doi: 10.1093/occmed/kqab142] [Medline: 34693972]

16. Goetzel RZ, Hawkins K, Ozminkowski RJ, Wang S. The health and productivity cost burden of the "top 10" physical and mental health conditions affecting six large U.S. employers in 1999. J Occup Environ Med 2003 Jan;45(1):5-14. [doi: 10.1097/00043764-200301000-00007] [Medline: 12553174]
17. Magnavita N, Meraglia I, Riccò M. Anxiety and depression in healthcare workers are associated with work stress and poor work ability. AIMS Public Health 2024;11(4):1223-1246. [doi: 10.3934/publichealth.2024063] [Medline: 39802561]
18. Tziner A, Rabenu E, Radomski R, Belkin A. Work stress and turnover intentions among hospital physicians: the mediating role of burnout and work satisfaction. Revista de Psicología del Trabajo y de las Organizaciones 2015 Dec;31(3):207-213. [doi: 10.1016/j.rpto.2015.05.001]
19. Çelmeçe N, Menekay M. The effect of stress, anxiety and burnout levels of healthcare professionals caring for COVID-19 patients on their quality of life. Front Psychol 2020 Nov 23;11:597624 [FREE Full text] [doi: 10.3389/fpsyg.2020.597624] [Medline: 33329264]
20. Phuekphan P, Aungsuroch Y, Yunibhand J. A model of factors influencing intention to leave nursing in Thailand. Pacific Rim International Journal of Nursing Research 2021;25(3):407-420.
21. The NHSO's five measures to reduce medical staff workload. National Health Security Office. 2023 Aug 22. URL: https://eng.nhso.go.th/view/1/Secretary-General/The-NHSOs-five-measures-to-reduce-medical-staff-workload/552/EN-US [accessed 2025-08-23]
22. Stehman CR, Testo Z, Gershaw RS, Kellogg AR. Burnout, drop out, suicide: physician loss in emergency medicine, part I. West J Emerg Med 2019 May;20(3):485-494 [FREE Full text] [doi: 10.5811/westjem.2019.4.40970] [Medline: 31123550]
23. Anchala R, Kannuri NK, Pant H, Khan H, Franco OH, Di Angelantonio E, et al. Hypertension in India: a systematic review and meta-analysis of prevalence, awareness, and control of hypertension. J Hypertens 2014 Jun;32(6):1170-1177 [FREE Full text] [doi: 10.1097/HJH.0000000000000146] [Medline: 24621804]
24. Lu Y, Wang R, Zhang Y, Su H, Wang P, Jenkins A, et al. Ecosystem health towards sustainability. Ecosyst Health Sustain 2017 Jun 20;1(1):1-15 [FREE Full text] [doi: 10.1890/ehs14-0013.1]
25. Tawfik DS, Scheid A, Profit J, Shanafelt T, Trockel M, Adair KC, et al. Evidence relating health care provider burnout and quality of care: a systematic review and meta-analysis. Ann Intern Med 2019 Oct 15;171(8):555-567 [FREE Full text] [doi: 10.7326/M19-1152] [Medline: 31590181]
26. Bahadirli S, Sagaltici E. Burnout, job satisfaction, and psychological symptoms among emergency physicians during COVID-19 outbreak: a cross-sectional study. Psychiatry Clin Psychopharmacol 2021 Mar 12;31(1):67-76. [doi: 10.5152/pcp.2021.20180] [Medline: 39619354]
27. Ma Y, Chen F, Xing D, Meng Q, Zhang Y. Study on the associated factors of turnover intention among emergency nurses in China and the relationship between major factors. Int Emerg Nurs 2022 Jan;60:101106. [doi: 10.1016/j.ienj.2021.101106] [Medline: 34864323]
28. Assawarungruang W. Factors influencing intention to quit of physician from healthcare facility in Bangkok. University of the Thai Chamber of Commerce. 2017. URL: https://doi.nrct.go.th/ListDoi/Download/629854/ee196d84fec9dcd443d10a30b255f473?Resolve_DOI=10.14456/fbms.2018.30 [accessed 2025-08-23]
29. Global strategy on human resources for health: Workforce 2030. World Health Organization. 2020 Jul 07. URL: https://www.who.int/publications/i/item/9789241511131 [accessed 2025-08-23]
30. Arigi LAG, Mustika R, Greviana N. How to deal with burnout during online learning in medical education? A systematic review. Jurnal Pendidikan Kedokteran Indonesia 2023 Jul 03;12(2):203. [doi: 10.22146/jpki.75898]
31. Vasan A, Mabey DC, Chaudhri S, Brown Epstein H, Lawn SD. Support and performance improvement for primary health care workers in low- and middle-income countries: a scoping review of intervention design and methods. Health Policy Plan 2017 Apr 01;32(3):437-452 [FREE Full text] [doi: 10.1093/heapol/czw144] [Medline: 27993961]
32. Schwerdtle P, Morphet J, Hall H. A scoping review of mentorship of health personnel to improve the quality of health care in low and middle-income countries. Global Health 2017 Oct 03;13(1):77 [FREE Full text] [doi: 10.1186/s12992-017-0301-1] [Medline: 28974233]
33. Framework for action on interprofessional education and collaborative practice. World Health Organization. 2010. URL: https://www.who.int/publications/i/item/framework-for-action-on-interprofessional-education-collaborative-practice [accessed 2025-08-23]
34. IPEC Core Competencies for Interprofessional Collaborative Practice: Version 3. Interprofessional Education Collaborative. 2023 Nov 20. URL: https://ipec.memberclicks.net/assets/core-competencies/IPEC_Core_Competencies_Version_3_2023.pdf [accessed 2025-08-23]
35. Costello M, Rusell K, Coventry T. Examining the average scores of nursing teamwork subscales in an acute private medical ward. BMC Nurs 2021 May 31;20(1):84 [FREE Full text] [doi: 10.1186/s12912-021-00609-z] [Medline: 34059037]
36. Rosen MA, DiazGranados D, Dietz AS, Benishek LE, Thompson D, Pronovost PJ, et al. Teamwork in healthcare: key discoveries enabling safer, high-quality care. Am Psychol 2018;73(4):433-450 [FREE Full text] [doi: 10.1037/amp0000298] [Medline: 29792459]
37. Pedersen AHM, Rasmussen K, Grytnes R, Nielsen KJ. Collaboration and patient safety at an emergency department – a qualitative case study. JHOM 2018 Jan 22;32(1):25-38. [doi: 10.1108/jhom-09-2016-0174]

38. Ajeigbe DO, McNeese-Smith D, Leach LS, Phillips LR. Nurse-physician teamwork in the emergency department: impact on perceptions of job environment, autonomy, and control over practice. J Nurs Adm 2013 Mar;43(3):142-148. [doi: 10.1097/NNA.0b013e318283dc23] [Medline: 23425911]
39. Alsabri M, Boudi Z, Lauque D, Dias RD, Whelan JS, Östlundh L, et al. Impact of teamwork and communication training interventions on safety culture and patient safety in emergency departments: a systematic review. J Patient Saf 2020 Sep 9;18(1):e351-e361. [doi: 10.1097/pts.0000000000000782]
40. Burström L, Letterstål A, Engström ML, Berglund A, Enlund M. The patient safety culture as perceived by staff at two different emergency departments before and after introducing a flow-oriented working model with team triage and lean principles: a repeated cross-sectional study. BMC Health Serv Res 2014 Jul 09;14:296 [FREE Full text] [doi: 10.1186/1472-6963-14-296] [Medline: 25005231]
41. Kemp S, Brewer M. Early stages of learning in interprofessional education: stepping towards collective competence for healthcare teams. BMC Med Educ 2023 Sep 22;23(1):694 [FREE Full text] [doi: 10.1186/s12909-023-04665-8] [Medline: 37740200]
42. Gellis ZD, Kim E, Hadley D, Packel L, Poon C, Forciea MA, et al. Evaluation of interprofessional health care team communication simulation in geriatric palliative care. Gerontol Geriatr Educ 2019;40(1):30-42. [doi: 10.1080/02701960.2018.1505617] [Medline: 30160623]
43. O'Neil-Pirozzi TM, Musler JL, Carney M, Day L, Hamel PC, Kirwin J. Impact of early implementation of experiential education on the development of interprofessional education knowledge and skill competencies. J Allied Health 2019;48(2):e53-e59. [Medline: 31167019]
44. Hood K, Cant R, Baulch J, Gilbee A, Leech M, Anderson A, et al. Prior experience of interprofessional learning enhances undergraduate nursing and healthcare students' professional identity and attitudes to teamwork. Nurse Educ Pract 2014 Mar;14(2):117-122. [doi: 10.1016/j.nepr.2013.07.013] [Medline: 23937910]
45. Corrêa CPS, Lucchetti ALG, da Silva Ezequiel O, Lucchetti G. Short and medium-term effects of different teaching strategies for interprofessional education in health professional students: a randomized controlled trial. Nurse Educ Today 2022 Oct;117:105496. [doi: 10.1016/j.nedt.2022.105496] [Medline: 35914346]
46. Hedges AR, Johnson HJ, Kobulinsky LR, Estock JL, Eibling D, Seybert AL. Effects of cross-training on medical teams' teamwork and collaboration: use of simulation. Pharmacy (Basel) 2019 Jan 19;7(1):1 [FREE Full text] [doi: 10.3390/pharmacy7010013] [Medline: 30669460]
47. Collin K, Paloniemi S, Mecklin J. Promoting inter‐professional teamwork and learning – the case of a surgical operating theatre. Journal of Education and Work 2010 Jan 21;23(1):43-63. [doi: 10.1080/13639080903495160]
48. Lakkala S, Turunen TA, Kangas H, Pulju M, Kuukasjärvi U, Autti H. Learning inter-professional teamwork during university studies: a case study of student-teachers' and social work students' shared professional experiences. Journal of Education for Teaching 2017 Jun 22:1-13. [doi: 10.1080/02607476.2017.1342051]
49. van Diggele C, Roberts C, Burgess A, Mellis C. Interprofessional education: tips for design and implementation. BMC Med Educ 2020 Dec 03;20(Suppl 2):455 [FREE Full text] [doi: 10.1186/s12909-020-02286-z] [Medline: 33272300]
50. Franz S, Muser J, Thielhorn U, Wallesch CW, Behrens J. Inter-professional communication and interaction in the neurological rehabilitation team: a literature review. Disabil Rehabil 2020 Jun;42(11):1607-1615. [doi: 10.1080/09638288.2018.1528634] [Medline: 30457016]
51. Stadick J. The relationship between interprofessional education and health care professional's attitudes towards teamwork and interprofessional collaborative competencies. Journal of Interprofessional Education & Practice 2020 Jun;19:100320 [FREE Full text] [doi: 10.1016/j.xjep.2020.100320]
52. Tasselli S. Social networks and inter-professional knowledge transfer: the case of healthcare professionals. Organization Studies 2015 Mar 17;36(7):841-872. [doi: 10.1177/0170840614556917]
53. Grand-Guillaume-Perrenoud JA, Cignacco E, MacPhee M, Carron T, Peytremann-Bridevaux I. How does interprofessional education affect attitudes towards interprofessional collaboration? A rapid realist synthesis. Adv Health Sci Educ Theory Pract 2025 Jun 23;30(3):879-933. [doi: 10.1007/s10459-024-10368-6] [Medline: 39313601]
54. Sezgin M, Bektas H. Effectiveness of interprofessional simulation-based education programs to improve teamwork and communication for students in the healthcare profession: a systematic review and meta-analysis of randomized controlled trials. Nurse Educ Today 2023 Jan;120:105619. [doi: 10.1016/j.nedt.2022.105619] [Medline: 36343420]
55. Marion-Martins A, Pinho D. Interprofessional simulation effects for healthcare students: a systematic review and meta-analysis. Nurse Educ Today 2020 Nov;94:104568 [FREE Full text] [doi: 10.1016/j.nedt.2020.104568] [Medline: 32932058]
56. Guraya S, Barr H. The effectiveness of interprofessional education in healthcare: a systematic review and meta-analysis. Kaohsiung J Med Sci 2018 Mar;34(3):160-165 [FREE Full text] [doi: 10.1016/j.kjms.2017.12.009] [Medline: 29475463]
57. Medina-Córdoba M, Cadavid S, Espinosa-Aranzales A, Aguía-Rojas K, Bermúdez-Hernández PA, Quiroga-Torres D, et al. The effect of interprofessional education on the work environment of health professionals: a scoping review. Adv Health Sci Educ Theory Pract 2024 Sep 01;29(4):1463-1480. [doi: 10.1007/s10459-023-10300-4] [Medline: 38038831]

58. Brashers V, Erickson JM, Blackhall L, Owen JA, Thomas SM, Conaway MR. Measuring the impact of clinically relevant interprofessional education on undergraduate medical and nursing student competencies: a longitudinal mixed methods approach. J Interprof Care 2016 Jul 07;30(4):448-457. [doi: 10.3109/13561820.2016.1162139] [Medline: 27269441]
59. Frenk J, Chen L, Bhutta ZA, Cohen J, Crisp N, Evans T, et al. Health professionals for a new century: transforming education to strengthen health systems in an interdependent world. Lancet 2010 Dec 04;376(9756):1923-1958. [doi: 10.1016/S0140-6736(10)61854-5] [Medline: 21112623]
60. Williams R, Jenkins DA, Ashcroft DM, Brown B, Campbell S, Carr MJ, et al. Diagnosis of physical and mental health conditions in primary care during the COVID-19 pandemic: a retrospective cohort study. The Lancet Public Health 2020 Oct;5(10):e543-e550. [doi: 10.1016/s2468-2667(20)30201-2]
61. Alexander M, Hall MN, Pettice YJ. Cinemeducation: an innovative approach to teaching psychosocial medical care. Fam Med 1994;26(7):430-433. [Medline: 7926359]
62. Rueb M, Rehfuess EA, Siebeck M, Pfadenhauer LM. Cinemeducation: a mixed methods study on learning through reflective thinking, perspective taking and emotional narratives. Med Educ 2024 Jan;58(1):63-92. [doi: 10.1111/medu.15166] [Medline: 37525520]
63. Rueb M, Siebeck M, Rehfuess EA, Pfadenhauer LM. Cinemeducation in medicine: a mixed methods study on students' motivations and benefits. BMC Med Educ 2022 Mar 12;22(1):172 [FREE Full text] [doi: 10.1186/s12909-022-03240-x] [Medline: 35279156]
64. Ber R, Alroy G. Twenty years of experience using trigger films as a teaching tool. Acad Med 2001 Jun;76(6):656-658. [doi: 10.1097/00001888-200106000-00022] [Medline: 11401816]
65. McCann E, Huntley-Moore S. Madness in the movies: an evaluation of the use of cinema to explore mental health issues in nurse education. Nurse Educ Pract 2016 Nov;21:37-43 [FREE Full text] [doi: 10.1016/j.nepr.2016.09.009] [Medline: 27716595]
66. Farré M, Bosch F, Roset P, Baños JE. Putting clinical pharmacology in context: the use of popular movies. J Clin Pharmacol 2004 Jan;44(1):30-36 [FREE Full text] [doi: 10.1177/0091270003260679] [Medline: 14681339]
67. Cambra-Badii I, Francés MDL, Videla S, Farré M, Montané E, Blázquez F, et al. Cinemeducation in clinical pharmacology: using cinema to help students learn about pharmacovigilance and adverse drug reactions. Eur J Clin Pharmacol 2020 Dec 04;76(12):1653-1658. [doi: 10.1007/s00228-020-02985-y] [Medline: 32886177]
68. Rosenstock J. Beyond a beautiful mind: film choices for teaching schizophrenia. Acad Psychiatry 2003;27(2):117-122. [doi: 10.1176/appi.ap.27.2.117] [Medline: 12824113]
69. Bhagar H. Should cinema be used for medical student education in psychiatry? Med Educ 2005 Sep;39(9):972-973 [FREE Full text] [doi: 10.1111/j.1365-2929.2005.02252.x] [Medline: 16150042]
70. Webster CR, Valentine LC, Gabbard GO. Film clubs in psychiatric education: the hidden curriculum. Acad Psychiatry 2015 Oct 05;39(5):601-604. [doi: 10.1007/s40596-014-0252-2] [Medline: 25476226]
71. Furst BA. Bowlby Goes to the Movies: film as a teaching tool for issues of bereavement, mourning, and grief in medical education. Academic Psychiatry 2007 Oct 01;31(5):407-410. [doi: 10.1176/appi.ap.31.5.407]
72. Fleming MZ, Piedmont RL, Hiam CM. Images of madness: feature films in teaching psychology. Teaching of Psychology 1990 Oct 01;17(3):185-187 [FREE Full text] [doi: 10.1207/s15328023top1703_12]
73. Alexander M, Lenahan P, Pavlov A. Cinemeducation : a comprehensive guide to using film in medical education. San Diego, CA: Radcliffe Publishing; 2005.
74. Cambra-Badii I, González-Caminal G, Gomar-Sancho C, Piqué-Buisan J, Guardiola E, Baños JE. The Value of Cinemeducation in Health Sciences Education. In: Varsou O, editor. Teaching, Research, Innovation and Public Engagement. Cham, Switzerland: Springer International Publishing; 2023:29-40.
75. The World Health Report 2001: Mental Disorders affect one in four people. World Health Organization. 2001 Sep 28. URL: https://www.who.int/news/item/28-09-2001-the-world-health-report-2001-mental-disorders-affect-one-in-four-people [accessed 2025-08-23]
76. Matorin S. Stigma as a barrier to recovery. Psychiatr Serv 2002 May;53(5):629-30; author reply 630. [doi: 10.1176/appi.ps.53.5.629-a] [Medline: 11986518]
77. Wahl OF. Stigma as a barrier to recovery from mental illness. Trends Cogn Sci 2012 Jan;16(1):9-10. [doi: 10.1016/j.tics.2011.11.002] [Medline: 22153582]
78. Søvold LE, Naslund JA, Kousoulis AA, Saxena S, Qoronfleh MW, Grobler C, et al. Prioritizing the mental health and well-being of healthcare workers: an urgent global public health priority. Front Public Health 2021 May 7;9:679397 [FREE Full text] [doi: 10.3389/fpubh.2021.679397] [Medline: 34026720]
79. Phelan SM, Salinas M, Pankey T, Cummings G, Allen JP, Waniger A, et al. Patient and health care professional perspectives on stigma in integrated behavioral health: barriers and recommendations. Ann Fam Med 2023 Feb;21(Suppl 2):S56-S60 [FREE Full text] [doi: 10.1370/afm.2924] [Medline: 36849477]
80. Abo Shereda HM, Alqhtani SS, ALYami AH, ALGhamdi HM, Ahmed MIO, ALSalah NA, et al. Exploring the relationship between compassion fatigue, stigma, and moral distress among psychiatric nurses: a structural equation modeling study. BMC Nurs 2025 Feb 12;24(1):163 [FREE Full text] [doi: 10.1186/s12912-025-02802-w] [Medline: 39939960]

81. Mehta SS, Edwards ML. Suffering in silence: mental health stigma and physicians' licensing fears. AJP Residents' Journal 2018 Nov 01;13(11):2-4. [doi: 10.1176/appi.ajp-rj.2018.131101]
82. Favre S, Bajwa NM, Dominicé Dao M, Audétat Voirol MC, Nendaz M, Junod Perron N, et al. Association between burnout and stigma in physicians. PLoS One 2023 Apr 5;18(4):e0283556 [FREE Full text] [doi: 10.1371/journal.pone.0283556] [Medline: 37018317]
83. Bianchi R, Verkuilen J, Brisson R, Schonfeld I, Laurent E. Burnout and depression: label-related stigma, help-seeking, and syndrome overlap. Psychiatry Res 2016 Nov 30;245:91-98 [FREE Full text] [doi: 10.1016/j.psychres.2016.08.025] [Medline: 27529667]
84. Zeppegno P, Gramaglia C, Feggi A, Lombardi A, Torre E. The effectiveness of a new approach using movies in the training of medical students. Perspect Med Educ 2015 Oct;4(5):261-263 [FREE Full text] [doi: 10.1007/s40037-015-0208-6] [Medline: 26346496]
85. Rehl D, Mangapora M, Love M, Love C, Shaw K, McCarthy J, et al. Feasibility of a cinematic-virtual reality training program about opioid use disorder for osteopathic medical students: a single-arm pre-post study. J Osteopath Med 2024 Nov 01;124(11):509-516 [FREE Full text] [doi: 10.1515/jom-2023-0188] [Medline: 38965036]
86. Kiraly D. A Social Constructivist Approach to Translator Education: Empowerment from Theory to Practice. London, England: Routledge; 2000.
87. Kontos P, Grigorovich A, Dupuis SL, Colobong R, Gray J, Jonas-Simpson C, et al. Projecting a critique of stigma associated with dementia on screen: the impact of a Canadian film on the importance of relational caring in the community. Gerontologist 2024 Feb 01;64(2):1. [doi: 10.1093/geront/gnad045] [Medline: 37067944]
88. Hawke LD, Michalak EE, Maxwell V, Parikh SV. Reducing stigma toward people with bipolar disorder: impact of a filmed theatrical intervention based on a personal narrative. Int J Soc Psychiatry 2014 Dec;60(8):741-750. [doi: 10.1177/0020764013513443] [Medline: 24351967]
89. Linton S, Hankir A, Anderson S, Carrick FR, Zaman R. Harnessing the power of film to combat mental health stigma. A University College London Psychiatry Society event. Psychiatr Danub 2017 Sep;29(Suppl 3):300-306. [Medline: 28953782]
90. Hoy MB. MOOCs 101: an introduction to massive open online courses. Med Ref Serv Q 2014 Feb 14;33(1):85-91. [doi: 10.1080/02763869.2014.866490] [Medline: 24528267]
91. Baturay M. An overview of the world of MOOCs. Procedia - Social and Behavioral Sciences 2015 Feb;174:427-433 [FREE Full text] [doi: 10.1016/j.sbspro.2015.01.685]
92. de Freitas SI, Morgan J, Gibson D. Will MOOCs transform learning and teaching in higher education? Engagement and course retention in online learning provision. British Journal of Educational Technology 2015;46(3):455-471 [FREE Full text] [doi: 10.1111/BJET.12268]
93. Liu C, Zou D, Chen X, Xie H, Chan WH. A bibliometric review on latent topics and trends of the empirical MOOC literature (2008–2019). Asia Pacific Educ. Rev 2021 Apr 17;22(3):515-534. [doi: 10.1007/s12564-021-09692-y]
94. Shah D. By The Numbers: MOOCs in 2020. Class Central. 2020 Nov 30. URL: https://www.classcentral.com/report/mooc-stats-2020 [accessed 2025-08-23]
95. Lu DW, Dresden S, McCloskey C, Branzetti J, Gisondi MA. Impact of burnout on self-reported patient care among emergency physicians. West J Emerg Med 2015 Dec;16(7):996-1001 [FREE Full text] [doi: 10.5811/westjem.2015.9.27945] [Medline: 26759643]
96. Ministry OHES. Thai MOOC Academy. Thailand Massive Open Online Course. URL: https://thaimooc.ac.th [accessed 2025-08-23]
97. Liyanagunawardena TR, Williams SA. Massive open online courses on health and medicine: review. J Med Internet Res 2014 Aug 14;16(8):e191 [FREE Full text] [doi: 10.2196/jmir.3439] [Medline: 25123952]
98. Hernández R, Morales M, Mota J, Teixeira A. Promoting Engagement in MOOCs Through Social Collaboration: Common Lessos from the Pedagogical Models of Universidad Galileo and Universidade Aberta. In: Teixeira AM, Szucks A, Wagner A, editors. Challenges for Research into Open & Distance Learning - Book of Abstracts -Research Workshop. Oxford, UK: EDEN; 2014:40-41.
99. Downes S. Places to go: connectivism and connective knowledge. Innovate: Journal of Online Education 2008;5(1):6 [FREE Full text]
100. Rodriguez O. The concept of openness behind c and x-MOOCs (massive open online courses). Open Praxis 2013;5(1):67-73. [doi: 10.5944/openpraxis.5.1.42]
101. Rodriguez CO. MOOCs and the AI-Stanford like courses: two successful and distinct course formats for massive open online courses. European Journal of Open, Distance and E-Learning 2012:1-13.
102. Chan T, Thoma B, Lin M. Creating, curating, and sharing online faculty development resources. Academic Medicine 2015;90(6):785-789. [doi: 10.1097/acm.0000000000000692] [Medline: 1888]
103. Liyanagunawardena TR, Aboshady OA. Massive open online courses: a resource for health education in developing countries. Glob Health Promot 2018 Sep 30;25(3):74-76. [doi: 10.1177/1757975916680970] [Medline: 28134014]
104. Eccleston C, Doherty K, Bindoff A, Robinson A, Vickers J, McInerney F. Building dementia knowledge globally through the Understanding Dementia Massive Open Online Course (MOOC). NPJ Sci Learn 2019 Apr 10;4(1):3 [FREE Full text] [doi: 10.1038/s41539-019-0042-4] [Medline: 30993003]

105. Jones J, Johnston JS, Ndiaye NY, Tokar A, Singla S, Skinner NA, et al. Health care workers' motivations for enrolling in massive open online courses during a public health emergency: descriptive analysis. JMIR Med Educ 2024 Jun 19;10:e51915-e51915 [FREE Full text] [doi: 10.2196/51915] [Medline: 38904474]
106. Peterson K, Mundo W, McGladrey L, Aagaard LM, Stalder S, Cook PF. Stress impact and care for COVID-19: pilot education and support course decreases burnout among nursing students. J Am Psychiatr Nurses Assoc 2023;29(5):363-374. [doi: 10.1177/10783903231186997] [Medline: 37534666]
107. Ricker M, Brooks AJ, Bodine S, Lebensohn P, Maizes V. Well-being in residency: impact of an online physician well-being course on resiliency and burnout in incoming residents. Fam Med 2021 Feb 3;53(2):123-128. [doi: 10.22454/fammed.2021.314886]
108. Hayes SC, Strosahl KD, Wilson KG. Acceptance and Commitment Therapy: The Process and Practice of Mindful Change. New York, NY: The Guilford Press; 2012.
109. Fonseca S, Trindade IA, Mendes AL, Ferreira C. The buffer role of psychological flexibility against the impact of major life events on depression symptoms. Clinical Psychologist 2020 Nov 09;24(1):82-90. [doi: 10.1111/cp.12194]
110. Pakenham KI, Landi G, Boccolini G, Furlani A, Grandi S, Tossani E. The moderating roles of psychological flexibility and inflexibility on the mental health impacts of COVID-19 pandemic and lockdown in Italy. J Contextual Behav Sci 2020 Jul;17:109-118 [FREE Full text] [doi: 10.1016/j.jcbs.2020.07.003] [Medline: 32834969]
111. Arslan G, Yıldırım M, Tanhan A, Buluş M, Allen K. Coronavirus stress, optimism-pessimism, psychological inflexibility, and psychological health: psychometric properties of the coronavirus stress measure. Int J Ment Health Addict 2021;19(6):2423-2439 [FREE Full text] [doi: 10.1007/s11469-020-00337-6] [Medline: 32837425]
112. Farchi M, Hirsch-Gornemann MB, Whiteson A, Gidron Y. The SIX Cs model for immediate cognitive psychological first aid: from helplessness to active efficient coping. Int J Emerg Ment Health 2018;20(2):1. [doi: 10.4172/1522-4821.1000395]
113. Hobfoll SE, Watson P, Bell CC, Bryant RA, Brymer MJ, Friedman MJ, et al. Five essential elements of immediate and mid-term mass trauma intervention: empirical evidence. Psychiatry 2007;70(4):283-315; discussion 316. [doi: 10.1521/psyc.2007.70.4.283] [Medline: 18181708]
114. Swanwick T, Forrest K, O'Brien B. Understanding Medical Education: Evidence, Theory, and Practice. Chichester, England, UK: Wiley-Blackwell; 2019.
115. Zapletal A, Hoppe M, Van Oss T, Baird J. Clinical Simulation for Healthcare Professionals. New York, NY: Routledge; 2022.
116. Raurell-Torredà M, Rascón-Hernán C, Malagón-Aguilera C, Bonmatí-Tomás A, Bosch-Farré C, Gelabert-Vilella S, et al. Effectiveness of a training intervention to improve communication between/awareness of team roles: a randomized clinical trial. J Prof Nurs 2021;37(2):479-487 [FREE Full text] [doi: 10.1016/j.profnurs.2020.11.003] [Medline: 33867108]
117. Uslu-Sahan F, Terzioglu F. Interprofessional simulation-based training in gynecologic oncology palliative care for students in the healthcare profession: a comparative randomized controlled trial. Nurse Educ Today 2020 Dec;95:104588. [doi: 10.1016/j.nedt.2020.104588] [Medline: 32980608]
118. Wu J, Chen H, Chiu Y, Chen Y, Kang Y, Hsu Y, et al. Comparison of simulation-based interprofessional education and video-enhanced interprofessional education in improving the learning outcomes of medical and nursing students: a quasi-experimental study. Nurse Educ Today 2022 Nov;118:105535. [doi: 10.1016/j.nedt.2022.105535] [Medline: 36084448]
119. Gilbert JHV, Yan J, Hoffman SJ. A WHO report: framework for action on interprofessional education and collaborative practice. J Allied Health 2010;39 Suppl 1:196-197. [Medline: 21174039]
120. Barton L, Lackie K, Miller SG. Scoping review: interprofessional simulation as an effective modality to teaching interprofessional collaborative competencies in the emergency department. Journal of Research in Interprofessional Practice and Education 2023 Mar 31;13(1):1. [doi: 10.22230/jripe.2023v13n1a349]
121. Gosa L, Davis A, Heyer J, Taft L, Gill L. Interprofessional education collaborative: a pilot simulation project. Teaching and Learning in Nursing 2024 Apr;19(2):e324-e329 [FREE Full text] [doi: 10.1016/j.teln.2023.12.008]
122. Krielen P, Meeuwsen M, Tan ECTH, Schieving JH, Ruijs AJEM, Scherpbier ND. Interprofessional simulation of acute care for nursing and medical students: interprofessional competencies and transfer to the workplace. BMC Med Educ 2023 Feb 11;23(1):105 [FREE Full text] [doi: 10.1186/s12909-023-04053-2] [Medline: 36774481]
123. Williams D, Stephen L, Causton P. Teaching interprofessional competencies using virtual simulation: a descriptive exploratory research study. Nurse Educ Today 2020 Oct;93:104535. [doi: 10.1016/j.nedt.2020.104535] [Medline: 32717697]
124. Kiessling A, Amiri C, Arhammar J, Lundbäck M, Wallingstam C, Wikner J, et al. Interprofessional simulation-based team-training and self-efficacy in emergency medicine situations. J Interprof Care 2022 Mar 27;36(6):873-881 [FREE Full text] [doi: 10.1080/13561820.2022.2038103] [Medline: 35341425]
125. Verkuyl M, Violato E, Harder N, Southam T, Lavoie-Tremblay M, Goldsworthy S, et al. Virtual simulation in healthcare education: a multi-professional, pan-Canadian evaluation. Adv Simul (Lond) 2024 Jan 10;9(1):3 [FREE Full text] [doi: 10.1186/s41077-023-00276-x] [Medline: 38200615]
126. Cunningham S, Foote L, Sowder M, Cunningham C. Interprofessional education and collaboration: a simulation-based learning experience focused on common and complementary skills in an acute care environment. J Interprof Care 2018 May;32(3):395-398. [doi: 10.1080/13561820.2017.1411340] [Medline: 29265889]

127. Zhan Y, Zhao S, Yuan J, Liu H, Liu Y, Gui L, et al. Prevalence and influencing factors on fatigue of first-line nurses combating with COVID-19 in China: a descriptive cross-sectional study. Curr Med Sci 2020 Aug;40(4):625-635 [FREE Full text] [doi: 10.1007/s11596-020-2226-9] [Medline: 32767264]
128. Umoren R, Scott P, Sweigart L, Gossett E, Hodson-Carlton K, Johnson M, et al. A comparison of teamwork attitude changes with virtual TeamSTEPPS ® simulations in health professional students. Journal of Interprofessional Education & Practice 2018 Mar;10:51-55 [FREE Full text] [doi: 10.1016/j.xjep.2017.12.001]
129. Weile J, Nebsbjerg MA, Ovesen SH, Paltved C, Ingeman ML. Simulation-based team training in time-critical clinical presentations in emergency medicine and critical care: a review of the literature. Adv Simul (Lond) 2021 Jan 20;6(1):3 [FREE Full text] [doi: 10.1186/s41077-021-00154-4] [Medline: 33472706]
130. Ferri P, Rovesti S, Barbieri A, Giuliani E, Vivarelli C, Panzera N. Interprofessional High-Fidelity Simulation on Nursing Students' Collaborative Attitudes: A Quasi-experimental Study Using a Mixed-Methods Approach. In: Kubincová Z, Lancia L, Popescu E, Nakayama M, Scarano V, Gil A, editors. Methodologies and Intelligent Systems for Technology Enhanced Learning, 10th International Conference. Workshops. MIS4TEL 2020. Advances in Intelligent Systems and Computing, vol 1236. Cham, Switzerland: Springer; 2021:99-110.
131. Alzahrani KH, Abutalib RA, Elsheikh AM, Alzahrani LK, Khoshhal KI. The need for non-technical skills education in orthopedic surgery. BMC Med Educ 2023 Apr 19;23(1):262 [FREE Full text] [doi: 10.1186/s12909-023-04196-2] [Medline: 37076848]
132. Gordon M, Fell CWR, Box H, Farrell M, Stewart A. Learning health 'safety' within non-technical skills interprofessional simulation education: a qualitative study. Med Educ Online 2017;22(1):1272838 [FREE Full text] [doi: 10.1080/10872981.2017.1272838] [Medline: 28178920]
133. Alexandrino H, Martinho B, Ferreira L, Baptista S. Non-technical skills and teamwork in trauma: from the emergency department to the operating room. Front Med (Lausanne) 2023;10:1319990 [FREE Full text] [doi: 10.3389/fmed.2023.1319990] [Medline: 38116034]
134. Patterson MD, Geis GL, Falcone RA, LeMaster T, Wears RL. In situ simulation: detection of safety threats and teamwork training in a high risk emergency department. BMJ Qual Saf 2013 Jun 20;22(6):468-477. [doi: 10.1136/bmjqs-2012-000942] [Medline: 23258390]
135. Rudolph JW, Raemer DB, Simon R. Establishing a safe container for learning in simulation: the role of the presimulation briefing. Simul Healthc 2014 Dec;9(6):339-349. [doi: 10.1097/SIH.0000000000000047] [Medline: 25188485]
136. Palaganas JC, Epps C, Raemer DB. A history of simulation-enhanced interprofessional education. J Interprof Care 2014 Mar 30;28(2):110-115. [doi: 10.3109/13561820.2013.869198] [Medline: 24372044]
137. Liaw SY, Choo T, Wu LT, Lim WS, Choo H, Lim SM, et al. Wow, woo, win - healthcare students' and facilitators' experiences of interprofessional simulation in three-dimensional virtual world: a qualitative evaluation study. Nurse Educ Today 2021 Oct;105:105018. [doi: 10.1016/j.nedt.2021.105018] [Medline: 34175564]
138. Hood RJ, Maltby S, Keynes A, Kluge MG, Nalivaiko E, Ryan A, et al. Development and pilot implementation of TACTICS VR: a virtual reality-based stroke management workflow training application and training framework. Front Neurol 2021 Nov 11;12:665808 [FREE Full text] [doi: 10.3389/fneur.2021.665808] [Medline: 34858305]
139. Elendu C, Amaechi DC, Okatta AU, Amaechi EC, Elendu TC, Ezeh CP, et al. The impact of simulation-based training in medical education: a review. Medicine (Baltimore) 2024 Jul 05;103(27):e38813 [FREE Full text] [doi: 10.1097/MD.0000000000038813] [Medline: 38968472]
140. Schram A, Jensen HI, Gamborg M, Lindhard M, Rölfing J, Kjaergaard-Andersen G, et al. Exploring the relationship between simulation-based team training and sick leave among healthcare professionals: a cohort study across multiple hospital sites. BMJ Open 2023 Oct 29;13(10):e076163 [FREE Full text] [doi: 10.1136/bmjopen-2023-076163] [Medline: 37899150]
141. Wayne DB, Butter J, Siddall VJ, Fudala MJ, Wade LD, Feinglass J, et al. Mastery learning of advanced cardiac life support skills by internal medicine residents using simulation technology and deliberate practice. J Gen Intern Med 2006 Mar;21(3):251-256 [FREE Full text] [doi: 10.1111/j.1525-1497.2006.00341.x] [Medline: 16637824]
142. McGaghie WC, Issenberg SB, Cohen ER, Barsuk JH, Wayne DB. Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. Acad Med 2011 Jun;86(6):706-711 [FREE Full text] [doi: 10.1097/ACM.0b013e318217e119] [Medline: 21512370]
143. Liaw SY, Chan SW, Chen F, Hooi SC, Siau C. Comparison of virtual patient simulation with mannequin-based simulation for improving clinical performances in assessing and managing clinical deterioration: randomized controlled trial. J Med Internet Res 2014 Sep 17;16(9):e214 [FREE Full text] [doi: 10.2196/jmir.3322] [Medline: 25230684]
144. van Soeren M, Devlin-Cop S, Macmillan K, Baker L, Egan-Lee E, Reeves S. Simulated interprofessional education: an analysis of teaching and learning processes. J Interprof Care 2011 Nov;25(6):434-440. [doi: 10.3109/13561820.2011.592229] [Medline: 21899398]
145. Forstater A, Sicks S, Collins L, Schmidt E. Team SAFE: A large-scale interprofessional simulation-based TeamSTEPPS® curriculum. Journal of Interprofessional Education & Practice 2019 Sep;16:100221 [FREE Full text] [doi: 10.1016/j.xjep.2018.12.002]

146. Morrell BLM, Cecil KA, Nichols AM, Moore ES, Carmack JN, Hetzler KE, et al. Interprofessional Education Week: the impact of active and passive learning activities on students' perceptions of interprofessional education. J Interprof Care 2021;35(5):799-802. [doi: 10.1080/13561820.2020.1856798] [Medline: 33451254]
147. Freeman S, Eddy SL, McDonough M, Smith MK, Okoroafor N, Jordt H, et al. Active learning increases student performance in science, engineering, and mathematics. Proc Natl Acad Sci U S A 2014 Jun 10;111(23):8410-8415 [FREE Full text] [doi: 10.1073/pnas.1319030111] [Medline: 24821756]
148. Boedeker P, Schlingmann T, Kailin J, Nair A, Foldes C, Rowley D, et al. Active versus passive learning in large-group sessions in medical school: a randomized cross-over trial investigating effects on learning and the feeling of learning. Med Sci Educ 2025 Feb;35(1):459-467. [doi: 10.1007/s40670-024-02219-1] [Medline: 40144125]
149. Reime MH, Johnsgaard T, Kvam FI, Aarflot M, Engeberg JM, Breivik M, et al. Learning by viewing versus learning by doing: a comparative study of observer and participant experiences during an interprofessional simulation training. J Interprof Care 2017 Jan;31(1):51-58. [doi: 10.1080/13561820.2016.1233390] [Medline: 27849424]
150. Ying Y, Yacob M, Khambati H, Seabrook C, Gerridzen L. Does being in the hot seat matter? Effect of passive vs active learning in surgical simulation. Am J Surg 2020 Sep;220(3):593-596 [FREE Full text] [doi: 10.1016/j.amjsurg.2020.01.052] [Medline: 32057411]
151. Jiang J, Fu S, Ma Y, Wang J, Koo M. Comparative impact of active participation and observation in simulation-based emergency care education on knowledge, learning effectiveness, and satisfaction among undergraduate nursing students. Teaching and Learning in Nursing 2024 Jul;19(3):e566-e573 [FREE Full text] [doi: 10.1016/j.teln.2024.04.003]
152. Martella AM, Martella RC, Yatcilla JK, Newson A, Shannon EN, Voorhis C. How rigorous is active learning research in STEM education? An examination of key internal validity controls in intervention studies. Educ Psychol Rev 2023 Nov 04;35(4):1. [doi: 10.1007/s10648-023-09826-1]
153. Couarraze S, Saint Jean M, Decormeille G, Houze Cerfon C, Minville V, Fourcade O, et al. Short term effects of simulation training on stress, anxiety and burnout in critical care health professionals: before and after study. Clinical Simulation in Nursing 2023 Feb;75:25-32 [FREE Full text] [doi: 10.1016/j.ecns.2022.12.001]
154. Yu T, Webster CS, Weller JM. Simulation in the medical undergraduate curriculum to promote interprofessional collaboration for acute care: a systematic review. BMJ Simul Technol Enhanc Learn 2016;2(3):90-96 [FREE Full text] [doi: 10.1136/bmjstel-2016-000103] [Medline: 35519428]
155. Shamputa IC, Kek B, Waycott L, Fournier T, McCarville S, Doucet J, et al. Exploring the efficacy of a virtual first year interprofessional education event. Healthcare (Basel) 2022 Aug 14;10(8):1 [FREE Full text] [doi: 10.3390/healthcare10081539] [Medline: 36011196]
156. Haerling KA. Cost-utility analysis of virtual and mannequin-based simulation. Simul Healthc 2018 Feb;13(1):33-40. [doi: 10.1097/SIH.0000000000000280] [Medline: 29373382]
157. Chávez-Valenzuela P, Kappes M, Sambuceti C, Díaz-Guio DA. "Challenges in the implementation of inter-professional education programs with clinical simulation for health care students: A scoping review". Nurse Educ Today 2025 Mar;146:106548 [FREE Full text] [doi: 10.1016/j.nedt.2024.106548] [Medline: 39740591]
158. Duffull S, Peterson A, Chai B, Cho F, Opoku J, Sissing T, et al. Exploring a scalable real-time simulation for interprofessional education in pharmacy and medicine. MedEdPublish (2016) 2020;9:240 [FREE Full text] [doi: 10.15694/mep.2020.000240.1] [Medline: 38058926]
159. Wetzlmair L, Kitema GF, O'Carroll V, El-Awaisi A, Power A, Owens M, et al. The impact of COVID-19 on the delivery of interprofessional education: it's not all bad news. British Journal of Midwifery 2021 Dec 02;29(12):699-705. [doi: 10.12968/bjom.2021.29.12.699]
160. Liaw SY, Sutini, Chua WL, Tan JZ, Levett-Jones T, Ashokka B, et al. Desktop virtual reality versus face-to-face simulation for team-training on stress levels and performance in clinical deterioration: a randomised controlled trial. J Gen Intern Med 2023 Jan;38(1):67-73 [FREE Full text] [doi: 10.1007/s11606-022-07557-7] [Medline: 35501626]
161. Goldsworthy S, Goodhand K, Baron S, Button D, Hunter S, McNeill L, et al. Co-debriefing virtual simulations: an international perspective. Clinical Simulation in Nursing 2022 Feb;63:1-4 [FREE Full text] [doi: 10.1016/j.ecns.2021.10.007]
162. Coggins A, Zaklama R, Szabo RA, Diaz-Navarro C, Scalese RJ, Krogh K, et al. Twelve tips for facilitating and implementing clinical debriefing programmes. Med Teach 2021 May;43(5):509-517. [doi: 10.1080/0142159X.2020.1817349] [Medline: 33032476]
163. Kumar P, Paton C, Simpson HM, King CM, McGowan N. Is interprofessional co-debriefing necessary for effective interprofessional learning within simulation-based education? International Journal of Healthcare Simulation 2021 Sep 01;1(1):49-55. [doi: 10.54531/INRX6536]
164. Holmes C, Mellanby E. Debriefing strategies for interprofessional simulation-a qualitative study. Adv Simul (Lond) 2022 Jun 18;7(1):18 [FREE Full text] [doi: 10.1186/s41077-022-00214-3] [Medline: 35717254]
165. Badowski D, Wells-Beede E. State of prebriefing and debriefing in virtual simulation. Clinical Simulation in Nursing 2022 Jan;62:42-51. [doi: 10.1016/j.ecns.2021.10.006]
166. Dale-Tam J, Thompson K, Dale L. Creating psychological safety during a virtual simulation session. Clinical Simulation in Nursing 2021 Aug;57:14-17. [doi: 10.1016/j.ecns.2021.01.017]

167. Violato E, MacPherson J, Edwards M, MacPherson C, Renaud M. The use of simulation best practices when investigating virtual simulation in health care: a scoping review. Clinical Simulation in Nursing 2023 Jun;79:28-39. [doi: 10.1016/j.ecns.2023.03.001]
168. Levin H, Cheng A, Catena H, Chatfield J, Cripps A, Bissett W, et al. Debriefing frameworks and methods. In: Clinical Simulation: Education, Operations and Engineering. New York, NY: Academic Press; 2019:483-505.
169. Cheng A, Palaganas J, Eppich W, Rudolph J, Robinson T, Grant V. Co-debriefing for simulation-based education: a primer for facilitators. Simul Healthc 2015 Apr;10(2):69-75. [doi: 10.1097/SIH.0000000000000077] [Medline: 25710318]
170. Boet S, Bould MD, Bruppacher HR, Desjardins F, Chandra DB, Naik VN. Looking in the mirror: self-debriefing versus instructor debriefing for simulated crises. Crit Care Med 2011 Jun;39(6):1377-1381. [doi: 10.1097/CCM.0b013e31820eb8be] [Medline: 21317645]
171. Sawyer T, Eppich W, Brett-Fleegler M, Grant V, Cheng A. More than one way to debrief: a critical review of healthcare simulation debriefing methods. Simul Healthc 2016 Jun;11(3):209-217. [doi: 10.1097/SIH.0000000000000148] [Medline: 27254527]
172. Fanning RM, Gaba DM. The role of debriefing in simulation-based learning. Simul Healthc 2007;2(2):115-125. [doi: 10.1097/SIH.0b013e3180315539] [Medline: 19088616]
173. Salik I, Paige JT. Debriefing the Interprofessional Team in Medical Simulation. Treasure Island, FL: StatPearls Publishing; 2023.
174. Andersen P, Coverdale S, Kelly M, Forster S. Interprofessional simulation: developing teamwork using a two-tiered debriefing approach. Clinical Simulation in Nursing 2018 Jul;20:15-23 [FREE Full text] [doi: 10.1016/j.ecns.2018.04.003]
175. Tannenbaum SI, Cerasoli CP. Do team and individual debriefs enhance performance? A meta-analysis. Hum Factors 2013 Feb;55(1):231-245. [doi: 10.1177/0018720812448394] [Medline: 23516804]
176. Comer M. Rethinking reflection-in-action: what did Schön really mean? Nurse Educ Today 2016 Jan;36:4-6 [FREE Full text] [doi: 10.1016/j.nedt.2015.08.021] [Medline: 26385252]
177. Schön DA. The Reflective Practitioner: How Professionals Think In Action. New York, NY: Basic Books; 1984.
178. Kolb DA. Experiential Learning: Experience as the Source of Learning and Development. Essex, England: Financial Times Prentice Hall; 2014.
179. Poore JA, Dawson JC, Dunbar D, Parrish K. Debriefing interprofessionally: a tool for recognition and reflection. Nurse Educ 2019;44(1):25-28. [doi: 10.1097/NNE.0000000000000518] [Medline: 29538051]
180. Brown DK, Wong AH, Ahmed RA. Evaluation of simulation debriefing methods with interprofessional learning. J Interprof Care 2018 Jul 19;32(1):779-781. [doi: 10.1080/13561820.2018.1500451] [Medline: 30024297]
181. Boet S, Pigford A, Fitzsimmons A, Reeves S, Triby E, Bould MD. Interprofessional team debriefings with or without an instructor after a simulated crisis scenario: an exploratory case study. J Interprof Care 2016 Nov;30(6):717-725. [doi: 10.1080/13561820.2016.1181616] [Medline: 27309589]
182. Boet S, Bould MD, Sharma B, Revees S, Naik VN, Triby E, et al. Within-team debriefing versus instructor-led debriefing for simulation-based education: a randomized controlled trial. Ann Surg 2013 Jul;258(1):53-58. [doi: 10.1097/SLA.0b013e31829659e4] [Medline: 23728281]
183. Gunasingam N, Burns K, Edwards J, Dinh M, Walton M. Reducing stress and burnout in junior doctors: the impact of debriefing sessions. Postgrad Med J 2015 Apr;91(1074):182-187. [doi: 10.1136/postgradmedj-2014-132847] [Medline: 25755266]
184. Colville GA, Smith JG, Brierley J, Citron K, Nguru NM, Shaunak PD, et al. Coping with staff burnout and work-related posttraumatic stress in intensive care. Pediatr Crit Care Med 2017 Jul;18(7):e267-e273. [doi: 10.1097/PCC.0000000000001179] [Medline: 28459762]
185. Bakker AB, Demerouti E. Job demands-resources theory: taking stock and looking forward. J Occup Health Psychol 2017 Jul;22(3):273-285. [doi: 10.1037/ocp0000056] [Medline: 27732008]
186. Hawes K, Goldstein J, Vessella S, Tucker R, Lechner BE. Providing support for neonatal intensive care unit health care professionals: a bereavement debriefing program. Am J Perinatol 2022 Mar;39(4):401-408. [doi: 10.1055/s-0040-1716481] [Medline: 32894870]
187. Azizoddin DR, Vella Gray K, Dundin A, Szyld D. Bolstering clinician resilience through an interprofessional, web-based nightly debriefing program for emergency departments during the COVID-19 pandemic. J Interprof Care 2020;34(5):711-715. [doi: 10.1080/13561820.2020.1813697] [Medline: 32990108]
188. Traylor AM, Tannenbaum SI, Thomas EJ, Salas E. Helping healthcare teams save lives during COVID-19: insights and countermeasures from team science. Am Psychol 2021 Jan;76(1):1-13 [FREE Full text] [doi: 10.1037/amp0000750] [Medline: 33119329]
189. Copeland D, Liska H. Implementation of a post-code pause: extending post-event debriefing to include silence. J Trauma Nurs 2016;23(2):58-64. [doi: 10.1097/JTN.0000000000000187] [Medline: 26953532]
190. Burt L, Clark L, Park C. Stronger together: learner reactions on a team-based, interprofessional first death simulation experience. J Interprof Care 2024 Jan 02;38(1):95-103. [doi: 10.1080/13561820.2023.2232408] [Medline: 37422861]

191. Rose SC, Ashari NA, Davies JM, Solis L, O'Neill TA. Interprofessional clinical event debriefing-does it make a difference? Attitudes of emergency department care providers to INFO clinical event debriefings. CJEM 2022 Nov;24(7):695-701. [doi: 10.1007/s43678-022-00361-6] [Medline: 36138325]
192. Kam AJ, Gonsalves CL, Nordlund SV, Hale SJ, Twiss J, Cupido C, et al. Implementation and facilitation of post-resuscitation debriefing: a comparative crossover study of two post-resuscitation debriefing frameworks. BMC Emerg Med 2022 Sep 02;22(1):152 [FREE Full text] [doi: 10.1186/s12873-022-00707-4] [Medline: 36056328]
193. Schmidt M, Haglund K. Debrief in emergency departments to improve compassion fatigue and promote resiliency. J Trauma Nurs 2017;24(5):317-322. [doi: 10.1097/JTN.0000000000000315] [Medline: 28885522]
194. Meichenbaum D. Stress inoculation training: A preventative and treatment approach. In: Lehrer PM, Woolfolk RL, Sime WE, editors. Principles and Practice of Stress Management. New York, NY: The Guilford Press; 2007:497-516.
195. Skegg E, McElroy C, Mudgway M, Hamill J. Debriefing to improve interprofessional teamwork in the operating room: a systematic review. J Nurs Scholarsh 2023 Nov;55(6):1179-1188 [FREE Full text] [doi: 10.1111/jnu.12924] [Medline: 37452720]
196. Hitchner L, Yore M, Burk C, Mason J, Sawtelle Vohra S. The resident experience with psychological safety during interprofessional critical event debriefings. AEM Educ Train 2023 Apr;7(2):e10864 [FREE Full text] [doi: 10.1002/aet2.10864] [Medline: 37013133]
197. McElroy C, Skegg E, Mudgway M, Murray N, Holmes L, Weller J, et al. Psychological safety and hierarchy in operating room debriefing: reflexive thematic analysis. J Surg Res 2024 Mar;295:567-573 [FREE Full text] [doi: 10.1016/j.jss.2023.11.054] [Medline: 38086257]
198. Servotte J, Welch-Horan TB, Mullan P, Piazza J, Ghuysen A, Szyld D. Development and implementation of an end-of-shift clinical debriefing method for emergency departments during COVID-19. Adv Simul (Lond) 2020 Nov 11;5(1):32 [FREE Full text] [doi: 10.1186/s41077-020-00150-0] [Medline: 33292850]
199. Umoren R, Kim S, Gray MM, Best JA, Robins L. Interprofessional model on speaking up behaviour in healthcare professionals: a qualitative study. BMJ Lead 2022 Mar;6(1):15-19. [doi: 10.1136/leader-2020-000407] [Medline: 35537012]
200. Britt TW, Shuffler ML, Pegram RL, Xoxakos P, Rosopa PJ, Hirsh E, et al. Job demands and resources among healthcare professionals during virus pandemics: a review and examination of fluctuations in mental health strain during COVID-19. Appl Psychol 2021 Jan;70(1):120-149 [FREE Full text] [doi: 10.1111/apps.12304] [Medline: 33362329]
201. Kaiser S, Patras J, Adolfsen F, Richardsen AM, Martinussen M. Using the job demands–resources model to evaluate work-related outcomes among Norwegian health care workers. Sage Open 2020 Aug 01;10(3):1. [doi: 10.1177/2158244020947436]
202. Turcotte M, Etherington C, Rowe J, Duong A, Kaur M, Talbot Z, et al. Effectiveness of interprofessional teamwork interventions for improving occupational well-being among perioperative healthcare providers: a systematic review. J Interprof Care 2023 Nov 02;37(6):904-921. [doi: 10.1080/13561820.2022.2137116] [Medline: 36373205]
203. Amorøe TN, Rystedt H, Oxelmark L, Dieckmann P, Andréll P. How theories of complexity and resilience affect interprofessional simulation-based education: a qualitative analysis of facilitators' perspectives. BMC Med Educ 2023 Oct 02;23(1):717 [FREE Full text] [doi: 10.1186/s12909-023-04690-7] [Medline: 37784048]
204. Chen J, Bamberger PA, Song Y, Vashdi DR. The effects of team reflexivity on psychological well-being in manufacturing teams. J Appl Psychol 2018 Apr;103(4):443-462. [doi: 10.1037/apl0000279] [Medline: 29239644]
205. Diver S, Buccheri N, Ohri C. The value of healthcare worker support strategies to enhance wellbeing and optimise patient care. Future Healthc J 2021 Mar;8(1):e60-e66 [FREE Full text] [doi: 10.7861/fhj.2020-0176] [Medline: 33791478]
206. Browning ED, Cruz JS. Reflective debriefing: a social work intervention addressing moral distress among ICU nurses. J Soc Work End Life Palliat Care 2018;14(1):44-72. [doi: 10.1080/15524256.2018.1437588] [Medline: 29488856]
207. Sorensen D, Cristancho S, Soh M, Varpio L. Team stress and its impact on interprofessional teams: a narrative review. Teach Learn Med 2024 Jan 10;36(2):163-173. [doi: 10.1080/10401334.2022.2163400] [Medline: 36625564]
208. Weaver JL, Bowers CA, Salas E. Stress and Teams: Performance Effects and Interventions. In: Hancock PA, Desmond PA, editors. Stress, Workload, and Fatigue. Boca Raton, FL: CRC Press; 2001:83-106.
209. Razinskas S, Hoegl M. A multilevel review of stressor research in teams. J Organ Behavior 2019 Dec 15;41(2):185-209 [FREE Full text] [doi: 10.1002/job.2420]
210. Lazarus RS, Folkman S. Stress, Appraisal, and Coping. New York, NY: Springer Publishing Company; 1984.
211. Goodnite PM. Stress: a concept analysis. Nurs Forum 2014;49(1):71-74. [doi: 10.1111/nuf.12044] [Medline: 24456555]
212. Zhang H, Wu C, Yan J, Liu J, Wang P, Hu M, et al. The relationship between role ambiguity, emotional exhaustion and work alienation among chinese nurses two years after COVID-19 pandemic: a cross-sectional study. BMC Psychiatry 2023 Jul 18;23(1):516 [FREE Full text] [doi: 10.1186/s12888-023-04923-5] [Medline: 37464335]
213. Kivimäki M, Vanhala A, Pentti J, Länsisalmi H, Virtanen M, Elovainio M, et al. Team climate, intention to leave and turnover among hospital employees: prospective cohort study. BMC Health Serv Res 2007 Oct 23;7:170 [FREE Full text] [doi: 10.1186/1472-6963-7-170] [Medline: 17956609]
214. Cullati S, Bochatay N, Maître F, Laroche T, Muller-Juge V, Blondon KS, et al. When team conflicts threaten quality of care: a study of health care professionals' experiences and perceptions. Mayo Clin Proc Innov Qual Outcomes 2019 Mar;3(1):43-51 [FREE Full text] [doi: 10.1016/j.mayocpiqo.2018.11.003] [Medline: 30899908]

215. Tsai C, Kung P, Wang S, Tsai T, Tsai W. The association between the workload of emergency physicians and the outcomes of acute myocardial infarction: a population-based study. Sci Rep 2023 Dec 01;13(1):21212 [FREE Full text] [doi: 10.1038/s41598-023-48150-0] [Medline: 38040727]
216. Teng C, Shyu Y, Chiou W, Fan H, Lam S. Interactive effects of nurse-experienced time pressure and burnout on patient safety: a cross-sectional survey. Int J Nurs Stud 2010 Nov;47(11):1442-1450 [FREE Full text] [doi: 10.1016/j.ijnurstu.2010.04.005] [Medline: 20472237]
217. Savelsbergh C, Gevers JMP, van der Heijden BIJM, Poell RF. Team role stress. Group & Organization Management 2012 Jan 17;37(1):67-100 [FREE Full text] [doi: 10.1177/1059601111431977]
218. Kamphuis W, Delahaij R, de Vries TA. Team coping: cross-level influence of team member coping activities on individual burnout. Front Psychol 2021 Nov 4;12:711981 [FREE Full text] [doi: 10.3389/fpsyg.2021.711981] [Medline: 34803799]
219. Rodriguez RM, Medak AJ, Baumann BM, Lim S, Chinnock B, Frazier R, et al. Academic emergency medicine physicians' anxiety levels, stressors, and potential stress mitigation measures during the acceleration phase of the COVID-19 pandemic. Acad Emerg Med 2020 Aug 21;27(8):700-707 [FREE Full text] [doi: 10.1111/acem.14065] [Medline: 32569419]
220. Alharbi J, Jackson D, Usher K. Personal characteristics, coping strategies, and resilience impact on compassion fatigue in critical care nurses: a cross-sectional study. Nurs Health Sci 2020 Mar;22(1):20-27. [doi: 10.1111/nhs.12650] [Medline: 31670474]
221. Sangal RB, Wrzesniewski A, DiBenigno J, Reid E, Ulrich A, Liebhardt B, et al. Work team identification associated with less stress and burnout among front-line emergency department staff amid the COVID-19 pandemic. leader 2020 Oct 27;5(1):51-54. [doi: 10.1136/leader-2020-000331]
222. Sasangohar F, Jones SL, Masud FN, Vahidy FS, Kash BA. Provider burnout and fatigue during the COVID-19 pandemic: lessons learned from a high-volume intensive care unit. Anesth Analg 2020 Jul;131(1):106-111 [FREE Full text] [doi: 10.1213/ANE.0000000000004866] [Medline: 32282389]
223. Martínez-López JÁ, Lázaro-Pérez C, Gómez-Galán J, Fernández-Martínez MDM. Psychological impact of COVID-19 emergency on health professionals: burnout incidence at the most critical period in Spain. J Clin Med 2020 Sep 20;9(9):3029 [FREE Full text] [doi: 10.3390/jcm9093029] [Medline: 32962258]
224. Yörük S, Güler D. The relationship between psychological resilience, burnout, stress, and sociodemographic factors with depression in nurses and midwives during the COVID-19 pandemic: a cross-sectional study in Turkey. Perspect Psychiatr Care 2021 Jan;57(1):390-398. [doi: 10.1111/ppc.12659] [Medline: 33103773]
225. Yehia AC, Moreira J, Premaor MO. Burnout syndrome in resident physicians: a study after the third COVID-19 wave in two tertiary hospitals of southeastern Brazil. PLoS One 2025 Apr 7;20(4):e0321443 [FREE Full text] [doi: 10.1371/journal.pone.0321443] [Medline: 40193375]
226. Mercado M, Wachter K, Schuster RC, Mathis CM, Johnson E, Davis OI, et al. A cross-sectional analysis of factors associated with stress, burnout and turnover intention among healthcare workers during the COVID-19 pandemic in the United States. Health Soc Care Community 2022 Sep;30(5):e2690-e2701. [doi: 10.1111/hsc.13712] [Medline: 35037346]
227. Duarte I, Teixeira A, Castro L, Marina S, Ribeiro C, Jácome C, et al. Burnout among Portuguese healthcare workers during the COVID-19 pandemic. BMC Public Health 2020 Dec 07;20(1):1885 [FREE Full text] [doi: 10.1186/s12889-020-09980-z] [Medline: 33287794]
228. Kelly D, Schroeder S, Leighton K. Anxiety, depression, stress, burnout, and professional quality of life among the hospital workforce during a global health pandemic. J Rural Health 2022 Sep;38(4):795-804 [FREE Full text] [doi: 10.1111/jrh.12659] [Medline: 35315126]
229. Batanda I. Prevalence of burnout among healthcare professionals: a survey at fort portal regional referral hospital. Npj Ment Health Res 2024 May 06;3(1):16 [FREE Full text] [doi: 10.1038/s44184-024-00061-2] [Medline: 38710834]
230. Mańkowska B. Burnout phenomenon still unresolved. The current state in theory and implications for public interest. Front. Organ. Psychol 2025 Mar 24;3:1. [doi: 10.3389/forgp.2025.1549253]
231. Bandura A. Social learning theory. Upper Saddle River, NJ: Prentice Hall; 1977.
232. Datta V. Madness and the movies: an undergraduate module for medical students. Int Rev Psychiatry 2009 Jun;21(3):261-266. [doi: 10.1080/09540260902748001] [Medline: 19459103]
233. Fritz GK, Poe RO. The role of a cinema seminar in psychiatric education. Am J Psychiatry 1979 Feb;136(2):207-210. [doi: 10.1176/ajp.136.2.207] [Medline: 760551]
234. Darbyshire D, Baker P. Cinema in medical education – has it penetrated the mainstream? J Med Mov 2011;7(1):8-14 [FREE Full text]
235. Kalra G. Talking about stigma towards mental health professionals with psychiatry trainees: a movie club approach. Asian J Psychiatr 2012 Sep;5(3):266-268. [doi: 10.1016/j.ajp.2012.06.005] [Medline: 22981056]
236. Kalra GS. Lights, camera and action: learning necrophilia in a psychiatry movie club. J Forensic Leg Med 2013 Apr;20(3):139-142. [doi: 10.1016/j.jflm.2012.06.001] [Medline: 23472790]
237. Batistatou A, Doulis EA, Tiniakos D, Anogiannaki A, Charalabopoulos K. The introduction of medical humanities in the undergraduate curriculum of Greek medical schools: challenge and necessity. Hippokratia 2010 Oct;14(4):241-243 [FREE Full text] [Medline: 21311630]

238. Sánchez JC, Gutiérrez JC, Morales MD. Cinema and theater as training tools for health students. Fam Med 2010 Jun;42(6):398-399. [Medline: 20526905]
239. Alexander M, Pavlov A, Lenahan P. Lights, camera, action: using film to teach the ACGME competencies. Fam Med 2007 Jan;39(1):20-23. [Medline: 17186442]
240. Salajegheh M, Sohrabpour AA, Mohammadi E. Exploring medical students' perceptions of empathy after cinemeducation based on Vygotsky's theory. BMC Med Educ 2024 Jan 29;24(1):94 [FREE Full text] [doi: 10.1186/s12909-024-05084-z] [Medline: 38287370]
241. Charon R. Knowing, seeing, and telling in medicine. Lancet 2021 Dec 04;398(10316):2068-2070 [FREE Full text] [doi: 10.1016/S0140-6736(21)02656-8] [Medline: 34863343]
242. Charon R. The patient-physician relationship. Narrative medicine: a model for empathy, reflection, profession, and trust. JAMA 2001 Oct 17;286(15):1897-1902. [doi: 10.1001/jama.286.15.1897] [Medline: 11597295]
243. Kemp SJ, Day G. Teaching medical humanities in the digital world: affordances of technology-enhanced learning. Med Humanit 2014 Dec 16;40(2):125-130. [doi: 10.1136/medhum-2014-010518] [Medline: 25031422]
244. Yerkes R, Dodson J. The relation of strength of stimulus to rapidity of habit - formation. J. Comp. Neurol. Psychol 2004 Oct 07;18(5):459-482 [FREE Full text] [doi: 10.1002/cne.920180503]
245. Rao D, Elshafei A, Nguyen M, Hatzenbuehler ML, Frey S, Go VF. A systematic review of multi-level stigma interventions: state of the science and future directions. BMC Med 2019 Feb 15;17(1):41 [FREE Full text] [doi: 10.1186/s12916-018-1244-y] [Medline: 30770756]
246. Thornicroft G, Mehta N, Clement S, Evans-Lacko S, Doherty M, Rose D, et al. Evidence for effective interventions to reduce mental-health-related stigma and discrimination. The Lancet 2016 Mar;387(10023):1123-1132. [doi: 10.1016/s0140-6736(15)00298-6]
247. Waqas A, Malik S, Fida A, Abbas N, Mian N, Miryala S, et al. Interventions to reduce stigma related to mental illnesses in educational institutes: a systematic review. Psychiatr Q 2020 Sep;91(3):887-903 [FREE Full text] [doi: 10.1007/s11126-020-09751-4] [Medline: 32372401]
248. Hauke C, Alister I. Jung and Film: Post-Jungian Takes on the Moving Image. New York, NY: Routledge; 2001.
249. Chang HM, Ivonin L, Díaz M, Català A, Chen W, Rauterberg GWM. From mythology to psychology: identifying archetypal symbols in movies. Technoetic Arts 2013;11(2):99-113. [doi: 10.1386/tear.11.2.99_1]
250. Jung CG. The Practice of Psychotherapy: Essays on the Psychology of the Transference and Other Subjects. New York, NY: Princeton University Press; 1985.
251. Hillman J. The Myth of Analysis: Three Essays in Archetypal Psychology. Evanston, IL: Northwestern University Press; 1998.
252. Gramaglia C, Jona A, Imperatori F, Torre E, Zeppegno P. Cinema in the training of psychiatry residents: focus on helping relationships. BMC Med Educ 2013 Jun 21;13(1):90 [FREE Full text] [doi: 10.1186/1472-6920-13-90] [Medline: 23800186]
253. Prince MJ, Felder RM. Inductive teaching and learning methods: definitions, comparisons, and research bases. Journal of Engineering Education 2006;95(2):123-138 [FREE Full text] [doi: 10.1002/j.2168-9830.2006.tb00884.x]
254. Persico L, Ramakrishnan S, Catena R, Charnetski M, Fogg N, Jones M, et al. The impact of prebriefing on simulation learning outcomes – a systematic review protocol. Clinical Simulation in Nursing 2024 Apr;89:101507. [doi: 10.1016/j.ecns.2023.101507]
255. Wood D, Bruner JS, Ross G. The role of tutoring in problem solving. J Child Psychol Psychiatry 1976 Apr 07;17(2):89-100. [doi: 10.1111/j.1469-7610.1976.tb00381.x] [Medline: 932126]
256. Lecomte F, Jaffrelot M. Chapter 33 - Prebriefing and Briefing. In: Chiniara G, editor. Clinical Simulation: Education, Operations and Engineering. New York, NY: Academic Press; 2019:471-482.
257. Sawyer TL, Deering S. Adaptation of the US Army's After-Action Review for simulation debriefing in healthcare. Simul Healthc 2013 Dec;8(6):388-397. [doi: 10.1097/SIH.0b013e31829ac85c] [Medline: 24096913]
258. Raymond-Dufresne É. Chapter 29 - Simulation for Critical Care. In: Chiniara G, editor. Clinical Simulation: Education, Operations and Engineering. New York, NY: Academic Press; 2019:419-430.
259. Windle G. What is resilience? A review and concept analysis. Rev. Clin. Gerontol 2010 Dec 21;21(2):152-169. [doi: 10.1017/s0959259810000420]
260. Bergström J, van Winsen R, Henriqson E. On the rationale of resilience in the domain of safety: a literature review. Reliability Engineering & System Safety 2015 Sep;141:131-141. [doi: 10.1016/j.ress.2015.03.008]
261. Gunderson L. Ecological resilience—in theory and application. Annu. Rev. Ecol. Syst 2000 Nov;31(1):425-439 [FREE Full text] [doi: 10.1146/annurev.ecolsys.31.1.425]
262. Denckla CA, Cicchetti D, Kubzansky LD, Seedat S, Teicher MH, Williams DR, et al. Psychological resilience: an update on definitions, a critical appraisal, and research recommendations. Eur J Psychotraumatol 2020 Nov 10;11(1):1822064 [FREE Full text] [doi: 10.1080/20008198.2020.1822064] [Medline: 33244362]
263. Lovell LP, Atherley AEN, Watson HR, King RD. An exploration of burnout and resilience among emergency physicians at three teaching hospitals in the English-speaking Caribbean: a cross-sectional survey. Lancet Reg Health Am 2022 Nov;15:100357 [FREE Full text] [doi: 10.1016/j.lana.2022.100357] [Medline: 36778072]

264. Luthar SS, Cicchetti D, Becker B. The construct of resilience: a critical evaluation and guidelines for future work. Child Dev 2000;71(3):543-562 [FREE Full text] [doi: 10.1111/1467-8624.00164] [Medline: 10953923]
265. Bonanno GA. Loss, trauma, and human resilience: have we underestimated the human capacity to thrive after extremely aversive events? Am Psychol 2004 Jan;59(1):20-28. [doi: 10.1037/0003-066X.59.1.20] [Medline: 14736317]
266. Fletcher D, Sarkar M. Psychological resilience: a review and critique of definitions, concepts, and theory. European Psychologist 2013;18(1):12-23. [doi: 10.1027/1016-9040/a000124]
267. Johnson J. Resilience: The bi-dimensional framework. In: Wood AM, Johnson J, editors. The Wiley Handbook of Positive Clinical Psychology. Hoboken, NJ: John Wiley & Sons Ltd; 2016:73-88.
268. Blanchard EE, Trost Z, Brown MR, Shum C, Meese M. Combining stress inoculation with virtual reality simulation training of malignant hyperthermia. Adv Simul (Lond) 2024 Aug 16;9(1):35 [FREE Full text] [doi: 10.1186/s41077-024-00308-0] [Medline: 39152517]
269. Varker T, Devilly GJ. An analogue trial of inoculation/resilience training for emergency services personnel: proof of concept. J Anxiety Disord 2012 Aug;26(6):696-701. [doi: 10.1016/j.janxdis.2012.01.009] [Medline: 22464031]
270. Feng X, Long T, Han P. Fostering team resilience through stressor exposure: an identity-based model and the role of charismatic leadership. Group & Organization Management 2025 Feb 19:1. [doi: 10.1177/10596011251322544]
271. Peña CJ. Early-life stress sensitizes response to future stress: evidence and mechanisms. Neurobiol Stress 2025 Mar;35:100716 [FREE Full text] [doi: 10.1016/j.ynstr.2025.100716] [Medline: 40134543]
272. Alliger GM, Cerasoli CP, Tannenbaum SI, Vessey WB. Team resilience. Organizational Dynamics 2015 Jul;44(3):176-184. [doi: 10.1016/j.orgdyn.2015.05.003]
273. Knowles MS, Holton III EF, Swanson RA. The Adult Learner: The definitive classic in adult education and human resource development. New York, NY: Routledge; 2015.
274. Pereira-Lima K, Mata DA, Loureiro SR, Crippa JA, Bolsoni LM, Sen S. Association between physician depressive symptoms and medical errors: a systematic review and meta-analysis. JAMA Netw Open 2019 Nov 01;2(11):e1916097 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.16097] [Medline: 31774520]
275. Park C, Holtschneider M. Interprofessional simulation: creative ways to integrate education and goal setting in the practice environment. J Nurses Prof Dev 2016;32(2):102-104. [doi: 10.1097/NND.0000000000000253] [Medline: 26985755]
276. Adams P. Exploring social constructivism: theories and practicalities. Education 3-13 2006 Oct;34(3):243-257. [doi: 10.1080/03004270600898893]
277. Vygotsky LS. Mind in Society: The Development of Higher Psychological Processes. Cambridge, MA: Harvard University Press; 1978.
278. Mezirow J. Learning as Transformation: Critical Perspectives on a Theory in Progress. San Francisco, CA: Jossey-Bass Inc Pub; 2000.
279. Philpott J, Batty H. Learning best together: social constructivism and global partnerships in medical education. Med Educ 2009 Sep;43(9):923-924 [FREE Full text] [doi: 10.1111/j.1365-2923.2009.03436.x] [Medline: 19709017]
280. Kriz WC. A systemic-constructivist approach to the facilitation and debriefing of simulations and games. Simulation & Gaming 2008 Jun 20;41(5):663-680. [doi: 10.1177/1046878108319867]
281. Fleming T. Mezirow and the Theory of Transformative Learning. In: Wang V, editor. Critical Theory and Transformative Learning: Information Science Reference. New York, NY: IGI Global Scientific Publishing; 2018:120-136.
282. Mezirow J. How critical reflection triggers transformative learning. Fostering critical reflection in adulthood 1990;1(20):1-6 [FREE Full text]
283. Edmondson A, Lei Z. Psychological safety: the history, renaissance, and future of an interpersonal construct. Annu. Rev. Organ. Psychol. Organ. Behav 2014 Mar 21;1(1):23-43 [FREE Full text] [doi: 10.1146/annurev-orgpsych-031413-091305]
284. Kolbe M, Eppich W, Rudolph J, Meguerdichian M, Catena H, Cripps A, et al. Managing psychological safety in debriefings: a dynamic balancing act. BMJ Simul Technol Enhanc Learn 2020;6(3):164-171 [FREE Full text] [doi: 10.1136/bmjstel-2019-000470] [Medline: 35518370]
285. Ganley BJ, Linnard-Palmer L. Academic safety during nursing simulation: perceptions of nursing students and faculty. Clinical Simulation in Nursing 2012 Feb;8(2):e49-e57. [doi: 10.1016/j.ecns.2010.06.004]
286. de Carvalho Filho MA, Schaafsma ES, Tio RA. Debriefing as an opportunity to develop emotional competence in health profession students: faculty, be prepared!. Scientia Medica 2018 Jan 26;28(1):28805. [doi: 10.15448/1980-6108.2018.1.28805]
287. Frazier ML, Fainshmidt S, Klinger RL, Pezeshkan A, Vracheva V. Psychological safety: a meta‐analytic review and extension. Personnel Psychology 2016 Oct 14;70(1):113-165. [doi: 10.1111/peps.12183]
288. Lateef F. Maximizing learning and creativity: understanding psychological safety in simulation-based learning. J Emerg Trauma Shock 2020;13(1):5-14 [FREE Full text] [doi: 10.4103/JETS.JETS_96_19] [Medline: 32395043]
289. Seelandt JC, Walker K, Kolbe M. "A debriefer must be neutral" and other debriefing myths: a systemic inquiry-based qualitative study of taken-for-granted beliefs about clinical post-event debriefing. Adv Simul (Lond) 2021 Mar 04;6(1):7 [FREE Full text] [doi: 10.1186/s41077-021-00161-5] [Medline: 33663598]
290. Cheng A, Grant V, Huffman J, Burgess G, Szyld D, Robinson T, et al. Coaching the debriefer: peer coaching to improve debriefing quality in simulation programs. Simul Healthc 2017 Oct;12(5):319-325. [doi: 10.1097/SIH.0000000000000232] [Medline: 28538446]

291. Palaganas JC, Charnetski M, Dowell S, Chan AKM, Leighton K. Cultural considerations in debriefing: a systematic review of the literature. BMJ Simul Technol Enhanc Learn 2021;7(6):605-610 [FREE Full text] [doi: 10.1136/bmjstel-2020-000857] [Medline: 35520973]
292. van Schaik S, Plant J, O'Brien B. Challenges of interprofessional team training: a qualitative analysis of residents' perceptions. Educ Health (Abingdon) 2015;28(1):52-57. [doi: 10.4103/1357-6283.161883] [Medline: 26261115]
293. Oriot D, Alinier G, Alinier G. Pocket book for simulation debriefing in healthcare. New York, NY: Springer; 2018.
294. Dahl R. The concept of power. Syst. Res 2007 Jan 17;2(3):201-215 [FREE Full text] [doi: 10.1002/bs.3830020303]
295. Bynum WE, Sukhera J. Perfectionism, power, and process: what we must address to dismantle mental health stigma in medical education. Acad Med 2021 May 01;96(5):621-623. [doi: 10.1097/ACM.0000000000004008] [Medline: 33885411]
296. Looman N, van Woezik T, van Asselt D, Scherpbier-de Haan N, Fluit C, de Graaf J. Exploring power dynamics and their impact on intraprofessional learning. Med Educ 2022 Apr;56(4):444-455 [FREE Full text] [doi: 10.1111/medu.14706] [Medline: 34841565]
297. Janss R, Rispens S, Segers M, Jehn KA. What is happening under the surface? Power, conflict and the performance of medical teams. Med Educ 2012 Sep 15;46(9):838-849. [doi: 10.1111/j.1365-2923.2012.04322.x] [Medline: 22891905]
298. Weller J, Boyd M, Cumin D. Teams, tribes and patient safety: overcoming barriers to effective teamwork in healthcare. Postgrad Med J 2014 Mar;90(1061):149-154. [doi: 10.1136/postgradmedj-2012-131168] [Medline: 24398594]
299. Robertson K, Ju M, O'Brien BC, van Schaik SM, Bochatay N. Exploring the role of power during debriefing of interprofessional simulations. J Interprof Care 2022 Feb 02:1-9. [doi: 10.1080/13561820.2022.2029371] [Medline: 35109751]
300. French Jr JRP, Raven B. The bases of social power. In: Cartwright D, editor. Studies in Social Power. Ann Arbor, MI: University of Michigan; 1959:150-167.
301. Raven BH. The bases of power and the power/interaction model of interpersonal influence. Anal Soc Iss & Public Policy 2008 Nov 24;8(1):1-22. [doi: 10.1111/j.1530-2415.2008.00159.x]
302. Joyce B, Carr D, Smart A, Armour D, Gormley GJ. Learning better together? A scoping review of in-person interprofessional undergraduate simulation. Adv Simul (Lond) 2025 Apr 29;10(1):24 [FREE Full text] [doi: 10.1186/s41077-025-00351-5] [Medline: 40301989]
303. Gotwals BA, Scholtz S. Video-enhanced simulation in pediatric end-of-life care. Nurs Educ Perspect 2016;37(6):360-362. [doi: 10.1097/01.nep.0000000000000077]
304. Lotrakul M, Sumrithe S, Saipanish R. Reliability and validity of the Thai version of the PHQ-9. BMC Psychiatry 2008 Jun 20;8(1):1. [doi: 10.1186/1471-244x-8-46]
305. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
306. Prober CG, Heath C. Lecture halls without lectures — a proposal for medical education. N Engl J Med 2012 May 03;366(18):1657-1659. [doi: 10.1056/nejmp1202451]
307. Kadeangadi DM, Mudigunda SS. Cinemeducation: using films to teach medical students. J Sci Soc 2019;46(3):73-74. [doi: 10.4103/jss.JSS_1_20]
308. Ayikoru M, Park HY. Films and critical pedagogy in management education: a tourism studies context. Academy of Management Learning and Education 2019 Sep;18(3):414-432 [FREE Full text] [doi: 10.5465/amle.2015.0134]
309. Ignacio J, Dolmans D, Scherpbier A, Rethans J, Chan S, Liaw SY. Stress and anxiety management strategies in health professions' simulation training: a review of the literature. BMJ Simul Technol Enhanc Learn 2016 Apr 06;2(2):42-46 [FREE Full text] [doi: 10.1136/bmjstel-2015-000097] [Medline: 35518196]
310. Kolbe M, Grande B, Spahn DR. Briefing and debriefing during simulation-based training and beyond: content, structure, attitude and setting. Best Pract Res Clin Anaesthesiol 2015 Mar;29(1):87-96. [doi: 10.1016/j.bpa.2015.01.002] [Medline: 25902470]
311. Fraser KL, Meguerdichian MJ, Haws JT, Grant VJ, Bajaj K, Cheng A. Cognitive Load Theory for debriefing simulations: implications for faculty development. Adv Simul (Lond) 2018 Dec 29;3(1):28 [FREE Full text] [doi: 10.1186/s41077-018-0086-1] [Medline: 30619626]
312. Hughes PG, Hughes KE. Briefing Prior to Simulation Activity. Treasure Island, FL: StatPearls Publishing; 2023.
313. Sharoff L. Simulation: pre-briefing preparation, clinical judgment and reflection. What is the connection? J Contemp Med 2015;5(2):1. [doi: 10.16899/CTD.49922]
314. Rutherford-Hemming T, Lioce L, Breymier T. Guidelines and essential elements for prebriefing. Simul Healthc 2019 Dec;14(6):409-414. [doi: 10.1097/SIH.0000000000000403] [Medline: 31804425]
315. Tyerman J, Luctkar-Flude M, Graham L, Coffey S, Olsen-Lynch E. Pre-simulation preparation and briefing practices for healthcare professionals and students: a systematic review protocol. JBI Database System Rev Implement Rep 2016 Aug;14(8):80-89. [doi: 10.11124/JBISRIR-2016-003055] [Medline: 27635748]
316. McDermott DS, Ludlow J, Horsley E, Meakim C. Healthcare simulation standards of Best PracticeTM prebriefing: preparation and briefing. Clinical Simulation in Nursing 2021 Sep;58:9-13. [doi: 10.1016/j.ecns.2021.08.008]

317. Kolbe M, Marty A, Seelandt J, Grande B. How to debrief teamwork interactions: using circular questions to explore and change team interaction patterns. Adv Simul (Lond) 2016;1:29 [FREE Full text] [doi: 10.1186/s41077-016-0029-7] [Medline: 29449998]
318. Brennan B, Hintz W, Zacher R. Prebriefing in simulation from the nursing student perspective: a qualitative descriptive study. Clinical Simulation in Nursing 2024 Dec;97:101634 [FREE Full text] [doi: 10.1016/j.ecns.2024.101634]
319. Nascimento JDSG, Nascimento KGD, Alves MG, Braga FTMM, Regino DDSG, Dalri MCB. Effectiveness of co-debriefing to develop clinical skills in basic life support: randomized pilot study. Rev Gaucha Enferm 2022;43(spe):e20220032 [FREE Full text] [doi: 10.1590/1983-1447.2022.20220032.en] [Medline: 36383828]
320. Sweller J. Cognitive load theory, learning difficulty, and instructional design. Learning and Instruction 1994;4(4):295-312 [FREE Full text] [doi: 10.1016/0959-4752(94)90003-5]
321. Alahmari DM, Alsahli FM, Alghamdi SA, Alomair OI, Alghamdi A, Alsaadi MJ. Assessment of patient knowledge level towards MRI safety before the scanning in Saudi Arabia. Int J Gen Med 2022;15:6289-6299 [FREE Full text] [doi: 10.2147/IJGM.S368652] [Medline: 35924179]
322. Kristensen TS, Borritz M, Villadsen E, Christensen KB. The Copenhagen Burnout Inventory: a new tool for the assessment of burnout. Work & Stress 2005 Jul;19(3):192-207. [doi: 10.1080/02678370500297720]
323. Zsido AN, Teleki SA, Csokasi K, Rozsa S, Bandi SA. Development of the short version of the spielberger state-trait anxiety inventory. Psychiatry Res 2020 Sep;291:113223 [FREE Full text] [doi: 10.1016/j.psychres.2020.113223] [Medline: 32563747]
324. Helton WS, Näswall K. Short Stress State Questionnaire: factor structure and state change assessment. European Journal of Psychological Assessment 2015 Jun;31(1):20-30. [doi: 10.1027/1015-5759/A000200]
325. Piperac P, Todorovic J, Terzic-Supic Z, Maksimovic A, Karic S, Pilipovic F, et al. The validity and reliability of the Copenhagen Burnout Inventory for examination of burnout among preschool teachers in Serbia. Int J Environ Res Public Health 2021 Jun 24;18(13):6805 [FREE Full text] [doi: 10.3390/ijerph18136805] [Medline: 34202911]
326. Phuekphan P, Aungsuroch Y, Yunibhand J, Chan SWC. Psychometric properties of the Thai version of Copenhagen Burnout Inventory (T-CBI) in Thai nurses. J Health Res 2016;30(2):135-142 [FREE Full text]
327. Radu C, Dabu A, Tudor C, Teleanu D. Efficacy of deep brain stimulation: a comparative analysis of STN and GPI targets and our clinic's experience with movement disorders. Romanian Neurosurgery 2024 Nov 14:120-121 [FREE Full text] [doi: 10.33962/roneuro-2024-123]
328. StataCorp. Spatial Autoregressive Models Reference Manual. College Station, TX: Stata Press; 2017.
329. Matthews G, Szalma J, Panganiban A, Neubauer C, Warm J. Profiling Task Stress with the Dundee Stress State Questionnaire. In: Cavalcanti L, Azevedo S, editors. Psychology of Stress: New Research. Hauppauge, NY: Nova Science Pub Inc; 2013:49-91.
330. Salcedo J, Lackey S, Badillo-Urquiola K. Leveraging Stress and Intrinsic Motivation to Assess Scaffolding During Simulation-Based Training. In: Shumaker R, Lackey S, editors. Virtual, Augmented and Mixed Reality. VAMR 2015. Lecture Notes in Computer Science(), vol 9179. Cham, Switzerland: Springer International Publishing; 2015:309-320.
331. Taylor JM. Psychometric analysis of the Ten-Item Perceived Stress Scale. Psychol Assess 2015 Mar;27(1):90-101. [doi: 10.1037/a0038100] [Medline: 25346996]
332. Celik B, Cagiltay K. The undervalued variable in massive open online course (MOOC) research: an analysis and conceptualization of readiness for online learning in MOOCs. Educ Inf Technol 2023 Feb 22;28(9):11569-11588. [doi: 10.1007/s10639-023-11662-3]
333. Baños J, Bosch F. Using feature films as a teaching tool in medical schools. Educación Médica 2015 Oct;16(4):206-211 [FREE Full text] [doi: 10.1016/j.edumed.2015.09.001]
334. Gonzalez-Caminal G, Gomar-Sancho C, Mastandrea PB, Arrebola-Trias X, Baños J, Cambra-Badii I. Combining simulation and cinemeducation to teach patient safety: a pilot study. Innovations in Education and Teaching International 2021 Oct 28;60(1):80-90. [doi: 10.1080/14703297.2021.1989322]
335. Kolb A, Kolb D. Experiential Learning Theory: A Dynamic, Holistic Approach to Management Learning, Education and Development. In: Armstrong SF, Fukami C, editors. The SAGE Handbook of Management Learning, Education and Development. New York, NY: SAGE Publications Ltd; 2011:42-68.
336. Baukal C, Ausburn F, Ausburn L. A proposed multimedia cone of abstraction: updating a classic instructional design theory. Journal of Educational Technology 2013 Mar 15;9(4):15-24. [doi: 10.26634/jet.9.4.2129]
337. Champoux JE. Film as a teaching resource. Journal of Management Inquiry 1999 Jun 01;8(2):206-217. [doi: 10.1177/105649269982016]
338. Shrivastava SR, Shrivastava PS. Incorporating movies and cinema in the medical education delivery: a curricular innovation. Medical Journal of Dr. D.Y. Patil Vidyapeeth 2022;15(5):662-665. [doi: 10.4103/mjdrdypu.mjdrdypu_296_21]
339. Anderson V, Gifford J, Wildman J. An evaluation of social learning and learner outcomes in a massive open online course (MOOC): a healthcare sector case study. Human Resource Development International 2020 Feb 03;23(3):208-237. [doi: 10.1080/13678868.2020.1721982]
340. Chaw LY, Tang CM. Driving high inclination to complete massive open online courses (MOOCs): motivation and engagement factors for learners. The Electronic Journal of e-Learning 2019 Jun 01;17(2):118-130. [doi: 10.34190/jel.17.2.05]

341. Connolly F, De Brún A, McAuliffe E. A narrative synthesis of learners' experiences of barriers and facilitators related to effective interprofessional simulation. J Interprof Care 2022;36(2):222-233 [FREE Full text] [doi: 10.1080/13561820.2021.1880381] [Medline: 33818255]
342. Watts P, McDermott D, Alinier G, Charnetski M, Ludlow J, Horsley E, et al. Healthcare simulation standards of Best PracticeTM simulation design. Clinical Simulation in Nursing 2021 Sep;58:14-21 [FREE Full text] [doi: 10.1016/j.ecns.2021.08.009]
343. Banks S, Stanley MJ, Brown S, Matthew W. Simulation-based interprofessional education: a nursing and social work collaboration. J Nurs Educ 2019 Feb 01;58(2):110-113. [doi: 10.3928/01484834-20190122-09] [Medline: 30721312]
344. Maddock A. The relationships between stress, burnout, mental health and well-being in social workers. The British Journal of Social Work 2024;54(2):668-686. [doi: 10.1093/bjsw/bcad232]
345. Koplow S, Morris M, Rone-Adams S, Hettrick H, Litwin B, Soontupe L, et al. Student experiences with engagement in a nursing and physical therapy interprofessional education simulation. IJAHSP 2020;18(1):1. [doi: 10.46743/1540-580x/2020.1842]
346. Chipman ML, Schreiber CM, Fey JM, Lane SJ, DiLisio C, Mallory LA. Engagement across professions: a mixed methods study of debriefing after interprofessional team training. Simul Healthc 2024 Aug 01;19(4):228-234. [doi: 10.1097/SIH.0000000000000736] [Medline: 37440428]
347. Piette AE, Attoe C, Humphreys R, Cross S, Kowalski C. Interprofessional simulation training for community mental health teams: findings from a mixed methods study. J Interprof Care 2018 Nov;32(6):762-770. [doi: 10.1080/13561820.2018.1511524] [Medline: 30142281]
348. Beverly EA, Miller S, Love M, Love C. Feasibility of a cinematic-virtual reality program educating health professional students about the complexity of geriatric care: pilot pre-post study. JMIR Aging 2025 Feb 12;8:e64633-e64633 [FREE Full text] [doi: 10.2196/64633] [Medline: 39937111]
349. Astbury J, Ferguson J, Silverthorne J, Willis S, Schafheutle E. High-fidelity simulation-based education in pre-registration healthcare programmes: a systematic review of reviews to inform collaborative and interprofessional best practice. J Interprof Care 2021 Jun 12;35(4):622-632. [doi: 10.1080/13561820.2020.1762551] [Medline: 32530344]
350. Liaw S, Ooi S, Mildon R, Ang E, Lau T, Chua W. Translation of an evidence-based virtual reality simulation-based interprofessional education into health education curriculums: an implementation science method. Nurse Educ Today 2022 Mar;110:105262 [FREE Full text] [doi: 10.1016/j.nedt.2021.105262] [Medline: 35063778]
351. Souchet AD, Lourdeaux D, Pagani A, Rebenitsch L. A narrative review of immersive virtual reality's ergonomics and risks at the workplace: cybersickness, visual fatigue, muscular fatigue, acute stress, and mental overload. Virtual Reality 2022 Jul 16;27(1):19-50. [doi: 10.1007/s10055-022-00672-0]
352. Parvez A, Rao S, Khan A. Physiological Responsiveness to VR and non-VR Environments2023. 2024 Presented at: International Conference on Computational Science and Computational Intelligence (CSCI); December 13-15, 2023; Las Vegas, NV. [doi: 10.1109/csci62032.2023.00095]
353. Chheang V, Weston BT, Cerda RW, Au B, Giera B, Bremer PT, et al. A Virtual Environment for Collaborative Inspection in Additive Manufacturing. 2024 Presented at: CHI Conference on Human Factors in Computing Systems; May 11-16, 2024; Honolulu, HI. [doi: 10.1145/3613905.3650730]

Abbreviations

CBI: Copenhagen Burnout Inventory
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
cMOOC: connectivist massive open online course
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
DSSQ: Dundee Stress State Questionnaire
EM: emergency medicine
ER-VIPE: Emergency Room Virtual Simulation Interprofessional Education
GEE: generalized estimating equation
HCW: health care worker
HP: health point
iMOOC: integrated massive open online course
IPC: interprofessional collaboration
IPE: interprofessional education
ISBAR: Identify, Situation, Background, Assessment, and Recommendation
ITT: intention-to-treat
JD-R: Job Demands-Resources
MANOVA: multivariate analysis of variance
MOOC: massive open online course

PEARLS: Promoting Excellence and Reflective Learning in Simulation
PHQ-9: Patient Health Questionnaire-9
PP: per-protocol
PPE: personal protective equipment
RCT: randomized controlled trial
SIMBIE: simulation-based interprofessional education
SIT: stress inoculation training
SMD: standardized mean difference
STAI: State-Trait Anxiety Inventory
TeamSTEPPS: Team Strategies and Tools to Enhance Performance and Patient Safety
VR: virtual reality
WHO: World Health Organization
xMOOC: massive open online course that is an extension of something else

Edited by B Lesselroth; submitted 31.12.24; peer-reviewed by R Nooripour; comments to author 25.02.25; revised version received 12.07.25; accepted 14.08.25; published 24.09.25.

Please cite as:
Srikasem S, Seephom S, Viriyopase A, Phutrakool P, Khowinthaseth S, Narajeenron K, ER-VIPE Study Group
Comparing the Effectiveness of Multimodal Learning Using Computer-Based and Immersive Virtual Reality Simulation–Based Interprofessional Education With Co-Debriefing, Medical Movies, and Massive Online Open Courses for Mitigating Stress and Long-Term Burnout in Medical Training: Quasi-Experimental Study
JMIR Med Educ 2025;11:e70726
URL: https://mededu.jmir.org/2025/1/e70726
doi:10.2196/70726
PMID:

Original Paper

Correlation Between Electroencephalogram Brain-to-Brain Synchronization and Team Strategies and Tools to Enhance Performance and Patient Safety Scores During Online Hexad Virtual Simulation-Based Interprofessional Education: Cross-Sectional Correlational Study

Atthaphon Viriyopase[1], PhD; Khuansiri Narajeenron[1], MD, MSc, MHPE, CHSE

Department of Emergency Medicine, Faculty of Medicine, Chulalongkorn University, King Chulalongkorn Memorial Hospital, Bangkok, Thailand

Corresponding Author:
Khuansiri Narajeenron, MD, MSc, MHPE, CHSE
Department of Emergency Medicine
Faculty of Medicine
Chulalongkorn University, King Chulalongkorn Memorial Hospital
1873 MDCU Faculty, Rama IV Road
Pathumwan
Bangkok, 10330
Thailand
Phone: 66 0855054209
Email: khuansiri.n@chula.ac.th

Abstract

Background: Team performance is crucial in crisis situations. Although the Thai version of Team Strategies and Tools to Enhance Performance and Patient Safety (TeamSTEPPS) has been validated, challenges remain due to its subjective evaluation. To date, no studies have examined the relationship between electroencephalogram (EEG) activity and team performance, as assessed by TeamSTEPPS, during virtual simulation-based interprofessional education (SIMBIE), where face-to-face communication is absent.

Objective: This study aims to investigate the correlation between EEG-based brain-to-brain synchronization and TeamSTEPPS scores in multiprofessional teams participating in virtual SIMBIE sessions.

Methods: This single-center study involved 90 participants (15 groups of 6 simulated professionals: 1 medical doctor, 2 nurses, 1 pharmacist, 1 medical technologist, and 1 radiological technologist). Each group completed two 30-minute virtual SIMBIE sessions focusing on team training in a crisis situation involving COVID-19 pneumonia with a difficult airway, resulting in 30 sessions in total. The TeamSTEPPS scores of each participant across 5 domains were independently assessed by 2 trained raters based on screen recording, and their average values were used. The scores of participants in the same session were aggregated to generate a group TeamSTEPPS score, representing group-level performance. EEG data were recorded using wireless EEG acquisition devices and computed for total interdependence (TI), which represents brain-to-brain synchronization. The TI values of participants in the same session were aggregated to produce a group TI, representing group-level brain-to-brain synchronization. We investigated the Pearson correlations between the TI and the scores at both the group and individual levels.

Results: Interrater reliability for the TeamSTEPPS scores among 12 raters indicated good agreement on average (mean 0.73, SD 0.18; range 0.32-0.999). At the individual level, the Pearson correlations between the TI and the scores were weak and not statistically significant across all TeamSTEPPS domains (all adjusted $P \geq .05$). However, strongly negative, statistically significant correlations between the group TI and the group TeamSTEPPS scores in the alpha frequency band (8-12 Hz) of the anterior brain area were found across all TeamSTEPPS domains after correcting for multiple comparisons (mean −0.87, SD 0.06; range −0.93 to −0.8).

Conclusions: Strong negative correlations between the group TI and the group TeamSTEPPS scores were observed in the anterior alpha activity during online hexad virtual SIMBIE. These findings suggest that anterior alpha TI may serve as an objective metric for assessing TeamSTEPPS-based team performance.

(JMIR Med Educ 2025;11:e69725) doi:10.2196/69725

KEYWORDS

brain-to-brain synchronization; EEG; electroencephalogram; emergency medicine; hyperscanning; interprofessional education; simulation; team communication; TeamSTEPPS; teamwork; virtual simulation

Introduction

The COVID-19 pandemic has highlighted the crucial importance of effective interprofessional collaboration in health care, especially in high-pressure environments such as emergency and critical care settings. Beyond individual cognitive abilities, medical professionals in these settings must work as a team, requiring shared goals, clear role understanding, and continuous communication [1]. Team Strategies and Tools to Enhance Performance and Patient Safety (TeamSTEPPS) has become a widely recognized framework for improving teamwork, communication, and overall clinical performance in such contexts. Supported by a robust body of evidence, TeamSTEPPS has been integrated into various simulation-based training programs, including online virtual simulations, which have become prominent during and after COVID-19 [2]. Although TeamSTEPPS includes 5 specific domains with validated psychometric evidence, the assessment process remains complex. Challenges include rater training, interrater reliability, and the subjective nature of evaluations. Additionally, the time and cost required for training raters add further complications [3]. Despite the availability of validated tools to measure teamwork (eg, a one-shot public goods game [4]), there are few objective measures—such as total interdependence (TI) (Multimedia Appendix 1) [5], debiased weighted phase lag index [6], and intersite phase clustering [7]—that directly link team dynamics with real-time electroencephalogram (EEG) neurological responses during collaborative tasks. An advantage of these objective measures is that they tend to predict team performance better than validated tools that are inherently subjective [8].

Hyperscanning—a method of simultaneously measuring brain activity in 2 or more individuals—has been used to explore the neural basis of social dynamics [9,10] using various neuroimaging techniques, including functional magnetic resonance imaging [11,12], functional near-infrared spectroscopy [13], EEG [14-17]. Different experimental paradigms have been used to investigate social dynamics, including movement synchronization [18], social decision-making [19], joint attention [20], team problem solving [8], team coordination [21], team communication [17], and classroom engagement [14]. Recent advancements in affordable wireless EEG technologies make EEG well suited for studying brain-to-brain synchronization, which refers to the coordinated brain activity between 2 or more individuals [22], during collaborative tasks in less controlled environments.

Reinero et al [8] demonstrated that brain-to-brain synchronization, measured by TI, predicts team performance on problem-solving tasks better than traditional self-report measures with real-time brain activity monitored using affordable wireless EEG devices. Dikker et al [14] further identified that TI-based brain-to-brain synchronization predicts class engagement and social dynamics, while Bevilacqua et al [23] extended Dikker's findings to show that social factors, such as perceived closeness, can predict cognitive outcomes, including academic performance. These studies highlight that EEG-based brain-to-brain synchronization reflects neural alignment and provides insights into the cognitive and emotional processes underlying teamwork. However, to our knowledge, previous studies focused on face-to-face interactions, and no study has explored the correlation between EEG-based brain-to-brain synchronization and TeamSTEPPS performance in virtual simulation–based interprofessional education (SIMBIE), where face-to-face communication is absent. In additional, Guttmann et al [24] introduced stricter sample size methodologies for EEG studies, whereas Asaad and Sheth [25] emphasized the need to balance rigor and practicality, ensuring ethical and scientific standards with meaningful interpretations using efficient sample sizes. We incorporated these considerations into this study's sample size justification to enhance research quality.

This study aims to address this gap by investigating the correlations between TI-based EEG brain-to-brain synchronization and TeamSTEPPS scores, from both individual and team perspectives, in 6 multiprofessional student groups during online virtual SIMBIE sessions. On the basis of previous studies, we hypothesized that these correlations would likely not be strongly positive owing to the absence of face-to-face communication in online virtual SIMBIE sessions. The findings may support the development of an objective, evidence-based approach to assessing team dynamics in the absence of face-to-face communication. This approach has the potential to enhance the validity and reliability of TeamSTEPPS evaluations and to facilitate automatic, real-time feedback.

Methods

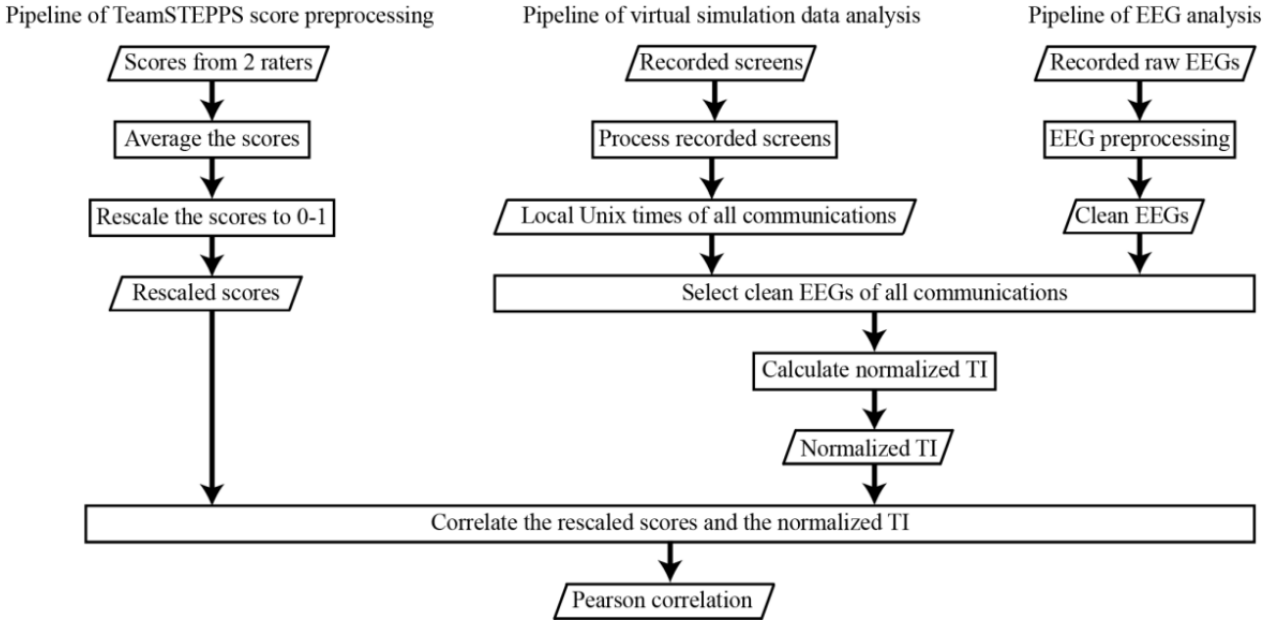
Study Design and Setting

This study was part of a larger study [26,27] conducted at the Chulalongkorn Healthcare Advanced Multi-Profession Simulation Center in a single university hospital from August 2022 to September 2023. It was designed as a cross-sectional correlational study using quantitative approaches. The variables of interest were TeamSTEPPS scores, which measure participants' teamwork and communication constructs of participants, and EEG features of brain activity capturing brain-to-brain synchronization; further details on TeamSTEPPS scoring are available elsewhere [28]. Although the scores and EEGs are applicable in free environments, we conducted the study during the typical hours of 6 PM to 9 PM in a controlled laboratory environment within the simulation center, where we controlled potential confounding variables (eg, temperature, visual and auditory noise, arousal confounds, motor confounds, and gameplay fluidity) [29].

This study included 30 sessions of gameplay in a virtual simulation, in which 6 participants each assumed one of 6 unique professions: (1) radiological technologist, (2) medical technologist, (3) medical doctor, (4) pharmacist, (5) circulation nurse, and (6) airway nurse. In the Emergency Room–Virtual Interprofessional Education platform—Thailand's pioneering virtual reality (VR) system for medical interprofessional training designed and developed by Dhanakoses et al [30]—participants were represented as avatars and interacted using microphones and speakers. Multimedia Appendix 2 provides detailed descriptions of the virtually simulated scenario. Each participant attended 2 sessions; the study thus required 90 participants to complete 30 gameplay sessions. Of the 30 sessions, 10 were conducted with participants wearing VR headsets (HTC VIVE Cosmos [31]; Figure S1D in Multimedia Appendix 3), and 20 without. Participants performed cooperative tasks on their personal computers (PCs) in the laboratory. In this scenario, a male patient (aged 70 years) with chronic obstructive pulmonary disease, a history of hypertension, diabetes mellitus, and ceftriaxone allergy arrived at an emergency department with his wife. Recently discharged from an intensive care unit, his vital parameters and laboratory tests suggested hyperkalemia and COVID-19 pneumonia with acute respiratory failure. Each of the 6 participants assumed a distinct role corresponding to a fully qualified, licensed profession within the simulation scenario. The 6 simulated professions were required to collaborate to diagnose the patient and implement a treatment plan. The virtually simulated scenario emphasized the development of interprofessional communication and clinical reasoning skills and provided an ideal controlled environment to study correlations between brain-to-brain synchronization within a team and TeamSTEPPS scores.

During the sessions, we recorded each participant's EEG signal and gameplay, while the corresponding TeamSTEPPS scores were obtained from the larger study [26]. Figure 1 provides an overview of the analysis pipelines for the 3 types of data. The EEG signal was preprocessed to mitigate artifacts, resulting in a clean EEG signal [32], while video recording of gameplay were preprocessed to extract Unix times [33] of verbal communications. The Unix times were used to segment the clean EEG signals corresponding to the communications. The segmented EEG signals were then used to compute normalized values of TI, an EEG feature measuring brain-to-brain synchronization [34]. The normalized TIs were correlated with the scaled TeamSTEPPS scores obtained by scaling the TeamSTEPPS scores to values between 0 and 1.

Figure 1. Flowchart of data analysis pipelines showing the Team Strategies and Tools to Enhance Performance and Patient Safety (TeamSTEPPS) score (left), virtual simulation data (middle), and electroencephalogram (EEG) data (right). Clean EEGs were partitioned by local Unix times of all communications, then used to calculate normalized total interdependence (TI), which was correlated with the rescaled scores.



Study Population

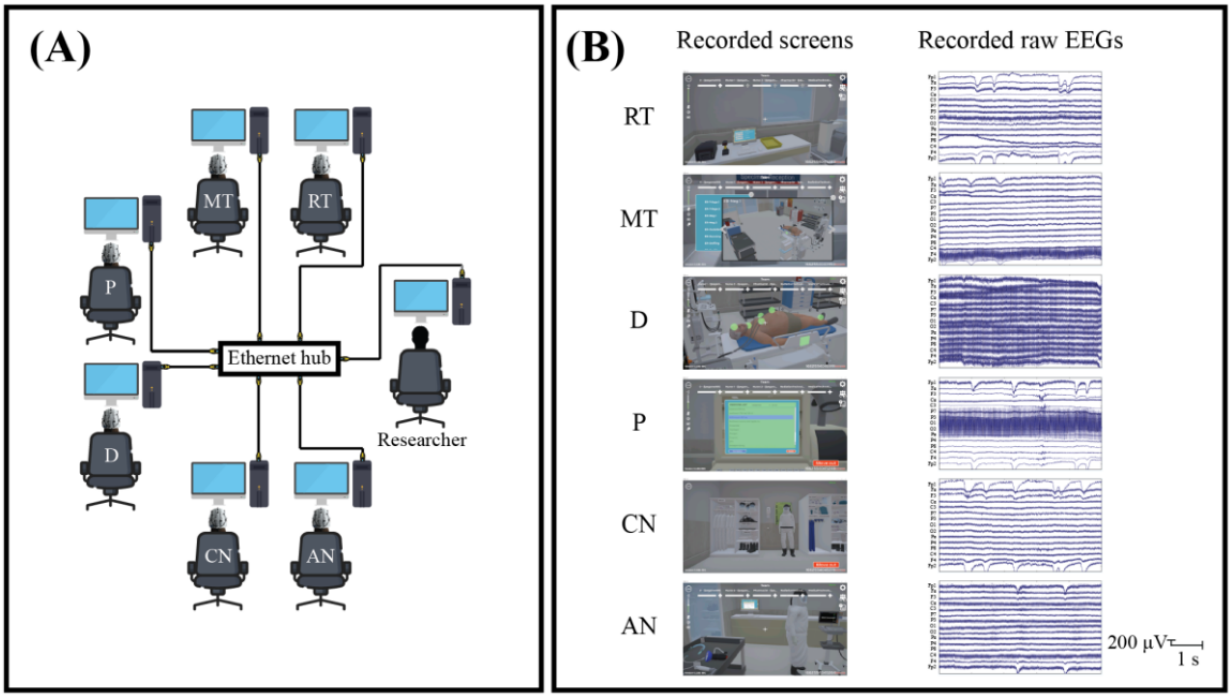
The target population comprised healthy adults aged 18-25 years who did not meet any of the following exclusion criteria: (1) predisposition to major depression, defined as a score ≥9 on the 9-item self-reported Patient Health Questionnaire-9 (PHQ-9) [35]; (2) history of substance use disorder, including cigarette smoking and alcohol addiction; (3) history of neurological or psychiatric disorders, including epilepsy; and (4) periodic use of antidepressants during the 2 weeks before the experiment. The study population, drawn from this target population, consisted of students in radiological technology, medical technology, medicine, pharmacy, and nursing from a university-affiliated hospital. We used convenience sampling by advertising the study through posters and social media on campuses, targeting 5th- or 6th-year medical students and pharmacy students, 4th-year medical technology students, as well as 3rd- or 4th-year radiological technology and nursing students. Students registered for the study through online forms without knowledge of the inclusion or exclusion criteria. Participants who were eligible and completed the experiment received monetary compensation, as indicated in the advertisement. Of 147 potential participants, 85 were eligible for the study, and 5 met the exclusion criteria but joined for practical reasons; we excluded these 5 participants from our

analysis. Note that they joined the sessions without knowing that their EEG activity would be excluded from analysis. Because virtual simulators can induce simulator sickness [36], participants experiencing symptoms were administered a 10 mg oral dose of dimenhydrinate.

Laboratory Setting

The experiments were conducted in a controlled 7×7 m room with 6 participants and a researcher, each using high-performance PCs connected through a low-latency network (Figure 2A). Non-VR sessions used monitors and audio equipment for communication, while VR sessions involved VR headsets and EEG caps, which were set up by research assistants to enhance immersion and reduce simulator sickness. The researcher's PC managed the simulation and EEG data to ensure smooth operation (Figure 2A). Multimedia Appendix 4 provides additional details.

Figure 2. Experiment setup and recorded data. (A) 6 participants' and the researcher's personal computers connected in a star network topology with the ethernet hub as the center node. Collected data include participants' video screens (B, left) and electroencephalogram (EEG) brain activity (B, right). AN: airway nurse; CN: circulation nurse; D: medical doctor; MT: medical technologist; P: pharmacist; RT: radiological technologist.



Data Sources

TeamSTEPPS Scores

The TeamSTEPPS framework was used to assess teamwork and communication skills on a 5-point Likert scale across five topics: (1) team structure, (2) communication, (3) leadership, (4) situation monitoring, and (5) mutual support (as verified and validated in a pervious study [28]). A participant's performance in each topic was quantitatively measured by a total score, calculated as the sum of scores ranging from 1 (worst) to 5 (best) across different aspects representing that topic. Each participant's TeamSTEPPS scores was determined by the 5 scores from the 5 topics. The TeamSTEPPS scores for this study were obtained from the larger study [26].

EEG Signals

EEG signals were recorded simultaneously from 6 participants during the virtual simulation using mobile EEG devices (OpenBCI; Figure S1A in Multimedia Appendix 3) with sintered-electrode caps (Figure S1 in Multimedia Appendix 4). We used 15 passive, gel-based electrodes placed according to the international 10-20 system [37] at the following locations: Fp1, Fp2, F3, Fz, F4, C3, Cz, C4, P7, P3, Pz, P4, P8, O1, and O2 (Figure S1B in Multimedia Appendix 3). The EEG signals were initially referenced to CPz and later rereferenced to an average reference [38] during offline analysis, effectively yielding a net zero potential across the scalp for components derived from independent component analysis, thereby facilitating manual inspection. To ensure signal quality, electrode impedances were maintained below 50 kΩ on average (mean 45.59, SD 116.41 kΩ) [39], validated using OpenBCI software (version 5.1.0) [40]. EEG data was digitized at 125 Hz with high precision, and participants adjusted the sound from the virtual simulation using in-ear speakers to a comfortable level.

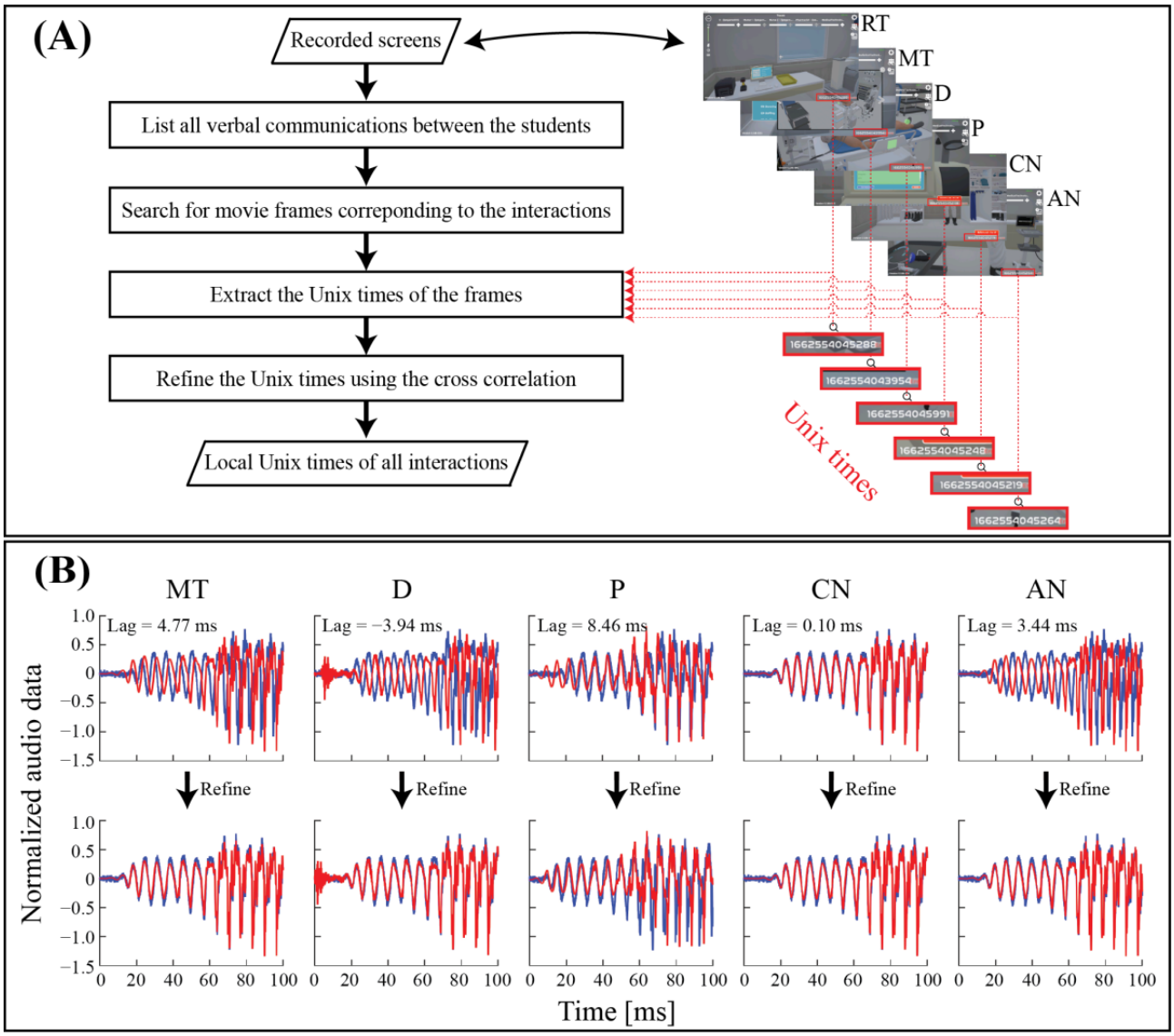
To monitor EEG signals in real time, we developed custom C# software that transmitted data from the participants' PCs to the researcher's PC (Figure S1 in Multimedia Appendix 4 provides a schematic). The process was as follows: first, EEG signals were transferred from the biosensing board of the mobile EEG device to the participant's PC using BrainFlow software (version 4.9.0 [41]). Next, every second (125 samples), a local Unix timestamp [33] was assigned to each sample on the participant's PC. The software then transmitted these 125-sample chunks through Ethernet using Lab Streaming Layer software (version 1.15.2 [42]) to the researcher's PC. EEG data from all 6 participants were received and stored in a proprietary file format for offline analysis (Figure 2B, right column, shows examples

of recorded EEG signals). Additionally, we performed multiple sanity checks to ensure that the data received by the researcher met strict quality standards (Multimedia Appendix 5).

Video Recordings of Virtually Simulated Scenarios

Participants' PC screens during the virtually simulated scenario were recorded at 60 frames per second with a resolution of 2560×1440 pixels, and audio was captured at 48,000 Hz using NVIDIA ShadowPlay software (version 3.25.1.27 [43]; Figure 2B, left column, shows examples of the 6 PC screens). In addition to capturing participants' actions, the screen recordings displayed local Unix timestamps [33], which were obtained using the DateTime function of C# to precisely time each action (Figure 3A, red boxes, shows examples). These video and audio recordings were saved locally on each participant's PC in MP4 format. The recorded screen footage was later used for offline synchronization between the simulated scenario actions and corresponding EEG signals.

Figure 3. Temporal alignment of verbal social interactions. (A) Extraction of temporal information from recorded screens, indicated by Unix times in red boxes. (B) Adjustment of time lags between participants' audio using cross-correlation, with RT as reference (blue) and other participants (red) before (top) and after (bottom) compensation. AN: airway nurse; CN: circulation nurse; D: medical doctor; MT: medical technologist; P: pharmacist; RT: radiological technologist.



Data Preparation

Preprocessing of the TeamSTEPPS Scores

Two trained raters independently rated the TeamSTEPPS scores for each profession (Table 1, "Rater 1" and "Rater 2" rows [26]). We averaged the scores from the 2 raters and scaled them using the minimum and maximum values (Table 1, "Range" rows), resulting in scaled TeamSTEPPS scores with values between 0 and 1 (Table 1, "Scaled AVG" rows). Additionally, we created an overall TeamSTEPPS score by summing all 5 topic scores to summarize each participant's teamwork and communication skills (Table 1, "Overall" column). Analogously, we defined a scaled overall TeamSTEPPS score (Table 1, "Scaled AVG" rows intersecting the "Overall" column). We evaluated the performance of a team using a group TeamSTEPPS score, calculated by averaging the scaled overall TeamSTEPPS scores of all 6 professions in the team.

Table 1. Descriptive statistics of the Team Strategies and Tools to Enhance Performance and Patient Safety (TeamSTEPPS) scores (N=30).

Professions	TeamSTEPPS scores					
	Team structure	Communication	Leadership	Situation monitoring	Mutual support	Overall
Radiological technologist (n=22)						
Range	2-10	4-20	5-25	5-25	4-20	20-100
Rater 1, mean (SD)	5.86 (2.88)	8.45 (2.65)	10.73 (2.35)	12.5 (2.52)	6.27 (1.55)	43.82 (9.73)
Rater 2, mean (SD)	5.64 (3.03)	9.27 (2.45)	10.95 (3.43)	12.59 (2.36)	5.77 (1.15)	44.23 (10.44)
ICC[a]	0.94	0.74	0.73	0.63	0.35	0.89
Scaled AVG[b], mean (SD)	0.47 (0.36)	0.30 (0.15)	0.29 (0.14)	0.38 (0.11)	0.13 (0.07)	0.30 (0.12)
Medical technologist (n=30)						
Range	2-10	4-20	5-25	4-20	3-15	18-90
Rater 1, mean (SD)	5.87 (3.34)	10.23 (3.26)	13.4 (4.38)	11.73 (2.05)	5.70 (3.11)	46.93 (12.74)
Rater 2, mean (SD)	5.87 (3.35)	10.00 (3.36)	13.43 (4.40)	11.67 (2.12)	5.40 (2.77)	46.3 (12.54)
ICC	>0.99	0.96	>0.99	0.99	0.94	0.99
Scaled AVG, mean (SD)	0.48 (0.42)	0.38 (0.20)	0.42 (0.22)	0.48 (0.13)	0.21 (0.24)	0.40 (0.18)
Medical doctor (n=30)						
Range	4-20	4-20	6-30	5-25	4-20	23-115
Rater 1, mean (SD)	12.23 (4.62)	13.13 (4.24)	17.43 (5.68)	16.80 (3.47)	10.33 (4.41)	69.93 (20.53)
Rater 2, mean (SD)	14.07 (2.99)	13.63 (3.77)	21.30 (5.51)	16.67 (3.54)	11.07 (3.41)	76.73 (17.71)
ICC	0.64	0.76	0.64	0.61	0.73	0.81
Scaled AVG, mean (SD)	0.57 (0.22)	0.59 (0.24)	0.56 (0.22)	0.59 (0.16)	0.42 (0.23)	0.55 (0.20)
Pharmacist (n=30)						
Range	2-10	4-20	5-25	2-10	4-20	17-85
Rater 1, mean (SD)	5.90 (3.14)	15.07 (3.95)	18.73 (4.29)	6.10 (2.16)	11.57 (3.88)	57.37 (13.88)
Rater 2, mean (SD)	4.77 (3.51)	13.90 (3.66)	15.33 (3.84)	6.97 (2.65)	8.43 (3.49)	49.40 (14.13)
ICC	0.78	0.56	0.66	0.43	0.51	0.75
Scaled AVG, mean (SD)	0.42 (0.40)	0.66 (0.21)	0.60 (0.20)	0.57 (0.26)	0.38 (0.21)	0.54 (0.20)
Circulation nurse (n=28)						
Range	3-15	4-20	5-25	5-25	4-20	21-105
Rater 1, mean (SD)	9.14 (3.76)	11.75 (3.45)	13.32 (3.67)	12.71 (2.57)	6.43 (2.18)	53.36 (12.89)
Rater 2, mean (SD)	8.71 (3.49)	11.82 (4.11)	12.54 (3.43)	11.75 (2.03)	5.96 (1.40)	50.79 (11.81)
ICC	0.89	0.84	0.65	0.66	0.32	0.92
Scaled AVG, mean (SD)	0.49 (0.29)	0.49 (0.23)	0.40 (0.17)	0.36 (0.11)	0.14 (0.09)	0.37 (0.15)
Airway nurse (n=30)						
Range	3-15	4-20	5-25	5-25	4-20	21-105
Rater 1, mean (SD)	9.13 (4.22)	11.43 (4.20)	15.80 (5.17)	16.37 (4.79)	11.63 (4.32)	64.37 (21.36)
Rater 2, mean (SD)	8.30 (3.68)	11.63 (3.55)	15.97 (3.70)	16.27 (3.51)	12.33 (4.18)	64.50 (16.26)
ICC	0.89	0.77	0.81	0.73	0.75	0.88
Scaled AVG, mean (SD)	0.48 (0.32)	0.47 (0.23)	0.54 (0.21)	0.57 (0.19)	0.50 (0.25)	0.52 (0.22)
Group (n=20)						
Scaled AVG, mean (SD)	0.49 (0.32)	0.49 (0.19)	0.47 (0.17)	0.50 (0.12)	0.31 (0.15)	0.45 (0.17)
Complete group (n=9)						
Scaled AVG, mean (SD)	0.50 (0.40)	0.47 (0.23)	0.47 (0.22)	0.49 (0.13)	0.31 (0.18)	0.45 (0.21)

[a]ICC: intraclass correlation coefficient.

[b]Scaled AVG: scaled average.

EEG Preprocessing

Overall, EEG preprocessing removed noise, corrected artifacts, rereferenced signals, and reduced dimensionality. Cleaned EEG data were filtered (1-40 Hz), grouped into 3 brain regions, and prepared for independent component analysis and TI analysis to ensure high-quality results (details provided in Multimedia Appendix 6).

Preprocessing of the Video Recordings: Aligning EEG Signals With the Recordings

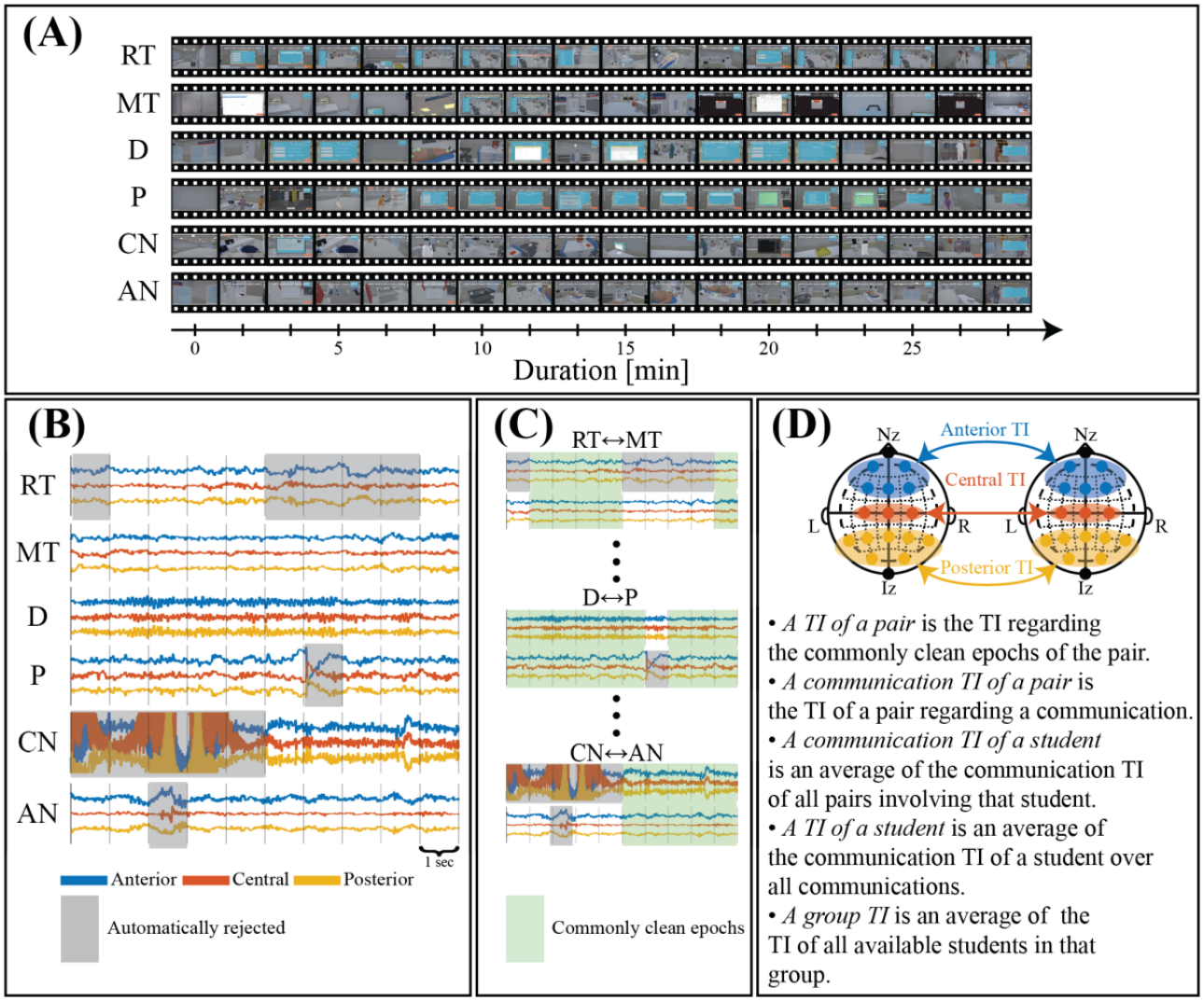
To conduct our hyperscanning study in the virtual simulation, we focused on synchronizing verbal social interactions recorded from participants' screens. Figure 3A (left side) illustrates the flowchart for processing these recordings to obtain Unix times corresponding to the interactions. First, we imported the recordings of 6 participants from the same session into Adobe Premiere Pro (version 22.2.0) and marked the start and end frames of each verbal interaction by visually matching the auditory stimuli of both the sender and receiver (Figure 3B, top row). After identifying these frames, we extracted the corresponding Unix timestamps (Figure 3A, red boxes). To enhance the accuracy of the timestamps, we adjusted for any lag using cross-correlation analysis of the average auditory stimuli from both left and right channels (Figure 3B, bottom row). This refined timing was then aligned with the clean EEG signal timestamps, allowing us to isolate periods of EEG data associated with specific verbal interactions. These selected time periods were crucial for calculating brain-to-brain synchronization among participants.

Computation of Normalized TI

In this study, we used the TI method to assess brain-to-brain synchronization among participants during the virtual simulation. TI is a technique that analyzes the temporal relationships between two signals, extending beyond zero-lag [34]. For our analysis, we focused on 2 specific frequency ranges: (1) all frequencies between 1-20 Hz defined as all frequency bands; and (2) the alpha frequency band (8-12 Hz), given its putative role in attention regulation [44,45] and cognitive control [46,47]. Further mathematical details regarding the TI calculation can be found in Multimedia Appendix 1.

TI calculation involved aligning participants' recorded screens (Figure 4A) and segmenting EEG signals into 1-second epochs (Figure 4B). Epochs exceeding 6 times the golden SDs—nonbiased estimates of variability derived from clean EEG signals (Multimedia Appendix 6)—were rejected (Figure 4B, gray boxes). The remaining commonly clean epochs (Figure 4C, green boxes) were used to compute TIs for the anterior, central, and posterior brain regions (Figure 4D). Because TI depends on duration, normalization was necessary (Multimedia Appendix 1 provides details on normalizing TI). We applied this procedure to calculate the communication TI for all possible pairs of participants. We defined the communication TI of a student as the average of the communication TIs of all pairs including that participant, and the student TI as the average of the communication TI across all valid communications for that student. We then derived a group TI by averaging all student TIs in that group. TIs were computed separately for the anterior, central, and posterior brain regions (Figure 4D). To our knowledge, this is the first study to apply the TI normalization technique.

Figure 4. Steps for calculating total interdependence (TI). (A) Alignment of recorded screens from 6 participants during a simulated scenario. (B) Aligned clean electroencephalogram (EEG) signals epoched into 1-second periods (vertical bars), with rejected epochs (gray boxes) exceeding 6 times the golden SDs. (C) Commonly clean epochs (green) for participant pairs, used to calculate TIs across anterior (blue), central (red), and posterior (yellow) brain areas (D). AN: airway nurse; CN: circulation nurse; D: medical doctor; Iz: Inion; L: left; MT: medical technologist; Nz: Nasion; P: pharmacist; R: right; RT: radiological technologist.



Statistical Analysis

Because the TeamSTEPPS scores for each profession were independently assessed by 2 trained raters, we evaluated interrater reliability using the intraclass correlation coefficient (ICC; [48]). We selected a 2-way random-effects model, which assumes that both raters and participants were randomly chosen, to generalize the reliability estimates to any trained raters. Given the resource-intensive nature of rater training, we were focused on estimating the reliability of scores if only one rater was used. Accordingly, we adopted the "single rater" type and the "absolute agreement" definition, following the reporting guidelines proposed by Koo and Li [49]. The ICC was calculated using the *icc* function from the *irr* package (version 0.84.1 [50]) in R software (version 4.3.2; R Foundation for Statistical Computing), with the selected model, type, and definition options, while other parameters were set to default. To aid interpretation of ICC values, we applied Cicchetti's [51] thresholds, categorizing reliability as excellent (ICC=0.75-1.00), good (0.60-0.74), fair (0.40-0.59), and poor (<0.40). We also note that Koo and Li [49] proposed an alternative system in which ICC values <0.5, 0.5-0.75, 0.75-0.9, and >0.9 correspond to poor, moderate, good, and excellent reliability, respectively.

Since the study involved 2 variables (TeamSTEPPS and TI), analyses were performed with multivariate outliers, which were identified using the Minimum Covariance Determinant [52] using the *MASS* package (version 7.3.60.0.1 [53]) in R software (version 4.3.2) with 75% of the samples regarded as the minimum number of "good" samples and an α level of .001. Correlation between both variables was assessed using the Pearson correlation for 36 cases: a total of 6 for the TeamSTEPPS scores (5 topics and 1 overall), 3 for the brain areas (anterior, central, and posterior), and 2 for the frequency bands (all frequency bands and the alpha frequency band). To control for Type I error, the Benjamini and Hochberg procedure [54,55] was applied for multiple comparison with a combined threshold of $P<.05$.

Sample Size Justification

As part of the larger study [26,27], our sample size of 30 sessions aligned with Reinero's [8] study, which included a

total of 44 groups. A total of 30 sessions in this study yielded 180 observations (90 participants × 2 sessions), which is comparable to the 176 observations (4 participants × 44 groups) reported in Reinero's [8] study. With a desired statistical power of 0.8 and a sample size of 30 sessions, we calculated the minimum detectable effect size for correlation to be 0.49 using the formula provided elsewhere [56].

Ethical Considerations

The study was approved by the Institutional Review Board of the Faculty of Medicine at Chulalongkorn University (COA No 1085/2022). All participants provided written informed consent to participate. Participants who completed the experiment received THB 1000 (US $30.27).

Results

Demographic

We included a total of 85 students, with a mean age of 21.87 (SD 1.17) years, a mean grade point average of 3.21 (SD 0.38), and a mean PHQ-9 score of 3.49 (SD 2.54). Of these, there were 11 radiological technologists, 15 medical technologists, 15 medical doctors, 15 pharmacists, 14 circulation nurses, and 15 airway nurses. A total of 20 sessions had complete EEG activities of all 6 simulated professions. Table 2 presents the unweighted demographic characteristics of the included participants for each simulated profession. Of the 59 women (69% of the total sample), 8 (14%) were radiological technologists, 12 (20%) were medical technologists, 4 (7%) were medical doctors; 7 (12%) were pharmacists; 13 (22%) were circulation nurses; and 15 (25%) were airway nurses. A total of 13 (15%) students were administered a 10 mg dose of oral dimenhydrinate.

Table 2. Demographic characteristics.

Characteristics	Professions (N=85)						
	Total	Radiological technologist (n=11)	Medical technologist (n=15)	Medical doctor (n=15)	Pharmacist (n=15)	Circulation nurse (n=14)	Airway nurse (n=15)
Age (years), mean (SD)	21.87 (1.17)	21.36 (0.50)	21.27 (0.46)	22.47 (0.74)	23.47 (1.19)	21.36 (1.01)	21.13 (0.64)
Sex, n							
Female	59	8	12	4	7	13	15
Male	26	3	3	11	8	1	—[a]
Academic year							
3	13	—	—	—	—	6	7
4	44	11	15	—	2	8	8
5	18	—	—	12	6	—	—
6	10	—	—	3	7	—	—
Grade point average, mean (SD)	3.21 (0.38)	3.23 (0.4)	3.11 (0.38)	3.34 (0.44)	3.20 (0.46)	3.21 (0.27)	3.19 (0.31)
PHQ-9[b] score, mean (SD)	3.49 (2.54)	4.36 (2.16)	4.73 (2.55)	4.20 (2.62)	3.60 (2.59)	2.29 (2.43)	1.93 (1.71)
Dimenhydrinate							
Female	8	1	2	—	—	2	3
Male	5	—	1	4	—	—	—

[a]Not applicable.

[b]PHQ-9: Patient Health Questionnaire-9.

Professions Tend to Perform Best in Situation Monitoring and Worst in Mutual Support

Table 1 presents descriptive statistics of the TeamSTEPPS scores, with each simulated profession assessed independently by 2 raters (Rater 1 and Rater 2). A total of 12 raters were involved in the study. The statistics were calculated from 22 sessions of 11 radiological technologists, 30 sessions of 15 medial technologists, 30 sessions of 15 medical doctors, 30 sessions of 15 pharmacists, 30 sessions of 15 airway nurses, and 28 sessions of 14 circulation nurses. To enable comparisons across professions and to compute the group TeamSTEPPS score, participants' scores were scaled to a range between 0 and 1, since the score ranges varied by topic and profession (Table 1, "Range" rows). This resulted in the Scaled AVG values in Table 1, where higher values indicate better performance.

The average interrater agreement for the TeamSTEPPS scores, as measured by the ICC, was good across the 5 topics and 6 professions (mean 0.73, SD 0.18; range 0.32-0.999), with excellent agreement for the overall score (mean 0.87, SD 0.09; range 0.75-0.99). The lowest ICC was 0.32 for mutual support within the circulation nurse profession, whereas the highest was 0.999 for leadership within the medical technologist profession.

Team structure yielded the highest average agreement across professions (mean 0.86, SD 0.13; range 0.64-0.997), while mutual support had the lowest (mean 0.60, SD 0.25; range 0.32-0.94). Among professions, medical technologists had the highest overall agreement (mean 0.98, SD 0.03; range 0.94-0.999), and pharmacists had the lowest (mean 0.59, SD 0.14; range 0.43-0.78). For radiological technologists, team structure showed the highest ICC and mutual support the lowest. For medical technologists, leadership scored highest and mutual support lowest. For medical doctors, communication had the highest agreement and situation monitoring the lowest. For pharmacists, circulation nurses, and airway nurses, team structure showed the highest ICC, while situation monitoring (pharmacists and airway nurses) and mutual support (circulation nurses) were lowest.

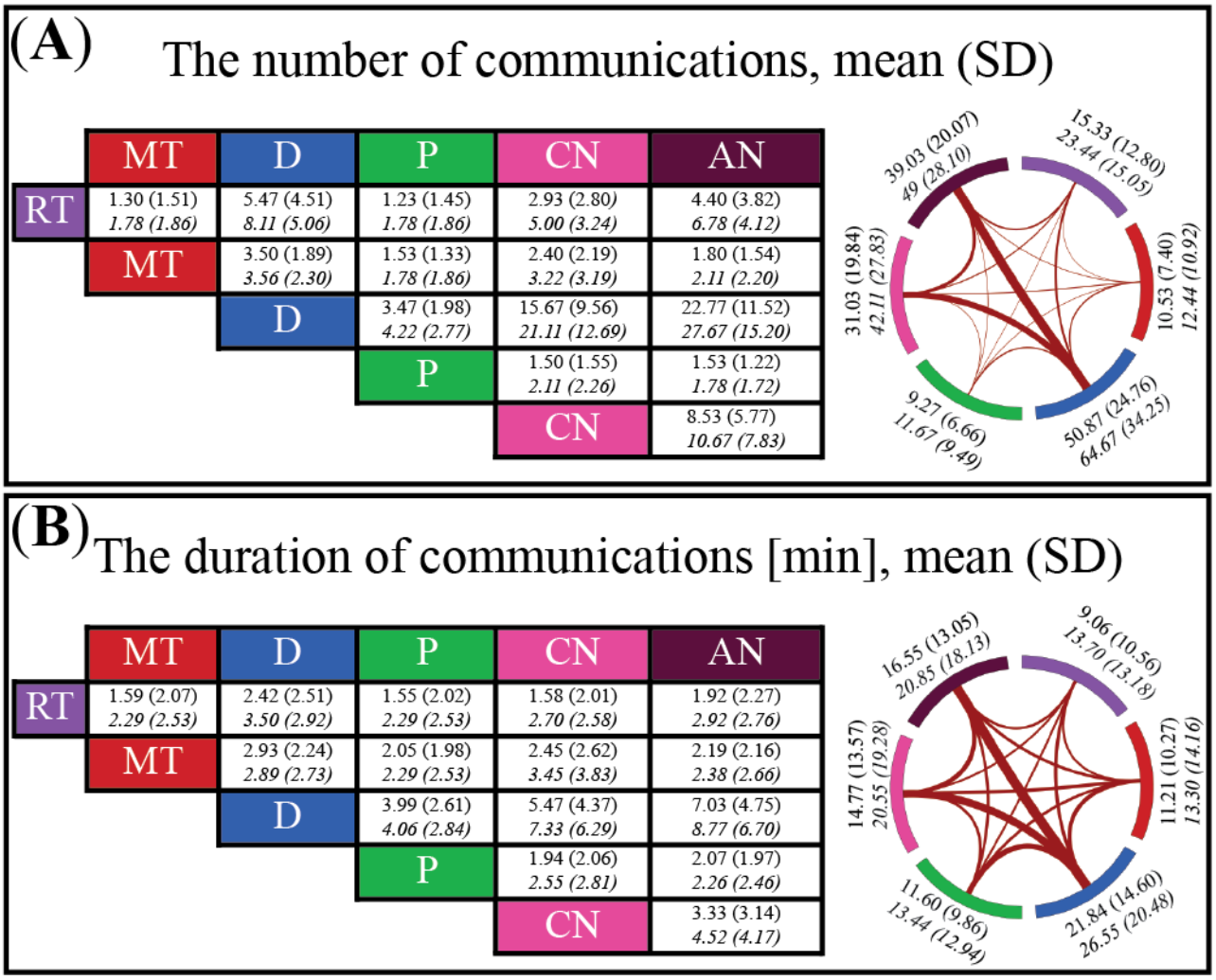
The results of the TeamSTEPPS performance comparison among professions across topics were summarized in the Scaled AVG rows of Table 1. Medical doctors demonstrated the highest performance in both team structure and situation monitoring compared to other professions. Pharmacists excelled in communication and leadership, while airway nurses performed best in mutual support. Notably, all professions exhibited their lowest Scaled AVG scores in mutual support, except for airway nurses, who had the lowest in communication. Each profession showed distinct strengths in different topics: radiological technologists and circulation nurses excelled in team structure; medical technologists, medical doctors, and airway nurses in situation monitoring, and pharmacists in communication.

Table 1 presents the group TeamSTEPPS scores for the 6 simulated professions across all 20 sessions and for the subset of 9 sessions in which all 6 student TIs were present. In both analyses, the lowest group Scaled AVG value was observed in mutual support, while the highest was observed in situation monitoring for the 20 sessions and in team structure for the 9 sessions. This indicated that the group performed worst in mutual support and best in situation monitoring across the 20 sessions.

Scenario Dominated by Communication Between Medical Doctors and Airway Nurse Professions

We examined the characteristics of communications between professions during the virtually simulated scenario, focusing on frequency and duration. As shown in Figure 5, the medical doctor and airway nurse professions communicated most frequently, with an average of 22.77 (SD 11.52) communications across 30 sessions and 27.67 (SD 15.20) across 9 sessions. Their communications also lasted the longest, averaging 7.03 minutes (SD 4.75 minutes) for the 30 sessions and 8.77 minutes (SD 6.70 minutes) for the 9 sessions. In contrast, communications between the radiological technologist and pharmacist professions were the least frequent (mean 1.23, SD 1.45) and shortest (mean 1.55 minutes, SD 2.02 minutes) across 30 sessions. In the 9 sessions, the shortest communication duration occurred between pharmacist and airway nurse professions (mean 2.26 minutes, SD 2.46 minutes), while the least frequent communications were observed between the following pairs: radiological technologist and medical technologist, radiological technologist and pharmacist, medical technologist and pharmacist, and pharmacist and airway nurse (all mean 1.78, SD 1.86).

Figure 5. Descriptive statistics of communications between participants. (A) The number of communications among 6 professions (left) and a connectogram (right) where thicker lines indicate more frequent communications. (B) Communication duration in minutes (left) and a connectogram (right) where thicker lines indicate longer durations. Numbers around connectograms indicate totals per profession. Italicized numbers denote complete groups in which all 6 student total interdependence (TI) were present. AN: airway nurse; CN: circulation nurse; D: medical doctor; MT: medical technologist; P: pharmacist; RT: radiological technologist.



The Alpha-Band Communication TI Deviates More From Baseline Than All Frequency Bands

Table 3 presents descriptive statistics of the communication TI for students across all frequency bands and specifically for the alpha frequency band, in the anterior, central, and posterior brain areas. Data are reported for each profession over the 30 sessions (total sessions) and for the 9 sessions (complete subset), representing the sessions in which the student TI of all 6 simulated professions was fully available.

The absolute values of the alpha-band communication TI of a student were larger than those of all frequency bands in most cases: 78% (14/18) during the 30 sessions and 67% (12/18) during the 9 sessions across the 6 professions and 3 brain areas. In the anterior brain area, 83% (5/6) of the professions showed higher alpha-band values during the 30 sessions, and 67% (4/6) during the 9 sessions. In the central brain area, this trend persisted for 83% of the professions in both sets, whereas in the posterior brain area, it was 67% for the 30 sessions and 50% for the 9 sessions. These results suggest that the alpha-band communication TI of a student deviated more from baseline and conveyed more information about brain-to-brain synchronization than the all frequency-bands communication TI.

We compared the absolute values of the alpha-band communication TI of a student across the 3 brain areas for each profession. In the 30 sessions, the central brain area exhibited larger values than the anterior and posterior areas in 67% (4/6) of professions, while the posterior area showed larger values in 33% (2/6) of professions, and no profession showed larger values in the anterior area. In the 9 sessions, larger values were observed in the anterior, central, and posterior brain areas in 33% (2/6), 17% (1/6), and 50% (3/6) of professions, respectively. The 30-session results suggest that the alpha-band communication TI in the anterior brain area may least represent brain-to-brain synchronization, while the 9-session results suggest a moderate representation in the anterior brain area.

Table 3. Descriptive statistics of the communication total interdependence of a student. Please note that, statistics were calculated after removing outliers.

Brain areas	The communication TI[a] of a student, mean (SD), n					
	RT[b] (N=22)	MT[c] (N=30)	D[d] (N=30)	P[e] (N=30)	CN[f] (N=28)	AN[g] (N=30)
Anterior						
All frequency bands (total sessions)	−0.01 (0.21), 21	0.09 (0.43), 28	0.002 (0.12), 22	−0.002 (0.45), 25	0.01 (0.20), 21	0.003 (0.22), 26
All frequency bands (complete subset)[a]	0.04 (0.23), 9	0.04 (0.42), 9	−0.0001 (0.23), 9	0.03 (0.25), 8	−0.08 (0.17), 9	0.09 (0.24), 9
Alpha band (total sessions)	0.06 (0.32), 21	0.07 (0.48), 27	0.02 (0.18), 24	−0.08 (0.31), 24	−0.02 (0.24), 22	0.03 (0.24), 26
Alpha band (complete subset)[a]	−0.05 (0.11), 8	0.26 (0.74), 9	0.02 (0.17), 9	−0.19 (0.21), 8	0.06 (0.16), 8	0.04 (0.25), 9
Central						
All frequency bands (total sessions)	0.06 (0.30), 21	−0.04 (0.41), 28	−0.02 (0.18), 24	−0.09 (0.41), 24	0.03 (0.30), 22	−0.04 (0.17), 26
All frequency bands (complete subset)[a]	−0.007 (0.21), 9	0.05 (0.44), 9	−0.06 (0.17), 9	−0.13 (0.22), 8	−0.03 (0.27), 9	0.03 (0.15), 9
Alpha band (total sessions)	−0.08 (0.29), 21	0.09 (0.51), 28	0.03 (0.19), 24	−0.20 (0.27), 22	−0.004 (0.31), 21	0.10 (0.20), 24
Alpha band (complete subset)[a]	−0.15 (0.27), 9	0.13 (0.52), 9	0.03 (0.15), 9	−0.19 (0.25), 8	−0.05 (0.20), 9	0.12 (0.21), 8
Posterior						
All frequency bands (total sessions)	0.07 (0.26), 20	0.03 (0.41), 28	−0.01 (0.16), 24	−0.01 (0.29), 25	0.0005 (0.24), 20	−0.07 (0.12), 24
All frequency bands (complete subset)[a]	0.11 (0.31), 8	−0.22 (0.21), 7	−0.04 (0.18), 9	−0.01 (0.25), 9	−0.11 (0.19), 9	0.01 (0.07), 8
Alpha band (total sessions)	−0.002 (0.27), 21	0.07 (0.52), 27	0.05 (0.22), 24	−0.07 (0.23), 22	−0.03 (0.32), 22	0.07 (0.18), 23
Alpha band (complete subset)[a]	−0.006 (0.15), 9	−0.06 (0.39), 8	0.07 (0.05), 7	−0.11 (0.19), 8	−0.07 (0.32), 9	0.14 (0.12), 9

[a]This corresponds to the complete subsets, in which all 6 student TIs were present.

[b]RT: radiological technologist.

[c]MT: medical technologist.

[d]D: medical doctor.

[e]P: pharmacist.

[f]CN: circulation nurse.

[g]AN: airway nurse.

Group TI Correlates With Group TeamSTEPPS Score

In this section, we present the correlations between TeamSTEPPS and TI after removing outliers using the Minimum Covariance Determinant method and correcting for multiple comparisons with the Benjamini and Hochberg procedure. The TeamSTEPPS scores included 5 topics—team structure, communication, leadership, situation monitoring, and mutual support—along with an overall score. The TI considered here comprised both the student TI and the group TI.

Initially, we examined the correlations between the TeamSTEPPS scores and the student TI. For all 6 simulated professions across both the all frequency and alpha bands, the correlations were weak and not statistically significant (all adjusted $P \geq .05$; Figures S2-S7 in Multimedia Appendix 7), except for the medical technologist simulated profession, which showed a significant negative correlation with the team structure topic in the alpha band at the anterior brain area (r=–0.76 adjusted P<.001). Combining data from all 6 simulated professions also resulted in weak and not statistically significant correlations for both frequency bands (adjusted P>.99; Figure S1 in Multimedia Appendix 7).

Second, we examined the correlations between the group TeamSTEPPS scores and the group TI, with results presented in Figure 6. This figure shows scatter plots illustrating the associations between the group TeamSTEPPS scores and the group TI for all frequency bands (blue) and the alpha frequency band (red) in the anterior (top row), central (middle), and posterior (bottom row) brain areas. The scatter plots are overlaid with the best-fit lines calculated using the least-squares method (*polyfit* function of MATLAB, The MathWorks, Inc). Out of the 30 sessions, a total of 9 sessions had complete student TI

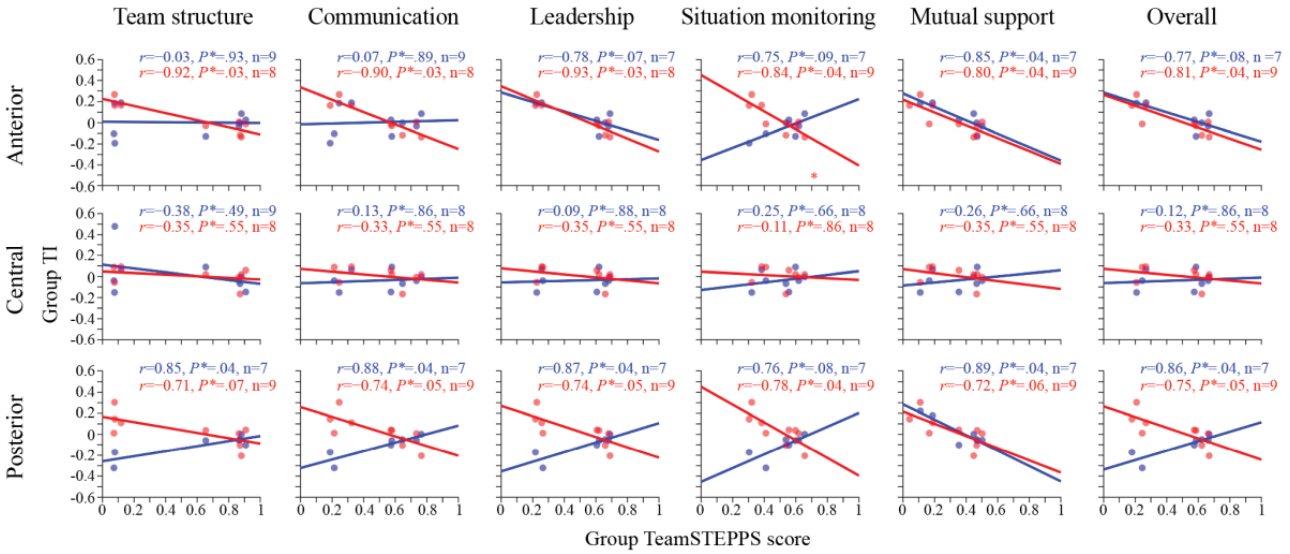
data for all 6 simulated professions, resulting in well-defined group TIs. Note that the number of sessions shown in the scatter plots may be lower than 9 due to the outlier screening procedure.

In the anterior brain area (Figure 6, top row), the alpha frequency band demonstrated strongly negative and statistically significant correlations with all 5 TeamSTEPPS topics, including the overall score: r=–0.92, adjusted P=.03 for team structure; r=–0.90, adjusted P=.03 for communication; r=–0.93, adjusted P=.03 for leadership; r=–0.84, adjusted P=.04 for situation monitoring; r=–0.80, adjusted P=.04 for mutual support; and r=–0.81, adjusted P=.04 for the overall score. In contrast, all frequency bands yielded strongly negative but nonstatistically significant correlations for the other topics, except for mutual support: r=–0.03, adjusted P=.93 for team structure; r=0.07, adjusted P=.89 for communication; r=–0.78, adjusted P=.07 for leadership; r=0.75, adjusted P=.09 for situation monitoring; r=–0.85, adjusted P=.04 for mutual support; and r=–0.77, adjusted P=.08 for the overall score.

In the central brain area (Figure 6, middle row), both all frequency bands and the alpha frequency band exhibited weak, nonstatistically significant correlations for all 5 topics including the overall score, with minimum adjusted P=.49 and a maximum absolute r=0.38.

In the posterior brain area (Figure 6, bottom row), all frequency bands displayed strongly positive and statistically significant correlations for team structure (r=0.85, adjusted P=.04), communication (r=0.88, adjusted P=.04), leadership (r=0.87, adjusted P=.04), and the overall score (r=0.86, adjusted P=.04). There were also strongly negative and statistically significant correlations for mutual support (r=–0.89, adjusted P=.04) and a strongly positive but nonstatistically significant correlation for situation monitoring (r=0.76, adjusted P=.08). The alpha frequency band in the posterior brain area showed strongly negative and statistically significant correlations for situation monitoring (r=–0.78, adjusted P=.04), but strongly negative and nonstatistically significant correlations for team structure (r=–0.71, adjusted P=.07), communication (r=–0.74, adjusted P=.05), leadership (r=–0.74, adjusted P=.05), mutual support (r=–0.72, adjusted P=.06), and the overall score (r=–0.75, adjusted P=.05).

Figure 6. Anterior group total interdependence (TI) in the alpha band negatively correlated with group Team Strategies and Tools to Enhance Performance and Patient Safety (TeamSTEPPS) scores. Scatter plots show relationships between group TI and TeamSTEPPS scores for anterior (top), central (middle), and posterior (bottom) brain areas, across all frequency bands (blue) and only the alpha band (red). Columns represent TeamSTEPPS domains, with the last column showing their sum. Legends display Pearson correlation (r), adjusted P value (P*), and sample size after excluding outliers (n). Best-fit least-squares lines are shown.



Discussion

Principal Findings

This study explored the correlation between brain-to-brain synchronization, measured by the TI, and team performance, evaluated using the TeamSTEPPS scores, in an online virtual SIMBIE setting where face-to-face communication was absent. At the individual level, no significant correlations were found between participants' synchronization with their team members (measured by the student TI) and their TeamSTEPPS scores (all adjusted P≥.05). These findings are consistent with previous studies [57,58], which similarly reported an absence of brain-to-brain synchronization when face-to-face interactions—crucial mechanisms such as eye contact, important for establishing trust and enhancing synchrony [59]—were missing.

However, at the team level, we identified strongly negative, statistically significant correlations between the group TIs and the group TeamSTEPPS scores, particularly in the anterior brain region (mean −0.87, SD 0.06; range −0.93 to −0.8). These findings are counterintuitive and contrast with our initial hypothesis, as higher team performance is generally associated with greater brain-to-brain synchronization [8]. To better understand this result, we examined the assessment method: team performance was evaluated using the TeamSTEPPS scores, which heavily rely on effective verbal communication [28]. Our findings tentatively suggest that a high degree of brain-to-brain synchronization does not necessarily correspond to improved

team performance. Conversely, at the team level, strongly positive, statistically significant correlations were observed between the group TIs based on all frequency bands and the group TeamSTEPPS scores—including the overall score and most topics—particularly in the posterior brain region.

Comparison With Previous Work

To our knowledge, no previous studies have examined the correlation between EEG-based brain-to-brain synchronization and TeamSTEPPS scores in an online hexad virtual SIMBIE setting without face-to-face communication. This study provides a novel contribution by incorporating a broader set of 6 participants and addressing an important gap in the literature on team-based interactions, using a cost-effective EEG device for analysis.

Exploring the Discrepancy: Negative Correlation of Anterior Alpha Band With TeamSTEPPS Performance and the Role of Intermittent Synchronization in Hexad Multi-Person Teams

Previous studies generally reported a positive correlation between brain-to-brain synchronization and team performance using basic social interaction tasks, such as Lego assembly, typically conducted in face-to-face dyadic groups and using high-cost methods, including functional near-infrared spectroscopy, near-infrared spectroscopy–based hyperscanning, and resource-intensive EEG studies [8,18,60-63]. In contrast, this study observed strongly negative, statistically significant correlations between team brain-to-brain synchronization—measured by the group TIs based on the anterior alpha band—and team performance assessed using the group TeamSTEPPS scores, including all 5 individual domains and their overall sum. By comparison, group TIs in the posterior frequency bands showed strongly positive, statistically significant correlations with team performance in nearly all individual domains and the overall sum.

We speculate that the discrepancy between the negative correlations observed in this study based on the anterior alpha band and the positive correlations reported in previous studies may result from the complexity of more-than-two-person communication in this study and the nature of the TeamSTEPPS scores, in contrast to previous studies focused on dyadic interactions [61-63]. For instance, compared to face-to-face dyads, groups with more than 2 individuals exhibit distinct patterns in verbal communication and nonverbal behaviors (eg, eye gaze) [64,65], which require intermittent synchronization—the brain's ability to shift focus among individuals—to maintain effective communication and coordination [66,67]. Intermittent synchronization in groups of more than 2 members was also observed in Stevens et al's [17,21] studies, which used mannequin-based group simulations with face-to-face interactions. These studies reported that triad groups exhibiting high TeamSTEPPS scores demonstrated elevated Shannon entropy of neurodynamic symbols derived from neural activity at the 10-Hz brain rhythm. Because the 10-Hz rhythm falls within the alpha frequency band and is associated with attention regulation, high Shannon entropy of the corresponding neurodynamic symbols likely reflects the waxing and waning of group members' attention during the simulation.

In our experimental setting, where 6 participants worked as a team, the brain's ability to engage in intermittent synchronization was crucial for effective communication and coordination [66,67]. Over synchronization within a team of more than 2 members may be detrimental to team performance, potentially leading to groupthink, a phenomenon in which members prioritize harmony and consensus over critical thinking and individual opinions [67,68].

The TeamSTEPPS scores rely heavily on effective verbal communication to establish a shared mental model of team tasks, strategies, and goals, which is critical for enhancing situational awareness and overall team performance [28,69-71]. TeamSTEPPS promotes an ideal team typology that requires intermittent synchronization—the ability to shift focus and support shared decision-making [72]. However, over synchronization, or an inability to shift focus among more-than-two team members, may disrupt communication and lead to groupthink [67,68]. In such cases, directive leadership and hierarchical structures within health care teams can prioritize consensus at the expense of independent critical thinking, potentially resulting in negative outcomes and impeding shared decision-making [68]. Groupthink, or cohesive typology, relies on a lower cognitive level than a shared mental model or facilitated team structure [72]. This may explain why higher brain-to-brain synchronization, or over-synchronization, as measured by group TIs, does not necessarily translate into optimal team performance as assessed by the TeamSTEPPS scores.

Anterior Alpha Band Synchronization: A Key Indicator for TeamSTEPPS Performance and Shared Attention in Hexad Teams Without Face-to-Face Interaction

The statistically significant correlations between brain-to-brain synchronization at the team level and TeamSTEPPS performance were primarily observed in the anterior brain areas, particularly in the alpha band. In other words, the anterior electrodes provided the most reliable measurements of TeamSTEPPS performance both overall and across each of the 5 domains, compared to the central and posterior electrodes. These findings align with previous studies that emphasize the critical role of the dorsomedial prefrontal cortex, which is associated with the perception of intention [73,74] and theory of mind, the ability to interpret others' mental states [75,76], as well as the frontopolar region, a part of the prefrontal cortex, during social interactions [77,78]. However, our results differ from other studies that identified different brain regions, such as the right temporo-parietal junction, as being highly activated during social interaction [57,79]. This discrepancy may arise from variations in the experimental conditions and the nature of tasks used [10].

Unlike the alpha band (8-12 Hz), the delta (1.5-4 Hz [80]), theta (4-8 Hz [81]), and beta (14-30 Hz [82]) bands were not analyzed individually due to limited theoretical support, as their primary functions are less relevant in studies focused on teamwork and

attention. Supporting our findings, previous research has identified a relationship between brain-to-brain synchronization in the alpha band and interactional synchrony during social interactions [61,63,83-87]. While some studies suggested that the gamma band (30-80 Hz) may also reflect brain-to-brain synchronization and shared intentions [88], the sampling rate of our EEG device (125 Hz), with a Nyquist frequency of 62.5 Hz limited our analysis to frequencies up to the beta band (14-30 Hz), thus preventing examination of the gamma band. Our emphasis on the alpha band is substantiated by its established roles in attention regulation [44,45,89,90] and cognitive control [46,47]. The alpha band has been extensively studied in contexts such as relaxation [91], inhibitory control [92], emotional processing [93], mental health [94], and sleep [95], making it a versatile marker for various cognitive processes. Our findings further highlight the value of the alpha band as an indicator of fluctuations in attention and focus during team-based activities as measured by TeamSTEPPS.

Overcoming Subjectivity in TeamSTEPPS Assessment: The Potential of Anterior Alpha Band EEG Synchronization in SIMBIE Scenarios

Figure 5 highlights differences in communication duration among participants, with medical doctors having the longest duration, which may explain their higher TeamSTEPPS scores. In contrast, radiological technologists, with the shortest communication duration, may have lacked closed-loop communication and sufficient identification, impacting critical information sharing as emphasized by the "Introduction, Situation, Background, Assessment, Recommendation" guidelines [96], and resulting in lower TeamSTEPPS scores. Interestingly, despite limited communication duration, the pharmacists achieved a relatively high TeamSTEPPS score. This may be due to the scenario design, where the pharmacist's role involved gathering medication history from a nonplayer character acting as a family member, thereby boosting TeamSTEPPS scores despite minimal communication with the team. However, no significant correlation was observed between the communication TI of a pair and communication duration, except for a low-value correlation in the central brain area within the alpha frequency band (Figure S1 in Multimedia Appendix 8).

The TeamSTEPPS scores in this study were derived from both verbal communications and nonverbal behaviors. The ICC values for the situation monitoring and mutual support domains were the lowest among the 5 domains (Table 1), likely due to their reliance on nonverbal behaviors and the need for experienced raters to assess them accurately. Examples include monitoring and checking patient's vital signs before, during, and after intubation, during X-ray procedures, and before intensive care unit transfer, as well as mutual assistance tasks such as raising bed rails and assisting with donning personal protective equipment. These findings indicate that assessing situation monitoring and mutual support with human raters is inherently limited by the subjectivity required to evaluate nonverbal behaviors [97]. However, EEG brain-to-brain synchronization, particularly in the anterior alpha band, offers a promising alternative method to support and enhance the prediction of TeamSTEPPS performance across all domains (Figure 6). The inherent subjectivity of human raters in evaluating situation monitoring and mutual support domains presents a challenge that could be intriguingly addressed by leveraging measures from the anterior alpha band. Previous studies [98-101] have used alpha-band EEG to measure individual situation awareness in real time. In contrast, our findings emphasize the potential of anterior alpha-band synchronization as a group-level metric, offering a fresh perspective on assessing team dynamics and collaborative performance.

Posterior All-Band Synchronization as a Potential Marker for Certain TeamSTEPPS Domains and Task-Dependent Roles in Hexad Teams Without Face-to-Face Interaction

Our results revealed strong positive and statistically significant correlations between the posterior group TIs based on all frequency bands (1-20 Hz, encompassing partial delta, theta, alpha, and partial beta) and the group TeamSTEPPS scores across most domains (r=0.76-0.88). Notably, the mutual support domain deviated from this pattern, showing a strong negative correlation (r=−0.89). The positive correlations between team performance and brain-to-brain synchronization across all frequency bands align with previous studies [8,14], which hypothesized that brain-to-brain synchronization is modulated by shared attention among team members. However, our results based on the posterior alpha band, which has a putative role in attention [44,45,89,90], showed the negative correlations that contrasted both with the posterior all frequency band results in this study and with previous studies based on all frequency bands.

Grounded in previous findings, the delta, theta, and beta frequency bands have well-known functional roles beyond attention—for example, levels of consciousness for the delta band [102], cognitive performance for the theta band [103], and motor control and execution for the beta band [104]. Our results tentatively suggest that in the complex interactions among the 6 team members in this study, brain-to-brain synchronization across all frequency bands cannot be explained solely by the shared attention mechanism. Therefore, we recommended that future research investigate brain-to-brain synchronization for each frequency band separately to understand their functional roles during cooperation and social interaction in teams with a larger number of participants, as in this study.

A recent study by Reinero et al [8] explored the relationship between team performance on problem-solving tasks and interbrain synchrony within a group of 4 participants. They found that higher interbrain synchrony, reflecting stronger connections among teammates, led to better performance on economic games and most problem-solving tasks. In contrast, this study found that higher interbrain synchrony correlated with lower TeamSTEPPS performance. Several factors may explain this difference. First, this study was conducted in a virtual environment without face-to-face interaction, which is known to reduce interbrain synchrony [8,57,58]. Second, our tasks required critical thinking and rapid, shared decision-making under time constraints, making effective verbal

communication essential. With a larger team of 6—exceeding the arguably optimal team size of 4 [105]—intermittent synchronization, or the brain's ability to shift focus among teammates, was crucial. Over interbrain synchrony may have disrupted this focus shifting, impeding team performance. In the study by Reinero et al [8], however, the team had fewer time pressures and communicated through private online chat, making intermittent synchronization less critical and allowing high interbrain synchrony to enhance performance. Our findings, together with those of Kikuchi et al [58] and Czeszumski et al [10], suggest that both the context of brain-to-brain synchronization and task characteristics are important factors in the relationship between brain-to-brain synchronization and team performance.

Strengths and Limitations

Excessive EEG noise during gameplay resulted in group TIs being available for only 45% (9/20) of sessions. Although correlations between group TIs and the TeamSTEPPS scores were statistically significant after correction for multiple comparisons, a minimum of 30 sessions is typically recommended to achieve meaningful correlations [106]. To account for potential data loss of up to 55% under similar conditions, we recommend that researchers plan for at least 67 sessions (100×30/45).

This study focused on healthy participants by excluding individuals with neurological, psychiatric, or substance use disorders, which are known to alter social cognition [107], emotional processing [108], or attentional processing [109]—mechanisms fundamental to brain-to-brain synchronization [110]. Several studies have reported that such disorders can influence brain-to-brain synchronization between individuals during various tasks. For instance, Deng et al [111] conducted an EEG hyperscanning study examining brain-to-brain synchronization during an emotional processing task in 25 parent-adolescent pairs with social anxiety, and found that the adolescents' level of social anxiety modulated the synchronization. Similarly, Wang et al [112] reported that the severity of autism spectrum disorder in children affected brain-to-brain synchronization during a cooperative task with their parent, although Kruppa et al [113] found no such modulation. Therefore, the applicability of our findings should be interpreted with caution in practical settings where screening for such disorders may not be feasible.

Time accuracy and precision are critical in brain-to-brain synchronization research. While our methodology produced statistically significant correlations and passed sanity checks, there is potential for further refinement. The first improvement is the Bluetooth connection between the EEG data-acquisition device and the PC. A higher baud rate can reduce transmission time but it also increases vulnerability to errors from electromagnetic interference. In this study, we set the baud rate to 115,200 bits per second, which theoretically allows a 24-bit data point to be transferred in 0.2083 milliseconds. However, factors such as distance, interference, Bluetooth overhead, and system delays in Windows 10 can extend transmission times to 0.6-10 milliseconds [114]. Despite these latency variations, BrainFlow software (version 4.9.0 [41]) maintained the correct order of data points. The second improvement involves better alignment of the 6 recorded screen footages. In this study, we manually aligned the footages using commercial software, which introduced the risk of human error, potentially causing partial loss of verbal communication data and related EEG information. We recommend real-time audio logging for both senders and receivers during virtual simulations to accurately capture team audio profiles. The third improvement focuses on enhancing the accuracy of the local Unix time. We used the C# DateTime function, which can differ from real time by approximately 10 milliseconds [115]. Implementing dedicated external hardware to synchronize time across EEG devices and team PCs would enhance both accuracy and precision.

In this study, we used a fully manual method to reject artifact portions of EEG signals. This approach poses several challenges: (1) potential data bias, (2) time consumption for large-scale studies, (3) reliance on EEG specialists, and (4) limited practicality for real-world applications. Artifact rejection remains a significant bottleneck in EEG data preprocessing, as unaddressed noise and artifacts can impair subsequent analysis. Despite its importance, there has been no consensus within the EEG community on effective artifact management. Recently, various tools and techniques for semiautomatic artifact rejection during offline analysis have emerged, such as the Clean Rawdata EEGLAB plug-in [116,117], Autoreject [118], FASTER [119], the Riemannian Potato technique [120], and the Robust Regression technique [121]. However, these methods often require manual parameter tuning, which can vary depending on the dataset.

Simulator sickness is a common issue in virtual simulations [122]. In this study, of the 90 participants, 13 (14%) took dimenhydrinate before the experiments to prevent nausea, vomiting, and dizziness associated with simulator sickness [123]. While dimenhydrinate can cause side effects such as drowsiness and hyperactivity, which may potentially impact EEG signals [29], Hu [124] reported no changes in peak frequency or the percentage distribution of the alpha band. Therefore, we speculate that our primary finding—the correlation between the alpha group TIs and the group TeamSTEPPS scores—was likely unaffected by dimenhydrinate, though it may have influenced results based on all frequency bands.

Our analysis approach evaluates the normalized TI immediately after each communication, rather than waiting until the end of the entire session. This allows for near real-time computation, making the method practical for real-world applications. This approach requires extensive computation, involving 1000 permutations to construct the TI empirical distribution for normalization. A small percentage of these distributions (175/4224, 4.14%) did not pass the Kolmogorov-Smirnov normality test, suggesting that a higher number of permutations may be needed, which would further increase computational demands.

This study was conducted in a single-room lab without electromagnetic shielding, where all 6 participants shared the same space due to facility limitations. This setup introduced unavoidable environmental factors, such as shared sounds,

visuals, and potential interference from external electromagnetic sources, which may have affected EEG signals [125-127]. To reduce spurious brain-to-brain synchronization from these external influences, we recommend that future studies place each participant in a separate, electromagnetically shielded room.

Future Research

Building on this study and considering the non-face-to-face context of the online hexadic virtual SIMBIE, we recommended using alpha-band activity in the anterior brain regions to evaluate TeamSTEPPS scores. This approach showed significant associations with all 5 TeamSTEPPS domains and provides a valuable framework for understanding shared attention and the development of shared mental models. Additionally, focusing on the anterior brain region is practical because affordable EEG devices with anterior electrodes are readily available, making this method cost-effective and widely accessible for broader applications.

Can we enhance the correlation between EEG brain-to-brain synchronization and TeamSTEPPS scores in virtual simulations or real-life practice to address the challenges of subjective TeamSTEPPS evaluations by raters? Our current methodology relies on verbal communication within a team of at least 2 collaborators to compute TIs. However, the TeamSTEPPS scores also encompass nonverbal components, such as early stages of situation monitoring (eg, visual attention) and mutual support in actions that are not captured by our TI calculations. Future studies could improve TIs by incorporating screen footage segments representing situation monitoring and mutual support, providing a more comprehensive analysis of both verbal and nonverbal components of TeamSTEPPS scores. In real-life clinical practice, it may be feasible to estimate the TeamSTEPPS scores in real-time using affordable, compact EEG devices positioned near the anterior brain region, enabling broader generalizability and practical application across various settings.

Conclusions

This study found no correlations between individual brain-to-brain synchronization, as measured by the student TI, and individual TeamSTEPPS performance during online virtual SIMBIE sessions without face-to-face communication. However, at the group level, strongly negative, statistically significant correlations were primarily observed in the alpha band in the anterior brain region between the group TI and the group TeamSTEPPS scores among teams of 6 multiprofessional students. These findings highlight the potential of EEG-based brain-to-brain synchronization analysis as an emerging tool for a more objective measure of TeamSTEPPS dynamics than subjective human evaluations, offering a novel approach to support TeamSTEPPS assessments with trend-based, quantitative, real-time feedback—ultimately enhancing team performance and patient safety.

Acknowledgments

This research was funded by the Second Century Fund, Chulalongkorn University. The authors thank the TeamSTEPPS raters for their invaluable contributions: Chulaluk Jaipang, Jiraphan Ritsamdang, Dr Khrongwong Musikatavorn, Dr Kanyarat Susantitaphong, Dr Navaporn Worasilchai, Noppawan Boonbumrong, Dr Nuntaree Chaichanawongsaroj, Dr Sararas Khongwirotphan, Sawitree Suayod, Dr Sunisa Seephom, and Sirikanyawan Srikasem. Their dedication and expertise were instrumental to the success of this study. The authors also thank the Chulalongkorn Healthcare Advanced Multi-Profession Simulation Center, Surachai Pianpetchlert, and research assistants—Chaiwat Takkanat, Chayanit Trakulpipat, Kitnipat Boonydhammakul, Nuttarin Panswad, Sirisopha Suwanchinda, Sutasinee Chaidej, Thanes Yunirundorn, Thanyared Sangsawad, Wallop Boonkua, and Werasak Boonwong—for their support in data collection and IT assistance. Special thanks to Dr Zarina Sadad for help with rater training, Dr Chaipat Chunharas and Dr Solaphat Hemrungrojn for advice on EEG devices, and Dr Mary Kay Smith for TeamSTEPPS training. Finally, the authors thank Dr Diego Reinero and Dr Jay Van Bavel for their valuable discussions.

During the preparation of this study, the authors used ChatGPT-4o and ChatGPT-5 to check and correct grammatical errors, refine language, and suggest alternative reader-friendly synonyms during the manuscript writing process. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

A preliminary version of this study was presented at the Society for Simulation in Europe conference on June 19, 2024 in Prague, Czech Republic. This study provides a comprehensive reanalysis with significant updates.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Total Interdependence.
[DOCX File , 244 KB - mededu_v11i1e69725_app1.docx]

Multimedia Appendix 2
Virtual Simulation Content Overview.

[DOCX File , 3973 KB - mededu_v11i1e69725_app2.docx]

Multimedia Appendix 3
Device specifications.
[DOCX File , 540 KB - mededu_v11i1e69725_app3.docx]

Multimedia Appendix 4
Experimental settings.
[DOCX File , 682 KB - mededu_v11i1e69725_app4.docx]

Multimedia Appendix 5
Sanity Checks of the EEG Acquisition System.
[DOCX File , 783 KB - mededu_v11i1e69725_app5.docx]

Multimedia Appendix 6
EEG Preprocessing.
[DOCX File , 759 KB - mededu_v11i1e69725_app6.docx]

Multimedia Appendix 7
Correlation between Student TI and the TeamSTEPPS Scores.
[DOCX File , 3053 KB - mededu_v11i1e69725_app7.docx]

Multimedia Appendix 8
Correlation between the Communication TI of a Pair and Characteristics of Communications.
[DOCX File , 352 KB - mededu_v11i1e69725_app8.docx]

References

1. Morey JC, Simon R, Jay GD, Wears RL, Salisbury M, Dukes KA, et al. Error reduction and performance improvement in the emergency department through formal teamwork training: evaluation results of the MedTeams project. Health Serv Res 2002;37(6):1553-1581. [doi: 10.1111/1475-6773.01104] [Medline: 12546286]
2. Liaw SY, Ooi SW, Rusli KDB, Lau TC, Tam WWS, Chua WL. Nurse-physician communication team training in virtual reality versus live simulations: randomized controlled trial on team communication and teamwork attitudes. J Med Internet Res 2020;22(4):e17279 [FREE Full text] [doi: 10.2196/17279] [Medline: 32267235]
3. Dietz AS, Pronovost PJ, Benson KN, Mendez-Tellez PA, Dwyer C, Wyskiel R, et al. A systematic review of behavioural marker systems in healthcare: what do we know about their attributes, validity and application? BMJ Qual Saf 2014;23(12):1031-1039. [doi: 10.1136/bmjqs-2013-002457] [Medline: 25157188]
4. Wills J, FeldmanHall O, PROSPEC Collaboration NYU, Meager MR, Van Bavel JJ, PROSPEC Collaboration NYU. Dissociable contributions of the prefrontal cortex in group-based cooperation. Soc Cogn Affect Neurosci 2018;13(4):349-356 [FREE Full text] [doi: 10.1093/scan/nsy023] [Medline: 29618117]
5. Geweke J. Measurement of linear dependence and feedback between multiple time series. J Am Stat Assoc 1982;77(378):304. [doi: 10.2307/2287238]
6. Vinck M, Oostenveld R, van Wingerden M, Battaglia F, Pennartz CMA. An improved index of phase-synchronization for electrophysiological data in the presence of volume-conduction, noise and sample-size bias. Neuroimage 2011;55(4):1548-1565. [doi: 10.1016/j.neuroimage.2011.01.055] [Medline: 21276857]
7. Cohen MX. Analyzing Neural Time Series Data: Theory and Practice. Cambridge: The MIT press; 2014.
8. Reinero DA, Dikker S, Van BJJ. Inter-brain synchrony in teams predicts collective performance. Soc Cogn Affect Neurosci 2021;16(1-2):43-57. [doi: 10.31234/osf.io/k2ft6]
9. Dumas G. Towards a two-body neuroscience. Commun Integr Biol 2011;4(3):349-352 [FREE Full text] [doi: 10.4161/cib.4.3.15110] [Medline: 21980578]
10. Czeszumski A, Eustergerling S, Lang A, Menrath D, Gerstenberger M, Schuberth S, et al. Hyperscanning: a valid method to study neural inter-brain underpinnings of social interaction. Front Hum Neurosci 2020;14:39 [FREE Full text] [doi: 10.3389/fnhum.2020.00039] [Medline: 32180710]
11. Koike T, Sumiya M, Nakagawa E, Okazaki S, Sadato N. What makes eye contact special? Neural substrates of on-line mutual eye-gaze: a hyperscanning fMRI study. eNeuro 2019;6(1) [FREE Full text] [doi: 10.1523/ENEURO.0284-18.2019] [Medline: 30834300]

12. Salazar M, Shaw DJ, Gajdoš M, Mareček R, Czekóová K, Mikl M, et al. You took the words right out of my mouth: dual-fMRI reveals intra- and inter-personal neural processes supporting verbal interaction. Neuroimage 2021;228:117697 [FREE Full text] [doi: 10.1016/j.neuroimage.2020.117697] [Medline: 33385556]
13. Osaka N, Minamoto T, Yaoi K, Azuma M, Shimada YM, Osaka M. How two brains make one synchronized mind in the inferior frontal cortex: fNIRS-based hyperscanning during cooperative singing. Front Psychol 2015;6:1811. [doi: 10.3389/fpsyg.2015.01811] [Medline: 26635703]
14. Dikker S, Wan L, Davidesco I, Kaggen L, Oostrik M, McClintock J, et al. Brain-to-brain synchrony tracks real-world dynamic group interactions in the classroom. Curr Biol 2017;27(9):1375-1380. [doi: 10.1016/j.cub.2017.04.002] [Medline: 28457867]
15. Chabin T, Tio G, Comte A, Joucla C, Gabriel D, Pazart L. The relevance of a conductor competition for the study of emotional synchronization within and between groups in a natural musical setting. Front Psychol 2019;10:2954 [FREE Full text] [doi: 10.3389/fpsyg.2019.02954] [Medline: 32010021]
16. Ciaramidaro A, Toppi J, Casper C, Freitag CM, Siniatchkin M, Astolfi L. Multiple-brain connectivity during third party punishment: an EEG hyperscanning study. Sci Rep 2018;8(1):6822 [FREE Full text] [doi: 10.1038/s41598-018-24416-w] [Medline: 29717203]
17. Stevens R, Galloway T, Gorman J, Willemsen-Dunlap A, Halpin D. Toward objective measures of team dynamics during healthcare simulation training. Proc Int Symp Hum Factors Ergon Healthc 2016;5(1):50-54. [doi: 10.1177/2327857916051010]
18. Pan Y, Cheng X, Zhang Z, Li X, Hu Y. Cooperation in lovers: an fNIRS-based hyperscanning study. Hum Brain Mapp 2017;38(2):831-841. [doi: 10.1002/hbm.23421] [Medline: 27699945]
19. Tang H, Zhang S, Jin T, Wu H, Su S, Liu C. Brain activation and adaptation of deception processing during dyadic face-to-face interaction. Cortex 2019;120:326-339. [doi: 10.1016/j.cortex.2019.07.004] [Medline: 31401400]
20. Saito DN, Tanabe HC, Izuma K, Hayashi MJ, Morito Y, Komeda H, et al. "Stay tuned": inter-individual neural synchronization during mutual gaze and joint attention. Front Integr Neurosci 2010;4:127 [FREE Full text] [doi: 10.3389/fnint.2010.00127] [Medline: 21119770]
21. Stevens R, Galloway T, Willemson-Dunlap A. Intermediate neurodynamic representations: a pathway towards quantitative measurements of teamwork. Proc Hum Factors Ergon Soc Annu Meet 2016;60(1):1996-2000. [doi: 10.1177/1541931213601454]
22. Montague PR, Berns GS, Cohen JD, McClure SM, Pagnoni G, Dhamala M, et al. Hyperscanning: simultaneous fMRI during linked social interactions. Neuroimage 2002;16(4):1159-1164. [doi: 10.1006/nimg.2002.1150] [Medline: 12202103]
23. Bevilacqua D, Davidesco I, Wan L, Chaloner K, Rowland J, Ding M, et al. Brain-to-brain synchrony and learning outcomes vary by student-teacher dynamics: evidence from a real-world classroom electroencephalography study. J Cogn Neurosci 2019;31(3):401-411. [doi: 10.1162/jocn_a_01274] [Medline: 29708820]
24. Guttmann-Flury E, Sheng X, Zhang D, Zhu X. A priori sample size determination for the number of subjects in an EEG experiment. Annu Int Conf IEEE Eng Med Biol Soc 2019;2019:5180-5183. [doi: 10.1109/EMBC.2019.8857482] [Medline: 31947025]
25. Asaad WF, Sheth SA. What's the n? On sample size vs subject number for brain-behavior neurophysiology and neuromodulation. Neuron 2024;112(13):2086-2090 [FREE Full text] [doi: 10.1016/j.neuron.2024.04.033] [Medline: 38781973]
26. Narajeenron K, Chintakovid T, Phutrakool P, Ritsamdang J, Viriyopase A, Musikatavorn K, ER-VIPE study group (Emergency Room – Virtual Interprofessional Education). Enhancing team strategies and tools to enhance performance and patient safety performance through medical movies, massive open online courses, and 3D virtual simulation-based interprofessional education: mixed methods double-blind quasi-experimental study. J Med Internet Res 2025;27:e67001 [FREE Full text] [doi: 10.2196/67001] [Medline: 40921059]
27. Srikasem S, Seephom S, Viriyopase A, Phutrakool P, Khowinthaseth S, Narajeenron K. Comparing the effectiveness of multimodal learning using computer-based and immersive virtual reality simulation-based interprofessional education with co-debriefing, medical movies, and massive online open courses for mitigating stress and long-term burnout in medical training: quasi-experimental study. JMIR Med Educ 2025;11:e70726. [doi: 10.2196/70726]
28. Kheawwan P, Thanomlikhit C, Narajeeenron K, Rojnawee S. Translation and psychometric validation of the Thai version of TeamSTEPPS® team performance observation tool. J Interprof Care 2024;38(3):573-582 [FREE Full text] [doi: 10.1080/13561820.2024.2307547] [Medline: 38343289]
29. Luck SJ. An Introduction to The Event-Related Potential Technique. Cambridge, MA: The MIT press; 2014.
30. Dhanakoses K, Pavavimol T, Issarasak S. The codesign process of virtual simulation games in medical education: a case study of the ER-VIPE platform. CoDesign 2025:1-20. [doi: 10.1080/15710882.2025.2515222]
31. HTC Corporation. 2019. URL: https://developer.vive.com/resources/hardware-guides/vive-cosmos-specs-user-guide/ [accessed 2024-09-15]
32. Klug M, Gramann K. Identifying key factors for improving ICA-based decomposition of EEG data in mobile and stationary experiments. Eur J Neurosci 2021;54(12):8406-8420. [doi: 10.1111/ejn.14992] [Medline: 33012055]
33. Butenhof DR. Programming with POSIX Threads. Japan: Addison-Wesley Professional; 1997.

34. Bastos AM, Schoffelen JM. A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. Front Syst Neurosci 2015;9:175 [FREE Full text] [doi: 10.3389/fnsys.2015.00175] [Medline: 26778976]
35. Lotrakul M, Sumrithe S, Saipanish R. Reliability and validity of the Thai version of the PHQ-9. BMC Psychiatry 2008;8(1):46. [doi: 10.1186/1471-244x-8-46]
36. Kim HK, Park J, Choi Y, Choe M. Virtual reality sickness questionnaire (VRSQ): motion sickness measurement index in a virtual reality environment. Appl Ergon 2018;69:66-73. [doi: 10.1016/j.apergo.2017.12.016] [Medline: 29477332]
37. Klem GH, Lüders HO, Jasper HH, Elger C. The ten-twenty electrode system of the international federation. The international federation of clinical neurophysiology. Electroencephalogr Clin Neurophysiol Suppl 1999;52:3-6. [Medline: 10590970]
38. Dien J. Issues in the application of the average reference: review, critiques, and recommendations. Behav Res Methods Instrum Comput 1998;30(1):34-43. [doi: 10.3758/bf03209414]
39. Garro F, Sappia MS, Costa HA. SSVEP-based brain-computer interface as an input device for an alternative communication system: Parameters assessment and case report of performance in a healthy and an ALS user. 2019 Presented at: 2019 IEEE International Conference on Systems, Man and Cybernetics; Oct 06, 2019; Bari, Italy. [doi: 10.1109/smc.2019.8914572]
40. OpenBCI. GitHub. 2023. URL: https://github.com/OpenBCI/OpenBCI_GUI [accessed 2023-12-21]
41. BrainFlow. GitHub. 2023. URL: https://github.com/brainflow-dev/brainflow/ [accessed 2023-12-21]
42. Lab Streaming Layer. GitHub. 2023. URL: https://github.com/labstreaminglayer/liblsl-Csharp [accessed 2023-12-21]
43. NVIDIA app. NVIDIA. 2023. URL: https://www.nvidia.com/en-us/geforce/geforce-experience/shadowplay/ [accessed 2023-12-22]
44. Klimesch W. α-band oscillations, attention, and controlled access to stored information. Trends Cogn Sci 2012;16(12):606-617 [FREE Full text] [doi: 10.1016/j.tics.2012.10.007] [Medline: 23141428]
45. Foxe JJ, Snyder AC. The role of alpha-band brain oscillations as a sensory suppression mechanism during selective attention. Front Psychol 2011;2:154 [FREE Full text] [doi: 10.3389/fpsyg.2011.00154] [Medline: 21779269]
46. Compton RJ, Arnstein D, Freedman G, Dainer-Best J, Liss A. Cognitive control in the intertrial interval: evidence from EEG alpha power. Psychophysiology 2011;48(5):583-590. [doi: 10.1111/j.1469-8986.2010.01124.x] [Medline: 20840195]
47. Clements GM, Bowie DC, Gyurkovics M, Low KA, Fabiani M, Gratton G. Spontaneous alpha and theta oscillations are related to complementary aspects of cognitive control in younger and older adults. Front Hum Neurosci 2021;15:621620 [FREE Full text] [doi: 10.3389/fnhum.2021.621620] [Medline: 33841114]
48. Fisher R. Statistical Methods for Research Workers. Edinburgh: Oliver and Boyd; 1950.
49. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016;15(2):155-163 [FREE Full text] [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
50. Package 'irr'. r-project. 2019. URL: https://cran.r-project.org/web/packages/irr/irr.pdf [accessed 2025-09-16]
51. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological Assessment 1994;6(4):284-290. [doi: 10.1037/1040-3590.6.4.284]
52. Rousseeuw PJ, van Driessen KV. A fast algorithm for the minimum covariance determinant estimator. Technometrics 1999;41(3):212. [doi: 10.2307/1270566]
53. Venables WN, Ripley BD. Modern Applied Statistics with S. New York: Springer; 2002.
54. Groppe D. fdr_bh. MATLAB Help Center. 2024. URL: https://www.mathworks.com/matlabcentral/fileexchange/27418-fdr_bh [accessed 2024-09-22]
55. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 2018;57(1):289-300. [doi: 10.1111/j.2517-6161.1995.tb02031.x]
56. Ngamjarus C, Pattanittum P. n4Studies Plus. Apple App Store. 2024. URL: https://apps.apple.com/th/app/n4studies-plus/id6450650727 [accessed 2024-10-25]
57. Tang H, Mai X, Wang S, Zhu C, Krueger F, Liu C. Interpersonal brain synchronization in the right temporo-parietal junction during face-to-face economic exchange. Soc Cogn Affect Neurosci 2016;11(1):23-32 [FREE Full text] [doi: 10.1093/scan/nsv092] [Medline: 26211014]
58. Kikuchi Y, Tanioka K, Hiroyasu T, Hiwa S. Interpersonal brain synchronization during face-to-face economic exchange between acquainted dyads. Oxf Open Neurosci 2023;2:kvad007 [FREE Full text] [doi: 10.1093/oons/kvad007] [Medline: 38596234]
59. Babiloni F, Astolfi L. Social neuroscience and hyperscanning techniques: past, present and future. Neurosci Biobehav Rev 2014;44:76-93 [FREE Full text] [doi: 10.1016/j.neubiorev.2012.07.006] [Medline: 22917915]
60. Cui X, Bryant DM, Reiss AL. NIRS-based hyperscanning reveals increased interpersonal coherence in superior frontal cortex during cooperation. Neuroimage 2012;59(3):2430-2437 [FREE Full text] [doi: 10.1016/j.neuroimage.2011.09.003] [Medline: 21933717]
61. Mu Y, Guo C, Han S. Oxytocin enhances inter-brain synchrony during social coordination in male adults. Soc Cogn Affect Neurosci 2016;11(12):1882-1893 [FREE Full text] [doi: 10.1093/scan/nsw106] [Medline: 27510498]
62. Szymanski C, Pesquita A, Brennan AA, Perdikis D, Enns JT, Brick TR, et al. Teams on the same wavelength perform better: inter-brain phase synchronization constitutes a neural substrate for social facilitation. Neuroimage 2017;152:425-436. [doi: 10.1016/j.neuroimage.2017.03.013] [Medline: 28284802]

63. Antonenko PD, Davis R, Wang J, Celepkolu M. On the same wavelength: exploring team neurosynchrony in undergraduate dyads solving a cyberlearning problem with collaborative scripts. Mind Brain and Education 2019;13(1):4-13. [doi: 10.1111/mbe.12187]
64. Solano CH, Dunnam M. Two's company: self-disclosure and reciprocity in triads versus dyads. Soc Psychol Q 1985;48(2):183-187. [doi: 10.2307/3033613]
65. Herrera D, Novick D, Jan D, Traum D. Dialog Behaviors across Culture and Group Size. Berlin: Springer Berlin Heidelberg; 2011:450-459.
66. Nowak AK, Vallacher RR, Praszkier R, Rychwalska A, Zochowski M. Synchronization in the emergence of social relations. In: Nowak AK, Vallacher RR, Praszkier R, Rychwalska A, Zochowski M, editors. Sync: The Emergence of Function in Minds, Groups and Societies. Cham: Springer International Publishing; 2020:87-112.
67. Nowak AK, Vallacher RR, Praszkier R, Rychwalska A, Zochowski M. Synchronization in groups and societies. In: Nowak AK, Vallacher RR, Praszkier R, Rychwalska A, Zochowski M, editors. Sync: The Emergence of Function in Minds, Groups and Societies. Cham: Springer International Publishing; 2020:113-136.
68. DiPierro K, Lee H, Pain KJ, Durning SJ, Choi JJ. Groupthink among health professional teams in patient care: a scoping review. Med Teach 2022;44(3):309-318 [FREE Full text] [doi: 10.1080/0142159X.2021.1987404] [Medline: 34641741]
69. Stout RJ, Cannon-Bowers JA, Salas E. In: Salas E, editor. The Role of Shared Mental Models in Developing Team Situational Awareness: Implications for Training. Routledge: Situational awareness; 2017:287-318.
70. Mohammed S, Hamilton K, Sánchez-Manzanares M, Rico R. In: Salas E, Rico R, Passmore J, editors. Team Cognition: Team Mental Models and Situation Awareness. USA: The Wiley Blackwell Handbook of the Psychology of Team Working and Collaborative Processes; 2017:369-392.
71. Fornander L, Garrido Granhagen M, Molin I, Laukkanen K, Björnström Karlsson K, Berggren P, et al. The use of specific coordination behaviours to manage information processing and task distribution in real and simulated trauma teamwork: an observational study. Scand J Trauma Resusc Emerg Med 2024;32(1):128 [FREE Full text] [doi: 10.1186/s13049-024-01287-x] [Medline: 39658788]
72. Brahler CJ, Donahoe-Fillmore B. Technology-enabled visualization of team typologies at a multi-institutional IPE event. Education Sciences 2023;13(10):981. [doi: 10.3390/educsci13100981]
73. Kang P, Moisa M, Lindström B, Soutschek A, Ruff CC, Tobler PN. Causal involvement of dorsomedial prefrontal cortex in learning the predictability of observable actions. Nat Commun 2024;15(1):8305 [FREE Full text] [doi: 10.1038/s41467-024-52559-0] [Medline: 39333062]
74. Mahmoodi A, Luo S, Harbison C, Piray P, Rushworth MFS. Human hippocampus and dorsomedial prefrontal cortex infer and update latent causes during social interaction. Neuron 2024;112(22):3796-3809.e9 [FREE Full text] [doi: 10.1016/j.neuron.2024.09.001] [Medline: 39353432]
75. Lizcano-Cortés F, Rasgado-Toledo J, Giudicessi A, Giordano M. Theory of mind and its elusive structural substrate. Front Hum Neurosci 2021;15:618630 [FREE Full text] [doi: 10.3389/fnhum.2021.618630] [Medline: 33762915]
76. Cristiano A, Finisguerra A, Urgesi C, Avenanti A, Tidoni E. Functional role of the theory of mind network in integrating mentalistic prior information with action kinematics during action observation. Cortex 2023;166:107-120. [doi: 10.1016/j.cortex.2023.05.009] [Medline: 37354870]
77. Mahmoodi A, Harbison C, Bongioanni A, Emberton A, Roumazeilles L, Sallet J, et al. A frontopolar-temporal circuit determines the impact of social information in macaque decision making. Neuron 2024;112(1):84-92.e6 [FREE Full text] [doi: 10.1016/j.neuron.2023.09.035] [Medline: 37863039]
78. Ferrucci L, Nougaret S, Ceccarelli F, Sacchetti S, Fascianelli V, Benozzo D, et al. Social monitoring of actions in the macaque frontopolar cortex. Prog Neurobiol 2022;218:102339 [FREE Full text] [doi: 10.1016/j.pneurobio.2022.102339] [Medline: 35963359]
79. Bilek E, Ruf M, Schäfer A, Akdeniz C, Calhoun VD, Schmahl C, et al. Information flow between interacting human brains: identification, validation, and relationship to social expertise. Proc Natl Acad Sci USA 2015;112(16):5207-5212 [FREE Full text] [doi: 10.1073/pnas.1421831112] [Medline: 25848050]
80. Nácher V, Ledberg A, Deco G, Romo R. Coherent delta-band oscillations between cortical areas correlate with decision making. Proc Natl Acad Sci USA 2013;110(37):15085-15090 [FREE Full text] [doi: 10.1073/pnas.1314681110] [Medline: 23980180]
81. Song K, Meng M, Chen L, Zhou K, Luo H. Behavioral oscillations in attention: rhythmic α pulses mediated through θ band. J Neurosci 2014;34(14):4837-4844 [FREE Full text] [doi: 10.1523/JNEUROSCI.4856-13.2014] [Medline: 24695703]
82. Engel AK, Fries P. Beta-band oscillations-signalling the status quo? Curr Opin Neurobiol 2010;20(2):156-165. [doi: 10.1016/j.conb.2010.02.015] [Medline: 20359884]
83. Goldstein P, Weissman-Fogel I, Dumas G, Shamay-Tsoory SG. Brain-to-brain coupling during handholding is associated with pain reduction. Proc Natl Acad Sci U S A 2018;115(11):E2528-E2537 [FREE Full text] [doi: 10.1073/pnas.1703643115] [Medline: 29483250]
84. Davidesco I, Laurent E, Valk H, West T, Milne C, Poeppel D, et al. The temporal dynamics of brain-to-brain synchrony between students and teachers predict learning outcomes. Psychol Sci 2023;34(5):633-643. [doi: 10.1177/09567976231163872] [Medline: 37053267]

85. Ahn S, Cho H, Kwon M, Kim K, Kwon H, Kim BS, et al. Interbrain phase synchronization during turn-taking verbal interaction-a hyperscanning study using simultaneous EEG/MEG. Hum Brain Mapp 2018;39(1):171-188. [doi: 10.1002/hbm.23834] [Medline: 29024193]
86. Jahng J, Kralik JD, Hwang DU, Jeong J. Neural dynamics of two players when using nonverbal cues to gauge intentions to cooperate during the prisoner's dilemma game. Neuroimage 2017;157:263-274. [doi: 10.1016/j.neuroimage.2017.06.024] [Medline: 28610901]
87. Wikström V, Saarikivi K, Falcon M, Makkonen T, Martikainen S, Putkinen V, et al. Inter-brain synchronization occurs without physical co-presence during cooperative online gaming. Neuropsychologia 2022;174:108316 [FREE Full text] [doi: 10.1016/j.neuropsychologia.2022.108316] [Medline: 35810882]
88. Barraza P, Pérez A, Rodríguez E. Brain-to-brain coupling in the gamma-band as a marker of shared intentionality. Front Hum Neurosci 2020;14:295 [FREE Full text] [doi: 10.3389/fnhum.2020.00295] [Medline: 32848670]
89. Haegens S, Händel BF, Jensen O. Top-down controlled alpha band activity in somatosensory areas determines behavioral performance in a discrimination task. J Neurosci 2011;31(14):5197-5204 [FREE Full text] [doi: 10.1523/JNEUROSCI.5199-10.2011] [Medline: 21471354]
90. Palva S, Palva JM. New vistas for alpha-frequency band oscillations. Trends Neurosci 2007;30(4):150-158. [doi: 10.1016/j.tins.2007.02.001] [Medline: 17307258]
91. Cahn BR, Polich J. Meditation states and traits: EEG, ERP, and neuroimaging studies. Psychol Bull 2006;132(2):180-211. [doi: 10.1037/0033-2909.132.2.180] [Medline: 16536641]
92. Jensen O. Distractor inhibition by alpha oscillations is controlled by an indirect mechanism governed by goal-relevant information. Commun Psychol 2024;2(1):36 [FREE Full text] [doi: 10.1038/s44271-024-00081-w] [Medline: 38665356]
93. Mouri FI, Valderrama CE, Camorlinga SG. Identifying relevant asymmetry features of EEG for emotion processing. Front Psychol 2023;14:1217178 [FREE Full text] [doi: 10.3389/fpsyg.2023.1217178] [Medline: 37663334]
94. van der Vinne N, Vollebregt MA, van Putten MJAM, Arns M. Frontal alpha asymmetry as a diagnostic marker in depression: fact or fiction? A meta-analysis. Neuroimage Clin 2017;16:79-87 [FREE Full text] [doi: 10.1016/j.nicl.2017.07.006] [Medline: 28761811]
95. Brancaccio A, Tabarelli D, Bigica M, Baldauf D. Cortical source localization of sleep-stage specific oscillatory activity. Sci Rep 2020;10(1):6976 [FREE Full text] [doi: 10.1038/s41598-020-63933-5] [Medline: 32332806]
96. Burgess A, van Diggele C, Roberts C, Mellis C. Teaching clinical handover with ISBAR. BMC Med Educ 2020;20(Suppl 2):459 [FREE Full text] [doi: 10.1186/s12909-020-02285-0] [Medline: 33272274]
97. Cooper S, Porter J, Peach L. Measuring situation awareness in emergency settings: a systematic review of tools and outcomes. Open Access Emerg Med 2013:1. [doi: 10.2147/oaem.s53679]
98. Feng C, Liu S, Wanyan X, Chen H, Min Y, Ma Y. EEG Feature analysis related to situation awareness assessment and discrimination. Aerospace 2022;9(10):546. [doi: 10.3390/aerospace9100546]
99. Kaur A, Chaujar R, Chinnadurai V. Effects of neural mechanisms of pretask resting EEG alpha information on situational awareness: a functional connectivity approach. Hum Factors 2020;62(7):1150-1170. [doi: 10.1177/0018720819869129] [Medline: 31461374]
100. Li Q, Ng KKH, Yu SCM, Yiu CY, Lyu M. Recognising situation awareness associated with different workloads using EEG and eye-tracking features in air traffic control tasks. Knowledge-Based Systems 2023;260:110179. [doi: 10.1016/j.knosys.2022.110179]
101. Li X, Kang Y, Chen W, Liu F, Jiao Y, Luo Y. Recognizing the situation awareness of forklift operators based on EEG techniques in a field experiment. Front Neurosci 2024;18:1323190 [FREE Full text] [doi: 10.3389/fnins.2024.1323190] [Medline: 38445257]
102. Frohlich J, Toker D, Monti MM. Consciousness among delta waves: a paradox? Brain 2021;144(8):2257-2277. [doi: 10.1093/brain/awab095] [Medline: 33693596]
103. Tan E, Troller-Renfree SV, Morales S, Buzzell GA, McSweeney M, Antúnez M, et al. Theta activity and cognitive functioning: integrating evidence from resting-state and task-related developmental electroencephalography (EEG) research. Dev Cogn Neurosci 2024;67:101404. [doi: 10.1016/j.dcn.2024.101404] [Medline: 38852382]
104. Schmidt R, Herrojo Ruiz M, Kilavik BE, Lundqvist M, Starr PA, Aron AR. Beta oscillations in working memory, executive control of movement and thought, and sensorimotor function. J Neurosci 2019;39(42):8231-8238. [doi: 10.1523/jneurosci.1163-19.2019]
105. Belbin RM. Chapter 9 - the art of building a team. In: Belbin RM, editor. Team Roles at Work. Milton Park: Routledge; 2022:91-100.
106. Fraenkel JR, Wallen NE, Hyun H. How to Design and Evaluate Research in Education. New York: McGraw Hill; 2022.
107. Lamblin M, Murawski C, Whittle S, Fornito A. Social connectedness, mental health and the adolescent brain. Neurosci Biobehav Rev 2017;80:57-68. [doi: 10.1016/j.neubiorev.2017.05.010] [Medline: 28506925]
108. Conrod PJ, Nikolaou K. Annual research review: On the developmental neuropsychology of substance use disorders. J Child Psychol Psychiatry 2016;57(3):371-394. [doi: 10.1111/jcpp.12516] [Medline: 26889898]
109. Novak A, Vizjak K, Rakusa M. Cognitive impairment in people with epilepsy. J Clin Med 2022;11(1):267 [FREE Full text] [doi: 10.3390/jcm11010267] [Medline: 35012007]

110. Mu Y, Cerritos C, Khan F. Neural mechanisms underlying interpersonal coordination: A review of hyperscanning research. Social & Personality Psych 2018;12(11):e12421. [doi: 10.1111/spc3.12421]
111. Deng X, Chen X, Zhang L, Gao Q, Li X, An S. Adolescent social anxiety undermines adolescent-parent interbrain synchrony during emotional processing: a hyperscanning study. Int J Clin Health Psychol 2022;22(3):100329 [FREE Full text] [doi: 10.1016/j.ijchp.2022.100329] [Medline: 36111264]
112. Wang Q, Han Z, Hu X, Feng S, Wang H, Liu T, et al. Autism symptoms modulate interpersonal neural synchronization in children with Autism spectrum disorder in cooperative interactions. Brain Topogr 2020;33(1):112-122. [doi: 10.1007/s10548-019-00731-x] [Medline: 31560088]
113. Kruppa JA, Reindl V, Gerloff C, Oberwelland Weiss E, Prinz J, Herpertz-Dahlmann B, et al. Brain and motor synchrony in children and adolescents with ASD-a fNIRS hyperscanning study. Soc Cogn Affect Neurosci 2021;16(1-2):103-116. [doi: 10.1093/scan/nsaa092] [Medline: 32685971]
114. Tosi J, Taffoni F, Santacatterina M, Sannino R, Formica D. Performance evaluation of Bluetooth low energy: a systematic review. Sensors (Basel) 2017;17(12) [FREE Full text] [doi: 10.3390/s17122898] [Medline: 29236085]
115. Precision and accuracy of date time. Microsoft. 2010. URL: https://learn.microsoft.com/en-gb/archive/blogs/ericlippert/precision-and-accuracy-of-datetime [accessed 2024-08-13]
116. Kothe CA, Makeig S. BCILAB: a platform for brain-computer interface development. J Neural Eng 2013;10(5):056014. [doi: 10.1088/1741-2560/10/5/056014] [Medline: 23985960]
117. Miyakoshi M, Jurgiel J, Dillon A, Chang S, Piacentini J, Makeig S, et al. Modulation of frontal oscillatory power during blink suppression in children: effects of premonitory urge and reward. Cereb Cortex Commun 2020;1(1):tgaa046 [FREE Full text] [doi: 10.1093/texcom/tgaa046] [Medline: 34296114]
118. Jas M, Engemann DA, Bekhti Y, Raimondo F, Gramfort A. Autoreject: automated artifact rejection for MEG and EEG data. Neuroimage 2017;159:417-429 [FREE Full text] [doi: 10.1016/j.neuroimage.2017.06.030] [Medline: 28645840]
119. Nolan H, Whelan R, Reilly RB. FASTER: fully automated statistical thresholding for EEG artifact rejection. J Neurosci Methods 2010;192(1):152-162. [doi: 10.1016/j.jneumeth.2010.07.015] [Medline: 20654646]
120. Barachant A, Andreev A, Congedo M. The Riemannian Potato: an automatic and adaptive artifact detection method for online experiments using Riemannian geometry. TOBI Workshop lV 2013.
121. Diedrichsen J, Shadmehr R. Detecting and adjusting for artifacts in fMRI time series data. Neuroimage 2005;27(3):624-634 [FREE Full text] [doi: 10.1016/j.neuroimage.2005.04.039] [Medline: 15975828]
122. Rebenitsch L, Owen C. Review on cybersickness in applications and visual displays. Virtual Real 2016;20(2):101-125. [doi: 10.1007/s10055-016-0285-9]
123. Dimenhydrinate. National Library of Medicine. 2022. URL: https://medlineplus.gov/druginfo/meds/a607046.html [accessed 2022-01-15]
124. Hu S. Effects of dimenhydrinate on electroencephalographic activity. Percept Mot Skills 1997;84(3 Pt 1):1105-1106. [doi: 10.2466/pms.1997.84.3.1105] [Medline: 9172229]
125. Hasson U, Nir Y, Levy I, Fuhrmann G, Malach R. Intersubject synchronization of cortical activity during natural vision. Science 2004;303(5664):1634-1640. [doi: 10.1126/science.1089506] [Medline: 15016991]
126. Madsen J, Parra LC. Cognitive processing of a common stimulus synchronizes brains, hearts, and eyes. PNAS Nexus 2022;1(1):pgac020 [FREE Full text] [doi: 10.1093/pnasnexus/pgac020] [Medline: 36712806]
127. Loughran SP, Verrender A, Dalecki A, Burdon CA, Tagami K, Park J, et al. Radiofrequency electromagnetic field exposure and the resting EEG: exploring the thermal mechanism hypothesis. Int J Environ Res Public Health 2019;16(9):1505 [FREE Full text] [doi: 10.3390/ijerph16091505] [Medline: 31035391]

Abbreviations

EEG: Electroencephalogram
ICC: Intraclass correlation coefficient
PC: Personal computer
PHQ-9: Patient Health Questionnaire-9
SIMBIE: Simulation-based interprofessional education
TeamSTEPPS: Team Strategies and Tools to Enhance Performance and Patient Safety
TI: Total interdependence
VR: Virtual reality

Edited by B Lesselroth; submitted 06.12.24; peer-reviewed by B Senst, M Hellaby; comments to author 08.04.25; revised version received 16.08.25; accepted 09.09.25; published 20.10.25.

Please cite as:
Viriyopase A, Narajeenron K
Correlation Between Electroencephalogram Brain-to-Brain Synchronization and Team Strategies and Tools to Enhance Performance and Patient Safety Scores During Online Hexad Virtual Simulation-Based Interprofessional Education: Cross-Sectional Correlational Study
JMIR Med Educ 2025;11:e69725
URL: https://mededu.jmir.org/2025/1/e69725
doi:10.2196/69725
PMID:

Transforming Medical Education to Make Patient Safety Part of the Genome of a Modern Health Care Worker

Peter Lachman[1*], BA, MBBCH, MMed, MPH, MD; John Fitzsimons[1,2*], BMedSci, MBBCh, MSc

[1]Quality Improvement Department, Royal College of Physicians of Ireland, 19 South Frederick Street, Dublin, Ireland

[2]Children's Health Ireland at Temple Street, Dublin, Ireland

[*]all authors contributed equally

Corresponding Author:
Peter Lachman, BA, MBBCH, MMed, MPH, MD
Quality Improvement Department, Royal College of Physicians of Ireland, 19 South Frederick Street, Dublin, Ireland

Related Article:

https://mededu.jmir.org/2024/1/e64125

Abstract

Medical education has not traditionally recognized patient safety as a core subject. To foster a culture of patient safety and enhance psychological safety, it is essential to address the barriers and facilitators that currently impact the development and delivery of medical education curricula. The aim of including patient safety and psychological safety competencies in education curricula is to insert these into the genome of the modern health care worker.

(JMIR Med Educ 2025;11:e68046) doi:10.2196/68046

KEYWORDS

patient safety; psychological safety; medical curriculum; professional competence; clinical competence

It has been over 25 years since the beginning of the active development of the patient safety movement, during which the theories and methods of patient safety science have evolved. We now understand the key drivers and practices for safer care [1], and while there have been many achievements and successes in implementation, transfer to different contexts, reliability, and sustainability remain challenging.

One of the underlying problems is that the health care workforce has limited training in the theories and methods of patient safety and insufficient training in improvement or implementation science. Developing sustainable changes in the way we approach patient safety requires radically rethinking how we educate the health care workforce of the future, so that patient safety becomes integrated into the way they work, that is, part of the genome of a modern health care worker.

In the past, patient safety was presumed to be synonymous with being a professional, so it was implicit that on completing medical or nursing education, one would be safe. This clearly is not the case. The paper by Carillo et al [2] approaches this challenge from the perspective of the critical practice of psychological safety in the workforce as the foundation for safer care. They first assessed the current status of training programs for medical students and trainees across Europe, with regard to the acquisition of knowledge, skills, and attitudes about patient safety. This was followed by the development of a suggested set of competencies for psychological safety that should be acquired during training programs. The curriculum is a valuable addition to our understanding of foundations for safer care. The focus on psychological safety as a key competency of patient safety training, rather than a mere focus on knowledge, is a novel approach to the curriculum. Additionally, it can form the basis of a better response to the second victim following an adverse event. Several themes arose from the paper that should be considered.

First, the focus on psychological safety changes the focus of education, shifting away from concentrating only on knowledge acquisition, as in other curricula such as the World Health Organization (WHO) [3] patient safety curriculum, which is under review. A modern patient safety curriculum needs to specify the learner outcomes of knowing what to do and how to generate feelings of being psychologically safe when applying safety science in the workplace to create safer clinical teams. This will be essential for the delivery of the WHO Patient Safety Global Action Plan [4].

Second, a patient safety curriculum cannot stand alone outside the wider concepts of quality in health care. This is an ongoing debate, but many other domains of quality impact the safe delivery of care. Therefore, a patient safety curriculum should be part of a comprehensive set of competencies that facilitate the implementation of patient safety improvement initiatives. Knowledge and skills of improvement methodology and implementation science are essential, and there are examples

of frameworks that achieve this goal for comprehensive quality in health care [5]. Equally, psychological safety influences the success of improvement efforts, implementation efforts, and innovation, all of which depend on being able to speak openly and share new ideas without fear.

Third, it should be determined whether one can engender psychological safety via a training program alone and whether a program is the sole foundation on which a safe system can be built. Organizational culture is fundamental for psychological safety. Psychological safety can only thrive within a team or organization that has a culture of safety that includes a focus on communication, feedback, respect, and trust [6]. Although this is part of the program suggested, there is often a disconnect between the theoretical classroom and trainees' lived experiences. Psychological safety requires positive team and organizational relationships that facilitate team members being safe [7]. The proposed framework for psychological safety includes structural, interpersonal, and individual factors that extend beyond education and depend heavily on leadership. Applying this in practice is challenging [8].

Fourth, to create a strategy for safer care delivery, we need to consider the reasons why medical education has not made patient safety an integral part of the curriculum, despite growing evidence of interventions that decrease harm and create a safer health care environment. Most academic institutions remain hierarchical and are steeped in the traditional medical model of teaching. Reasons for the reluctance to incorporate patient safety include lack of awareness of the emerging science, lack of leadership prioritization, curriculum overload, and competition with other emerging sciences [9]. For the proposed curriculum to succeed, these challenges need to be addressed head-on, and a radical rethink of medical education is required.

Finally, we need to consider the efficacy of patient safety training to make a difference. Two systematic reviews indicate the heterogeneity of papers that assess the effectiveness of patient safety education programs. The link between education and improved clinical outcomes is not strong [10]. There appears to be a disconnect between undergraduate patient safety training and what happens in the clinical setting [11]. This indicates the need for training programs to be integrated into postgraduate and undergraduate programs. It also suggests the need for early evaluation of any new program to ensure that what is imagined is being achieved.

In conclusion, Carillo et al [2] have shown a way forward for patient safety training. The challenge will be implementation within traditional medical education curricula. Perhaps the solution to this could be coproduced by educators, trainees, and patients rather than created by experts and mentors alone. Even though the goal must be transformation in how patient safety is considered within medical education, we can start to create the conditions for these competencies to thrive at our next classroom meeting, simulation session, team huddle, or handover. Imagine the power of a senior clinician openly sharing their vulnerability of not being able to know everything that is required to be safe, inviting respectful dissent, and graciously embracing difficult news. That is the improvement way!

Conflicts of Interest

None declared.

References

1. Lachman P, Runnacles J, Jayadev A, Brennan J, Fitzsimons J, editors. Oxford Professional Practice: Handbook of Patient Safety: Oxford University Press; 2022. [doi: 10.1093/med/9780192846877.001.0001]
2. Carrillo I, Skoumalová I, Bruus I, et al. Psychological safety competency training during the clinical internship from the perspective of health care trainee mentors in 11 pan-european countries: mixed methods observational study. JMIR Med Educ 2024 Oct 7;10:e64125. [doi: 10.2196/64125] [Medline: 39374073]
3. Patient safety curriculum guide: multi-professional edition. World Health Organization. 2011 Jul 6. URL: https://www.who.int/publications/i/item/9789241501958 [accessed 2025-01-06]
4. Global patient safety action plan 2021-2030. World Health Organization. 2021. URL: https://www.who.int/publications/b/57613 [accessed 2025-01-06]
5. Schrimmer K, Williams N, Mercado S, Pitts J, Polancich S. Workforce competencies for healthcare quality professionals: leading quality-driven healthcare. J Healthc Qual 2019;41(4):259-265. [doi: 10.1097/JHQ.0000000000000212] [Medline: 31283704]
6. Hallam KT, Popovic N, Karimi L. Identifying the key elements of psychologically safe workplaces in healthcare settings. Brain Sci 2023 Oct 11;13(10):1450. [doi: 10.3390/brainsci13101450] [Medline: 37891818]
7. Edmondson AC. The Fearless Organization: Creating Psychological Safety in the Workplace for Learning, Innovation, and Growth: John Wiley & Sons, Inc; 2019.
8. Ito A, Sato K, Yumoto Y, Sasaki M, Ogata Y. A concept analysis of psychological safety: further understanding for application to health care. Nurs Open 2022 Jan;9(1):467-489. [doi: 10.1002/nop2.1086] [Medline: 34651454]
9. Wu AW, Busch IM. Patient safety: a new basic science for professional education. GMS J Med Educ 2019 Mar 15;36(2):Doc21. [doi: 10.3205/zma001229] [Medline: 30993179]
10. Kirkman MA, Sevdalis N, Arora S, Baker P, Vincent C, Ahmed M. The outcomes of recent patient safety education interventions for trainee physicians and medical students: a systematic review. BMJ Open 2015 May 20;5(5):e007705. [doi: 10.1136/bmjopen-2015-007705] [Medline: 25995240]

11. Sheehan P, Joy A, Fleming A, Vosper H, McCarthy S. Human factors and patient safety in undergraduate healthcare education: a systematic review. Hum Factors Health 2022 Dec;2(2772-5014):100019. [doi: 10.1016/j.hfh.2022.100019]

Abbreviations

WHO: World Health Organization

Edited by B Lesselroth; submitted 27.10.24; this is a non–peer-reviewed article;accepted 07.12.24; published 17.01.25.

Please cite as:
Lachman P, Fitzsimons J
Transforming Medical Education to Make Patient Safety Part of the Genome of a Modern Health Care Worker
JMIR Med Educ 2025;11:e68046
URL: https://mededu.jmir.org/2025/1/e68046
doi:10.2196/68046

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org